Data-driven Toolkit To Combat Cyberbullying amongst Teens

Kanishk Verma

BEng, MSc.

A dissertation submitted in fulfilment of the requirements for the award of Doctor of Philosophy (PhD)

to the



School of Computing, Dublin City University

Supervisors:

Dr Brian Davis, Dublin City University

Dr Tijana Milosevic, University College Dublin

Dr Rebecca Umbach, Google

March, 2026

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of *Doctor of Philosophy (PhD)* is entirely my own work, and that I have exercised reasonable care to ensure that the work is original and has conformed to the regulations on the use and declaration of Generative AI, and does not, to the best of my knowledge, breach any law of copyright, and has not been taken from the work of others, save to the extent that such work has been cited and acknowledged within the text of my work.

Signed: **Kanishk Verma**

ID No.: 21270321

Date: March 27, 2026

Generative AI Declaration

Generative AI (genAI) is used in this study for: (i) paraphrasing, spell checks and to verify grammar; and (ii) part of research methodology and extensive benchmarking experiments with both open- and closed-source genAI tools (See Chapters 5, 6, 7).

Tools (T): For textual support, we used the DCU-whitelisted tool Google Gemini 2.5-Flash. For the methodological and benchmarking purposes, we evaluated multiple open- and closed-source LLMs, including Claude-2, Llama-2.7B-Chat, GPT-2, Gemini-2.5-Flash, and GPT-4o (details and version references are provided within the relevant chapters).

Outputs (O): The outputs generated for proofreading consisted of paraphrased text, grammar suggestions, and spelling corrections. These were manually reviewed, edited, and partially discarded to preserve authorial intent. In the research methodology, genAI tools generated different outputs depending on the tasks (details are provided within the relevant chapters).

Prompts (P): Proofreading prompts were short task-specific queries such as "*Improve grammar while preserving meaning*". Methodological prompts followed standardised task definitions and experimental templates (details appear within the relevant chapters).

Iterations (I): Textual outputs were typically refined through 1–2 iterative edits using human judgment. Experimental benchmarking involved multiple iterations with different prompts, hyperparameters, and test configurations (details appear within the relevant chapters).

Critical Reflection (C): All genAI outputs were critically reviewed for content accuracy, potential hallucination, bias, and suitability. The tools did not contribute conceptual content or interpretations. Please note that analytical reasoning, synthesis, and argumentation remain the author's own. The use of genAI is consistent with DCU's academic integrity guidelines.

Acknowledgements

This research was primarily supported by Research Ireland (formerly the Irish Research Council) through the Enterprise Partnership Scheme with Google for Online Content Safety (grant number EPSPG/2021/161). Additional support was provided by the Disruptive Technologies Innovation Fund (DTIF) under the project "*Cilter: Protecting Children Online*" (grant number DT 2021 0362) from the Department of Enterprise, Trade and Employment in Ireland, administered by Enterprise Ireland. This work was also conducted with the financial support of the ADAPT Centre, the Research Ireland Centre for AI-Driven Digital Content Technology at Dublin City University (grant 13/RC/2106_P2).

I would first like to thank Dr Brian Davis for providing me the opportunity to work as a Research Assistant and start my research career at the ADAPT Centre, Dublin City University. Working with Dr Tijana Milosevic on the Facebook Content Policy Award Phase 2 project introduced me to the field of online safety research and shaped my research direction. I am thankful for their continued support. I am particularly grateful to Dr Rebecca Umbach for her invaluable guidance in understanding the complexities of trust and safety. My thanks extend to Dr Kolawole Adebayo and Dr Joachim Wagner for their supportive discussions throughout this journey.

I want to acknowledge Prof. James O'Higgins Norman, Dr Darragh McCashin, Dr Eileen Culloty, and the entire Anti-Bullying Centre team for their encouragement and collaborative spirit. I am grateful to Alexandros Poulis at TransPerfect and Jane McGarrigle at Webwise for their support and insights. Special thanks to Sri Balaaji Natarajan Kalaivendan for his contributions to various projects we worked on together.

I deeply appreciate the support of my peers, Michela Lorandi, Tytti Rintamaki, Liang Xu (Charlie), Mohammed Sabry, Dr Lauren Cassidy, and Dr Simone Mille, who made this challenging journey more enjoyable and provided a wonderful support network. My heartfelt gratitude goes to Kinjalk Verma, Swarushi Vikhrol and my partner Rachana Talekar, for their unwavering support and understanding throughout this process.

Finally, I want to mention my mother, Rashmi Ravija, who unexpectedly passed away this year. Although you are not here to see this Maa, I thank you most of all, and wish you were here. This work is dedicated to your memory.

Contents

I Prelude **xix**

1 Introduction **1**
1.1 Motivation and Problem Description 1
1.2 Research Questions 10
1.2.1 Research Question 1. (RQ.1) 10
1.2.2 Research Question 2. (RQ.2) 11
1.2.3 Research Question 3. (RQ.3) 11
1.3 Contributions 12
1.4 Thesis Structure 14
1.5 Publications 17
1.5.1 Computational Science Publications 17
1.5.2 Social Science Publications 17
1.5.3 Human-computer Interaction Publications 17

II Foundations **19**

2 Background and Related Work **20**
2.1 Literature Review Methodology 21
2.2 Contextual and Theoretical Landscape of Social Scientific Explorations 23
2.2.1 Contrasting cyberbullying and physical bullying 24
2.2.2 Theoretical Background 27
2.3 Youth and Adult Participation in CB studies 30
2.3.1 Youth perspectives on cyberbullying detection 32
2.3.2 Statistical Methods for Survey and Annotation Studies 38
2.3.3 Understanding and Annotating Cyberbullying in NLP 44
2.4 Foundational Landscape of Computational Investigations 51
2.4.1 Automatic Evaluation Metrics in CB detection 58
2.4.2 Binary Cyberbullying Detection 61
2.4.3 Fine-grained Cyberbullying Detection 74
2.4.4 Multi-task Cyberbullying Detection 80
2.4.5 Multi-modal Cyberbullying Detection 82
2.4.6 Synthetic data generation and related applications 94
2.4.7 Understanding Obfuscations and Perturbation in NLP 102
2.5 Analytical Conclusion and Research Gaps 105

III Contributions 116

3 Youth Participatory Methodology and Framework 117
3.1 Overview of SDLC Phases . . . 118
3.2 Instruments Informing Participatory Approach . . . 120
3.2.1 Cyberbullying in a Computational Linguistic Frame . . . 120
3.2.2 Cyberbullying through Routine Activities Theory . . . 122
3.2.3 Participatory Software Development Life Cycle . . . 123
3.2.4 Conceptualising Engagement, Witnessing Harm, and Reporting in Cyberbullying Contexts . . . 124
3.2.5 Youth Agency as Co-researchers from a Rights-based Perspective . . . 127
3.3 SDLC Phases . . . 130
3.3.1 Phase-1: Requirement Gathering . . . 130
3.3.2 Phase-2: High-level Design . . . 137
3.3.3 Phase-3: Detailed Dataset Design . . . 140
3.4 Young Research Assistant Programme . . . 144
3.5 Summary . . . 146

4 Youth Centred Insights into Cyberbullying 149
4.1 Phase-1: Requirement Gathering Survey Results . . . 150
4.1.1 Social Media and Platform Usage . . . 150
4.1.2 Platform Features and Engagement Styles . . . 152
4.1.3 Cyberbullying Exposure and Statistical Analysis: . . . 153
4.1.4 Forms of cyberbullying witnessed and content modality . . . 154
4.1.5 Response strategies . . . 157
4.2 Phase-2: High-Level Design Results . . . 160
4.2.1 Scoping Review Results . . . 163
4.2.2 Scenario Evaluation . . . 166
4.2.3 Scenario Adaptation . . . 172
4.2.4 Scenario Emergence . . . 175
4.3 Phase-3: Detailed Design Results . . . 177
4.3.1 Scenario Modifications . . . 177
4.3.2 Youth Dataset Co-creation . . . 178
4.3.3 Descriptive Statistics of Youth Co-created Dataset . . . 182
4.4 Summary . . . 183

5 Curating Cyberbullying Data 187
5.1 Data Augmentation Strategies and Insights . . . 188
5.1.1 Human Translation (HT) ($\mathcal{D}^{HT}_{IT\rightarrow EN}$) . . . 188
5.1.2 Machine Translation (MT) ($\mathcal{D}^{MT}_{IT\rightarrow EN}$) . . . 194
5.1.3 Human-aided Machine Translation (H-MT) ($\mathcal{D}^{\text{HMT}}_{FR\rightarrow EN}$) . . . 197
5.1.4 LLM-aided generation ($\mathcal{D}_{\text{LLM}}$) . . . 201
5.2 Dataset Profiling . . . 213
5.2.1 Descriptive Statistics . . . 214
5.2.2 Lexical Diversity . . . 218
5.2.3 Turn-taking Analysis . . . 220
5.2.4 Conclusion and Comparative Insights . . . 229
5.3 Intrinsic Evaluation and Results . . . 230

5.3.1 Objectives and Dimensions . . . 232
5.3.2 Study Setup . . . 235
5.3.3 Results . . . 243
5.3.4 Conclusion . . . 251
5.4 Young Adult Annotation and Results . . . 252
5.4.1 Objectives and Tasks . . . 252
5.4.2 Ethical Considerations, Pilot Study and Feedback . . . 256
5.4.3 Main Study . . . 262
5.4.4 Results . . . 265
5.4.5 Conclusion and Insights . . . 277
5.5 Summary . . . 279

6 Insights into Binary, Multi-task Cyberbullying Detection, and Assessing Classifier Robustness 282
6.1 Binary Benchmarking with Appropriate Cyberbullying Datasets . . . 283
6.1.1 Problem Formulation . . . 284
6.1.2 Experiment Setup - Cross-platform benchmarking . . . 285
6.1.3 Experiment Setup - Evaluating Translation Strategies . . . 290
6.1.4 Results and Discussion . . . 294
6.2 Moving Towards Explainable Cyberbullying Detection . . . 300
6.2.1 Related Work . . . 300
6.2.2 Experiment Setup . . . 303
6.2.3 Results and Discussion . . . 306
6.3 Towards Multi-task Classification - Is it hurtful? By whom? . . . 313
6.3.1 Multi-class Role Identification . . . 313
6.3.2 Multi-task Hurtful + Role Identification . . . 317
6.3.3 Results and Discussion . . . 320
6.4 Assessing Model Robustness . . . 328
6.4.1 Related Work . . . 328
6.4.2 Adversarial Sample Curation . . . 329
6.4.3 Experiment Setup . . . 333
6.4.4 Results and Discussion . . . 340
6.5 Summary . . . 346

7 Vision-Language Models for Meme-Based Harm Detection 350
7.1 Methodology . . . 351
7.1.1 Phase-1: Unseen Data Curation . . . 353
7.1.2 Phase-2: Baseline Benchmarking . . . 357
7.1.3 Phase-3: Pilot Human Preference Dataset Curation . . . 363
7.2 Results and Discussion . . . 366
7.2.1 Data Collection . . . 366
7.2.2 Data Annotation . . . 367
7.2.3 Vision-language Benchmarking . . . 374
7.2.4 Pilot Human Preference Data Analysis . . . 383
7.3 Summary . . . 390

IV Conclusion **393**

8 Conclusion and Future Work **394**
8.1 Contributions . . . 395
8.2 Research Question Revisited . . . 397
8.3 Limitations . . . 402
8.4 Open Questions and Future Work . . . 406

A **476**
A.1 Young Research Assistant Programme . . . 476
A.2 Cyberbullying Exposure Sensitivity Analysis . . . 478
A.3 Descriptive Patterns in Scenario Evaluation Task . . . 483
A.4 Descriptive Statistics of Youth Co-created Dataset . . . 484
A.5 Individual message-level H-MT Analysis . . . 491
A.6 Pilot study: LLM-aided generation . . . 494
A.7 Phase-1: Plain Language Statement, Consent and Assent Forms . . . 496
A.7.1 Plain Language Statement . . . 496
A.7.2 Parent/Guardian Consent . . . 498
A.7.3 Teenager Assent . . . 499
A.8 Phase-1: Detailed Results . . . 500
A.9 Phase-2: High Level Design Scenario Lists . . . 503
A.10 LLM-aided Generation - Pilot Study . . . 525
A.11 LLM-aded Generation - Phase-2 Llama3.3 . . . 525
A.12 Sequential Turn Analysis Details . . . 530
A.13 Human Annotation Disagreement Pairs . . . 534
A.14 Technical Explanation . . . 535
A.15 Integrated Gradient Axioms . . . 535
A.16 Optimal Hyperparameters Binary Classification . . . 538
A.17 Phase-2: LLM Prompt Template . . . 538
A.18 Phase-2: Benchmarking Hyperparameter Configuration . . . 538
A.19 Phase-3: Pilot Prompt Template . . . 540
A.20 Optimal Hyperparameters Vision-language Benchmarking . . . 540

List of Figures

1.1 Heat-map raw counts of verbal abuse type (x-axis) and role of participant (y-axis) for both (A) (Sprugnoli et al., 2018) and (B) (Ollagnier et al., 2022) 5
1.2 Different multi-modal cyberbullying scenarios on different social media platforms. Please note, all figures are created using the Figma design tool. Figures (A) & (B) are adapted with permission from Milosevic et al. (2023) 6
1.3 Structural overview of research contributions in Part III 16

2.1 Distribution of labels in Hosseinmardi et al. (2015); Rafiq et al. (2015) 84

3.1 Overview of the youth participatory methodology framework across different phases 118

4.1 Distribution of age and gender of respondents for Phase-1 Requirement Gathering Survey 151
4.2 Platform Usage Frequency for Requirement Gathering Survey (Phase-1) 151
4.3 Percentage breakdown of Online Harms across all features collected via Requirement Gathering Phase-1 157
4.4 Heatmap showing the percentage of unique respondents reporting each type of online harm across platform features. Cell annotations are presented as x/y, where x = number of unique respondents reporting the specific harm type within that feature, and y = total number of respondents reporting any cyberbullying in that feature. Darker red indicates higher percentages, while lighter yellow indicates lower percentages. 158
4.5 Co-occurrence matrix of different forms of cyberbullying in response to Q4 of Phase-1: Requirement Gathering 159
4.6 Distribution of all content modalities that respondents have witnessed different forms of cyberbullying (Q5 in Table 3.2) 159
4.7 Distribution of content modalities that respondents have witnessed different forms of cyberbullying for specific features 160
4.8 Response Type associated with different social media features collected via Requirement Gathering Phase-1 161
4.9 Why would they ignore reporting as per different social media features collected via Requirement Gathering Phase-1 161
4.10 Likert-scale responses across Scenario categories by Cohort_1 169
4.11 Likert-scale responses across Scenario categories by Cohort_2 169

4.12 Likert-scale responses across Scenario categories by both Cohort_1 and Cohort_2 across six scenarios . . . 170
4.13 Percentage breakdown of social media features adapted for scenarios by Cohort_1 categories by cyberbullying types . . . 174

5.1 Relationship between different text descriptive statistics across all datasets . . . 216
5.2 Text feature comparison across youth datasets with Amica i.e., (Van Hee et al., 2018) and IG i.e., (Hosseinmardi et al., 2015) . . . 217
5.3 Text feature comparison across synthetic datasets with Amica i.e., (Van Hee et al., 2018) and IG i.e., (Hosseinmardi et al., 2015) . . . 217
5.4 Type-Token-Ration (TTR) across all datasets and also across Bully/Enabler texts and Victim/Defender texts across all datasets . . . 219
5.5 Shannon Entropy Distribution at both Word (Token) or Character Level across datasets . . . 220
5.6 Network graph depicting direction of Bi-gram transition with probabilities as weights across all datasets . . . 227
5.7 Gender and Ethnicity breakdown of the recruited annotators . . . 265
5.8 Age distribution of recruited annotators and their experience counts with witnessing different types of harm . . . 266
5.9 Heatmap depicting shift in majority labels across both limited and extended context conditions for Task B. . . . 273
5.10 Heatmap depicting shift in majority labels across both limited and extended context conditions for Task C. . . . 278

6.1 Imbalance in Dataset (0: `not-hurtful`; 1: `hurtful`) . . . 287
6.2 Pearson's Correlation between Mean Attention Weights and Mean Gradient-Importance Score for all LMs and Datasets . . . 309
6.3 BERT+ Mean-attention weights per Layer POS-tag wise. Please note, *Bullying* samples mean *Hurtful* samples and *Non-Bullying* samples mean *Not-Hurtful* samples. . . . 310
6.4 HateBERT+ Mean-attention weights per Layer POS-tag wise. Please note, *Bullying* samples mean *Hurtful* samples and *Non-Bullying* samples mean *Not-Hurtful* samples. . . . 310
6.5 Attention per layer for Negative Sentiment words. Please note X-axis denotes layers 0-11 and y-axis denotes range of attention scores. . . . 311
6.6 Attention per layer for Positive Sentiment words Please note X-axis denotes layers 0-11 and y-axis denotes range of attention scores. . . . 311
6.7 Attention analysis for Positive Sentiment Word . . . 312
6.8 Attention analysis for Negative Sentiment Word . . . 312
6.9 Hierarchical Pair-wise Classification Method . . . 316
6.10 BERT and GPT2 multi-task results with different configurations . . . 324
6.11 F1 score drop when measuring performance of different *encoder* and *decoder* transformers on $\mathcal{D}_{\text{Test}}$ to $\mathcal{D}_{\text{Gold}}$ . . . 325
6.12 F1 scores across different *encoder* and *decoder* transformers on $\mathcal{D}_{\text{Test}}$ and $\mathcal{D}_{\text{Gold}}$ . . . 326
6.13 Architectural Flow of Char-CNN . . . 334
6.14 Architectural Flow of Char-LSTM . . . 335
6.15 Architectural Flow of Character-CNN-BERT . . . 338

7.1 Methodological Framework for **RQ2.2** Section mall*Note: '*' indicates values differing from (Kiela et al., 2020), as original test-set labels are unavailable (total 9,000 images). 80% is used for training (90% of original training), 10% of training plus half of the dev set for validation, and the remaining 500 samples for testing.* 352
7.2 Image Collected from Imgur as part of the dataset 356
7.3 Unimodal baseline architectures includes (a) image-encoder based classifiers ($\mathcal{U}_I$): ViT and CLIP$_I$; (b) text-encoder base classifiers ($\mathcal{U}_T$): BERT 360
7.4 Multimodal fusion baseline architectures includes (a) Late Fusion: Fine-tuning CLIP with frozen and unfrozen layers and concatenation; (b) Early Fusion: Fine-tuning ViLT with a classification head 361
7.5 Prompting Large Vision-Language Models (LVLMs) such as Qwen-V2-VL and Llava-COT for binary classification 362

A.1 Diverging bar chart of mean Likert-scale responses across 15 scenarios for Cohort$_1$. Responses are averaged over three evaluative dimensions (Real, Report, Serious) and mapped to a numeric scale ranging from –3 (strong disagreement) to +3 (strong agreement). The central vertical line (0) represents neutrality, with shaded regions denoting agreement (green) and disagreement (red) zones. 485
A.2 Diverging bar chart of mean Likert-scale responses across 6 scenarios for Cohort$_2$. Responses are averaged over three evaluative dimensions (Real, Report, Serious) and mapped to a numeric scale ranging from –3 (strong disagreement) to +3 (strong agreement). The central vertical line (0) represents neutrality, with shaded regions denoting agreement (green) and disagreement (red) zones. 486
A.3 Diverging bar chart of mean Likert-scale responses across 6 scenarios for both Cohort$_1$ and Cohort$_2$. Responses are averaged over three evaluative dimensions (Real, Report, Serious) and mapped to a numeric scale ranging from –3 (strong disagreement) to +3 (strong agreement). The central vertical line (0) represents neutrality, with shaded regions denoting agreement (green) and disagreement (red) zones. 487
A.4 Comparative Character Length Distribution across messages for Youth Co-created dataset (Left top), Italian dataset by (Sprugnoli et al., 2018) (Right top), French dataset by Ollagnier et al. (2022); Ollagnier (2024), and real-life cyberbullying dataset by (Van Hee et al., 2018) 489
A.5 Distribution of BLEU, chrF, and COMET-QE scores of Company's MT + post Editing Outputs across 876 messages 493
A.6 Distribution of Alignment Agreement and Disagreement with translated (EN) and source (FR) corpora by categories of scenarios 493
A.7 Snippet of cyberbullying-like conversations generated by Claude-2 Anthropic (2023) and Llama2Chat$_{7B}$ Touvron et al. (2023). (See Table A.13 in Appendix A.10 for detailed response) 497
A.8 Network graph depicting direction of Bi-gram transition with probabilities as weights across all datasets 532

List of Tables

1.1 Mean role labels per question-answering thread in the English dataset by Van Hee et al. (2018). Note: Van Hee et al. (2018) refers to `Enabler` as *Bystander-assistants* and `Defender` as *Bystander-defenders.* 4
1.2 Computational Science Publications 17
1.3 Social Science Publications 18
1.4 Human Computer Interaction Publications 18

2.1 2×2 contingency table 41
2.2 List of studies that claim to detect "cyberbullying" but use out-of-domain datasets 69
2.3 Dataset Studies grouped by source platform. Note: Disagr. → Disagreement Analysis; IAA → Inter-Annotator Agreement; EN → English; NL → Dutch; IT → Italian; 72
2.4 Multi-class cyberbullying label distribution in Van Hee et al. (2018) . 77

3.1 Table indicating Platform-Type and associated Features 131
3.2 Survey Questions with conditional display of questions based on participant responses 133
3.3 Engagement with Platform Feature and Response Choices modified and adapted from Shaw et al. (2022) 133
3.4 Categorisation of online harm adopted by Rodríguez-de Dios et al. (2018); Stoilova et al. (2021) 134

4.1 Descriptive Statistics for Platform Usage and Cyberbullying Witnessing Behaviour by Platform Feature. Please note, usage and witness rate and counts are calculated by dichotomising as "No (0)" (Never/Hardly Ever) and "Yes (1)" (Sometimes/Often/Very Often) . 155
4.2 *Method A (Sequential Survey Logic)* Fisher's Exact Test Results for association - Feature Engagement Style v/s Witnessing Cyberbullying 155
4.3 Age and gender distribution of Cohort_1 and Cohort_2 162
4.4 Timeline and distribution of Cohort_1 - *young research assistants* in this study 163
4.5 Distribution of broad cyberbullying categories of scenarios and their modes of occurrence, indicating whether these scenarios were shortlisted for inclusion in the study. Please note: "`None`" *category refers to two scenarios from (Sprugnoli et al., 2018) that describe events themselves rather than individuals' reactions, meaning no specific form of cyberbullying is directly applicable* 166
4.6 Mapping of scenarios as per scenario-number, their sources, broad cyberbullying categories, and interaction modes. 167

4.7 Inter-rater agreement between both cohorts for each dimension presented as Krippendorff's Alpha (α) . . . 168
4.8 Fisher's exact test results examining associations between the three dimensions (*realism*, *seriousness*, and *report-worthiness*) across Cohort_1, Cohort_2, and the combined cohorts. Significant associations ($p < 0.05$) are highlighted in bold, with corresponding odds ratios (OR) and corrected odds ratios (COR) indicating the strength of the relationships. . . . 172
4.9 Overview of modified scenarios adapted by young research assistants. Text highlighted in blue indicates participant-introduced modifications. . . . 179
4.10 Distribution of messages across 10 different scenarios in youth co-created dataset . . . 182
4.11 Distribution of roles across youth co-created datasets . . . 183
4.12 Distribution of roles across three datasets of which (Van Hee et al., 2018) is a real-life cyberbullying dataset. Also (Ollagnier et al., 2022) has another role label conciliators that have 12.68% distribution . . . 183

5.1 Cohen's Kappa score for individual annotators with original annotations. . . . 191
5.2 Translation ambiguity, original labelling and domain expert error analysis. . . . 192
5.3 Number of segments (sentences) in training and test data for MT systems . . . 196
5.4 Comparison of Italian-English systems by automatic evaluation scores BLEU and chrF. . . . 197
5.5 Distribution of messages across different files and scenario-types as created by (Ollagnier et al., 2022; Ollagnier, 2024). Note '*bis*', '*third*', denote the use of the same scenario characters and situation but with messages role-played by different youth participants. . . . 199
5.6 Distribution of messages across different files and scenario-types as created by (Ollagnier et al., 2022; Ollagnier, 2024) used for Student Translations. . . . 199
5.7 Mean (M) and standard deviation (σ) of evaluation metrics at the atomic level (876 messages). BLEU/chrF are shown with company and student translations as reciprocal references, and COMET-QE provides reference-free quality estimates. . . . 200
5.8 Breakdown of conversations generated by LLMs across 10 problem input prompts . . . 210
5.9 Distribution of number of messages by participant role generated across both phases of LLM generation. Note: P1 is phase-1 and P2 is phase-2 . . . 210
5.10 Self-BLEU 1-4gram (token) scores for LLM-generated and youth-created datasets, where lower values indicate greater diversity and reduced redundancy within the dataset. . . . 211
5.11 Comparison of synthetic and real texts for Enabler and Defender roles using Fréchet BERT Distance (FBD) and Cosine Similarity, benchmarked against both Youth Co-created dataset and real life dataset by (Van Hee et al., 2018). . . . 212

5.12 Characteristics of datasets, including number of scenarios, messages, and words. Only subsets relevant to bullying labels are reported for Ask.fm and Instagram.***Note:*** *(Van Hee et al., 2018) original dataset for English after cleaning has 79314 messages across 6056 conversations, and (Hosseinmardi et al., 2015) original dataset for English has 42,259 comments across 769 posts (image+comment). In this table only the ones that are labelled as bullying for (Hosseinmardi et al., 2015) and role-labels are not negative for (Van Hee et al., 2018) are presented* 215
5.13 Mean and standard deviation of Type Token Ratio (TTR) across all datasets 219
5.14 Summary of datasets with total turns, number of conversations, average messages and turns per conversation, and turn-taking rate. . . 224
5.15 Maximum consecutive turn per role across all conversations 225
5.16 Average consecutive turn per role across all conversations 225
5.17 Average proportion of turns (percentages) taken by different roles across opening, middle, and closing phases in the Italian (Sprugnoli et al., 2018), French (Ollagnier et al., 2022; Ollagnier, 2024), youth co created, and LLM generated datasets. 228
5.18 Guideline for intrinsic human evaluation with definitions of core-dimensions 239
5.19 Guideline for intrinsic human evaluation with definitions of cyberbullying social dynamics with response choices. For detailed examples refer Table 5.20. 240
5.20 Cyberbullying Social Dynamics Dimension Examples with Positive and Negative Instances 242
5.21 Raw agreement percentages between five evaluators for all nine dimensions across the five datasets. 244
5.22 Inter-rater agreement scores for all data sources combined 245
5.23 IAA Fliess Kappa and PABAKK-OS across all dimensions **split by data-sources** 247
5.24 IAA Fliess Kappa and PABAKK-OS across all dimensions across **Social Scientist Group** and **Social Scientist Group with Adult Teen** 248
5.25 IAA Fliess Kappa and PABAKK-OS across all dimensions across **Content Moderator Group** and **Content Moderator Group with Adult Teen** 248
5.26 Pilot Study Inter-rater Relatability Scores 258
5.27 Annotation Category Examples and Non-Examples 261
5.28 Inter-rater reliability scores presented as Krippendorff's α values for all annotation tasks 266
5.29 Count of disagreement pairs for Task A: Aggression Type Labelling . 267
5.30 confusion matrix showing the results of classification with limited context versus extended context. 269
5.31 Distribution of majority and collapsed annotation labels across limited and extended context conditions for Sub-task B. 270
5.32 Count of disagreement pairs for Task B: Harm Type Labelling. Further details are presented in Table A.15 in Appendix A.13 271

5.33 Distribution of majority and collapsed annotation labels across limited and extended context conditions for Sub-task C. 274
5.34 Count of disagreement pairs for Task C: Intent Type Labelling. Further details are presented in Table A.16 in Appendix A.13 274

6.1 Overview of datasets used for cyberbullying detection, including the number of sentences, proportion of bullying labels, and average sentence length (in tokens) per platform. 286
6.2 Dataset Split 292
6.3 Results for Cross-platform Benchmarking (§6.1.2) 295
6.4 Original and Replicated Classification Results of training and testing GRU on the Italian Corpus 297
6.5 F1 scores for classification systems on both Human Translated (HT §5.1.1) and Machine Translated (MT §5.1.2) datasets discussed in §6.1.3 on **Hold-out Test set** 299
6.6 F1 scores for classification systems on both Human Translated (HT §5.1.1) and Machine Translated (MT §5.1.2) datasets discussed in §6.1.3 on **Scenario-C** with Expert labels 301
6.7 Dataset Description 304
6.8 Classification Results 308
6.9 Role Label Distribution across different synthetic datasets and youth co-created $\mathcal{D}_{\text{Gold}}$ 314
6.10 Distribution of role labels across Train, Validation, and Test sets. Percentages indicate the relative proportion per split. 314
6.11 Distribution of binary hurtfulness labels across the multi-source dataset. $\mathcal{D}_{\text{LLM}}^{\text{Phase-2}}$ is excluded since it does not contain hurtfulness annotations. This distribution highlights class imbalance across the combined dataset. . . . 318
6.12 Distribution of role labels across the multi-source dataset for multi-task learning. $\mathcal{D}_{\text{LLM}}^{\text{Phase-2}}$ only contributes to role labels, while $\mathcal{D}_{\text{AMiCA}}$ includes both role and hurtfulness labels. 318
6.13 Top-3 baseline benchmarking on $\mathcal{D}_{\textbf{Gold}}$ 320
6.14 Detailed breakdown of baseline benchmarking on $\mathcal{D}_{\textbf{Gold}}$ 321
6.15 Most common error across best performing models on $\mathcal{D}_{\text{Gold}}$ 321
6.16 Model performance comparison on Gold and Test datasets. All Llama-3.3-70B models use Temperature=0.1, do_sample=True, max_tokens=50, top_p=50, top_k=50. RoBERTa and GPT2-medium models trained on $\mathcal{D}_{IT\rightarrow EN}^{HT} \bigcup \mathcal{D}_{FR\rightarrow EN}^{HMT} \bigcup \mathcal{D}_{\text{LLM}}^{\text{Phase-2}}$ with upsampling and focal loss (Gold) or combined data (Test). 323
6.17 Error rates (%) for multi-class classification using GPT-2 with Hard Parameter Sharing and Focal Loss (HP-FL). Rows represent predicted labels and columns represent true labels, shown for both the Test-Set and $\mathcal{D}_{\text{Gold}}$ youth co-created dataset. 327
6.18 Error rates (%) for multi-class classification using BERT with Hard Parameter Sharing and Cross-Entropy loss (HP-CE). Rows represent predicted labels and columns represent true labels, shown for both the Test-Set and $\mathcal{D}_{\text{Gold}}$ youth co-created dataset. 328

6.19 Distribution of obfuscation patterns employed by seven participants (N=1,005 total variations). Patterns are mutually exclusive with each variation assigned to exactly one category. 332
6.20 Performance comparison of models on held-out test set and youth co-created $\mathcal{D}_{\text{Gold}}$ dataset. FP: False Positive rate, FN: False Negative rate. 341
6.21 Performance comparison on original sentences ($n = 61$) and adversarially obfuscated variations ($n = 423$). Industry APIs (Perspective, OpenAI) are benchmarked against custom-trained models. 342
6.22 Per-class flip rates and robustness scores across models. Robustness score is computed as $1 - \text{Flip Rate}$ 343
6.23 Attack Success Rate (ASR) per class and overall, computed as the proportion of originally correct predictions that become incorrect after obfuscation. 344
6.24 Average confidence (probability) drop per class across models, computed as the mean absolute difference between original and obfuscated prediction probabilities. 345

7.1 Distribution of Training, Validation, and Test-set. Note: `0` denotes (not-hateful or not-harmful) and `1` denotes (hateful or harmful) . . . 358
7.2 IAA statistics for data annotation discussed in §7.1.1 368
7.3 Themes and sub-themes based on descriptive annotator responses . . 373
7.4 Distribution of annotator judgments across implicit, explicit, and benign themes, showing the majority consensus between two raters on whether the content should be limited for youth or adult audiences. "`Not-Sure`" indicates cases where annotators disagreed or expressed uncertainty. 374
7.5 Performance comparison of unimodal, multimodal fusion, and prompt-based vision-language models on *hateful* and *harmful* datasets. . . . 375
7.6 False positive rates (FPR) and false negative rates (FNR) for different model architectures evaluated on $\mathcal{D}_{Hate}$ and $\mathcal{D}_{Harm}$ datasets. Lower values indicate better performance for both metrics. 376
7.7 Performance of mixed dataset-trained classifiers on 212 annotator-agreed Imgur memes, evaluated separately for adult-restriction and youth-restriction tasks. `0` denotes non-restricted content, `1` denotes restricted content. 380
7.8 Error distribution of MF^{L}_{CLIP} predictions by theme and sub-theme for adult-restriction classification on Imgur dataset. Predicted=1, Adult-Label=0 indicates false positives; Predicted=0, Adult-Label=1 indicates false negatives. 381
7.9 Error distribution of MF^{L}_{CLIP} predictions by theme and sub-theme for youth-restriction classification on Imgur dataset. Predicted=1, Youth-Label=0 indicates false positives; Predicted=0, Youth-Label=1 indicates false negatives. 382
7.10 Thematic error distribution for GPT-4o and Gemini-2.5-Flash across youth and adult audience contexts. Errors are categorised into benign, explicit (E1-E5), and implicit (I2-I9) content themes, with sub-theme breakdowns showing model misclassification. 386

7.11 Dimension-wise Krippendorff's α . . . 387
7.12 Agreement Distribution for Preference Analysis . . . 387
7.13 Distribution of Contextual Understanding Response Combinations: Note: System A = Gemini and System B = GPT . . . 388
7.14 Distribution of Visual Interpretation Response Combinations: Note: System A = Gemini and System B = GPT . . . 388
7.15 Distribution of Textual Comprehension Response Combinations: Note: System A = Gemini and System B = GPT . . . 389
7.16 Distribution of Reasoning for Limiting Mixed Access Response Combinations: Note: System A = Gemini and System B = GPT . . . 389
7.17 Distribution of Reasoning for Limiting Youth Access Response Combinations: Note: System A = Gemini and System B = GPT . . . 390

8.1 Datasets created through this doctoral dissertation . . . 396

A.1 *Method A - Sequential Survey Logic* Fisher's Exact Test Results for association - Feature Engagement Style v/s Witnessing Cyberbullying . . . 480
A.2 Fisher's Exact Test Results for association - Feature Engagement Style v/s Witnessing Cyberbullying as per *Method B - Lenient Threshold* and *Method C - Strict Threshold* . . . 481
A.4 *Method C - Strict Threshold Logic* Fisher's Exact Test Results for association - Feature Engagement Style v/s Witnessing Cyberbullying . . . 481
A.3 *Method B - Lenient Threshold Logic* Fisher's Exact Test Results for association - Feature Engagement Style v/s Witnessing Cyberbullying . . . 482
A.5 Mean and median values of number of messages by roles across conversations in all four datasets. Note: For the purpose of profiling the (Van Hee et al., 2018)* denotes only messages within conversations annotated with role-based labels are included, while those tagged solely as "`negative`" are excluded. . . . 488
A.6 Average Length by characters in messages by Role and Dataset . . . 490
A.7 Evaluation metrics at the document level i.e., conversation thread level (5 conversations). BLEU/chrF are shown with company and student translations as reciprocal references, and COMET-QE provides reference-free quality estimates. . . . 492
A.8 Breakdown of feature usage by respondents in percentages collected via Requirement Gathering Phase-1 . . . 501
A.9 Cyberbullying Scenarios Compilation . . . 504
A.10 Response to Prompt-1 by Claude-2 and Llama2Chat. . . . 526
A.11 Response to Prompt-2 by Claude-2 and Llama2Chat. . . . 527
A.12 Response to Prompt-3 by Claude-2 and Llama2Chat. . . . 528
A.13 Response to Prompt-4 by Claude-2 and Llama2Chat. . . . 529
A.14 Detailed Prompt Structure for LLM conversational data generation (Full prompt description available `https://github.com/youthcodesign/emnl-industry-track-submission/blob/main/supplementary_material/prompt_template.pdf`) . . . 530
A.15 Count of disagreement pairs for Task B: Harm Type Labelling . . . 534

A.16 Count of disagreement pairs for Task C: Intent Type Labelling . . . 535
A.17 Optimal Hyper-parameters for every model with every dataset 537
A.18 Optimal hyperparameter settings for each model across datasets. . . . 541

Data-driven Toolkit To Combat Cyberbullying amongst Teens

Kanishk Verma

Abstract

Cyberbullying (CB), or online bullying, remains a pressing concern, particularly among youth. Yet current automated detection or moderation systems fall short in addressing the social, linguistic, and ethical complexities of this issue. This doctoral dissertation advances the methodological and ethical foundations for CB detection by foregrounding the lived experiences of young people in the development of Artificial Intelligence (AI) systems designed to safeguard them. It addresses **six key research gaps** in the field: (i) the lack of quality youth-centered datasets in English; (ii) the absence of ethical frameworks for involving minors in CB dataset creation; (iii) insufficient human evaluation standards that account for ecological and social dimensions of CB; (iv) limited use of multi-task learning for modelling different types of aggression and user-roles; (v) the vulnerability of natural language processing (NLP) systems to obfuscated and adversarial language used by youth; and (vi) the limited investigation into how vision–language models handle age-differentiated content moderation criteria. This interdisciplinary doctoral dissertation explores these gaps by leveraging computational and social science frameworks and involving minors, young adults and experts in bullying behaviour across different phases of this research. This results in (a) co-creating ecologically valid cyberbullying conversational datasets; (b) obfuscated or perturbed language resources for CB detection; and (c) an unseen image-text based "*meme*" dataset with age-differentiated labels. This dissertation also benchmarks different classification paradigms for detecting both `hurtfulness` as a precursor to CB detection and different `user-roles` (e.g., `enabler`, `defender`, `bully`, `victim`) in CB contexts. Thereby attempting to address **who** (`user-role`) **said what** (`is it hurtful?`), within youth cyberbullying conversations. Furthermore, this dissertation also evaluates open and closed source large language model (LLM) reasoning capabilities to extract *explicit* and *implicit* features within potentially harmful image-based content (*memes*). By integrating perspectives from computational linguistics, human-computer interaction, and the social sciences, this interdisciplinary doctoral research lays the groundwork for the ethical inclusion of youth perspectives to inform the development of context-aware cyberbullying detection systems.

Content Warning: *This doctoral dissertation contains examples and illustrations of antisocial online behaviour like cyberbullying that may be distressing and could potentially re-trigger or re-traumatize individuals. Reader discretion is advised.*

Part I

Prelude

Chapter 1

Introduction

1.1 Motivation and Problem Description

Cyberbullying (CB), or online bullying, is characterised by Patchin and Hinduja (2006) as "*repeated online harm intentionally directed at an individual through digital communication technologies*". (Menesini et al., 2013; Alipan et al., 2020) highlight that *intentionality* and *repetition* of online harm do not fully represent CB and call for the need to recognise the multifaceted nature of CB. Moving away from definitional debates, CB can be viewed as an evolving phenomenon that,

(i) is distinct yet interconnected with physical (traditional) bullying (Kowalski et al., 2014; McFarland and Ployhart, 2015),

(ii) is relational (Kofoed and Staksrud, 2019; Alipan et al., 2020),

(iii) involves transitions from online-to-offline settings or vice versa (O'Higgins Norman et al., 2023),

(iv) involves multiple participants (Leung et al., 2018; Song and Oh, 2018; Van Hee et al., 2018),

Extending beyond the `harm` vs. `no-harm` and `perpetrator` vs. `target` dichotomy, CB manifests through a range of distinct *nasty* and *hurtful* online behaviours like *denigration*, *exclusion*, *flaming*, *outing*, *stalking* (Willard, 2007; Slonje et al., 2013;

Kowalski et al., 2014; Bauman, 2015). For example, *denigration* refers to harmful behaviour intended to damage someone's reputation, whereas *flaming* involves aggressive behaviour marked by hostility or insults. CB incidents involve **bystanders** who adopt distinct roles: `enabler` who escalate the conflict, `defender` who support or defend the `victim` (Leung et al., 2018; Song and Oh, 2018; Van Hee et al., 2018), `conciliator` who attempts to resolve the situation (Ollagnier et al., 2022), or `silent bystanders` who remain passive observers (Song and Oh, 2018).

In 2023, Beresford et al. (2023) found that 40% of 930 youths aged 12 to 16 in Ireland experienced cyberbullying[1]. Notable gender disparities emerged: 43% girls and 30% boys reported being targeted, while non-binary individuals faced significantly higher rates (74%). Recently, CyberSafeKids (2024) reported that one in four children aged 8–12 (N = 5,874)[2] had experienced cyberbullying, and 46% of girls aged 12–14 (N = 692 girls)[3] reported similar incidents in Ireland. In addition to self-reported data, since the enforcement of European Union (EU) Digital Services Act (DSA), Article 10[4]. `https://www.eu-digital-services-act.com/Digital_Services_Act_Article_10.html` directives on Instagram (Meta Platforms, Inc., 2025) reveal that government requests concerning child online safety and online bullying account for a combined total of 24.08 % (with 14 % related to child safety and 9.25 % to online bullying). The report also sheds light that 8.75 % of both proactive and reactive moderation actions for online bullying, were appealed by users. These patterns draw attention to the platforms where young people most actively interact, and where moderation practices directly shape their online experiences.

Instagram, TikTok and other gaming platforms such as, Minecraft[5] and Roblox[6], have been identified by Ireland's National Advisory Council for Online Safety (2021)

[1] Beresford et al. (2023) conducted this anonymous survey across 13 secondary schools in Ireland
[2] CyberSafeKids (2024) conducted this anonymous survey across 84 primary schools in Ireland
[3] CyberSafeKids (2024) conducted this anonymous survey across 17 secondary schools in Ireland
[4] This article allows EU Member States to request very large online platforms (VLOPs) like Meta platforms, YouTube, TikTok, etc., to provide users' information and access to actions taken by VLOPs on illegal content, including cases such as cyberbullying or online harassment, which is a step towards law enforcement action
[5] `https://www.minecraft.net/en-us`
[6] `https://www.roblox.com/`

report as platforms commonly used by youth as they facilitate peer-to-peer communication via different modes (voice-chat, text-messages, emojis, etc). Such platforms rely on a combination of user reporting (*reactive moderation*) and Artificial Intelligence (AI) driven automated flagging (*proactive moderation*) to detect and manage harmful content (Milosevic et al., 2023). As these moderation strategies are increasingly algorithm-driven, questions arise about the transparency and reliability of such systems. Taken together, the prevalence rates of cyberbullying and the moderation statistics reported under DSA's Article 10 directives highlight the urgency of evaluating and scrutinising how these systems operate. Gorwa et al. (2020) highlight the need for studies evaluating how platforms report governmental requests and manage algorithmic moderation and content biases. While the DSA's Article 9[7] and 10 directives now mandates transparency reports, Urman and Makhortykh (2023) caution that reports from very large online platforms (VLOPs) and social media platforms (SMPs) often overstate moderation effectiveness, potentially "*white-washing*" actual performance. Especially they offer **little or no detail** on how proactive systems address child online safety concerns, such as cyberbullying.

Automated cyberbullying detection: Cyberbullying detection has traditionally relied on deep learning models such as RNNs, LSTMs, and GRUs (Agrawal and Awekar, 2018; Menini et al., 2019; Aggarwal et al., 2020; Talpur and O'Sullivan, 2020). However, these models struggle to capture long-range dependencies in text (Jozefowicz et al., 2015; Le and Zuidema, 2016; Chowdhury and Zamparelli, 2018). Transformer-based architectures built on Vaswani et al. (2017)'s self-attention mechanism, including BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), address this limitation by learning richer contextual representations and have achieved strong results in cyberbullying detection (Mali et al., 2025; Philipo et al., 2024; Paul and Saha, 2022). However, a closer examination of such studies reveal that they conflate cyberbullying with other forms of online harm by training models on datasets labelled for `hate-speech`, `offensive language` or `personal attack` by (Waseem

[7] `https://www.eu-digital-services-act.com/Digital_Services_Act_Article_9.html`

and Hovy, 2016; Wulczyn et al., 2017; Davidson et al., 2017; Founta et al., 2018; Chatzakou et al., 2019). This overlap blurs conceptual boundaries and complicates accurate classification. Moreover, language that appears "*nasty*" or "*hurtful*" in one context may be benign in another (e.g., banter), underscoring the importance of context and intent in interpreting online communication (Buglass et al., 2021). Misclassification often arises when models fail to account for this nuance.

C.1: ***Paucity of youth-centered cyberbullying datasets in the English language.***

Role	**Mean Role Labels per Thread**
`Bully`	2.18
`Enabler`	1.14
`Defender`	1.47
`Victim`	1.63
Negative	15.62

Table 1.1: Mean role labels per question-answering thread in the English dataset by Van Hee et al. (2018). Note: Van Hee et al. (2018) refers to `Enabler` as *Bystander-assistants* and `Defender` as *Bystander-defenders*.

Understanding cyberbullying from youth perspectives: Despite the paucity of minor-reflective datasets in natural language processing (NLP) domain, parallel research in adjacent computational domains like Human Computer Interaction (HCI) have used qualitative and quantitive methods to advance the understanding of youth online harms, particularly cyberbullying (Ashktorab and Vitak, 2016; Badillo-Urquiola et al., 2019; Fiani et al., 2024; Huang et al., 2020; Paul Benjamin Lowry and Chatterjee, 2017; Zou et al., 2024; Zhou and Farzan, 2021). These studies often use narratives or vignettes[8] to explore online bullying behaviour through surveys or participatory co-design with minors. Notably, they often study cyberbullying and other relational aggressions to create mitigation or detection strategies based on vignettes. Such vignettes are valuable not only for intervention strategies but also

[8] Vignettes are text-based descriptions of online bullying scenarios.

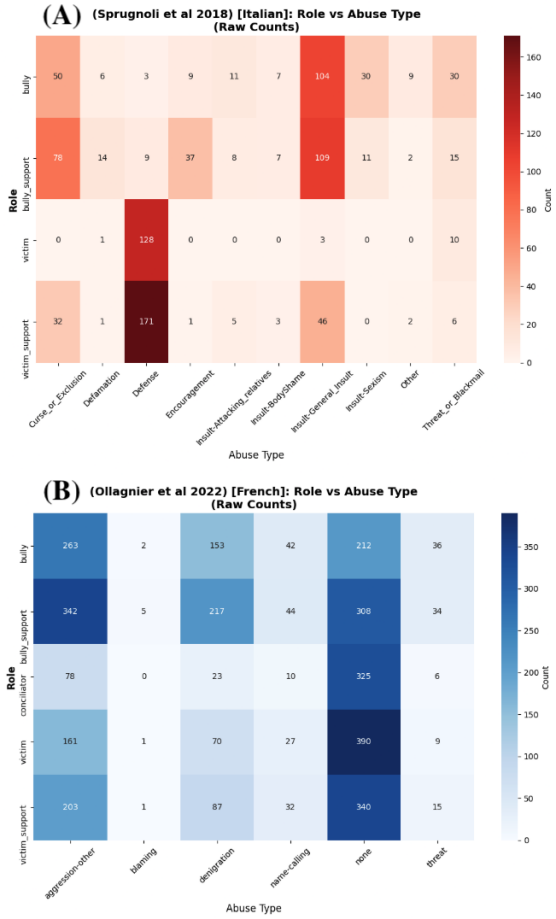
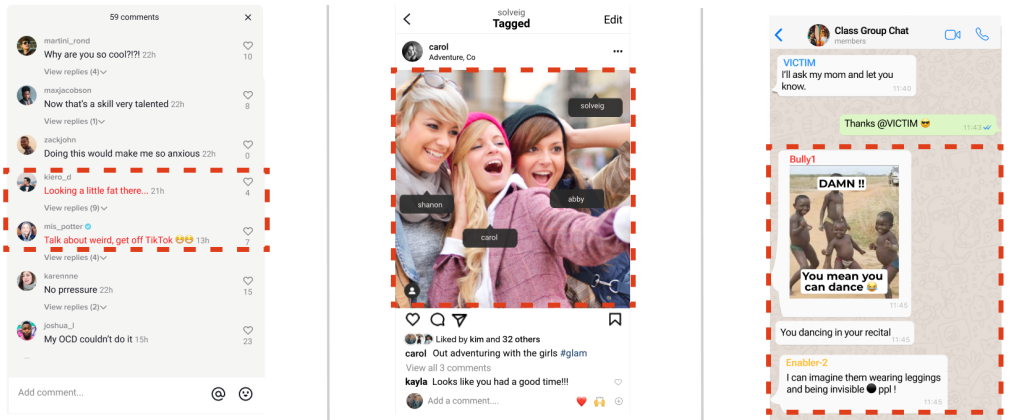


Figure 1.1: Heat-map raw counts of verbal abuse type (x-axis) and role of participant (y-axis) for both (A) (Sprugnoli et al., 2018) and (B) (Ollagnier et al., 2022)

as ethical and context-rich resources for creating youth-informed datasets. However, a challenge for researchers and content moderation systems alike is the pace of advancements in social media platforms (SMPs) and internet communication technologies (ICTs) (Singh et al., 2017; Kim et al., 2021a). In particular, the increase in diverse multimedia user-generated content (UGC) sharing through multiple SMP features tests the limitations of established detection and moderation systems, and quickly renders datasets outdated (Singh et al., 2017; Milosevic et al., 2022). These features allow UGC to be shared either permanently or for a specific duration of time or allow content to disappear once viewed by the recipient (Singh et al., 2017; Milosevic et al., 2023). Figure 1.2 demonstrates the various modalities through which different types of "online harms" associated with cyberbullying occur. Figure 1.2 (A) shows the malicious intent conveyed by a combination of text and emojis.

Similarly, Figure 1.2 (B) shows intentional social exclusion where users tag others in unoccupied spaces within images, conveying exclusion from activities. Additionally, Figure 1.2 (C) depicts a group chat where users share multi-modal content (such as images with text) to manifest hostility toward specific individuals. These examples underscore how cyberbullying today often unfolds through complex, multimodal, and context-dependent interactions that go beyond what current NLP classification systems detect. Consequently, researchers (Singh et al., 2017; Kim et al., 2021a; Milosevic et al., 2022) emphasise the importance of incorporating youth perspectives in the design and development of moderation systems, highlighting a critical challenge:

C.2: ***Limited reflection of adolescents' social media interactions and engagement patterns in youth-centered cyberbullying mitigation and detection approaches.***



(a) Mean comments with emojis

(b) Social Exclusion

(c) Mean multimodal group-chats

Figure 1.2: Different multi-modal cyberbullying scenarios on different social media platforms. Please note, all figures are created using the Figma design tool. Figures (A) & (B) are adapted with permission from Milosevic et al. (2023)

Acknowledging challenges in text-based cyberbullying detection: The frequent conflation of hate speech, offensive language, and cyberbullying in computational research, such as (Mali et al., 2025; Aggarwal et al., 2020; Menini et al.,

2019; Paul and Saha, 2022), have obscured the decision boundaries for many state-of-the-art classification techniques to classify text accurately as being "*cyberbullying*" related. For example, the sentence "***`nobody asked for your opinion, get a life`***", contains no profanities, slurs, or overt aggression, yet it constitutes cyberbullying by undermining or isolating the target. Systems trained on data that conflates these categories are therefore prone to misclassification (Salawu et al., 2017; Rosa et al., 2019; Ziems et al., 2020; Emmery et al., 2021). This conflation also neglects core characteristics that distinguish CB, such as, social dynamics, relational aggression, and repetitiveness (Salmivalli, 2014; Englander et al., 2017; Steer et al., 2020; Ziems et al., 2020). Although work by (Rathnayake et al., 2020; Jacobs et al., 2022; Ollagnier et al., 2023) shows promise by modelling speaker roles (e.g., `bully`, `victim`, `enabler`, `defender`), such approaches still struggle to capture contextual and role-based distinctions necessary for identifying **who** (*user-role*) said **what** (*harm type*). Transfer learning with pre-trained language models (PLMs) has enabled effective adaptation for sequence classification (Ruder et al., 2019), and prompting large language models (LLMs) has advanced text classification tasks (Brown et al., 2020; Radford et al., 2019). In CB research, approaches range from feature engineering with traditional classifiers (Bretschneider and Peters, 2016), to static embeddings with neural models (Agrawal and Awekar, 2018; Menini et al., 2019), to fine-tuning PLMs (Aggarwal et al., 2020; Paul and Saha, 2022). However, these techniques have largely been evaluated in isolation on data from a single social media platform (e.g., Twitter (Xu et al., 2012) or ASK.fm (Sprugnoli et al., 2018)) and remain under-explored in terms of their cross-platform generalisability. Moreover, they have yet to be assessed for their ability to capture the speaker's role (*who*) and the type of harm expressed (*what*). This presents a key challenge,

C.3: ***Limited benchmarking of classification approaches (e.g., traditional machine learning, pre-trained language models, role-aware models) for (i) detecting whether the content is `hurtful` and (ii) identifying the `user roles` in cyberbullying contexts.***

Acknowledging deliberate noise in social media text: In real-world scenarios, social media text is often noisy and unstructured, making it particularly difficult to identify offensive or hateful content (Kirk et al., 2021; Aggarwal and Zesch, 2022). This intentional or unintentional noise is frequently referred to as "*obfuscation*," where the communicator deliberately obscures the intended meaning to confuse the reader (Brunton and Nissenbaum, 2015). A related concept is "perturbation", which involves introducing deliberate changes or disturbances to alter the original form or content of a message (Qian et al., 2022). Both techniques aim to complicate the interpretation and understanding of the message, making it harder to discern the true intent behind the communication. Perpetrators often utilise techniques of obfuscation and perturbation to conceal their harmful intentions, further complicating the task of identifying abusive or offensive text (Aggarwal and Zesch, 2022). It is worth noting that these tactics represent evasive strategies developed in response to automated content moderation. These obfuscation techniques involve various methods including but not limited to, adding positive words to otherwise offensive or hateful text (e.g., "*i h8 u*" → "*i h8 u greatman*") (Gröndahl et al., 2018), masking vowels or specific characters in offensive words using a combination of punctuation marks or special characters (e.g., "*loser*" → "*l0$3R*") Aggarwal and Zesch (2022), as well as substituting nouns, verbs and other parts-of-speech words with emojis (e.g., "*punch him*" → "👊 🏿") (Kirk et al., 2021). These techniques can alter the meaning and disguise hateful content, making it difficult to identify and address instances of CB. Despite advancements in language modelling through machine and deep learning techniques, content moderation systems struggle to accurately identify harmful content when it is intentionally obfuscated or perturbed (Kirk et al., 2021; Aggarwal and Zesch, 2022; Nogara et al., 2025). This reveals another challenge:

C.4: ***Content altered through obfuscation or perturbation frequently bypasses moderation and limits the robustness of detection systems.***

Acknowledging the visibility challenge in hurtful multimodal content: The rapid evolution of user-generated content (UGC) on very large online platforms (VLOPs) such as Twitter, Instagram, and TikTok, has increased the prevalence of image-based or mixed-media communication, including memes, emojis, and text-over-image (Shifman, 2013; Du et al., 2020). Leerssen (2023) notes that content moderation now extends beyond removing content, encompassing interventions such as demotion, recommendation restrictions, age limits, or sanctions on users. Kaushal et al. (2024) analysed the EU's Digital Services Act transparency data[9] focusing on Statements of Reasons filed in November 2023 by VLOPs. They found that 93.4% of moderation actions targeted content visibility rather than removal, signalling that platforms increasingly regulate how content circulates rather than whether it remains online. In cyberbullying (CB) contexts, datasets from Instagram and Vine by (Hosseinmardi et al., 2015; Rafiq et al., 2015) treat CB as a post-level binary label, collapsing captions, comments, and visuals into one decision, which risks obscuring relational and contextual cues of hurtful intent. Similarly, the English–Hindi multimodal CB dataset (Maity et al., 2022) annotates emotions and sarcasm but still frames CB as a binary label. Recent work leverages LLM-guided abductive reasoning[10] (Lin et al., 2023) and dialectical debate[11] pipelines (Lin et al., 2024) to infer implicit premises and generate conflicting rationales from harmless and harmful perspectives. These methods, combining textual and visual features via cross-attention, outperform some prior systems on *Hateful* (Kiela et al., 2020) and *Harmful* (Pramanick et al., 2021a) meme detection tasks. Yet, challenges remain: models struggle to distinguish explicit cues (slurs, violent imagery) from implicit cues (stereotypes, contextual innuendo), and their predictions are not fully aligned with platform-relevant outcomes like nuanced visibility restrictions. This reveals another challenge:

[9] `https://transparency.dsa.ec.europa.eu/?lang=en`

[10] Reasoning to generate an initial plausible hypothesis (e.g., observing a wet lawn and hypothesising it rained).

[11] A method that involves presentation of opposing viewpoints (thesis and antithesis) and the use of rational exchange of arguments to arrive at a higher truth or a synthesis perspectives (Basseches, 1984)

C.5: ***Limited understanding of explicit (e.g., slurs, violent imagery) and implicit (e.g., cultural, contextual, or social cues) meanings in multimodal content which constrain the ability to determine age-appropriate visibility restrictions on `hurtful` multimodal content.***

1.2 Research Questions

Addressing the **five** challenges discussed in §1.1 and to bridge the disciplinary divide between natural language processing (NLP), human-computer interaction (HCI), and the social sciences, this inter-disciplinary doctoral study explores **three** key research questions. Each question is intended to advance the development of computational systems capable of meaningfully capturing the complex, dynamic, and often obfuscated nature of youth expression in cyberbullying contexts.

1.2.1 Research Question 1. (RQ.1)

RQ.1 How can we engineer datasets and test cases for building real-world cyberbullying detection systems in a safe and ethical manner?

This research question seeks to bridge the gaps identified in **C.1** and **C.2** by exploring **qualitative** (e.g., co-design), **quantitative** (e.g., surveys) methods to understand youth perspectives and **computational** methods (e.g., data augmentation) for engineering multimodal datasets and test-cases to support CB detection while adhering to safe and ethical practices. Central to this inquiry is the understanding of,

> *How adolescent social media behaviour, including engagement with platform-specific characteristics, bystander experiences, and reporting behaviours, can inform AI-driven cyberbullying detection?*

This research inquiry helps integrate youth experiences in cyberbullying contexts to design and engineer ecologically valid datasets that can help explore ***who*** (*user roles*), ***said what*** (*harm type*), and ***why*** (*intent*).

1.2.2 Research Question 2. (RQ.2)

RQ2.1 Which classification paradigms that leverage state-of-the-art pre-trained language models (PLM) are best suited for detecting cyberbullying in social media content created by youth?

RQ2.2 How can classification techniques be adapted to detect `hurtfulness` in cyberbullying contexts within obfuscated or perturbed textual social media content?

Addressing **C.3**, the research question **RQ2.1** helps examine within cyberbullying contexts the ability of different language models and classifiers to (a) detect `hurtfulness` as a binary task for understanding "*nasty or hurtful*" behaviour[12] as a precursor to detect CB and (b) identify one of the multiple `user-roles` in cyberbullying contexts. To address **C.4**, the research question **RQ2.2** helps *(a)* explore different strategies to improve the resilience of deep learning models so that they are more robust to deceptive techniques, and *(b)* enhance test-cases with perturbed and obfuscated samples to cover multiple types of online harms associated with cyberbullying. The findings of this research will contribute to creating safer online spaces by enabling accurate detection and mitigation of cyberbullying incidents.

1.2.3 Research Question 3. (RQ.3)

RQ.3 To what extent can vision-language models trained on `harmful` and `hateful` meme datasets generalise to age-differentiated content moderation decisions across varying levels of semantic complexity ranging from *explicit* (e.g., slurs, violent imagery) to *implicit* cues (e.g., cultural or contextual cues), to *benign* content (e.g., scenery, neutral humour)?

[12] This terminology draws on Ireland's (National Advisory Council for Online Safety, 2021) survey findings, which suggest that the phrasing phrasing 'nasty or hurtful' behaviour helps reduce confusion among young people regarding the multifaceted nature of the cyberbullying phenomenon. (See §2.3.1)

Addressing **C.5**, this research question evaluates whether vision-language models trained on binary hateful and harmful meme classification can serve as effective baselines for age-differentiated content moderation. This content moderation task requires nuanced judgments about audience appropriateness rather than absolute toxicity detection. Using an unseen test set of memes annotated with age-differentiated visibility judgments (mixed-age platforms vs. youth-heavy platforms), we benchmark performance across multiple architectural paradigms: unimodal baselines (image-only, text-only), multimodal fusion approaches (CLIP late fusion, ViLT early fusion), and large vision-language models (LLaVA-CoT, Qwen-VL, GPT-4o, Gemini-2.5). Through error analysis, we examine how model performance varies across a spectrum of semantic complexity ranging from *explicit* harmful cues that are visually or linguistically overt (e.g., slurs, violent imagery) to *implicit* meanings that require cultural or contextual interpretation (e.g., sarcasm, coded references) to *benign* content that poses no harm (e.g., scenery, neutral humour). Additionally, by generating a pilot human preference dataset from cases where closed-source models disagree, we identify boundary conditions of multimodal harm interpretation and establish foundations for future alignment research that can better capture the implicit semantics critical to responsible content moderation.

1.3 Contributions

This doctoral dissertation makes the following contributions in relation to specific challenges: Challenges **C.1** and **C.2** are addressed through **RQ.1**, leading to the following contributions:

A: Ethical frameworks through which youth participation in Trust & Safety research[13] can help inform and curate child-centered language datasets.

B: Machine generated yet realistic cyberbullying datasets curated for solely to

[13] Trust & safety research refers to the interdisciplinary study of how individuals misuse digital platforms to cause real-world harm, often through intended platform functionalities (Stanford Internet Observatory, 20)

serve as training data for classifiers ($\mathcal{D}_{\textbf{Silver}}$) and quality youth co-created dataset ($\mathcal{D}_{\textbf{Gold}}$) representative of online bullying characteristics in English language.

C: Human evaluation frameworks through which ecological validity can be assessed for cyberbullying datasets across key dimensions of cyberbullying such as, *power-dynamics conflict*, *support*, *seriousness*, *sarcasm*.

D: A three-level annotation framework capturing conversational context to identify (a) whether aggression is present and if it is `overt` or `covert`, (b) the type of online harm (e.g., `blackmail`, `insults`, `exclusion`, `encouragement of harassment`), and (c) message intent, including `attack`, `defense`, `abetment`, or `instigation`.

Challenge **C.3** is addressed through **RQ2.1**, and helps make the following contribution:

E: Benchmarking of natural language processing (NLP) classification paradigms to detect cyberbullying in a multi-task setting i.e., specifically identifying the presence of aggression, and its target.

Challenge **C.4** is addressed through **RQ2.2**, and helps make the following contributions:

F: Robust NLP classification systems capable of handling obfuscated or perturbed language commonly used by youth (e.g., "*die*" $\rightarrow$ "*unalived*")

Challenge **C.5** is addressed through **RQ3**, and helps make the following contribution:

G: A multimodal dataset annotated for content moderation, indicating whether content is suitable for adults or youth.

H: Benchmarking of vision-language classification paradigms to detect content that is hateful, harmful, or inappropriate for adults or young audiences, considering both implicit and explicit meanings.

1.4 Thesis Structure

This thesis is structured in **four parts**. ***Part I Prelude*** is already presented and discussed in Chapter 1.

Part II Foundations: Chapter 2 reviews relevant background and related work across computational linguistics, human–computer interaction (HCI), and the social sciences, with a particular emphasis on cyberbullying (CB) in youth contexts. It situates the thesis within existing research by identifying key knowledge gaps and establishing the interdisciplinary foundations for the research presented in the following chapters.

Part III Contributions: The multifaceted contributions of this thesis are presented in six key stages as depicted in Figure 1.3. Stage 1 (Chapters 3 and 4)a ddresses research question **RQ1** which explores young people's lived experiences to build a dataset reflecting real-world cyberbullying (CB) dynamics. Chapter 3 presents a youth participatory rights-based framework to examine how adolescents' social media practices shape engagement and witnessing behaviours. It positions youth as co-designers of a conversational dataset and outlines the study's ethical safeguards. Chapter 4 analyses the content and reports on the forms and spread of CB in adolescents' online interactions. It discusses **three** key outputs: (a) *ecologically valid vignettes of serious cyberbullying*, (b) *a youth co-created conversational dataset reflecting authentic online interactions*, and (c)*patterns of obfuscated offensive messages*. These contributions offer ethically grounded and contextually relevant resources for cyberbullying research.

Addressing research question **RQ1**, Stage 2 (See Figure 1.3) discussed in chapter 5 explores data augmentation strategies, including human translation (HT), machine translation (MT), human-aided MT (HMT), and LLM-assisted generation to aid synthetic dataset generation for CB detection. This chapter also presents two participatory protocols: (a) human annotation and (b) expert/youth evaluation, to create a (i) benchmarking dataset from youth co-created data and (ii) fine-tuning dataset from translated and synthetic content.

Stages 3 and 4 in Figure 1.3) address **RQ2.1** in 6 by re-framing cyberbullying detection as a multi-task problem, modelling *user roles* and *hurtfulness* as precursors to CB classification. It builds towards exploring the robustness of classification strategies for obfuscated content using a youth-developed test set of *hurtful* variations. This helps evaluate adversarial training and augmentation strategies and benchmark them against closed-source proprietary moderation systems.

Finally, Stage 5, depicted in Figure 1.3 addresses **RQ3** in Chapter 7 by benchmarking vision–language models for harmful meme detection under visibility-based moderation. It evaluates their ability to capture explicit and implicit *harmful* semantics and assesses whether LLM-guided reasoning improves detection and moderation alignment.

Part IV Conclusion: Chapter 8 synthesises findings across all research questions, discussing their implications for youth-centred AI design and ethical data practices. It also outlines the limitations of the work and provides recommendations for future research and policy focused on cyberbullying-related content moderation.

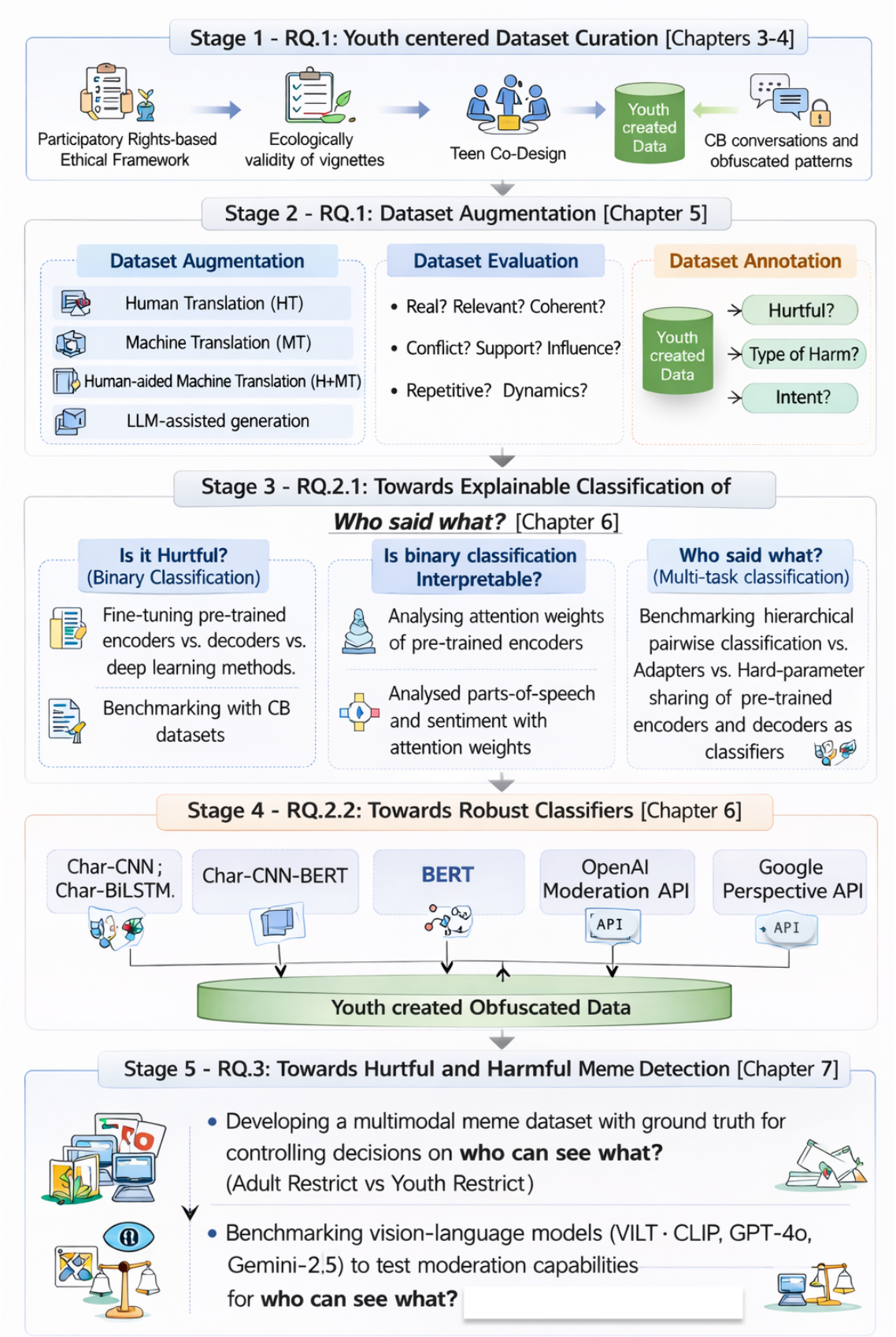


Figure 1.3: Structural overview of research contributions in Part III

1.5 Publications

1.5.1 Computational Science Publications

This includes 1 journal article, 2 main conference articles, and 3 specialised workshop articles as first author (See Table 1.2).

Title	Venue	Note
BullyBench: Youth & Experts-in-the-loop Framework for Intrinsic and Extrinsic Cyberbullying NLP Benchmarking	EMNLP Industry Track 2025	Represents the publishable output of **RQ1**, **RQ2.1** are discussed in Chapter 5 and 6.
Benchmarking Language Models for Cyberbullying Identification and Classification from Social-media Texts	LateRAISSE at LREC 2022	
Can Attention-based Transformers Explain or Interpret Cyberbullying Detection?	TRAC at COLING 2022	
DCU at SemEval-2023 Task 10: A Comparative Analysis of Encoder-only and Decoder-only Language Models with Insights into Interpretability	SemEval at ACL 2023	
Beyond Binary: Towards Embracing Complexities in Cyberbullying Detection and Intervention - a Position Paper	Joint Conference of LREC-COLING 2024	Represents the publishable output of **RQ.1** addressing Challenge **C.1**. Discussed in Chapter 5.
Leveraging Machine Translation for Cross-Lingual Fine-Grained Cyberbullying Classification amongst Pre-Adolescents.	Journal of Natural Language Engineering 2023	

Table 1.2: Computational Science Publications

1.5.2 Social Science Publications

Includes 1 journal article as second author and 2 conference articles as first author (See Table 1.3).

1.5.3 Human-computer Interaction Publications

Includes 1 conference article as first author (See Table 1.4)

Title	**Venue**	**Note**
Effectiveness of artificial intelligence–based cyberbullying interventions from youth perspective (Co-author).	Social Media+ Society	Represents the publishable output of **RQ.1** addressing Challenge **C.2**. Informs the research methodology discussed in Chapter 2 and 3.
Designing Ethical Artificial Intelligence (AI) Systems With Meaningful Youth Participation: Implications and Considerations	AoIR Selected Papers of Internet Research 2023	
Examining The Effectiveness of Artificial Intelligence-based Cyberbullying Moderation on Online Platforms: Transparency Implications	AoIR Selected Papers of Internet Research 2022	

Table 1.3: Social Science Publications

Title	**Venue**	**Note**
From Users to Co-Designers: Youth Participation in Understanding Cyberbullying	IDC 2025	This work received honourable mention at the Interaction Design Children Conference 2025. This work is discussed in Chapter 3 and 4

Table 1.4: Human Computer Interaction Publications

Part II

Foundations

Chapter 2

Background and Related Work

Abstract

To understand the background and related work on cyberbullying (CB) research from an interdisciplinary perspective, the present chapter is divided into **three fundamental strands**. Each strand offers insight into how CB is approached by specific disciplines: (i) social sciences, (ii) human–computer interaction, and (iii) computer science. Strand 1 (§2.2) highlights social scientific theories that frame CB as a social and criminogenic phenomenon, emphasising its motivations, consequences, and distinct contexts compared to offline bullying. Strand 2 (§2.3) examines empirical research with youth and adults, exploring lived experiences and their roles in annotating and evaluating CB datasets. Strand 3 (§2.4) focuses on computational detection methods, such as natural language processing, machine learning, and deep learning, that underpin automated identification and platform-led proactive moderation. Social science offers strong theoretical insight but limited scalability; computational approaches provide technical solutions yet risk overlooking socio-cultural nuance and conflating CB with other harms; and human–computer interaction seeks to bridge these divides but remains underdeveloped.

2.1 Literature Review Methodology

To ensure coherence and transparency in the review presented in this chapter, the following describes the approach adopted for identifying, selecting, and analysing the literature across the three strands: (i) social scientific foundations (§2.2), (ii) human–computer interaction and participatory research (§2.3), and (iii) computational detection methods (§2.4).

Search Strategy and Sources: Literature was identified through searches across academic databases: Google Scholar[14], Scopus[15], JSTOR[16], IEEE Xplore[17], and ACL Anthology[18], and supplementary searches were conducted on Semantic Scholar[19] and DBLP [20]. Given the interdisciplinary scope of the thesis, search terms were tailored to each strand. For the social scientific strand, searches combined terms such as "*cyberbullying, online bullying, digital harassment, power imbalance, routine activities theory, social balance theory, and cybercrime theory*". For the HCI and participatory strand, terms included "*participatory design, youth perspectives, annotation, inter-rater reliability, cyberbullying dataset, and evaluation*". For the computational strand, searches used combinations of "*cyberbullying detection, online abuse detection, benchmarking abuse detection systems, multi-modal, NLP, and deep learning*". Searches for the social scientific strand were not restricted to a fixed date range, as seminal foundational works, such as (Cohen and Felson, 1979) Routine Activities Theory or (Olweus, 1994) bullying criteria, remain theoretically central regardless of publication date. However, for computational methods, particular emphasis was placed on literature published from 2017 onwards, reflecting the shift toward transformer-based architectures following Vaswani et al. (2017). Where relevant, forward citation searches were conducted via Google Scholar to identify influential follow-on works that cited key papers. Reference lists of systematic lit-

[14] https://scholar.google.com/
[15] https://www.scopus.com/pages/home?display=basic#basic
[16] https://www.jstor.org/
[17] https://ieeexplore.ieee.org/Xplore/home.jsp
[18] https://aclanthology.org/
[19] https://www.semanticscholar.org/
[20] https://dblp.org/

erature reviews in the field, including (Rosa et al., 2019; Emmery et al., 2021; Kim et al., 2021b; Teng et al., 2024) among others, were also consulted to identify additional relevant studies and to ensure coverage of well-established benchmarks and datasets.

Overall inclusion and exclusion criteria: Studies were included if they met one or more of the following criteria: (i) they directly addressed cyberbullying or closely related online harms (hate speech, online aggression, harassment) in ways immediately relevant to the thesis's research questions; (ii) they introduced datasets, annotation frameworks, or detection architectures that have become reference points in the field; (iii) they contributed theoretical or methodological frameworks applicable to understanding or detecting cyberbullying in digital environments; or (iv) they examined youth perspectives, participation, or lived experience in relation to online harm. Studies were excluded if they addressed offline bullying exclusively without extending to the digital context, or if they were superseded by substantially updated versions that rendered the original contribution obsolete.

Strand-wise literature inclusion criteria: The social scientific literature reviewed in §2.2 were included as they explicitly compare cyberbullying and traditional/physical bullying across recurring definitional dimensions: intentionality, repetition, power imbalance, anonymity, accessibility, and publicity. These criteria were retained because they recur across independent systematic comparisons, supporting their cross-study validity (Olweus, 2013; McFarland and Ployhart, 2015; Thomas et al., 2015; Kowalski, 2018; Chan et al., 2019). Theoretical frameworks discussed in §2.2 were selected based on their explanatory value for observable behaviour in digital environments and their direct relevance to building tools for fair automated moderation. Because the thesis aims to design and evaluate computational systems for cyberbullying detection, priority was given to theories that model situational conditions, opportunity structures, and platform affordances rather than internal belief systems. Literature on youth and adult participation discussed in §2.3 was selected where studies directly involved children or adolescents in cyberbullying

research or where they introduced statistical measures used in annotation, evaluation and survey analysis. Participatory design studies were included when they engaged youth or experts in structured activities and reported sufficient methodological detail. Annotation and evaluation frameworks were reviewed to trace the evolution of cyberbullying operationalisation for natural language processing. Statistical measures are discussed only where they were directly employed in the reviewed studies, ensuring methodological exposition remained grounded in empirical practice. Computational studies discussed in §2.4 were organised to reflect substantive developments since earlier reviews (Salawu et al., 2017; Rosa et al., 2019; Emmery et al., 2021; Wang et al., 2024c). Inclusion criteria distinguished between binary, fine-grained, multi-task, and multimodal detection, as well as emerging work on synthetic data generation and robustness to textual obfuscation. Finally, studies were selected where they trained and evaluated models on cyberbullying or closely related harms using recognised evaluation frameworks, or where they addressed data scarcity and classifier vulnerability.

2.2 Contextual and Theoretical Landscape of Social Scientific Explorations

Cyberbullying (CB) mirrors physical bullying but amplifies its effects through digital means, involving repeated harm where the perpetrator holds power over the target (Tokunaga, 2010). This power imbalance may stem from social status, physical dominance, or group dynamics in physical bullying, and from technological advantage, online visibility, or social capital in CB (Hinduja and Patchin, 2013). Recently, (O'Higgins Norman et al., 2023) note the overlap between offline and online interactions in the cyberbullying phenomenon and highlights offline/physical bullying transitions to online spaces and vice-versa.

Despite these shared foundations, CB differs from physical bullying in significant ways. While repetition in physical bullying is enacted through recurring behaviours

across multiple encounters, repetition in CB is embedded in the persistence of digital content. As depicted in Figure 1.2 (Chapter 1), a single harmful post, message, or image can be disseminated widely, accessed repeatedly, and remain visible indefinitely, thus amplifying its psychological impact (Slonje et al., 2013). Smith and Steffgen (2013) argue that since the fundamental criteria of traditional bullying also apply to CB, power imbalances should be assessed in terms of technological expertise, network visibility, and digital platform affordances. They also emphasise that factors such as friendship in social networks, perceived popularity, social hierarchies, and marginalised group identities are crucial for understanding the roles of both perpetrators and victims across contexts.

To understand *cyberbullying* (CB) as a phenomenon, we first examine both the continuities and differences between physical bullying and CB. We then introduce the theoretical foundations, outlining key concepts and frameworks that help better understand **how cyberbullying manifests among adolescents**.

2.2.1 Contrasting cyberbullying and physical bullying

In the last decade, (Olweus, 2013; McFarland and Ployhart, 2015; Thomas et al., 2015; Kowalski, 2018; Chan et al., 2019), among others, have undertaken extensive research to compare and contrast cyberbullying with physical or traditional bullying. Across these studies, **six** fundamental criteria have been consistently identified as central to understanding both the perpetration of CB and the severity of its detrimental effects. These include,

1. **Intentionality**: Smith and Steffgen (2013) defined the concept of intentionality within the context of cyberbullying (CB) as purposeful intent to inflict harm. In traditional bullying, the intention to cause harm is apparent in tangible actions like physical aggression. In CB the intent to harm can be manifested through the creating, sending, or sharing of CB-related media by the perpetrator (Menesini et al., 2013). This intentionality is both direct and explicit in nature. It is important to remember that understanding inten-

tionality is not just about what the person doing the harm wants ("*wants to hurt*"), but also about how the person being hurt feels ("*felt hurt*") (Menesini et al., 2013; Alipan et al., 2020).

2. **Repetition**: In physical or traditional bullying, repetition often involves physically harming a person multiple times (Menesini et al., 2013; Slonje et al., 2013). In the context of CB, social media platforms make repetition simple, as users can view, respond to, or re-share the bullying content (Kazerooni et al., 2018; Alipan et al., 2020; Zito, 2022).

3. **Power imbalance**: In traditional or physical bullying, power imbalances usually manifest as a more influential individual directing their influence towards someone with diminished power Vaillancourt et al. (2008). In CB context, Smith and Steffgen (2013) states this phenomenon can be understood from two perspectives: a) *the dominating power of the bully*, and b) *lack of power held by the victim*. While this demonstrates the similarity, power imbalance in CB is also unique, for example, those with social capital can be trolled or face repetitive bullying (Barlett et al., 2022; Valenzuela-García et al., 2023). Additionally, on social media platforms the prevalence of power imbalance is enhanced further by a) the perpetrator's expertise with technological features intrinsic to the platform Englander et al. (2017), and b) the features offered by platforms for users to remain anonymous on Slonje et al. (2013). For example, anonymous profiles and Virtual Private Networks (VPNs) that hide identity and location can embolden users on digital platforms (Chan et al., 2021). The interconnectedness of online platforms fuels the spread of harmful content, leaving victims feeling powerless (Kane et al., 2014; Dredge et al., 2014).

4. **Anonymity**: Unlike traditional bullying, the identity of those perpetrating CB is comparatively easier to hide (Chan et al., 2021). Social Media Platforms (SMPs) today provide conducive environments for anonymity. SMPs

like ASK.fm, Reddit[21], 4Chan[22], Instagram[23], among others, characterised by their simplicity and image-based format, allow individuals to post comments and share images devoid of personal identifiers. Such SMPs afford perpetrators an avenue for anonymity where they can easily shroud their identity, erasing any trace of accountability.

5. **Accessibility**: Accessibility denotes the ease with which someone can be targeted. In traditional bullying, targeted individuals can safeguard themselves by seeking refuge in secure spaces (Nansel et al., 2001). However, on SMPs, CB can overcome the confines of face-to-face interactions and can be perpetrated through routine interactions on the platform (Cassidy et al., 2013; Chan et al., 2019). The limitless connectivity intrinsic to SMPs affords the perpetuation of bullying irrespective of time or location and regardless of the presence of victims (Bastiaensens et al., 2015).

6. **Publicity**: Traditional bullying often stays confined to specific settings like classrooms or workplaces (Sticca and Perren, 2013). In CB, SMPs provide perpetrators with avenues to amplify the reach of their bullying actions. SMPs enable bullies to widely share acts, upload images, and prompt users to view and comment on bullying content. Through hashtags[24], CB-related content can spread beyond involved users (Kazerooni et al., 2018; Alipan et al., 2020). Dynamics of publicising CB incidents include SMP features like "stories", where content disappears after a set time but gains visibility. Direct messaging, including disappearing messages, serves as a channel for publicising CB content. Sticca and Perren (2013) found online publicity makes CB slightly worse than traditional bullying.

[21] `https://www.reddit.com/`
[22] `https://4chan.org/`
[23] `https://instagram.com`
[24] Keywords with '#' to coordinate distributed discussions between users (Bruns and Burgess, 2011).

2.2.2 Theoretical Background

While a considerable body of research exists, researchers have observed that many studies exploring cyberbullying (CB) lack a strong theoretical foundation (Monks et al., 2009; Tokunaga, 2010; Ansary, 2020; O'Higgins Norman, 2020). Physical or traditional or offline bullying has been examined through well-established theoretical frameworks, such as evolutionary theories, social learning theories, and socio-cultural perspectives that explain the behaviours of both victims and perpetrators (Monks et al., 2009). Tokunaga (2010) highlights the importance of not only adopting established theories but also incorporating newly developed ones that account for emerging technologies and media. For the specific focus of this doctoral thesis on identifying CB in online spaces, we recognise the need to delve into theories that can offer insights into these digital dynamics. These include,

1. **Routine Activities Theory** (RAT): Cohen and Felson (1979) formulated RAT to help understand that criminal activities are more likely to occur when motivated offenders and vulnerable victims converge in an environment that lacks capable guardianship. This serves as a foundational criminological framework extensively employed to probe and understand deviant, delinquent, and criminal behaviours Osgood et al. (1996); Vazsonyi et al. (2002, 2018). With advancements in internet communication technologies (ICTs) and the connectivity offered by social media platforms, RAT has helped understand hurtful and anti-social online behaviours like cyber-crimes Kigerl (2012); Leukfeldt and Yar (2016), cyber-dating abuse Van Ouytsel et al. (2018), online sex crimes Navarro and Jasinski (2015), and CB among adolescents Navarro and Jasinski (2012); Arntfield (2015); Xiao et al. (2016); Kalia and Aleem (2017); Aizenkot (2022). As Kaluarachchi et al. (2020) emphasises, RAT, through its triadic interaction (offender, victim, absence of guardianship), helps understand that social media platforms can create environments that facilitate cyberbullying and highlights the need for effective monitoring and protective measures.

2. **Crime Opportunity Theory (COT)**: Felson and Clarke (1998) proposed the Crime Opportunity Theory to suggest that changes in society and technology create new "*opportunities*" for criminal behaviours and acts of deviance. Its relevance extends to various categories of criminal activities, irrespective of their inherent characteristics or severity levels. COT identifies two primary components contributing to criminal acts: the presence of "likely perpetrators", and "environmental conditions" offering criminogenic opportunities Felson and Clarke (1998). COT helped Zabyelina (2017) to gain insights into pharmaceutical scams and black markets and Quayle (2020) to understand child sexual exploitation and abuse. Lowry et al. (2016) used the lens of COT to conclude that the "anonymous environment" on social media platforms is an exploitable factor by bullies. Furthermore, COT has been instrumental in examining technology-related crimes, such as data breaches (Sen and Borle, 2015) and computer crimes (Willison and Backhouse, 2006), as well as various aspects of bullying behaviours, as explored by Chan et al. (2021). Kim et al. (2021b) explored how COT helps explain why people engage in CB and elucidates the triadic relationship between victims, perpetrators, and bystanders.

3. **Affordance Theory**: Drawing from the foundational principles of ecological psychology by (Gibson, 1966, 1977), the concept of "*affordances*" within an environment refers to the opportunities and limitations it offers individuals. This concept has been adapted across various domains such as information management systems (Leonardi, 2011; Strong et al., 2014), technology studies (Hutchby, 2001; Faraj and Azad, 2012), and the investigation of cyberbullying on social media platforms (SMPs) (Chan et al., 2019). Ronzhyn et al. (2023) conducted a thorough literature review on "*affordance*" theory across multiple discipline within the context of SMPs. They found that while this theory is commonly applied to investigate the communication opportunities facilitated by SMPs through their technological features, there is a notable conflation between "*technological*" and "*social media*" affordances. The con-

cept of "technological affordances", described by Faraj and Azad (2012), refers to the alignment between the intentions of individuals or actors and the capabilities of technology. This alignment allows for specific actions to be carried out effectively. Building on this, Ronzhyn et al. (2023) defines *"social media affordances"* as "*observable or envisioned characteristics of SMPs, resulting from technological, social, and contextual interplay, enabling or constraining platform use*". They state that it is imperative to note that certain concepts, such as 'content sharing', 'collaboration', 'evaluation and feedback', 'communication', and similar technological and platform functionalities, should not be classified as affordances but rather as specific outcomes of affordance actualisation. They go a step further and exemplify how this definition can be used to evaluate existing affordances studied in previous research, which are as follows,

- Visibility and Edit-ability: *Visibility* refers to content's ease of view, while *edit-ability* allows content modification or deletion after sharing (Treem and Leonardi, 2013).
- Accessibility and Immutability: Tied to the core concept of "persistence", both *accessibility* and *immutability* preserve content over time (Boyd, 2010; Treem and Leonardi, 2013).
- Search-ability or Information Retrieval (IR): Boyd (2010) notes this as an aspect of visibility, where SMPs enable users to retrieve or search for content.
- View-ability, Address-ability, and Reach-ability: *View-ability*, as per Mansour et al. (2013), SMPs enable sharing, viewing, and making things visible without necessarily implying the ability to edit or comment. *Address-ability*, mentioned by (Hennebert et al., 2021), refers to the possibility of SMPs to easily target specific groups or individuals. *Reach-ability*, as described by (Sharma et al., 2016), involves connecting with distant, potentially interested users, facilitated by adaptive automated algorithms.

- Associations: Studied by (Treem and Leonardi, 2013), this affordance involves two types: a) social ties, evident through connections on a social network, and b) connections between individuals and pieces of information, like contributing to a blog, wiki-page, or tagging an article.

2.3 Youth and Adult Participation in CB studies

To detect cyberbullying among young people, it is important to understand the social and developmental perspectives of the young people who are most affected by this phenomenon. Building on the principles of the United Nations Convention on the Rights of the Child (UNCRC) Article 12[25], the European Union's Joint Research Centre (JRC) report by (Charisi et al., 2022) emphasises that youth should be regarded not only as technology users but as active contributors whose insights must inform the design and governance of digital systems. Guided by this **rights-based perspective**, prior research on cyberbullying has developed along two interconnected paths. One emphasises understanding of young people's lived experiences and the dynamics of their online interactions. In contrast, the other focuses on the role of youth and adult participation in creating or evaluating content for the detection of online harms like cyberbullying, among others.

Studies in human-computer interaction and social sciences (e.g., (Ashktorab and Vitak, 2016; Badillo-Urquiola et al., 2019; Almeida et al., 2022)) have explored youth perspectives on cyberbullying and online harms through participatory design. In contrast, others (e.g., (Huang et al., 2020; Rodríguez-de Dios et al., 2018; Livingstone et al., 2015; Thorn and Benenson Strategy Group, 2021)) have used surveys and statistical analyses to examine patterns of perceived online risk (e.g., cyberbullying, sexual harassment), `seriousness` and `report-worthiness` of those risks among youth. Together, these participatory, qualitative, and quantitative methods can reveal aspects of youth experiences with cyberbullying that might otherwise be

[25] `https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child`

overlooked. In parallel, research has explored how datasets can reflect youth perspectives on cyberbullying in non-English languages (e.g., (Sprugnoli et al., 2018; Ollagnier et al., 2022)) and how human annotation practices shape what is considered harmful or abusive (e.g., (Van Hee et al., 2018; Ziems et al., 2020; Shannag et al., 2022; Kumar et al., 2022)). However, (Salawu et al., 2017; Rosa et al., 2019; Emmery et al., 2021) highlight challenges of reliability, bias, and subjectivity in categorising user-generated content (UGC) for cyberbullying. For example, Bonetti and Tonelli (2022) engaged youth as annotators using a gamified approach to re-label Sprugnoli et al. (2018)'s Italian dataset, finding that youth often have a higher threshold for labelling content as offensive than expert annotators. To manage subjectivity, Röttger et al. (2022) distinguishes between two human annotation paradigms: (a) a **descriptive paradigm**, which embraces annotator subjectivity to capture granular judgments, and (b) a **prescriptive paradigm**, which minimises subjectivity by instructing annotators to follow a fixed definition of harmful content. Röttger et al. (2022) note *prescriptive* paradigm instructs annotators to encode a single, predefined belief specified in annotation guidelines, minimising subjective variation to produce consistent labels aligned with platform policies. For example, annotators might be instructed: "Label content as `hateful` if and only if it attacks individuals based on protected characteristics (race, religion, gender)". In contrast, Röttger et al. (2022) note the *descriptive* paradigm embraces annotator subjectivity, allowing individual judgments to reflect diverse perspectives and contextual interpretations. Rather than enforcing uniform labels, descriptive annotations allow for capturing how different stakeholder groups perceive content by acknowledging that a UGC may be interpreted as satirical by adults but harmful by adolescents.

Accordingly, this section is organised into **three parts**. The first reviews studies that foreground youth perspectives and lived experiences to understand the dynamics of cyberbullying (Section 2.3.1). The second discusses the statistical methods leveraged in (i) quantitative survey studies to understand association and effect sizes and (ii) prescriptive annotation studies to understand reliability amongst annotators

(Section 2.3.2). The third discusses the broader process of recognising and annotating cyberbullying and also examines how adult and youth stakeholders participate differently in prescriptive annotation studies (Section 2.3.3).

2.3.1 Youth perspectives on cyberbullying detection

Understanding children and adolescents' perspectives is central to designing effective cyberbullying (CB) detection and prevention strategies on social media platforms (SMPs). Prior research has engaged youth in participatory design and related methodologies, revealing both the value of their contributions and the limitations of existing approaches.

Participatory design studies on cyberbullying: To understand the perspectives of children and youth on cyberbullying (CB) incidents on social media platforms (SMPs), Ashktorab and Vitak (2016) engaged 21 high school students over the course of five weeks through participatory design sessions. Conducted during their 45-minute free period in the school schedule, Ashktorab and Vitak (2016) recruited 14 students aged 14–15 and 7 students aged 17–18. Participants shared their personal experiences of CB and collaborated on potential solutions to detect and deter CB behaviour using visual aids (vignettes). Through participatory design theories like Mixing-of-Ideas (Guha et al., 2004), Co-operative Inquiry (Druin, 1999), Bags-of-Study (Yip et al., 2013), Ashktorab and Vitak (2016) studied five distinct types of CB-related online harms, such as *denigration*, *exclusion*, *outing*, *flaming*, and *harassment* across different SMPs. Despite their novel contributions, their framework primarily focused on victim perspectives in platform feature design, overlooking other variations of CB and the roles of various participants.

Work with younger children has highlighted platform-specific risks. For instance, Badillo-Urquiola et al. (2019) organised participatory design sessions with pre-adolescents aged 8-11, focusing on features of SMPs that enable strangers to re-

share content. They studied the SMP platform Musical.ly[26] (now TikTok[27]) and its feature "*duet*" that allows strangers to create collaborative videos by combining their own content with another user's video content. While this study placed little emphasis on CB, it focuses on specific CB scenarios, e.g., *re-sharing of content with mean and hurtful messages*. Through colourful vignettes and qualitative research, they used user interfaces to collect the victims' perspectives on how they would respond to such incidents.

Subsequent studies broadened the scope of participatory design. Almeida et al. (2022) conducted four participatory design sessions with high school students in Brazil. They went beyond the traditional bully-victim dichotomy by including bystanders in the design, aiming to gain a comprehensive understanding of various perspectives on cyberbullying prevention. The primary focus of their research remained on aspects of user interface and user experience, particularly the placement of response features like comment boxes and notification pop-ups within the application interface. While this is essential for empowering young individuals to respond effectively to CB incidents, it overlooks the diverse types of CB and focuses only on one modality of content, i.e., text-based comments.

In their study, Milosevic et al. (2023) leverages focus group interviews and one-on-one interviews with pre-adolescents and young adults to assess the effectiveness of hypothetical AI detection of CB. To understand *how youth perceive AI-enabled* ***proactive*** *moderation or supportive features*, this study examines a wide range of CB types and includes perspectives from bystanders. Despite its efforts, this work is limited in its ability to co-design the scenarios and response options with youth. In particular, Milosevic et al. (2023) focused solely on presenting hypothetical AI-enabled features, such as bystander helper notification for hypothetical CB scenarios. Despite this, the thematic analysis highlighted that pre-adolescents and teenagers have privacy concerns about AI algorithms scanning their personal profiles and shared content. However, some participants also recognised the importance of AI

[26] `https://en.wikipedia.org/wiki/Musical.ly`
[27] `https://www.tiktok.com/`

in detecting bullies and enabling proactive moderation to identify CB before users report it on SMPs.

Expanding youth roles and engagement online: Studies such as (Almeida et al., 2022; Milosevic et al., 2023) illustrate a growing recognition that youth experiences cannot be reduced to the simple victim–perpetrator dichotomy. Research such as (Livingstone and Helsper, 2007; Livingstone et al., 2015; Livingstone and Stoilova, 2021; Thorn and Benenson Strategy Group, 2021), among many others, has increasingly sought to capture the broader ecology of participation, including *engagement*, *witnessing*, and *reporting* behaviours that shape how adolescents encounter online harm. Early work often relied on an "active–passive" binary, distinguishing between users who generate content and those who observe (Verduyn et al., 2017). Later refinements have introduced distinctions between bidirectional (e.g., commenting, messaging) and unidirectional (e.g., liking, viewing) interactions (Meier and Reinecke, 2021). Shaw et al. (2022) proposed a triadic model separating *interactive* (content creation), *reactive* (emoji or like-based responses), and *passive* (consumption without response) behaviours. Their findings show that interactive users are more socially embedded, while reactive and passive users report weaker social connections. Such typologies are particularly relevant for CB detection, as different engagement styles influence exposure to harm and potential involvement in harmful exchanges.

Parallel research on witnessing CB has also expanded understanding. Large-scale projects such as EU Kids Online (EUKO)[28] and Global Kids Online (GKO)[29] highlight that children are not merely passive recipients of risk but exercise agency and resilience in navigating harmful content (Rodríguez-de Dios et al., 2018). These initiatives have mapped diverse online harms, emphasising that witnessing online harms, such as threats, exclusion, rumours, and discriminatory behaviour, is important for shaping our understanding of youth digital risks. Comparative studies

[28] https://www.lse.ac.uk/media-and-communications/research/research-projects/eu-kids-online

[29] https://globalkidsonline.net/

underscore that bystanders can escalate, defuse, or ignore harm, and that methodological clarity is needed to capture these dynamics consistently (Livingstone and Stoilova, 2021). National-level studies, such as Ireland's National Advisory Council for Online Safety (2021) survey, further stress that terminology critically shapes understanding. They note that the term "cyberbullying" often does not resonate with young people, who may instead describe it as "*repeated* ***nasty*** *or* ***hurtful*** *online behaviour*". This mismatch has important implications for promoting reporting behaviours and accurately assessing the prevalence of CB.

Thorn and Benenson Strategy Group (2021) designed youth-centred surveys that presented age-relevant scenarios like receiving unwanted messages, harassment, or pressure to share content. Their focus was on understanding how young people decide whether and how to disclose harmful experiences. Their work revealed that young people's choices about reporting are shaped by multiple considerations, including whom to disclose to (peers, parents, platform moderators), the perceived trustworthiness of reporting mechanisms, anticipated consequences, and fears of stigma or escalation. A key contribution of this research is recognising that formal platform-level reporting and informal disclosures to trusted adults or peers are distinct but equally valid strategies according to the youth surveyed by (Thorn and Benenson Strategy Group, 2021). Both reflect the agency of young people in navigating online risks, though each carries its own barriers and limitations. Extending these insights to the domain of cyberbullying foregrounds the need to critically evaluate whether current platform reporting tools and safeguarding practices are appropriately designed to meet young people's expectations and developmental needs.

Youth agency in research and design: Guishard and Tuck (2013) note that youth involvement in research is often superficial, limited to eliciting narratives or opinions without enabling young people to shape the framing, methods, or outcomes of a study. They state that such approaches risk reinforcing hierarchical researcher–participant dynamics and undermining the empowering goals of participatory research. More recent work, such as that of Clark et al. (2022a), has sought to

reconfigure this dynamic by advocating for models that position young people as co-researchers and co-creators, particularly in data-intensive and algorithmic contexts. Their work outlines strategies such as youth-led question development, collaborative analysis, and joint authorship, illustrating how youth agency can be integrated throughout the research lifecycle. Within the field of participatory design, a parallel body of work has advanced methodological tools for engaging children and adolescents as active design partners. Approaches such as Cooperative Inquiry (Druin, 1999) and Mixing-of-Ideas (Guha et al., 2004) operationalise principles of shared authorship in design processes, enabling young participants to contribute not only their experiential knowledge but also creative artefacts and design prototypes. These techniques have been successfully adapted in studies focused on online safety and cyberbullying. For instance, Ashktorab and Vitak (2016) employed cooperative inquiry and idea-mixing techniques with teenagers to co-design tools to address cyberbullying, demonstrating the value of youth perspectives in shaping interventions that resonate with their lived experiences.

Building on these foundations, recent empirical work has further explored youth-centred participatory methods in the context of cyberbullying, privacy, and online harm (Zou et al., 2024; Zhou and Farzan, 2021; Fiani et al., 2024). Collectively, these studies illustrate the methodological feasibility and ethical imperative of involving youth as co-researchers, while also pointing to persistent challenges in safeguarding, addressing power asymmetries, and translating youth input into tangible design and policy outcomes.

The importance of incorporating the viewpoints of children and young individuals in the development of AI systems has been highlighted by recent research by Charisi et al. (2022). While researchers have explored the perspectives of these demographics on CB through participatory studies, there has been a significant oversight in considering their input during the creation of content-based automated safeguards against CB. Although some studies have addressed aspects of user interface and user experience by involving feedback from young users, a notable limitation is the lack

of solicitation of opinions on content-based automated cyberbullying detection.

Youth agency in developing datasets: In recent years, there has been growing interest in involving young people directly in the creation of datasets for studying cyberbullying, cyber-hate, and online aggression. This shift aligns with participatory and user-centred research traditions, where adolescents are not merely treated as passive subjects of observation but engaged as active co-creators of knowledge and tools intended to protect them. Two notable studies exemplify this approach, Sprugnoli et al. (2018) in Italian and Ollagnier et al. (2022) in French. Both works employ structured role-playing simulations in school contexts, creating controlled, synthetic yet realistic datasets of harmful online interactions among youth.

Sprugnoli et al. (2018) developed their dataset within the CybeRbullying EffEcts Prevention (CREEP) project, an Italian initiative designed to monitor and prevent cyberbullying. Their methodology, embedded in the "*living lab*" framework by Dell'Era and Landoni (2014), emphasised ecological validity and participatory design. Their study involved 70 middle school students (aged 12–13), who engaged in staged WhatsApp conversations on four scenarios under carefully assigned roles such as `bully`, `enabler`, `victim`, or `defender`. To preserve realism, students were encouraged to use offensive language and enact cyberbullying dynamics, while teachers and researchers supervised the chats. The sessions unfolded over several weeks, integrating lectures, role-play experiments, and reflective discussions with students, parents, and teachers. The resulting corpus was anonymised and linguistically annotated, with the broader aim of supporting automated detection of harmful content and building preventive interventions.

Ollagnier et al. (2022), under the same CREEP project, extended this role-playing methodology through "*living labs*" to a larger cohort of 142 French high school students (aged 16–18). Unlike the Italian study, interactions occurred in chatrooms with anonymised participations, where participants used pseudo-usernames to remain unknown to classmates. The scenarios, developed with sociologists and education experts, reflected real-world themes of cyber-bullying and discrimination

based on ethnicity, religion, obesity, and sexual orientation, drawing from prior qualitative studies such as Blaya and Audrin (2019). They introduced a new role in cyberbullying or aggression contexts, "`conciliator`" i.e., someone who is a friend to both the bully and victim and tries to de-escalate or resolve the conflict. In this study, students role-played the roles of bully, enabler, defender, and conciliator. At the same time, researchers played victims to avoid directly exposing adolescents to simulated victimisation. Each session lasted 45 minutes, followed by reflective workshops and surveys to evaluate students' perceptions of cyber aggression and the role-play experience itself.

Together, these studies illustrate the promise of participatory simulations in generating cyberbullying datasets that are both structured and grounded in youth perspectives. They also underscore important differences in design: Sprugnoli et al. (2018) prioritised role-playing through a platform (WhatsApp) already embedded in pre-adolescent social lives. In contrast, Ollagnier et al. (2022) emphasised anonymity and ethical safeguards by researchers enacting the role of victims. However, both approaches share limitations. The simulations remain scripted and artificially constrained, raising questions about how well they capture the complexity and spontaneity of real-world online aggression and the social dynamics of cyberbullying. Despite their attempts to mitigate ethical risks through supervision or researchers enacting as victims, they still introduce ethical risks, including the potential for re-traumatisation, the normalisation of nasty or hurtful behaviours, or unintended reinforcement of negative attitudes among adolescents. These concerns highlight the delicate balance between methodological innovation and the responsibility to safeguard young participants in research on online harms.

2.3.2 Statistical Methods for Survey and Annotation Studies

Different statistical methods are used for survey and annotation studies. We first discuss statistical methods used in survey studies, followed by methods used in

annotation studies to assess reliability.

Association and Effect Sizes in Survey Studies: Large-scale (e.g., N >500) survey studies with youth to understand online risks (e.g., (Rodríguez-de Dios et al., 2018)), perceptions on `seriousness` (e.g., (Huang et al., 2020)), or to understand user engagement patterns (e.g., (Shaw et al., 2022)), among others, commonly employ parametric statistical methods. These include independent t-tests, one-way ANOVA for group comparisons, and factor analytic techniques such as Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) to validate measurement instruments and test hypotheses about group differences. Given large sample sizes, using such methods allows making specific assumptions about the distribution of the population from which the sample is drawn (Hoskin, 2012). In survey research, especially with categorical data (e.g., Classes, Likert-scales), measuring associations between variables requires statistical tests suited to the data and sample characteristics (Kim, 2017). When sample sizes are small or expected cell frequencies are low (e.g., fewer than 5 observations in any cell), parametric tests that help generalise the findings do not work as they fail to meet the assumption that the sampling distribution approximates a normal distribution (violation of asymptotic assumptions) (Agresti, 2002; Kim, 2017). This leads to un-reliable p-value[30] and increased Type I error[31] rates (Agresti, 2002; Kim, 2017). Addressing this in categorical data, (Kim, 2017) places Fisher's Exact Test (Fisher, 1922) as a suitable non-parametric test that calculates the exact probability of observing the data under the null hypothesis of independence (no change).

- **Fisher's Exact Test**: Proposed by (Fisher, 1922), Kim (2017) note that Fisher's Exact test can be used to assess the association between two categorical variables in a 2×2 contingency table (See Table 2.1) when expected cell counts are small. The test computes the exact probability of observing a table as extreme as (or more extreme than) the one obtained, under the null

[30]The probability of observing data as extreme as, or more extreme than the data actually observed, assuming that null hypothesis (no change/association) is true.

[31]False positive, also referred to as the error of rejecting the null hypothesis

hypothesis. For the contingency Table 2.1 it calculates probability (p-value) as,

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{2.1}$$

This helps determine if there is a statistically significant association between the categorical variables.

- **Odds Ratio**: A measure of association strength quantifying how strongly the presence or absence of one characteristic is associated with another. For categorical variables depicted in Table 2.1, Odds Ratio (OR) is calculated as,

$$OR = \frac{a \cdot d}{b \cdot c} \tag{2.2}$$

OR indicates how the odds of an outcome differ between two groups: OR =1 indicates no association, OR >1 suggests positive association (higher odds in the exposed group), and OR <1 indicates negative association (lower odds in the exposed group). For instance, OR=3.0 means the odds of the outcome are **three times higher** in one group compared to another (Szumilas, 2010). However, OR cannot be computed when any cell contains zero, as this leads to division by zero or undefined ratios.

- **Haldane-Anscombe Corrected OR**: When contingency tables contain zero cells (common in small sample or rare event studies), the standard OR becomes undefined (Haldane, 1956). The Haldane-Anscombe correction addresses this by adding 0.5 to each cell before calculating the OR (Anscombe, 1956; Agresti, 2002). This continuity correction enables OR computation and reduces the severe bias that affects maximum likelihood estimates in sparse tables (Jewell, 2003; Weber et al., 2020). However, the correction introduces its own bias toward the null hypothesis (OR=1), particularly in very small samples (J. Sweeting et al., 2004). Weber et al. (2020) note that despite this bias, the 0.5 correction remains most widely used due to its simplicity and adequate performance across typical scenarios

	B_1	B_2	**Row Total**
A_1	a	b	$a+b$
A_2	c	d	$c+d$
Column Total	$a+c$	$b+d$	

Table 2.1: 2×2 contingency table

Inter-Rater Reliability Measures in Annotation Studies: In annotation and evaluation studies, especially within the "*prescriptive paradigm*", multiple raters independently code or label data based on specific guidelines. Measuring consistency among raters is crucial to establishing the quality and trustworthiness of both the annotations and the guidelines. Inter-rater reliability (IRR) measures to quantify agreement beyond what would be expected by chance alone. We discuss the statistics behind some measures used in the studies discussed in Section 2.3.3 to provide a better understanding of statistical reliability measurement in annotation tasks.

- **Cohen's Kappa (κ)**: Designed for two raters classifying items into nominal categories, Cohen's Kappa (κ) (Cohen, 1960), adjusts observed agreement for chance agreement by calculating it as,

$$\kappa = \frac{P_O - P_E}{1 - P_E} \tag{2.3}$$

  Where P_O is observed agreement calculated as the ratio of number of items where both raters agree and total number of items, and P_E is expected agreement by chance, calculated as$\sum_{c=1}^{k} P_{A,c} \cdot P_{B,c}$, where k is number of categories, $P_{A,c}$ is the proportion of items rater A assigns category c, and $P_{B,c}$ is the proportion of items rater B assigns category c. Kappa (κ) values help measure "*how much raters agree beyond chance*". These values range from -1 to 1: Landis and Koch (1977) note, values 0.81-1.00 indicate almost perfect agreement, 0.61-0.80 $\rightarrow$ substantial agreement, 0.41-0.60 $\rightarrow$ moderate agreement, 0.21-0.40 $\rightarrow$ fair agreement, and below 0.20 $\rightarrow$ slight agreement. However, it is important to note that Cohen's Kappa (κ) is sensitive to the prevalence and

distribution of categories (Byrt et al., 1993). When categories are imbalanced (one category dominates), the expected agreement (P_E) becomes large, which can reduce the κ even if the observed agreement (P_O) is high.

- **Fleiss' Kappa**: Fleiss (1971) extends Cohen's kappa to handle three or more raters by calculating agreement across all possible rater pairs simultaneously. Instead of averaging pairwise Cohen's kappa values, Fleiss' kappa provides a single measure of agreement for the entire group of raters, making it more efficient for multi-rater scenarios. This can be calculated as,

$$\kappa = \frac{\bar{P} - \bar{P_E}}{1 - \bar{P_E}} \tag{2.4}$$

  Where $\bar{P}$ is average observed agreement across all items and $\bar{P_E}$ is expected agreement by chance across all categories. However, similar to Cohen's Kappa, Fleiss' Kappa is also sensitive to the prevalence and distribution of categories.

- **Krippendorff's Alpha (α)**: As Kappa measures (Cohen's, Fliess') helps measure how much raters agree *beyond* chance. (Krippendorff, 2004, 2011) propose a versatile reliability measure (α) that measures observed disagreement to *what would be expected* by chance. (Krippendorff, 2011, 2004) measure it as,

$$\alpha = 1 - \frac{D_O}{D_E}, \tag{2.5}$$

  where D_O is the observed disagreement, calculated as the average pairwise disagreement across all items. D_E is expected disagreement by chance, reflecting how much disagreement would occur if ratings were assigned randomly, given the overall distribution of ratings. This value ranges from 0 to 1 (though negative values are theoretically possible in cases of systematic disagreement (Krippendorff, 2004, 2011)). Thresholds for interpretation are: α >0.80 $\rightarrow$ highly reliable, 0.67-0.80 $\rightarrow$ tentatively reliable, and <0.67 $\rightarrow$ low reliability (Krippendorff, 2004). It's considered the most general-purpose IRR measure and is particularly valuable when not all raters code all items (Kumar et al.,

2022; Röttger et al., 2022; Kumar et al., 2024). Unlike kappa measures that assume fixed marginal totals, Krippendorff's alpha accounts for the observed distribution of categories, making it more robust to prevalence effects.

- **Dice Similarity Co-efficient**: To measure inter-annotator agreement for labels assigned to words or sub-phrases related to bullying in their adolescent cyberbullying dataset, Sprugnoli et al. (2018) used the Dice coefficient (Dice, 1945), a metric commonly employed in segmentation tasks and named-entity recognition. For binary classification scenarios (e.g., determining whether a token belongs to an entity), the this similarity coefficient is calculated as,

$$Dice = \frac{2|A \cap B|}{(|A| + |B|)} \tag{2.6}$$

A and B are the sets of items annotated by two raters, $|A \cap B|$ denotes the number of elements common to both sets, and $|A|$ and $|B|$ represent the total number of elements in set A and set B, respectively. Intuitively, the Dice coefficient measures how much the rater's positive annotations overlap relative to their total positive annotations. This ranges from 0 (no overlap) to 1 (perfect agreement) and emphasises agreement on positive cases. It is particularly useful when the presence of an annotation is more informative than its absence (Sprugnoli et al., 2018). Unlike kappa measures, the Dice coefficient does not adjust for chance agreement. It is more similar to precision-recall metrics in machine learning and is preferred when true negatives (items neither rater selected) are abundant and uninformative (Fidon et al., 2017; Müller et al., 2022).

- **PABAK**: Byrt et al. (1993) introduced Prevalence-Adjusted and Bias-Adjusted Kappa (PABAK) to correct for two key limitations of Cohen's Kappa: (1) category prevalence imbalance (i.e., when one category dominates) and (2) rater bias (i.e., when raters systematically favour different categories). PABAK removes the chance correction term (P_E) from Cohen's and Flies Kappa (Equa-

tions 2.3, 2.4), making it independent of marginal distributions. It is calculated as:

$$PABAK = 2P_O - 1, \tag{2.7}$$

where P_O is the observed proportion of agreement between two raters. PABAK values range from -1 to 1, with the same interpretation thresholds as Cohen's Kappa (Landis and Koch, 1977). By eliminating the chance correction, PABAK provides a more stable measure when category distributions are highly imbalanced. However, this simplification assumes equal prevalence and no bias, which may not hold in all scenarios.

- **PABAK-OS**: Parker et al. (2011) extended PABAK to handle ordinal scales by incorporating the degree of disagreement between categories. While PABAK treats all disagreements equally (a disagreement between adjacent categories is weighted the same as a disagreement between distant categories), PABAK-OS accounts for the ordinal nature of categories. It is calculated as:

$$PABAK\text{-}OS = 1 - \frac{D_O}{D_{max}}, \tag{2.8}$$

where D_O is the observed weighted disagreement (sum of absolute differences between raters' ordinal ratings), and D_{max} is the maximum possible disagreement. The weighting reflects the fact that disagreements between distant categories (e.g., rating 1 vs. 5) are more severe than those between adjacent categories (e.g., rating 1 vs. 2). PABAK-OS values range from -1 to 1, with higher values indicating greater agreement. This measure is beneficial when rating scales have inherent ordering (e.g., Likert scales, severity ratings) and when prevalence imbalances exist.

2.3.3 Understanding and Annotating Cyberbullying in NLP

Recognising cyberbullying dimensions and categories: As internet communication technologies (ICTs) and social media platforms (SMPs) continue to evolve,

the definition of cyberbullying (CB) established by Patchin and Hinduja (2006) and Smith and Steffgen (2013) remains widely acknowledged. This definition characterises CB as *"online harm" repeatedly* targeted at an individual through ICTs with the specific *intent* to cause harm. Moreover, Leung et al. (2018); Song and Oh (2018); Ollagnier et al. (2022); Chan et al. (2022) suggest CB incidents involve different types of individuals beyond the bully-victim dichotomy. Beyond the traditional roles of bully and victim, research indicates that CB often involves other participants (Leung et al., 2018; Song and Oh, 2018; Ollagnier et al., 2022; Chan et al., 2022), including:

1. *Bystander-**Enabler***: Those who encourage or reinforce the bullying behaviour.

2. *Bystander-**Defender***: Those who actively support or defend the victim or call out the bully.

3. ***Silent Bystanders***: Those who are aware of the bullying but remain passive.

4. ***Conciliator***: Introduced by Ollagnier (2024), those who mediate between bully and victim to resolve the conflict.

Furthermore, cyberbullying often spans offline and online contexts (O'Higgins Norman et al., 2023), and manifests in several distinct types, each differing in mechanism, intent, and social impact (Willard, 2007; Nadali et al., 2013; Slonje et al., 2013; Kowalski et al., 2014; Bauman, 2015). These include,

1. *Flaming*: Characterised as an open and hostile confrontation online with an intent to provoke immediate emotional response. This involves direct insults, name-calling, or hostile exchanges.

2. *Harassment*: It is characterised as persistent, targeted attacks over time with an intent to distress or intimidate the victim over an extended period. This involves repeated threats or offensive content.

3. *Denigration*: Characterised as actions aimed at harming reputation or self-esteem through rumours, mockery, or posting harmful content. Its sole intention is to damage reputation or social standing.

4. *Masquerading* (Impersonation): This involves online activities like creating fake accounts or sending messages under the victim's identity. The intent of pretending to be a victim is to confuse, embarrass, or harm the victim indirectly.

5. *Outing*: Characterised as publicly sharing private or sensitive information about someone with the intent to humiliate or embarrass the victim.

6. *Trickery*: Distinct from outing, this involves manipulating the victim into revealing private information. The intent is to deceive or coerce an individual into disclosing information, which is then shared publicly.

7. *Social Exclusion*: Characterised as making a victim feel unwanted or isolated through deliberate isolation from groups or interactions. This involves ignoring, blocking, or excluding someone from online communities or chats.

8. *Cyber-stalking*: This behaviour involves monitoring online activity or interactions with the sole intent to control, intimidate, or induce fear amongst victims.

Although these cyberbullying categorisations might seem to overlap, each differs in meaningful ways from the others. For instance, harassment involves repeated, targeted actions intended to distress or intimidate someone over time, making it distinct from flaming, which is usually a single hostile or insulting outburst. It also differs from denigration as harassment focuses on causing ongoing emotional harm while denigration aims to damage the victim's reputation or social standing.

Early Taxonomies of CB behaviour in NLP: Motivated by (Vandebosch and Van Cleemput, 2009), Van Hee et al. (2018) provided one of the first comprehen-

sive frameworks for identifying key categories of online harm related to cyberbullying (CB) in English and Dutch. They acknowledged the role of defense strategies and emphasised that cyberbullying exchanges involves aggressive bystander responses and victim resistance. Although broad and pioneering, Van Hee et al. (2018) grouped diverse behaviours under umbrella terms. For instance, *identity-based harassment* was subsumed under *insults*, potentially obscuring the distinct dynamics of race-, gender-, or religion-based abuse. This limitation highlighted the need for finer-grained distinctions to reflect the social realities of online abuse.

Recognising 'Body-shaming' as CB category: Sprugnoli et al. (2018) extended the early taxonomy by Van Hee et al. (2018) to youth contexts and the Italian language, and explicitly introduced "*body-shaming*" as a sub-category of *insults*. Their contribution underscored that harassment based on appearance is both prevalent and uniquely harmful, warranting its own label rather than being collapsed under generic insult. However, Sprugnoli et al. (2018)'s annotation schema remained narrow as they only added one new dimension of "*body-shaming*" to Van Hee et al. (2018)'s taxonomy rather than attempting to restructure the broader taxonomy.

Towards Unified Online Harm Typologies: Recognising the need for a unified view of online harms, Banko et al. (2020) consolidated categories from platforms, civil society guidance, and policy documents into a practical schema for moderation. This typology recognises *identity-attacks*, *identity-misrepresentation*, *insults*, *sexual aggression*, and *violent threats* as distinct categories, effectively aligning academic and regulatory perspectives. However, it tends to conflate bullying with general toxicity and focuses on surface behaviours, offering limited insight into the relational dynamics, such as *intent* or *escalation* that shape cyberbullying interactions. Urbaniak et al. (2022) distinguished direct ad hominem abuse from broader `identity-based harassment` in their work on understanding *personal attacks*. They highlight that aggression can target individuals as persons (e.g.,

through comparisons to animals or derogatory insinuations) without necessarily invoking group-based identities. This distinction was useful for disentangling types of harm and for distinguishing between "*personal attacks*" and "*identity-based*" attacks, which often co-occur in adolescent bullying.

Recognising Overt vs. Covert Online Aggression: Kumar et al. (2022) distinguishes between *overt* and *covert* aggression through their multi-lingual conversational datasets. *Overt aggression* includes explicit acts such as insults or threats, whereas *covert aggression* captures indirect hostility through sarcasm, rhetorical questions, or insincere politeness. This framework marked an important shift by recognising that online harm is often conveyed through subtext and social cues.

Recognising Intent in Harmful Conversations: Kumar et al. (2024); Ollagnier (2024) expanded the view of *overt* and *covert* by modelling aggression through conversational functions such as *attack*, *defend*, *counter-speech*, *gaslighting*, and *instigation*. Although Ollagnier (2024) adapted these categories for their French teen role-played dataset (Ollagnier et al., 2022), they were only able to label 36.8% of it with two annotators. Despite its conceptual value, this relational aggression intent schema has seen limited application in youth cyberbullying contexts.

Recognising Self-Harm Encouragement as CB category: When reviewing the recent UK Online Safety Act[32], Trengove et al. (2022) recognised that "*encouragement to self-harm*" is a pressing form of online abuse. While earlier taxonomies often collapsed such content under "*threats of violence*", Trengove et al. (2022) argued that "*self-harm incitement*" carries unique risks and social dynamics, especially in adolescent populations vulnerable to peer influence. This recognition exposed a critical gap in prior frameworks and underscored the need to make "*self-harm encouragement*" an explicit category in annotation schemes.

[32] https://www.legislation.gov.uk/ukpga/2023/50

Taken together, these studies illustrate a clear progression in cyberbullying research, ranging from broad, overarching taxonomies (Van Hee et al., 2018) to more incremental refinements (Sprugnoli et al., 2018; Urbaniak et al., 2022), policy-driven syntheses (Banko et al., 2020), distinctions grounded in conversational analysis (Kumar et al., 2022, 2024), cultural adaptations (Ollagnier, 2024), and legislative recognition of emerging harms (Trengove et al., 2022). However, despite their granularity, many of these approaches remain limited by collapsing or subsuming overlapping categories or by being constrained to specific datasets or contexts. Collectively, these shortcomings point to the need for a holistic framework that integrates multiple dimensions: (a) aggression type (`overt/covert`), (b) harm category (`threats`, `personal attacks`, `identity-based harassment`, `body-shaming`, `self-harm encouragement`, etc.), and (c) communicative intent (`attack`, `defend`, `gaslight`, `instigate`, etc.).

Adult or Youth Participation: A consistent trend across much of the cyberbullying annotation literature is reliance on adult participants, often drawn from linguistics, computational linguistics, social science, or professional annotation backgrounds. For instance, Van Hee et al. (2018) employed trained linguists, two for Dutch and three or more for English cyberbullying corpora. They report inter-rater reliability using Cohen's Kappa (Cohen, 1960) (See Equation 2.3) for Dutch (κ=0.69) and Fleiss' Kappa (Fleiss, 1971) (See Equation 2.4) for English (κ=0.59). Similarly, Sprugnoli et al. (2018) engaged two social science experts as annotators working on a web-based platform. They only measured the Dice coefficient (Dice, 1945) (See Equation 2.6) to assess the overlap of annotated expressions and the precision of the assignment of attributes at the message-level. Sprugnoli et al. (2018) report that by involving annotators while adopting (Van Hee et al., 2018)'s guideline and labelling one scenario-based conversation[33] yields 0.8 Dice similarity co-efficient. They attribute these improved scores to the direct involvement of

[33]Scenario - C: Your classmate is seen as an overachiever who keeps to themself [...], and after their mother asked teachers to increase homework for everyone, a careless disclosure has made several students angry with [...]

annotators in creating the guideline.

Adult expert-driven annotation is also evident in multilingual efforts. Kumar et al. (2022, 2024) recruited multiple adult annotators, applied majority-voting aggregation, and refined schemas through iterative annotation rounds assessed with Krippendorff's alpha (Krippendorff, 2011). In Kumar et al. (2022), nine annotators (5 Meitei, 3 Hindi, 2 Bangla) produced initial inter-rater agreement values between 0.28 and 0.48, depending on language and task. After guideline refinement, a second round with balanced annotator distribution (3 per language) yielded substantial improvements, with alpha scores ranging from 0.49 to 0.93. In their pilot study, Kumar et al. (2024) initially observed a very low alpha (0.25). This prompted revisions such as merging overlapping categories (e.g., combining race, ethnicity, and nationality under "Descent") and introducing new ones for greater nuance (e.g., additional categories under mood and modality). These adjustments, along with clearer guidelines, raised reliability to α=0.53. Likewise, (Ollagnier et al., 2022; Ollagnier, 2024) restricted their annotator pool to computational linguists, reporting exceptionally high agreement (κ>0.9) for categories like `hate speech` and `verbal abuse`.

This reliance on adult annotators extends beyond cyberbullying datasets to broader online safety corpora. A notable example is the Offense Language Identification Dataset (OLID) Zampieri et al. (2019) curated through crowdsourcing with adult annotators from an online recruitment platform. Using a structured multi-level scheme (e.g., offense vs. not offense, type, and target), each message in the dataset was labelled by at least two annotators., with a third resolving disagreements. Final labels required consensus across all levels, and inter-annotator agreement for the first layer reached Fleiss' Kappa of 0.83. Such frameworks emphasize that multi-level labelling, analysing redundancies, and strict agreement thresholds help to ensure reliability. However, these approaches risk neglecting age-specific perspectives, particularly in contexts like cyberbullying where interpretations of intent, harm, and offensive language often diverge between adults and youth.

Bonetti and Tonelli (2022) introduced an innovative youth-centred approach by

involving teenagers directly in annotation through a the "games with a purpose approach" developed by Lafourcade et al. (2015). Middle and high school students in Italy participated in a game-based activity where they identified and modified abusive content in dialogues and graffiti-like text. These texts were sourced from Sprugnoli et al. (2018) and were presented in Italian language. This method engaged youth in an accessible, motivating format, and produced 590 annotations across 199 sentences. Crucially, their results revealed marked discrepancies between annotations made by professional linguists in (Sprugnoli et al., 2018) and those collected from students. These differences may reflect the distinct ways youth perceive offensive language, as well as game-driven behaviours such as eagerness to alter sentences or occasional skipping. Nevertheless, the study underscores the value of integrating youth perspectives into cyberbullying research, providing insights into how the populations most affected by online abuse interpret harmful content.

2.4 Foundational Landscape of Computational Investigations

The detection of cyberbullying (CB), online harassment, online abuse, or hate speech is often tackled by treating it as a classification task, wherein Machine Learning (ML) or Deep Learning (DL) algorithms are leveraged to categorise user-generated content (UGC) into different classes or labels. These methods typically assign a probability score to each category, indicating the likelihood that the content belongs to that class. Many studies have explored both the diverse computational approaches and datasets related to the detection of textual CB (Salawu et al., 2017; Rosa et al., 2019; Vidgen and Derczynski, 2020; Bayari and Bensefia, 2021; Emmery et al., 2021; Kim et al., 2021b; Wang et al., 2024c). They organise CB research into relatively coarse categories such as **binary** (e.g., `bullying` vs. `non-bullying`), **fine-grained** (e.g., *`user-roles`*), and **multi-modal** (text, image, metadata).

Ever since Vaswani et al. (2017) revolutionised natural language processing

(NLP) by unifying the encoder-decoder paradigm (Sutskever et al., 2014) and the attention mechanism (Bahdanau et al., 2014) to propose "*self attention in transformers*", researchers have developed pre-trained language models (LM) based on this paradigm (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)). These language models, when adapted for sequence classification tasks such as binary text classification (e.g., `CB` vs `non-CB`), have outperformed traditional machine learning methods (e.g., Support Vector Machines (SVM)), deep learning methods (e.g., Long Short Term Memory (LSTM)) (Paul and Saha, 2022; Tripathy et al., 2022; Philipo et al., 2024). It is important first to understand the "*self-attention*" paradigm before delving into its applications in cyberbullying detection.

The "*self-attention*" mechanism by (Vaswani et al., 2017) drives today's language models (LMs) by computing contextual representations of words in a sentence based on their relationships to other words in the same sentence. For example, in the sentence "*the robbers robbed the bank next to the river bank*", the word '*bank*' appears twice but with different meanings. Self-attention helps recognise contextual clues such as the presence of the word '*robbed*' before the first time '*bank*' appears and '*river*' before the second time '*bank*' appears, thereby creating distinct representations of the same word depending on its context. This enables each input token (a word) in a sequence (sentence) to dynamically attend to other tokens in that sequence. When processing the first occurrence of '*bank*' (token at position i), it is represented as an embedding x_i[34]. Intuitively, as (Vaswani et al., 2017) explains, x_i is linearly projected into three vectors, (i) the *query* vector ($q_i = x_i W_Q$)[35], represents what information the current token i is seeking (e.g., the word "*robbed*" looks for a word *victim* or *crime*), (ii) the *key* vector ($k_j = x_j W_K$), represents what information each token j can provide (e.g., word "*bank*" offers either financial or geographical context), and (iii) the value vector ($v_j = x_j W_V$) contains the actual semantic content from token j. The dot product $q_i \cdot k_j^T$ measures the compatibility

[34] A high-dimensional floating point vector retrieved from a pre-trained embedding lookup table.
[35] Here, $W_Q W_K W_V$ are learned weight matrices containing trainable parameters (typically 32-bit floats).

of tokens i and j to determine the actual semantic content from token j that should influence token i. For example, a high score for $q_{\text{bank}}.k^T_{\text{robbed}}$ indicates strong compatibility because financial institutions are robbery targets. Whereas, $q_{bank} \cdot k^T_{river}$ would score lower for the first time '*bank*' appears. This mechanism by (Vaswani et al., 2017) then derives attention weights,

$$\alpha_{i,j} = \text{softmax}\left(\frac{q_i \cdot k_j^T}{\sqrt{d_k}}\right), \tag{2.9}$$

Where d_k refers to the dimensionality of the *key* vectors, and softmax (also known as the normalised exponential function or softargmax (Bridle, 1990)) helps compute a probability distribution of the input vector. The softmax is defined as a vector function, the i-th component of the output is calculated as:

$$\text{softmax}(z) = \frac{e^{z_j}}{\sum_{j=1}^{N} e^{z_j}} \quad \text{for } j = 1, \ldots, N \tag{2.10}$$

is applied to the vector of all scaled dot products for a single query (q_i) against all keys in the sequence. If input to softmax is $z_{i.j} = \frac{q_i \cdot k_j^T}{\sqrt{d_k}}$, then Equation 2.9 can be understood as

$$\alpha_{i,j} = \frac{e^{\frac{q_i \cdot k_j^T}{\sqrt{d_k}}}}{\sum_{m=1}^{L} e^{\frac{q_i \cdot k_j^T}{\sqrt{d_k}}}}$$

Here, m is the index used to sum over all possible key tokens ($m = 1$ to L) in the input sequence. This ensures the attention weights ($\alpha_{i,j}$) are normalised, i.e, they sum to 1. Finally, the attention output for token i is the weighted sum of the value vectors: $\sum_{j=1}^{L} \alpha_{i,j} v_j$. This allows language models to capture the contextual and semantic dependencies among tokens (e.g., how '*bank*' relates to '*robbed*' or '*river*' when given context). Computational researchers such as (Xie et al., 2017a; Mullenbach et al., 2018) have leveraged these attention weights ($\alpha_{i,j}$) to explain and interpret neural networks on how much token i "*looks at*" token j for different language modelling tasks.

Vaswani et al. (2017) proposed this self-attention-based architecture within Sutskever

et al. (2014)'s *encoder-decoder*, where the *encoder* processes the input sequence (e.g., a sentence in English) through multiple layers of self-attention and feed-forward networks[36] to create contextualised representations. While the *decoder* generates the output sequence (e.g., a French translation) using both self-attention over previously generated tokens and cross-attention to the encoder's representation. Unlike self-attention (where Query (q_i), Key (k_j), and Value (v_j) vectors all come from the same sequence), Vaswani et al. (2017) also introduce *cross-attention*, where the current decoder output (the sequence being generated) is the query (q_i) and its key (k_j) and value (v_j) vectors from the encoder's output (the input sequence for the decoder). This allows the decoder to "*look back*" and focus on the most relevant parts of the input sequence when generating each new output token.

Building on (Vaswani et al., 2017)'s self-attention transformers, Devlin et al. (2019) recognised that the entire *encoder-decoder* architecture is not required for text classification, named entity recognition tasks and developed Bi-directional Encoder Representation from Transformers (BERT), a language model that processes input through a multi-layer bi-directional transformer encoders. For the sentence "*the robbers robbed the bank next to the river bank*", the input representation at the input layer of BERT consists of three embedding types: (a) token embeddings (WordPiece[37] tokens), (b) segment embeddings (two distinct phrases or sentences), and (c) positional embeddings (to encode position the token in the sequence). In BERT, these embeddings then flow upward through multiple stacked transformer encoder layers, where each layer contains two sub-components: (i) a multi-head self-attention mechanism and (ii) a position-wise feed-forward network. The multi-head self-attention mechanism allows a computing attention to different aspects of the sequence. By splitting the embeddings into multiple subspaces (heads $h \in H$),

[36] A two-layer, position-wise fully-connected network applied independently and identically to each position (token) in the sequence. This introduces non-linearity to allow the model to refine the representations generated by the attention sub-layer. It helps to process each token's attention-infused vector individually.

[37] A sub-word level tokenisation algorithm which tokenises words (e.g., robbed → 'rob', '##b', '##ed', where # represents connection: See `https://huggingface.co/learn/llm-course/en/chapter6/6`)

self-attention for each head is computed as

$$head_h = Attention(XW_Q^h, XW_K^h, XW_V^h)$$

The position-wise feed-forward network introduces non-linear transformations to each token embedding, increasing the model's representational capacity. Both sublayers are wrapped with residual connections and layer normalisation. Specifically, the input to a sublayer is added to its output (residual connection) to facilitate gradient flow and allow the model to learn identity mappings if necessary. The sum is then normalised across the hidden dimension using layer normalisation, stabilising training by ensuring mean= 0 and variance= 1 in the token representations. Devlin et al. (2019) train this architecture with large text corpora (e.g., BookCorpus (Zhu et al., 2015)) on two unsupervised objectives, i.e., without providing target tokens or values during training. This included masked language modelling (MLM) and next sentence prediction (NSP). Within the MLM task, a random 15% of input tokens are masked, and the model must predict the tokens. For instance, in the sentence "*the robbers robbed the bank next to the river bank*" if the word '*bank*' at position 4 is replaced with '`[MASK]`', the model learns to predict using representation of words in both directions i.e., before and after '`[MASK]`'. Specifically, the final hidden state $h_{[\text{MASK}]}$ from the top encoder layer passes through a softmax classifier over the entire vocabulary as

$$P(\text{token}) = \text{softmax}(h_{[\text{MASK}]}W_{\text{vocab}}),$$

In this manner, the model learns to assign high probability to token '*bank*'. In the NSP task, two sentences are used as inputs (e.g., our example sentence as sentence A and a following sentence as B), and use the final hidden state of the special token '`[CLS]`' as $h_{[\text{CLS}]}$ is used to compute probability using softmax classifer as

$$P(\text{IsNext}) = \text{softmax}(h_{[\text{CLS}]}W_{\text{NSP}})$$

Through these tasks, BERT learns contextualised representations. For example, the

two instances of '*bank*' in our example develops distinct representations (a) h_{bank_1} to capture financial context via strong attention to '*robbed*' and (b) h_{bank_2} to capture geographical context via strong attention to '*river*'.

While BERT builds on Vaswani et al. (2017)'s self-attention in encoder-only architecture for understanding tasks, Radford et al. (2019); Brown et al. (2020), among others use only the decoder component of the original transformer architecture for generative tasks. The key distinction lies in the **autoregressive** nature of decoder-only models, i.e., they sequentially generate one token at a time, where each token can only attend to previously generated tokens through a causal (or masked) self-attention mechanism. For instance, if a decoder-only model is generating the sentence "*the robbers robbed the bank next to the river bank*", when predicting the first '*bank*' at position 4, the model's self-attention is restricted to only the preceding context: "*the robbers robbed the*". The query q_{bank} can compute attention scores $q_{\text{bank}} \cdot k^T_{\text{robbers}}$ and $q_{\text{bank}} \cdot k^T_{\text{robbed}}$. However, attention scores for future tokens like '*next*', '*river*', or the second '*bank*' are explicitly masked (e.g., set to ∞) to prevent information leakage from future positions. This masking ensures the attention weights ($\alpha_{i,j}$) are the same as Equation 2.9 if $i \geq j$ or 0 if $j > i$. During pre-training, decoder-only models use a causal language modelling (CLM) objective, i.e., given the prefix "*the robbers robbed the*", the model predicts the next token '*bank*' by computing the probability of the token as,

$$P(\text{token}_i|\text{token}_j) = softmax(h_i W_{\text{vocab}}),$$

where h_i is the hidden state encoding only leftward context and W. These decoder-only models are trained to maximise the likelihood of each token given all previous tokens:

$$\mathcal{L} = -\sum_{i=1}^{n} logP(\text{token}_i, \text{token}_1, ..., \text{token}_{i-1})$$

Unlike BERT's bidirectional attention, which allows the first '*bank*' to simultaneously attend to both '*robbed*' (left) and '*next*' (right), the decoder-only model rely

solely on leftward context when generating each token, making it naturally suited for text generation tasks like completion, translation, and dialogue, where future tokens are unknown at inference time.

Ruder et al. (2019) demonstrated how these pre-trained models can be adapted to downstream tasks, such as text classification, sentiment analysis, or question answering, or natural language inference, without needing to train a new model from scratch. Adapting these models to tasks such as classification involves adding a lightweight component, often referred to as a "*classification head*", on top of the pre-trained model (Ruder et al., 2019). This frequently involves a single layer that employs the softmax function/layers (Bridle, 1990) (Equation 2.10), is crucial as it converts the model's raw numerical outputs, known as *logits*, into a probability distribution where the probabilities of all possible categories (C) sum to one (LeCun et al., 2002).

Today, the advent of open-source and closed-source large language models (LLMs) or vision-language models (VLMs) which have been built on the unified foundations of *self-attention in encoder-decoder* paradigm by (Vaswani et al., 2017) has helped further explore many domains of computational research, such as information retrieval (Bonifacio et al., 2022), data augmentation through prompt-based or fine-tuned generation (Yoo et al., 2021; Anaby-Tavor et al., 2020) and synthetic data creation for low-resource tasks, such as harmful content detection (Hui et al., 2024) and clinical conversations (Wang et al., 2024a), among many others. Although their effectiveness in data generation compared to authentic data remains comparable (Li et al., 2023d), LLMs present a promising alternative to address data scarcity in low-resource tasks. Prompting LLMs in various settings has enabled researchers to even use them as moderators of harmful content (Nirmal et al., 2024; Bauer et al., 2024). Recognising the developments in NLP and evolving evaluation paradigms suggests that the coarse **three categories** (*binary*, *fine-grained*, *multimodal*) of computational investigations by (Salawu et al., 2017; Rosa et al., 2019; Vidgen and Derczynski, 2020; Bayari and Bensefia, 2021; Emmery et al., 2021; Wang et al., 2024c)

are no longer sufficient. Drawing on the foundational research, we expand the three categories further and organise them as (i) understanding automatic evaluation metrics in CB detection (§2.4.1), (ii) binary CB detection (§2.4.2), (iii) fine-grained CB detection (§2.4.3), (iv) multi-task CB detection (e.g., detecting `bullying` vs. `non-bullying` and *`user-roles`*) (§2.4.4), (v) multi-modal CB detection (§2.4.5), (vi) llm-aided generation (e.g., conversation stories) and related applications (§2.4.6), and (vii) understanding obfuscations and perturbation in cyberbullying detection (§2.4.7)

2.4.1 Automatic Evaluation Metrics in CB detection

In classification tasks, binary (e.g., `CB (1)` vs `not-CB (0)`), multi-class (e.g., `enabler`, `defender`, `victim`, `bully`), or multi-label (e.g., `insult`, `body-shame`, `personal attack`), the performance of a classifier is typically assessed through automated metrics that compare predicted labels with ground truth. However, automated classifiers output a probability (p_i^c) belonging to a specific class c for each instance i (Grandini et al., 2020). In a binary (two-class) classification, a threshold (e.g., 0.5) is typically applied to determine whether an instance qualifies as positive or negative. In multi-class settings, the model produces a probability score for each class or label. Either the highest probabilities are selected or they are normalised using a softmax function, and the class with the highest softmax score is selected as the predicted label (Grandini et al., 2020). Moreover, as discussed earlier in Section 1.1 of Chapter 1, existing CB datasets are imbalanced, i.e., positive instances (*CB-related*) are substantially outnumbered by negative (*not CB-related*) instances. As a result, only relying on **accuracy** can provide a misleading view of performance, since a classifier could achieve high accuracy by simply predicting the majority class (e.g., always predicting not CB-related) (van Aken et al., 2018). To that effect, we discuss some automatic evaluation metrics commonly employed in CB or other online abuse detection tasks.

Recall and Precision: Evaluating classifier performance requires understanding not only how often the model predicts the positive class, but how often those predictions are correct. **Recall** helps measure the proportion of actual positive instances correctly identified by the model, while **precision** helps measure the proportion of predicted positive instances that are genuinely positive. These metrics help reveal a classifier's behaviour in imbalanced scenarios (e.g., CB detection) where the minority (positive) class carries primary importance. Recall is formally defined as,

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{TP} + \text{False Negative (FN)}} \tag{2.11}$$

Where, True Positive and False Negative are calculated by comparing the predicted label $\hat{y}_i$ with the true label y_i for each instance i. True Positive: $\text{TP} = \sum_{i=1}^{N} 1(\hat{y}_i = 1 \ \ \& \ \ y_i = 1)$; False Negative $\text{TP} = \sum_{i=1}^{N} 1(\hat{y}_i = 0 \ \ \& \ \ y_i = 1)$. Similarly, **precision** is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{False Positive (FP)}}. \tag{2.12}$$

Where FP is computed as the number of instances where the model predicts the positive class but the true class is negative ($\text{FP} = \sum_{i=1}^{N} 1(\hat{y}_i = 1 \ \ \& \ \ y_i = 0)$). For multi-class classification tasks (e.g., `enabler`, `defender`, `victim`, `bully`), TP, FP, FN cannot be calculated directly for the whole task because there are multiple classes. Instead, they are calculated per-class by treating each class as "positive (`1`)" and all others as "negative (`0`)". So, for class c, $\text{TP}_c = \sum_{i=1}^{N} 1(\hat{y}_i = c \ \ \& \ \ y_i = c)$; $\text{FP}_c = \sum_{i=1}^{N} 1(\hat{y}_i = c \ \ \& \ \ y_i \neq c)$; $\text{FN}_c = \sum_{i=1}^{N} 1(\hat{y}_i \neq c \ \ \& \ \ y_i = c)$. Therefore, precision and recall are also calculated per class in multi-class classification. In a multi-label classification task where one instance can have more than one label, predictions are made for each label independently. For label c, these are computed as: $\text{TP}_c = \sum_i 1(c \in \hat{Y}_i \ \ \& \ \ c \in Y_i$; $\text{FP}_c = \sum_i 1(c \in \hat{Y}_i \ \ \& \ \ c \notin Y_i)$; $\text{FN}_c = \sum_i 1(c \notin \hat{Y}_i \ \ \& \ \ c \in Y_i)$. Similar to multi-class classification, precision and recall are also computed per label in multi-label classification.

Macro and Micro Average F scores: van Rijsbergen (1979) introduced the **F** measure to balance between **Precision** (P) and **Recall** (R) in information retrieval and binary classification tasks as harmonic mean of precision and recall

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

. To extend this metric to evaluate models in multi-class and multi-label classification tasks, two primary aggregation strategies are employed: (a) Micro Average F1 and (b) Macro Average F1 (Grandini et al., 2020). Micro Average F1 aggregates contributions of all classes ($c \in C$) by calculating global true positive, false positives, and false negatives as,

$$F1_{\text{micro}} = \frac{2 \cdot \sum_{c=1}^{C} TP_c}{2 \cdot \sum_{c=1}^{C} TP_c + \sum_{c=1}^{C} FP_c + \sum_{c=1}^{C} FN_c} \tag{2.13}$$

This metric weights each instance equally, making it dominated by the performance on the majority classes. For example, in a dataset with 1,000 instances of class A and 100 instances of class B, micro-F1 is heavily influenced by performance on class A. (Arango et al., 2019) note that studies reporting state-of-the-art performance of 94% F1-score in detecting cyberbullying, such as (Agrawal and Awekar, 2018), are reporting micro-average F1 scores, making it difficult to interpret true performance. We find this trend of reporting micro-average F1 even in vision-language ***harmful*** meme detection tasks such as (Kumar and Nandakumar, 2022). In imbalanced settings, Macro-averaged F1 computes the F1-score independently for each class C and then takes their unweighted mean:

$$\mathrm{F1}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} F1_c \tag{2.14}$$

Since the macro-averaged F1-score computes F1 independently for each class, it offers a clearer view of per-class performance, where a higher value indicates better prediction for that class.

2.4.2 Binary Cyberbullying Detection

A critical qualitative analysis by Kofoed and Staksrud (2019) underscores the limitations of traditional developmental psychology frameworks like the one developed by Olweus (1994), which often reduce CB to a simplistic victim-perpetrator dichotomy. In a similar vein, machine learning approaches aimed at automating the detection of CB on social media often adopt a binary classification schema, labelling user-generated textual content as `1 (CB)` or `0 (not-CB)` (Dadvar and Eckert, 2020; Zhao et al., 2020; Phanomtip et al., 2021; Tripathy et al., 2022; Paul and Saha, 2022; Ahmed et al., 2022). These approaches often train machine learning algorithms on annotated datasets, in which they learn to pair samples with binary labels, also referred to as *supervised learning*. This learning process is generally guided by a "*loss*" function to ensure the model improves over time by measuring how far the predictions of a message deviate from the true label.

Within this supervised learning framework, early research by (Yin et al., 2009; Reynolds et al., 2011; Dadvar et al., 2013; Sanchez and Kumar, 2011; Bayzick et al., 2011), predominantly relied on classical machine learning algorithms such as Naïve Bayes by (Zhang, 2004) or logistic regression (LR) by (Cox, 1958) or support vector machines (SVM) (Cortes and Vapnik, 1995) or decision trees (DT) (Quinlan, 1986). These methods largely depended on supervised learning techniques built around non-contextual text-based features like Term Frequency-Inverse Document Frequency (TF-IDF), n-grams, bag of words (BoW), profanity lexicons, and sentiment polarity scores (Salawu et al., 2017; Rosa et al., 2019; Vidgen and Derczynski, 2020; Bayari and Bensefia, 2021; Emmery et al., 2021). However, Emmery et al. (2021) observed that many of these studies rely on the CAW 2.0 (Content Analysis for Web 2.0)[38] datasets, which may be unsuitable as they fail to capture the nuances of cyberbullying. Alongside these feature-engineered approaches, rule-based systems like Serra and Venter (2011) and Bretschneider and Peters (2016), among others, were developed for predicting if a message as cyberbullying or not. For instance, Bretschnei-

[38] `http://www.ra.ethz.ch/cdstore/www2009/caw2.barcelonamedia.org/index.html`

der and Peters (2016) developed a binary CB classifier that incorporated rule-based features of text phrases such as word normalisation, named-entity-recognition for person-specific references, and multiple curse word dictionaries[39] and phrases of offensive language or slang from Millwood-Hargrave (2000). While these classical supervised learning and rule-based approaches established the feasibility of CB detection, their reliance on handcrafted features and surface-level text representations limited their effectiveness. Specifically, they struggled to account for linguistic ambiguity, evolving online slang, and the context-dependent nature of CB language (Rosa et al., 2019; Emmery et al., 2021; Kim et al., 2021b).

To overcome these limitations, subsequent research shifted toward deep learning methods, which could learn distributed text representations automatically. Techniques like recurrent neural networks (RNNs) (Elman, 1990), including long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), gated recurrent units (GRUs) (Cho et al., 2014), and Bi-directional LSTM (Bi-LSTM) (Huang et al., 2015) have been leveraged to model language in cyberbullying datasets and classify a message as `0 (victim)` or `1 (bully)` (Agrawal and Awekar, 2018; Menini et al., 2019; Aggarwal et al., 2020). Bi-LSTM with and without (Vaswani et al., 2017)'s attention mechanism have been reported by (Aggarwal et al., 2020; Agrawal and Awekar, 2018) to outperform traditional machine learning algorithms (e.g., SVM, LR, DT) on binary cyberbullying detection tasks. These Bi-LSTM networks consist of deep neural networks (LSTMs) to process text sequentially, i.e., reading one word (as a vector embedding) at a time and using internal gates and a shared *cell* state. This cell state in an LSTM provides a direct path for information flow from one time step to the next, with only linear interactions mimicking a conveyor-belt-like path. This additive, not multiplicative, path allows the signal that guides weight updates (gradient) to flow backwards without being repeatedly scaled down, thereby maintaining its magnitude and reducing the vanishing gradient problem[40]. The flow of

[39](1) `https://www.noswearing.com/`; (2) `https://www.bsa.govt.nz/assets/oldsite/assets/Research/Acceptability_of_Words_2013_WEB.pdf`

[40]The vanishing gradient problem occurs when the gradient (the signal updating model weights) shrink exponentially during back-propagation through time, making long-range dependencies dif-

information on this path is maintained by three interactive gates: (i) *forget* gate - determines which parts of the previous cell state should be retained or discarded; (ii) *input* gate - comprising a logical gate to determine which dimensions to update and a candidate vector of new content (usually passed through a tanh[41] layer) to be potentially written to the cell state, the *input* gate decides which new information to add to the shared cell state; (iii) *output* gate - it filters the cell-state content and then applies a non-linear transformation (e.g., tanh) to produce output activation for deciding which parts of the update cell state should be exposed to the next layer. The **Bi-directional** configuration enhances this by processing the text in both the forward and backward directions, allowing the model to incorporate context from both preceding and succeeding words before making a final classification.

As discussed earlier in this Section 2.4, Vaswani et al. (2017)'s self-attention mechanism in the transformer *encoder-decoder* architecture allows for parallel processing and generation of tokens (words) in a sequence (sentence) regardless of their position. Leveraging the pre-trained *encoder*-only architectures such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), (Mali et al., 2025; Philipo et al., 2024; Paul and Saha, 2022) among others *fine-tune* these pre-trained language models on sequence (sentence) classification tasks. They do so by attaching a classification head on top of the '`[CLS]`' token's final hidden state representation ($h_{[\text{CLS}]}$) and training this combined architecture using appropriate activation and loss functions for *supervised learning*.

The classification head architecture for modern Transformer models, such as those proposed by (Devlin et al., 2019), typically consists of **three** key components that work together to transform pre-trained representations (e.g., BERT embeddings) into task-specific predictions: (i) *intermediate transformation layers* that learn non-linear mappings from the pre-trained `[CLS]` representation to a task-relevant feature space; (ii) *stabilisation and regularisation mechanisms* that prevent

ficult for standard RNNs to learn (Bengio et al., 1994; Jozefowicz et al., 2015)

[41] A non-linear activation function that maps values to the range [-1, 1], enabling the network to model complex relationships and stabilising the scale of cell-state updates

over-fitting during fine-tuning by constraining model complexity; (iii) *an output layer* that produces class probabilities using task-appropriate activation functions.

(i) *Intermediate Transformation Layers*: For a an input (x_i), only the `[CLS]` representation from the final *encoder* layer passes through one or more fully connected (dense) layers with activation functions that introduce non-linearity essential for learning complex decision boundaries. In deep neural networks such as 12-24 layer transformers, the choice of activation function critically affects gradient flow during back-propagation (Vaswani et al., 2017). The input to an activation function (z) is the weighted sum ($z = Wx + b$), where b is the bias term. One of the activation functions, such as, sigmoid activation is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.15}$$

This has derivative $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ that approaches zero for large positive or negative inputs (e.g., when $z = 5$, $\sigma'(5) \approx 0.007$), causing gradients to diminish exponentially through multiple layers (vanishing gradient problem). This makes sigmoid unsuitable for hidden layers in deep architectures. In contrast, **ReLU (Rectified Linear Unit)** activation (Nair and Hinton, 2010), defined as

$$\text{ReLU}(z) = \max(0, z) = \begin{cases} z & \text{if } z > 0 \\ 0 & \text{if } z \leq 0 \end{cases} \tag{2.16}$$

Helps maintains a constant gradient of 1 for positive inputs ($\text{ReLU}'(z) = 1$ when $z > 0$), enabling efficient gradient flow and approximately and approximately 6× faster training compared to sigmoid networks (Krizhevsky et al., 2012). ReLU's computational efficiency (requiring only a simple thresholding operation rather than expensive exponential computations) and its promotion of sparse activations (by zeroing negative values) further improve both training speed and generalisation. The hidden layer transformation is therefore:

$$h_{\text{hidden}} = ReLU(h_{[\text{CLS}]}W_1 + b_1) \tag{2.17}$$

where W_1 and b_1 are learnable weight matrix and bias vector that adapt the pre-trained representation to the target task.

(ii) *Stabilisation and Regularisation Mechanisms*: In transfer learning, (Ruder et al., 2019; Howard and Ruder, 2018) highlight these dense layers incorporate multiple regularisation and optimising techniques to prevent overfitting, which is a critical concern when fine-tuning large pre-trained models on smaller task-specific datasets. **Dropout** (Srivastava et al., 2014) is one such regularisation mechanism that randomly sets neuron activations to zero during training with probability p (typically $p \in$[0.1,0.3]), effectively training an ensemble of sub-networks by forcing the model to learn distributed representations across many neurons rather than relying on specific pathways. Mathematically, during training, as (Srivastava et al., 2014) note, each activation h_i is multiplied by a Bernoulli random variable:

$$h_i^{\text{dropout}} = h_i \cdot \text{Bernoulli}(1-p) \tag{2.18}$$

At test time, activations are scaled by $(1-p)$ to maintain the expected value. This stochastic regularisation reduces co-adaptation between neurons and decreases variance in predictions without significantly affecting the mean, making the model more robust to input changes (Srivastava et al., 2014). **Layer normalisation** (Ba et al., 2016) is an optimisation technique that stabilizes training by normalising activations across the feature dimension for each instance independently,

$$\text{LayerNorm}(x) = \gamma \odot \frac{x-\mu}{\sqrt{\sigma^2+\epsilon}} + \beta \tag{2.19}$$

Here $\mu = \frac{1}{d}\sum_{i=1}^{d} x_i$ and $\sigma^2 = \frac{1}{d}\sum_{i=1}^{d}(x_i-\mu)^2$ are the mean and variance computed across the d features. ϵ (usually set to 10^{-5}) prevents division by zero. γ is a learnable scaling vector that allows the network to set the optimal variance for the features (override variance= 1 from regularisation) and β is a

learnable shifting vector that allows the network to set the optimal mean (bias) for the features (override mean= 0 from regularisation). Layer normalisation by re-centering and re-scaling activations accelerates convergence and allows higher learning rates, thereby stabilising the problem where the distribution of layer inputs changes during training as parameters in previous layers update (covariate shift problem) (Ba et al., 2016). Another regularisation mechanism is the **weight regularisation** or L2 weight decay (also called ridge regression in the context of linear models). It adds a penalty term to the loss function that shrinks weight magnitudes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \sum_i w_i^2 \tag{2.20}$$

Here λ controls the regularisation strength. This penalty encourages the model to prefer smaller weights, which corresponds to smoother decision boundaries and reduces the model's capacity to fit noise in the training data. The effect is to reduce the variance of predictions (making the model less sensitive to small changes in input) while maintaining the overall bias, thereby improving generalisation to unseen data. In the context of fine-tuning transformers, Ruder et al. (2019) recommend *discriminative fine-tuning*, where different learning rates are applied to different layers (lower rates for early layers that capture general features, higher rates for task-specific head layers), and *gradual unfreezing*, where layers are progressively unfrozen from top to bottom during training to prevent catastrophic forgetting of pre-trained knowledge.

(iii) *Output Layer*: The output layer maps the regularised hidden representation to class *logits*, which are then converted into probabilities using task-appropriate activation and loss functions. For binary classification tasks (e.g., `CB (1)` vs `not-CB (0)`), sigmoid activation $\hat{y} = \sigma(\text{out}) = \frac{1}{1+e^{-z_{out}}}$ produces a probability $\hat{y} \in (0, 1)$. Goodfellow et al. (2016) note, Cross Entropy (CE) measures the difference between the predicted probability distribution $\hat{y}$ and the true

distribution y over C classes, and for a single sample it is defined as

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log \hat{y}_c, \tag{2.21}$$

i.e., y_c is 1 if the sample belongs to class c (0 otherwise), and $\hat{y}_c$ is the predicted probability for class c. The negative log-likelihood formulation penalises incorrect predictions proportionally to the model's confidence, i.e., predicting $y_c \log \hat{y}_c = \hat{y}_{true} = 0.1$ for the correct class incurs loss $-\log(0.1) \approx 2.30$, while $\hat{y}_{true} = 0.9$, incurs a loss of only $-\log(0.9) \approx 0.11$, encouraging high-confidence correct predictions. For **multi-label classification** where instances can have multiple labels simultaneously (e.g., text containing both racist and sexist content), each of the C labels uses an independent sigmoid activation, with CE applied to each label:

$$\mathcal{L}_{\text{CE-ML}} = -\frac{1}{C}\sum_{c=1}^{C}[y_c \log(\hat{y}_c) + (1 - y_c)\log(1 - \hat{y}_c)].$$

Lin et al. (2020) highlight, it is important to note that CE penalises the model more when it assigns low probability to the correct class. However, as it treats all samples equally, in imbalanced datasets, the majority classes dominate training. To circumvent this, (Lin et al., 2020) introduces Focal Loss (FL), a modulating factor that down-weights well-classified, "*easy*" examples, thereby forcing the model to focus more on hard-to-classify, i.e., minority-class samples. For a single sample, (Lin et al., 2020) define it as,

$$\mathcal{L}_{FL} = -\alpha_t(1 - \hat{y}_t)^{\gamma} \log \hat{y}_t \tag{2.22}$$

Where $\hat{y}_t$ is the predicted probability of class t, $\alpha_t \in [0, 1]$ is a weighting factor for class t to address class imbalance, and $\gamma \geq 0$ is the focusing parameter that controls the down-weighting of easy examples. As (Lin et al., 2020), describe $\gamma = 0$, FL reduces to standard cross-entropy, and as γ increases, the loss for

easy examples ($\hat{y}_t \to 1$), the modulating factor $(1 - \hat{y}_t)^\gamma \to 0$ significantly down-weights the loss contribution. For hard examples or minority-class examples, ($\hat{y}_t \to 0.5$ or less) the modulating factor $(1 - \hat{y}_t)^\gamma \to 1$ leaves the loss contribution nearly unchanged. This property makes FL particularly useful for handling class imbalance. During fine-tuning with these loss functions, gradient-based optimisers such as AdamW (Loshchilov and Hutter, 2017) are leveraged. This decouples weight decay from gradient updates, improving generalisation and updates both the classification head parameters and the pre-trained encoder weights through back-propagation, enabling the model to learn task-specific representations while preserving generally useful features learned during pre-training.

Studies outlined in Table 2.2 leverage the "classification head" described earlier for *encoder-only* architectures (e.g., BERT, RoBERTa) and claim to detect "cyberbullying". Their proposed systems also out-perform classical machine learning and RNN-based deep learning classifiers. However, a closer examination reveals that almost all studies conflate "cyberbullying" with hate-speech, personal attacks, or abusive language. This conflation or flattening arises as they leverage binary datasets labelled for peripheral behaviours such as hate speech by (Waseem and Hovy, 2016; Davidson et al., 2017) or personal attacks by (Wulczyn et al., 2017), or abusive or offensive language by (Founta et al., 2018; Zampieri et al., 2019) to detect *cyberbullying*. By training these sophisticated transformer encoder architectures on such peripheral datasets, the models tend to capture general toxicity, but fail to reliably distinguish actual instances of cyberbullying from other forms of online harms. As cyberbullying exhibits unique relational and contextual characteristics that differ from other harmful content (See Section 2.2), these training approaches blur boundaries and complicate both the evaluation of model performance and the development of targeted interventions (Emmery et al., 2021; Ziems et al., 2020; Kim et al., 2021b).

In the past decade, Dinakar et al. (2011); Sanchez and Kumar (2011); Bayzick

Study	Dataset	Dataset Type	Harm Type
Phanomtip et al. (2021)	Waseem and Hovy (2016); Wulczyn et al. (2017); Reynolds et al. (2011)	Binary	Personal Attacks; Binary Bullying; Hate-speech
Zhao et al. (2020)			
Aggarwal et al. (2020)			
Paruchuri and Rajesh (2022)			
Yadav et al. (2020)	Agrawal and Awekar (2018); Davidson et al. (2017); Waseem and Hovy (2016); Chatzakou et al. (2019); Bretschneider and Peters (2016)	Binary	Hate-speech; Aggression; Offensive language
Pradhan et al. (2020)	Wulczyn et al. (2017); Reynolds et al. (2011); Waseem and Hovy (2016)	Binary	Hate-speech and Personal attacks
Fang et al. (2021)	Reynolds et al. (2011); Waseem and Hovy (2016)	Binary	Hate-Speech; Binary Bullying
Tripathy et al. (2022)	Davidson et al. (2017)	Binary	Hate-Speech
Muneer and Fati (2020)	Founta et al. (2018)	Binary	Hate-Speech
Behzadi et al. (2021)			
Paul and Saha (2022)	Waseem and Hovy (2016); Davidson et al. (2017)	Binary	Hate-Speech; Offensive language
Agrawal and Awekar (2018)			
Wang et al. (2020)			
Elsafoury et al. (2021)			
Bharti et al. (2021)			
Dadvar and Eckert (2020)	Waseem and Hovy (2016)	Binary	Hate-speech
Raj et al. (2021)			
Alotaibi et al. (2021)			
Menini et al. (2019)			
Lu et al. (2020)	Waseem and Hovy (2016); Wulczyn et al. (2017); Kaggle (2018)	Binary	Hate-speech; Offensive language
Jain et al. (2021)	Waseem and Hovy (2016); Wang et al. (2020)	Binary	Hate-Speech; Offensive language
Ahmed et al. (2021)	Wulczyn et al. (2017)	Binary	Personal attacks
Pericherla and Ilavarasan (2021)	Wulczyn et al. (2017)	Binary	Personal attacks
Chatzakou et al. (2019)	Custom (No Annotation Guideline)	Binary	Hate-speech & Abusive language
Ahmed et al. (2022)	Wang et al. (2020); Waseem and Hovy (2016)	Binary	Hate-speech & Offensive language

Table 2.2: List of studies that claim to detect "cyberbullying" but use out-of-domain datasets

et al. (2011) created the earliest binary CB datasets by gathering comments from SMPs like YouTube[42], Twitter[43], and defunct SMPs like MySpace[44] and Formspring.me[45]. Despite a failure to capture the multifaceted nuances of CB, these datasets hold significance as they were among the initial binary CB datasets to be annotated by two or more human annotators. Additionally, Bretschneider and Peters (2016) collected text data from discussion forums of two widely played online multiplayer games, World of Warcraft (WoW)[46] and League of Legends (LoL)[47] while building a rule-based cyberbullying detection system. In their study, this dataset underwent initial binary annotation by three human annotators to identify instances of harassment, and subsequently it was further annotated to establish the binary labels for cyberbullying roles like `0 (victim)` or `1 (bully)`. While these efforts laid the groundwork for automatic detection, the binary labelling scheme and limited contextual coverage continue to constrain model generalisation.

Nevertheless, binary-labelled datasets studied in systematic literature reviews (Rosa et al., 2019; Emmery et al., 2021; Kim et al., 2021b; Teng et al., 2024) continue to hold substantial value for advancing cyberbullying research, particularly when developed under rigorous annotation and validation protocols. Such datasets are most effective when: (i) annotations follow established cyberbullying definitions by (Patchin and Hinduja, 2006; Kowalski et al., 2012; Hinduja and Patchin, 2013; Smith et al., 2013), (ii) content is labelled by multiple annotators or social science experts, (iii) inter-annotator agreement is measured, or disagreement patterns are analysed and discussed. Adhering to these standards ensures that datasets accurately capture cyberbullying-specific nuances, rather than conflating it with general abusive language.

However, when constructing cyberbullying datasets, researchers have often not

[42] `https://www.youtube.com/`
[43] `https://www`
[44] `https://www.myspace.com`
[45] `https://en.wikipedia.org/wiki/Spring.me`
[46] `https://worldofwarcraft.com/en-gb/`
[47] `https://www.leagueoflegends.com/en-us/news/community/saying-farewell-to-boards/`

adhered to formal definitions of cyberbullying, resulting in diverse behaviours being grouped under a single "*cyberbullying*" label and highlighting the need for greater standardisation. Building on this and the systematic reviews by (Kim et al., 2021b; Emmery et al., 2021; Teng et al., 2024), existing datasets can be categorised based on the characteristics of the platforms from which they are collected. Platforms that enable users to comment or reply to multimodal content (e.g., Instagram) can be classified as *User-Comment* sources. Those that support anonymous question-and-answer interactions, such as the now-defunct ASK.fm[48] or Formspring.me[49], fall under *Question-Answering* sources. Meanwhile, platforms like Twitter (now X) and other micro-blogging services, which allow users to post and reply publicly, can be categorised as *Micro-blogging* sources. As illustrated in Table 2.3, of the 12 datasets, only eight datasets follow established cyberbullying definitions among their annotation schema. Moreover, three studies by (Reynolds et al., 2011), (Xu et al., 2012), and (Edwards et al., 2020) do not discuss or present any inter-annotator agreement or disagreement analysis. Furthermore, some datasets, such as those curated by Chatzakou et al. (2019) from X using offensive keywords for topics like #GamerGate, #BBCpay, and #NBA, employ role-based labelling (`bully`, `aggressor`, `normal`, `spammer`) that conflates abusive language with cyberbullying, highlighting the importance of precise annotation protocols (Rosa et al., 2019; Emmery et al., 2021; Kim et al., 2021b).

Among micro-blogging datasets, X (previously Twitter) remains a particularly significant source. Xu et al. (2012) collected 1,762 tweets encompassing a variety of bullying scenarios, including direct attacks, self-disclosures as victims, accusations, and reports of bullying. Each tweet was annotated for Task-A (`bullying` vs. `not-bullying`) and role-based labels (`Accuser (A)`, `Bully (B)`, `Reporter (R)`, `Victim (V)`) by multiple annotators. Rule-based classifiers using Naïve Bayes, Logistic Regression, and Support Vector Machines were trained via Weka (Frank et al., 2016). While role-based labelling yielded strong performance for bullies and accusers, victim

[48] `https://www.esafety.gov.au/key-topics/esafety-guide/askfm`
[49] `https://spring.me/`

Platform	Study	Def. Used	#Annotators	IAA / Disagr. Analysis	Language
User-Comment	(Hosseinmardi et al., 2015)	✓	5	Disagr.	EN
	(Rafiq et al., 2015)	✓	5	Disagr.	EN
	(Van Bruwaene et al., 2020)	✓	3	IAA (κ = 0.247)	EN
Question-Answering	(Van Hee et al., 2018)	✓	>=2	IAA (κ:0.59-0.69)	EN & NL
	(Reynolds et al., 2011)	✗	3	✗	EN
Micro-blogging	(Xu et al., 2012)	✓	Same as TREC 2011[a]	✗	EN
	(Chatzakou et al., 2019)	✗	5	Disagr.	EN
	(Ziems et al., 2020)	✓	3	IAA κ=0.18	EN
	(Salawu et al., 2021)	✓	17 young-adults	IAA (α=0.67)	EN
Online Gaming	(Bretschneider and Peters, 2016)	✓	3	IAA (κ=0.51)	EN
Messaging	(Edwards et al., 2020)	✗	Not clear	✗	EN
	(Sprugnoli et al., 2018)	✓	2 Experts	IAA (Dice Coefficent=0.8)	IT

Table 2.3: Dataset Studies grouped by source platform. Note: Disagr. → Disagreement Analysis; IAA → Inter-Annotator Agreement; EN → English; NL → Dutch; IT → Italian;

[a] `https://trec.nist.gov/data/microblog2011.html`

identification was moderate, particularly when victims were referenced implicitly through pronouns, illustrating the challenges in capturing nuanced cyberbullying behaviour.

A more recent large-scale effort by Salawu et al. (2021) introduced a dataset of 62,587 tweets in English covering bullying and offensive content such as insults, trolling, profanity, sarcasm, threats, pornography, and exclusion. Annotated by 17 social-media-frequent, high-school-educated annotators (three per tweet), the dataset achieved a Krippendorff's α of 0.67 for binary bullying labels. Salawu et al. (2021) evaluated downstream transformer-based classifiers, including BERT,

RoBERTa, DistilBERT (Sanh et al., 2019), and XLNet (Yang et al., 2019), finding that fine-tuned BERT achieved a macro-F1 score of 0.632. This demonstrates that pre-trained language models (PLMs) can effectively leverage high-quality binary-labelled datasets for cyberbullying detection, directly addressing the RQ regarding suitable classification paradigms.

Other studies have explored multi-source datasets and cross-platform generalisation. Edwards et al. (2020) combined SMS messages from youth aged 10–14, Formspring.me posts, and Twitter content, focusing on cyberbullying around the 2016 U.S. elections. Using machine learning models such as SVM, Decision Trees, Random Forest, and AdaBoost, systems achieved macro-F1 scores of 0.76–0.79. However, the narrow topical focus and reliance on a defunct platform limit its generalizability. To improve cross-platform applicability, Nikhila et al. (2020); Yi and Zubiaga (2022) investigated deep learning and PLM-based approaches for training on one source and testing on another. Nikhila et al. (2020) addressed class imbalance via oversampling strategies, including Synonym Replacement using Word2Vec or FastText embeddings, and General Adversarial Network (GAN) based data generation to synthesise realistic minority-class samples. When classified with a 1D CNN on stacked Flair and GloVe embeddings, GAN-based oversampling significantly improved performance. Meanwhile, Yi and Zubiaga (2022) proposed XP-CB, a transformer-based cross-platform framework with dual alignment between source and target embeddings, incorporating input length optimisation, hidden state selection, adaptive batch normalisation, and adversarial alignment using GAN and ADDA-based modules. Evaluated on Formspring, Twitter, and Wikipedia datasets in zero-shot settings, XP-CB demonstrated superior cross-platform generalisation compared to baseline models, highlighting the potential of PLM-based architectures for robust cyberbullying detection. However, the continued reliance of these sophisticated approaches on datasets originally curated for *hate speech*, *personal attacks*, or *offensive/abusive* language often blurs the boundaries, making it challenging to detect the nuanced behaviours specific to cyberbullying accurately.

Overall, several gaps emerge: (i) cyberbullying is often conflated with other forms of antisocial online behaviour, such as hate speech, personal attacks, and abusive or offensive language; (ii) there is limited benchmarking of different classification architectures (e.g., encoder- vs. decoder-only transformers) across platforms to assess cross-platform generalisation; and (iii) treating cyberbullying as a binary phenomenon oversimplifies its behavioural spectrum and obscures important distinctions in participant roles, intent, and types of online harm.

2.4.3 Fine-grained Cyberbullying Detection

Researchers have underscored the importance of examining the dynamics of social interaction in order to better understand cyberbullying (CB) behaviours (Lee and Shin, 2017; Van Royen et al., 2017; DeSmet et al., 2018; Leung et al., 2018; Song and Oh, 2018). For instance, a tweet such as "*@BernieSanders has come off the rails. He's unhinged. He's attempting to burn the house down over nothing. #ImWithHer*" from Chatzakou et al. (2019), while consists of offensive and critical in tone, the intent can be best understood as a political attack and does not demonstrate targeted or repeated harm towards an individual in the sense required to constitute CB. In contrast, the message "*your mum is a b***h she deserves the death penalty for giving birth to you. if u come back, u should get one too.*" from Van Hee et al. (2018) is a clear instance of cyberbullying as it directly targets a specific individual (`victim`), conveys an explicit intent to cause serious harm, and carries an implied threat of continued harassment.

Xu et al. (2012) introduced the notion of "*bully traces*" by examining the emotions expressed in tweets containing specific keywords. Their study approached the task at two levels: (i) a binary decision on whether a tweet was related to CB, and (ii) a more detailed classification of participant roles, including *accuser*, *bully*, *victim*, *reporter*, and others. To identify such traces, they relied on keywords such as `bully`, `bullied` and `bullying`. These were often flagged posts that describe or report incidents (e.g., "*I observed a girl being bullied at school today #bullyingisnotcool*")

labelled as *reporter*. This marked an important step towards moving beyond binary detection by foregrounding the social dynamics of interactions among users. Subsequent studies drew heavily on this approach, inspiring the creation of specialised datasets for CB analysis. However, two key limitations persist: (i) the dependence on a narrow set of keywords, and (ii) its insufficient attention to the diversity of cyberbullying forms, which constrains its ability to capture the full spectrum of CB behaviours discussed in §2.2.1 and §2.3.3.

To address these limitations, Van Hee et al. (2018) in their dataset collection process gathered message interactions from a now-defunct question-answering platform (ASK.fm) and annotated them for multiple participant roles and various forms of cyberbullying (CB) (see §2.3.3). Their research centered on curating datasets in both English and Dutch languages, featuring annotations that encompassed a range of participant roles, including `bully`, `victim`, `bystander-enabler` (`enabler`), `bystander-defender` (`defender`). Additionally, they covered different types of CB, such as `insults, threats, curses, defamation`, and `sexual talk`, among others. Although their English and Dutch language datasets span 113,698 and 78,387 posts respectively, as depicted in Table 2.4, the labels presented in the English dataset only capture a very small number of positive cases for different forms of CB, as low as 0.03% to 2%. Van Hee et al. (2018) also developed a baseline binary classifier using Support Vector Machines (SVM) on text-based features like n-gram, bag-of-words (BoW), and developed to identify bullying on the basis of aggressive language and profanity terms in messages. Through this feature and rule-based binary classification systems they achieved 64.32% and 58.72% F_1 score in detecting if a message is bullying or not. Despite the coarse categorisation of cyberbullying behaviour outlined in Table 2.4, Van Hee et al. (2018)'s English datasets provide only high-level participant-role labels. These role-based annotations enable researchers to move beyond simple `bully-victim` dichotomies and explore the more nuanced dynamics of online aggression. Building on this dataset, (Rathnayake et al., 2020) fine-tuned BERT (Devlin et al., 2019) to identify participant roles through a cascad-

ing classification model. Their approach first determines whether a message belongs to a bully or victim, and then applies a secondary classifier, i.e., messages identified as bullying are further distinguished as either bully or enabler, while victim-labelled messages are classified as either victim or defender. Using this hierarchical approach, they reported achieving F_1 scores ranging from 0.83 to 0.67 when distinguishing between bully–enabler and victim–defender roles. However, a closer inspection reveals a significant limitation. Their model consistently failed to identify enabler roles, with a recorded F_1 score of 0.00 for that class for their ensemble model (See Table 2 of (Rathnayake et al., 2020)). This highlights the inherent difficulty of recognising less explicitly defined roles in cyberbullying contexts, particularly those that involve indirect or subtle behaviours.

Similarly, Jacobs et al. (2022) addressed participant role classification but approached it as a multi-class classification task rather than a two-step cascading problem. Instead of first separating bully and victim messages, they directly treated role prediction as a downstream multi-label classification task. By fine-tuning and combining several pre-trained transformer models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019) into an ensemble framework, they demonstrated that this approach consistently outperformed baseline methods based on traditional machine learning algorithms such as SVM and logistic regression. Their ensemble model achieved an overall F_1 score of 0.5331, which, while still leaving room for improvement, underscores the advantages of leveraging transformer-based architectures over classical methods to capture the complex and context-dependent nature of the participant roles in cyberbullying. This suggests that future research should continue to refine transformer-based models, explore data augmentation strategies, and focus on improving the detection of under-represented roles such as enablers and defenders, which remain the most challenging to classify accurately.

In youth contexts, both the Italian (Sprugnoli et al., 2018) and French (Ollagnier et al., 2022) datasets, discussed in Section 2.3.1, were curated using the '*living lab*'

Category	Count	Percentage
General_insult	2521	2.43
Harmless_sexual_talk	1415	1.36
Sexual_harassment	625	0.60
Curse_Exclusion	376	0.36
Sexism	336	0.32
General_defense	274	0.26
Defamation	216	0.21
Racism	170	0.16
Threat_Blackmail	166	0.16
Attacking_relatives	138	0.13
Other	87	0.08
Assertive_selfdef	47	0.04
Powerless_selfdef	37	0.03
Participant Roles		
Harasser/Bully	3442	3%
Victim	1335	1.16%
Enabler	24	0.02
Defender	422	0.37%

Table 2.4: Multi-class cyberbullying label distribution in Van Hee et al. (2018)

role-play methodology. Within this method, participant roles in cyberbullying interactions were pre-defined. Building on this, Ollagnier et al. (2023) developed a system to automatically infer the sender's role within cyberbullying identification tasks. Their approach relied on a binary 'hate-speech' classifier in combination with a heuristic that considered two factors: (i) *repetition*, measured by how frequently a participant's messages were labelled as harmful, and (ii) *intention*, approximated by the proportion of harmful content within an individual's total messages. These were aggregated into a single metric expressed as `role-score` $= \frac{repetition+intention}{2}$. This formulation allowed the authors to approximate the bully–victim dynamics of online interactions, offering a useful first step toward role attribution in cyberbullying contexts. The trade-off is that by focusing primarily on direct harm signals, this approach overlooks the nuanced social and relational complexities inherent in cyberbullying. This makes their framework limited in its scope and applicability. For instance, online exchanges are often shaped by escalation patterns where initial teasing may intensify into sustained aggression or power asymmetries, which shift over time as participants gain or lose influence, and bystander behaviours, where peers may either reinforce or mitigate harmful acts through intervention or silence

(Hinduja and Patchin, 2013; Englander et al., 2017; Steer et al., 2020; Macaulay et al., 2022). Furthermore, elements such as sarcasm and humour that may downplay aggression play a critical role in shaping how participants perceive and respond to cyberbullying, yet remain absent from their heuristic approach.

Recent studies such as Salawu et al. (2021) and Ziems et al. (2020) have also attempted to move beyond binary classification by curating fine-grained datasets of abusive and bullying behaviours from Twitter. Salawu et al. (2021) proposed a multi-label scheme that included categories such as *bullying*, *insult*, *profanity*, *sarcasm*, *threat*, *exclusion*, *porn*, and *spam*. Their dataset was annotated by a large pool of annotators (N=17) from diverse backgrounds within the age group of 18-35. This yielded in an inter-rater reliability of Krippendorff's $\alpha = 0.67$. Beyond dataset construction, they benchmarked a range of models, from classical machine learning algorithms to transformer-based architectures including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019). Among these, BERT achieved the strongest performance with a macro F_1 score of 0.6318, underscoring the advantage of deep contextualised representations for this task. Salawu et al. (2021) note that minority categories such as *sarcasm*, *exclusion*, and *threat* are severely under-represented in their dataset, which limits model performance on precisely those behaviours most central to cyberbullying. This imbalance risks reinforcing the conflation of offensive language with bullying and since models learn to prioritise majority categories like *profanity* or *insult* and overlook key definitional aspects of cyberbullying highlighted in datasets by Van Hee et al. (2018) and Sprugnoli et al. (2018).

Ziems et al. (2020) approached this problem of paucity in cyberbullying datasets differently by seeking to operationalise core definitional elements of cyberbullying. They first collected 1.3 million unlabelled tweets from the Twitter Filter API and then used the hate speech and aggression classifier of Davidson et al. (2017) to filter this pool down to 9,803 tweets containing aggressive language. Their assumption was that aggression is a necessary precursor for cyberbullying and that enriching for

such content would increase the prevalence of bullying examples in the final dataset. They then developed an annotation scheme around five conceptual criteria, such as *aggression*, *repetition*, *harmful intent*, *visibility among peers*, and *power imbalance*. They employed three Amazon Mechanical Turks (MTurks) per item to label the filtered tweets. The resulting annotations revealed strong skew across categories. While aggression was present in 74.8% of cases, repetition and harmful intent appeared in only 6.6% and 16.1% of tweets, respectively. Inter-annotator agreement also varied considerably, from relatively low values for aggression (α=0.23) and repetition (α=0.18) to more moderate agreement for harmful intent (α=0.42) and peer visibility (α=0.51). Correlation with overall cyberbullying labels was strongest for harmful intent (0.68) but negligible for power and visibility measures, underscoring the difficulty of capturing complex social dynamics from isolated tweets. Despite their ambition to ground annotations in conceptual definitions of cyberbullying, two limitations stand out. First, their reliance on an existing hate-speech classifier to filter tweets implicitly biases the dataset toward overtly aggressive language, thereby excluding more subtle yet equally harmful forms of bullying such as *exclusion*, *relational aggression*, or *sarcasm*. Second, their annotation results show that while aggression is easy to capture, key definitional aspects of cyberbullying such as *repetition*, *power imbalance*, and *peer involvement*, were both rare in their sampled data and difficult for annotators to agree on. As such, this dataset risks reinforcing the conflation of aggression with cyberbullying, rather than disentangling the nuanced interplay of behaviours(Lee and Shin, 2017; Van Royen et al., 2017; DeSmet et al., 2018; Leung et al., 2018; Song and Oh, 2018).

Overall, datasets exclusively from Twitter (Xu et al., 2012; Salawu et al., 2021; Ziems et al., 2020) contain many annotated tweets that are now deleted and hinder the reproducibility and undermine their long-term utility for the research community. Their annotations often capture offensive or aggressive content without necessarily aligning with the definitional criteria of cyberbullying. The absence of demographic information from some of these datasets (Van Hee et al., 2018; Xu

et al., 2012; Salawu et al., 2017; Ziems et al., 2020) limits the capacity of such datasets to inform youth-centred understanding of online harm. In contrast, role-played approaches such as those documented in Sprugnoli et al. (2018); Ollagnier et al. (2022) demonstrate stronger alignment with conceptual definitions of cyber-bullying, but face their own challenges of class imbalance and ethical complexity.

2.4.4 Multi-task Cyberbullying Detection

In tasks such as cyber-bullying detection, there is growing interest in multi-task learning (MTL), i.e., using deep learning architectures to jointly model several related text-classification objectives (for example, binary hurt detection alongside multi-class role attribution) (e.g., (Xingyi and Adnan, 2024; Maity et al., 2022)). While there have been efforts to treat sarcasm detection as part of a multi-task setup in bullying contexts (e.g., (Guo and Gauch, 2024)), the authors themselves note that the paucity of annotated datasets means these setups cannot yet be thoroughly evaluated. However, in closely related work on online abuse detection (not strictly cyberbullying), the Semantic Evaluation (SemEval) 2020 OffensEval by Zampieri et al. (2020), defined a structured set of subtasks (binary offensive vs. non-offensive, multi-class target identification or type of offence) that invited MTL approaches. One such system (Dai et al., 2020) fine-tuned a single BERT encoder jointly on the hierarchical subtasks, with separate task heads on the shared representation. They found that sharing representations across the subtasks improves robustness on noisy social media text by allowing related supervision signals to regularise the learned representations.

A more recent challenge, SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023), presented a three-level hierarchical taxonomy of sexist content (Task-A: binary sexism detection; Task-B: four-class category of sexism; Task-C: 11 fine-grained sub-classes for Task-B) with a dataset of ≈20,000 social-media posts with fine-grained labels. While our traditional machine learning baselines discussed in (Verma et al., 2023a), which included k-Nearest Neighbour

(k-NN) and Logistic Regression (LR) classifiers trained on document embeddings derived from fine-tuned transformer models (encoder: HateBERT (Caselli et al., 2021); decoder: OPT (Zhang et al., 2022c)), achieved only modest results (Task B: 51.04 F1; Task C: 37.95), several top-performing systems in the EDOS challenge adopted more advanced modelling paradigms. These approaches incorporated joint or hierarchical architectures, task-adaptive pre-training, and adapter-based or multi-task learning frameworks (Kirk et al., 2023). In particular, (Pickard et al., 2023) demonstrated the effectiveness of AdapterFusion (Pfeiffer et al., 2021), an extension of adapter-based fine-tuning (Houlsby et al., 2019). Instead of updating all model parameters, small task-specific adapter modules are inserted within each transformer layer, allowing efficient fine-tuning without altering the shared base model (Pfeiffer et al., 2021). In online abuse detection tasks, two transfer learning approaches are particularly popular for multi-task learning, i.e., *hard-parameter sharing* and *adapter fusion* techniques.

Hard Parameter Sharing: As (Ruder, 2017) notes, hard parameter sharing is the canonical MTL formulation in deep learning: a single encoder g_θ: $\mathcal{X} \rightarrow \mathcal{H}$ (e.g., BERT) produces a representation $h = g_\theta(x)$, and task-specific heads $\{d_{\phi i}\}$ to map h to task outputs. Specifically, for two tasks (task-1: binary classification; task-2: multi-class classification), (Ruder, 2017) write,

$$\hat{y}_A = f_A(h) = softmax(W_A h + b_A); \quad \hat{y}_B = f_B(h) = softmax(W_B h + b_B);$$

With shared parameters θ and head parameters $\phi_A = \{W_A, b_A\}$; $\phi_B = \{W_B, b_B\}$. As (Ruder, 2017) notes, this hard sharing reduces over-fitting and parameter cost and lets auxiliary labels improve representation learning, but it can cause negative transfer when one task (often the binary task) dominates gradients and drowns signals needed for rarer, fine-grained classes (e.g., enabler/defender). As (Chen et al., 2024b) notes, this hard-parameter sharing or joint-parameter training is a quite popular approach, in tasks such as online abuse detection research. (Mishra

et al., 2021; Dai et al., 2020) among others, use this technique and achieve near state-of-the-art performances. However, as (Ruder, 2017) notes, this MTL-approach for binary and multi-class can cause negative transfer when one task (often the dominant binary task) overwhelms the gradients needed for rarer, fine-grained classes.

AdapterFusion (task adapters + learned fusion): Adapters originally proposed by (Houlsby et al., 2019), are small, bottleneck neural network modules inserted into a pre-trained transformer layer (e.g., after the Multi-Head Attention or Feed-Forward Network) that only contain a small fraction of the model's parameters. This allows for parameter efficient fine-tuning by training only these new modules and the task-specific head, while keeping the base model weights frozen. AdapterFusion (Pfeiffer et al., 2021) extends this by replacing full fine-tuning with these small, task-specific adapter modules inserted into a pre-trained transformer; after training adapters for several tasks, a lightweight fusion module learns how to combine adapter outputs for a target task without destructive update of the base model. Formally, for layer l the adapter for task t computes $a_t^{(l)} = Adapter_t^{(l)}(h_l)$. AdapterFusion learns fusion weights $w^{(l)}$ (often a small network) that compute a composition $a_{fused}^{(l)} = \sum_t \alpha_t^{(l)} a_t^{(l)}$, where $\alpha_t^{(l)}$ are learned attention-style weights and injects $a_{fused}^{(l)}$ into the transformer. This design permits non-destructive task composition and task reuse, making it attractive when combining diverse supervision (binary and multi-class). while avoiding catastrophic forgetting. AdapterFusion is effective and parameter-efficient in many transfer and MTL setups (Pfeiffer et al., 2021).

However, to date, there are no studies that have attempted to benchmark these two multi-task classification methods in cyberbullying contexts.

2.4.5 Multi-modal Cyberbullying Detection

Early studies by Hosseinmardi et al. (2015) and Rafiq et al. (2015) curated one of the first binary *multi-modal* cyberbullying datasets in the English language. In their

work, Hosseinmardi et al. (2015) extracted approximately 2,260 images and their associated comments from the image-sharing platform Instagram. Meanwhile, Rafiq et al. (2015) gathered nearly 959 videos and their associated comments from the now-defunct video-sharing platform Vine. Across both platforms, a "*post*" refers to an image or video uploaded by a user together with its accompanying textual caption. These datasets were annotated with binary labels for both cyberbullying and cyber-aggression, with Hosseinmardi et al. (2015) further including image-level labels (e.g., "*person*", "*text*") and Rafiq et al. (2015) introducing sentiment-based video annotations. As depicted in Figure 2.1, the image-based labels in the Hosseinmardi et al. (2015) Instagram dataset are dominated by the category "person", followed by "text" or "sports". This indicates that most images depict either individuals or textual content such as quotes and memes[50]. Likewise, Figure 2.1 illustrates that the emotion-sentiment labels in the Rafiq et al. (2015) Vine dataset are overwhelmingly skewed toward 'neutral' and 'joy', with only a small number categorised as 'fear' or 'love'. Both Hosseinmardi et al. (2015) and Rafiq et al. (2015) leveraged the then crowdsourcing platform CrowdFlower[51] to recruit annotators for content labelling. Posts and their associated comment threads were annotated in its entirety using a binary labelling schema: (a) bullying vs non-bullying, and (b) aggression vs non-aggression. Each item (post+comments) were labelled by at least three annotators, with the majority vote determining the final label.

Despite their novel efforts, neither Hosseinmardi et al. (2015) nor Rafiq et al. (2015) report the reliability between annotators, leaving the consistency and robustness of the annotations unclear. Furthermore, their labelling approach collapses potentially diverse interactions in posts and comments into one single binary label. This obscures the important distinctions between the post itself and the conversational dynamics that unfold in the comments. For instance, a benign image may attract hostile comment threads, or vice-versa where an aggressive caption may be

[50]Memes are humorous or interesting user generated content typically an image or video overlaid with text.

[51]This platform has changed over the years, it is now referred to as Figure Eight or Appen `https://www.appen.com/`

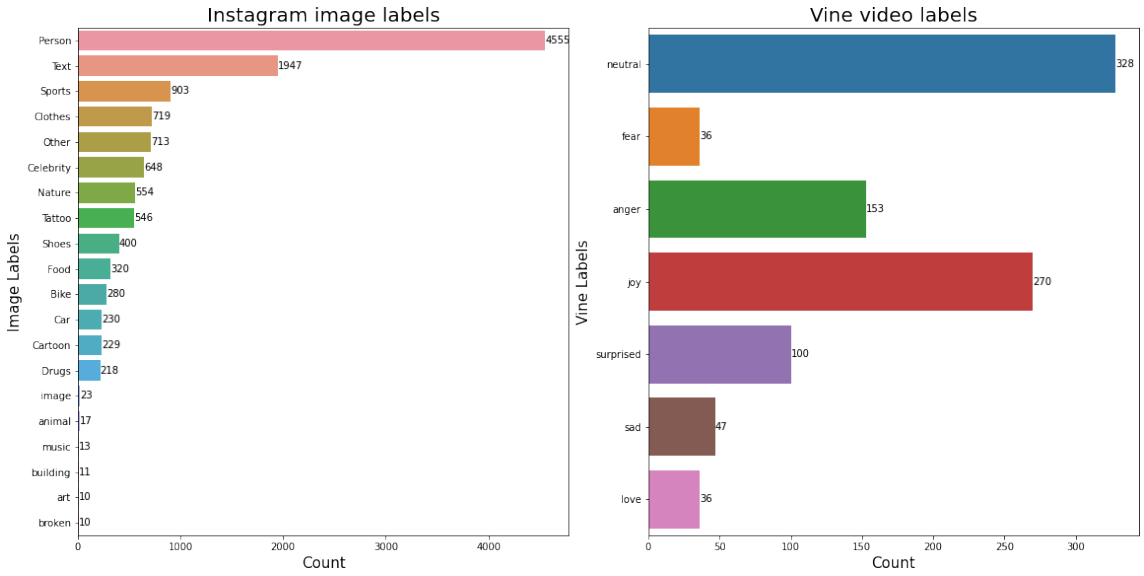


Figure 2.1: Distribution of labels in Hosseinmardi et al. (2015); Rafiq et al. (2015)

met with neutral or supportive responses. By reducing these layered interactions to a single label, the datasets risk oversimplifying the complex, multi-actor nature of cyberbullying and overlook exclusionary practices (e.g., tagging or group targeting) that occur within the comment structure rather than in the post content alone. Moreover, reproducibility of Hosseinmardi et al. (2015) is limited. When attempting to use the Instagram dataset, we were unable to align 1,476 out of the 2,259 posts (image+caption pairs) with their corresponding comments, an issue acknowledged by the dataset creators themselves (Hosseinmardi, 2022).

Despite these limitations,Cheng et al. (2019) and Chen and Li (2020) capitalised on the textual elements of this dataset and developed the Hierarchical Attention Networks for Cyberbullying Detection (HANCD) architecture and the HEterogeneous Neural Interaction Networks (HENIN) respectively to aid CB detection. Cheng et al. (2019) leveraged a bi-directional Gated Recurrent Unit (GRU) to capture contextual information from comments and hierarchical attention layers to assess the relevance of each comment within a social media post. Chen and Li (2020) leveraged a bi-directional GRU with an attention mechanism coupled with RNN self-attention networks to encode textual comments. While both Cheng et al. (2019); Chen and Li (2020) leverage user-post interactivity features like *user likes, user followers, user following* information for multi-modality, they do not leverage image features such

as image objects, image segments, or other image attributes to inform the decision-making process, thereby their approaches cannot be considered truly multimodal. However, Chen and Li (2020) introduced a novel post-comment co-attention mechanism that operates at the word embedding level, allowing for mutual interactions between the posted captions and comments to be learned through co-attention mechanism. Despite their novel approaches, both the performance and effectiveness of experiments carried out by Cheng et al. (2019); Chen and Li (2020) do not account for the *image* or *video* content and vary when applied to different social media platforms or datasets, as they are platform-specific.

Qiu et al. (2022) on multimodal cyberbullying detection leverages a combination of convolutional neural networks (CNNs) (Krizhevsky et al., 2012), Tensor Fusion (Zadeh et al., 2017), and VGG-19 networks (Simonyan and Zisserman, 2014), to construct a multi-layer perceptron model capable of fusing image and text features for classifying if content is `CB` or `not-CB`. However, their approach involves training the model on an abusive behaviour text-based dataset by Founta et al. (2018) and a sentiment labelled image-dataset by Vadicamo et al. (2017). This dependency on peripheral behaviour datasets to train deep learning architectures that claim to detect "cyberbullying" does not actually capture bullying itself, but rather instances of abusive or negative sentiment online behaviour. Maity et al. (2022) introduced "*MultiBully*", a multimodal meme dataset annotated with binary labels for bullying and sarcasm, three sentiment categories (positive, negative, neutral), and eight emotion categories, following the taxonomy of previous studies for both image and text content. The annotations were performed by three linguists, yielding Fleiss' Kappa (Fleiss, 1971) scores ranging from 0.69 to 0.82 across tasks, indicating moderate to substantial agreement. However, the study lacks reproducibility because the authors were unable to provide detailed annotator guidelines upon request (Maity, 2023). Furthermore, it is important to clarify that "MultiBully" is primarily an image dataset. Its characterisation as '*multimodal*' stems from the fact that many images contain overlaid text captions, rather than user-comments based on image

content as in (Hosseinmardi et al., 2015). This distinction has implications for how cross-modal interactions are interpreted in downstream modelling. However, Maity et al. (2022) developed a novel multimodal multi-task framework that leveraged contrasting learning to model for both text and image features. They developed a multi-task classifier with a combination of BERT with ResNET (He et al., 2016a) and CLIP (Radford et al., 2021). Leveraging CentralNet (Vielzeuf et al., 2018) as the fusion layer to embed both image and text features in a shared feature space, they were able to create task-specific classifier heads. On their dataset, this model architecture achieves an average F_1 of 0.7101 for binary bullying detection, 0.638 for sarcasm detection, 0.2332 for emotion recognition and 0.5783 for sentiment detection. However, as observed by Liu et al. (2024) in downstream tasks such as question-answering, such fusion architectures often rely primarily on text features, with VLMs frequently producing correct answers even in the absence of image inputs. This behaviour has been largely attributed to data artefacts (Wang et al., 2024d; Liu et al., 2024) or to the design of the encoders themselves. Fusion strategies like CentralNet and late-fusion BERT–ResNet combinations used by Maity et al. (2022) are prone to this modality imbalance. When one modality (typically text) provides clearer and more reliable cues for prediction or classification, it dominates the learning process, reducing the model's sensitivity to visual cues such as gestures, expressions, and symbolic imagery. Addressing these challenges requires carefully designed attention and fusion mechanisms that balance the contributions of each modality while preserving semantic depth across diverse input types.

Despite the rapid growth of multimodal user-generated content (UGC), current moderation systems often struggle to detect nuanced or context-dependent forms of harm in multimodal UGC (Gongane et al., 2022). Although existing "*cyberbullying*" datasets by (Hosseinmardi et al., 2015; Rafiq et al., 2015; Maity et al., 2023, 2022) face several key limitations, including (i) the use of overly simplistic binary classification labels (`0 CB` vs. `1 not-CB`) that fail to capture the nuanced nature of cyberbullying, (ii) ambiguous or insufficient annotation guidelines (e.g., (Maity et al., 2022)),

(iii) limited reproducibility due to restricted access to data and annotation protocols, missing configuration files (Hosseinmardi, 2022; Maity, 2023). This hinders validation and comparative analysis between studies claiming to detect "*cyberbullying*". However, (Sharma et al., 2022), in their systematic review on "***harmful meme detection***", highlight significant progress in curation of structured multimodal datasets. Notable examples include the "*Hateful Memes Challenge*" by Facebook AI (Kiela et al., 2020) or the `harm` vs. `non-harm` meme dataset with target labels by (Pramanick et al., 2021b), or `offensive` vs. `non-offensive` meme dataset by (Suryawanshi et al., 2020), among others. The "*Hateful Memes Challenge*" by (Kiela et al., 2020) introduced a purpose-built multimodal benchmark annotated by three independent annotators with binary labels `hate (1)` vs. `not-hate (0)`, and reported a Cohen's Kappa of 0.684, indicating moderate inter-annotator agreement (McHugh, 2012). Building on this line of work, Pramanick et al. (2021a) proposed a harmful meme dataset that defines `harmfulness` as "*a multimodal unit consisting of an image and embedded text that has the potential to cause harm to an individual, an organisation, a community, or society. Here, harm includes mental abuse, defamation, psycho-physiological injury, socio-economic damage, proprietary damage, emotional disturbance, and compromised public image*". This dataset comprises two topical subsets: (a) [`Harm-C`] COVID-19-related memes, and (b) [`Harm-P`] memes related to the 2020 U.S. presidential debate. This dataset was annotated for three levels of `harmfulness` (`harmful`, `partially harmful`, `not-harmful`), as well as for the target of harm (`individual`, `organisation`, `community`, or `society`). Pramanick et al. (2021a) reported inter-annotator agreement scores (Cohen's Kappa) of 0.683 for harmfulness and 0.782 for target identification in [`Harm-C`], and 0.675 for harmfulness and 0.790 for target identification in [`Harm-P`], indicating moderate agreement between annotators amongst different topics.

In the definitions proposed by Kiela et al. (2020) and (Pramanick et al., 2021a), the notions of *hateful* and *harmful* content exhibit substantial conceptual overlap, leading to their integration in subsequent multimodal studies such as Hate-CLIPper

(Kumar and Nandakumar, 2022). Both definitions encompass attacks on individuals or groups and acknowledge the ensuing psychological, emotional, and societal harm. While Kiela et al. (2020)'s *hateful meme* dataset requires content to be labelled as `hate` only if it targets specific protected characteristics (race, religion, gender, etc.), Pramanick et al. (2021a)'s *harmful* meme dataset does not require targeting identity-based characteristics to be the sole reason for `harm` label. Despite these definitional differences, both dataset authors acknowledge that content causing harm or distress to individuals or communities should be classified as either `harmful` or `hateful`. These datasets remain highly influential training resources because they enable models to learn subtle multimodal cues, especially how text and imagery interact to convey both explicit and implicit forms of harm (Sharma et al., 2022; Zhao et al., 2023). These datasets are valuable not only for their diverse meme templates and carefully designed annotation schemes but also for their reproducibility, which enables consistent and transparent benchmarking across studies (Zhao et al., 2023).

Recently, vision-language models (VLMs) have emerged as a powerful class of deep learning architectures capable of jointly processing visual and textual information. At their core, VLMs generally consist of two modality-specific encoders to encode representations of both image and text content. This includes an image encoder (e.g, convolutional or transformer-based) and a text encoder (usually a transformer encoder). These encoders are trained in tandem, with model architecture designs such as:

1. *Contrastive alignment*, this involves jointly training large image and text pairs to align their embeddings in a shared feature space, as seen in CLIP by (Radford et al., 2021), ALIGN by (Jia et al., 2021), De-CLIP (Li et al., 2021), among others.

2. *Masked language modelling with image-text matching*, involves predicting masked words in text conditioned on corresponding image regions and verifying the alignment between the image and text pair. This approach allows for modelling on fine-grained semantic association between the two modalities like the

works by LXMERT (Tan and Bansal, 2019), FLAVA (Singh et al., 2022), and BridgeTower (Xu et al., 2023).

3. *Multimodal fusion with cross-attention*, where visual features are directly injected into the language model's attention layers, enabling the model to condition text generation on image content, as seen in VisualGPT (Chen et al., 2022), VC-GPT (Luo et al., 2022), and Flamingo (Alayrac et al., 2022).

Building on these architectures, vision-language modelling approaches leveraged in *hateful* or *harmful* detection can be further categorised as,

(i) Unimodal Baseline: Early attempts at `hateful` or `harmful` meme detection established unimodal baselines that processed **visual** and **textual** modalities independently before classification (Kiela et al., 2020; Sharma et al., 2022). **Text-based** approaches leveraged pre-trained language models such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), while **vision-based** methods employed Convolutional Neural Networks (CNNs) like ResNet (He et al., 2016b) for feature extraction. Text-based baselines such as BERT have reported below 65% accuracy on the *Hateful Meme* dataset (Kiela et al., 2020) and approximately 70% accuracy on the *Harmful Meme* dataset (Pramanick et al., 2021a). For image-only models, early baselines relied on architectures such as VGG19 (Simonyan and Zisserman, 2015), DenseNet161 (Huang et al., 2017), ResNet-152 (He et al., 2016a), ResNeXt-101 (Xie et al., 2017b), achieving 68-70% accuracy on *Harmful Meme* dataset (Pramanick et al., 2021a); on the *Hateful Meme* detection task, Region-based detectors (e.g., Faster-RCNN (Ren et al., 2015) with ResNeXt152 (He et al., 2016a)) or grid-based image-encoder, i.e., Faster-RCNN (Ren et al., 2015) achieved accuracies of roughly 50–52%. More recently, unimodal image encoders have undergone substantial evolution. The Vision Transformer (ViT) (Dosovitskiy et al., 2021) recasts an image as a sequence of patches processed via self-attention, thereby capturing both local and global visual dependencies without relying on convolutional in-

ductive biases[52]. This flexibility allows ViTs to model long-range dependencies and complex global relationships between image regions, which can be particularly advantageous for meme detection (Zhao et al., 2023). Meanwhile CLIP (Radford et al., 2021) learns image embeddings via a large-scale contrastive training procedure against paired text, enabling strong zero- or low-shot image representations even without appearing paired during inference. The image encoder in CLIP can generate semantically meaningful features that capture concepts and relationships aligned with natural language allowing it to generalise to unseen images and tasks without requiring task-specific fine-tuning (Radford et al., 2021). Because these architectures significantly outperform older CNN-based image encoders in representational richness (Sharma et al., 2022; Kumar and Nandakumar, 2022; Zhao et al., 2023), they warrant designation as the contemporary minimal unimodal baseline in hurtful or nasty meme detection studies.

(ii) Multimodal Fusion Strategies: Building upon unimodal baselines, multimodal fusion strategies aim to combine visual and textual information in a unified model. One common approach is "**late fusion**", where each modality is processed independently via separate encoders, and their outputs are merged (via concatenation or weighted summation) before feeding into a downstream classifier (Zhao et al., 2023; Fan and Wu, 2022). For instance, ResNet features (He et al., 2016b) could be extracted from images while BERT embeddings (Devlin et al., 2019) were generated from text, with the resulting feature vectors concatenated and fed into a downstream classifier (e.g., multilayer perceptron or logistic regression) (Kiela et al., 2020; Fan and Wu, 2022; Zhao et al., 2023). More recently, models such as CLIP (Radford et al., 2021) have been employed for late fusion, where the image and text encoders are used to extract

[52] An inductive bias is a set of assumptions built into a model that guides learning. For example, convolutional neural networks assume that nearby pixels are more related than distant ones and that patterns can appear anywhere in an image (translation invariance). These assumptions help the network learn efficiently from limited data, but they also constrain the kinds of patterns the model can easily represent.

modality-specific embeddings. In practical setups, some layers of CLIP can be frozen to retain pretrained semantic representations, while the remaining layers and the classifier are trainable, allowing the model to adapt to the target dataset without over-fitting (Tsimpoukelli et al., 2021; Zhang et al., 2022b). For *hateful* or *harmful* meme detection, several studies (Huertas-Tato et al., 2024; El-Sayed and Nasr, 2024) have employed CLIP-based models, fine-tuning them for the downstream binary classification task while freezing selected pretrained layers to retain semantic knowledge. These approaches have demonstrated promising performance with relatively higher computational efficiency. Despite their efficiency and encouraging results, late fusion strategies continue to face limitations in capturing fine-grained cross-modal interactions, as the visual and textual modalities remain largely isolated until the final classification stage (Gomez et al., 2020). In contrast, "**early fusion**", approaches integrate textual and visual features at an earlier stage, enabling the model to capture fine-grained cross-modal interactions and context-dependent semantics more effectively (Fan and Wu, 2022; Zhao et al., 2023). For example, Vision-and-Language Transformer (ViLT) (Kim et al., 2021c) combines tokenized text embeddings with linearly projected image patch embeddings within a unified transformer architecture, allowing self-attention layers to model interactions between text tokens and image regions directly. These fused representations are then passed to a classifier, resulting in improved detection of *harmful* or *hateful* content (Ma et al., 2022; Zhang et al., 2020a). Overall, (Zhao et al., 2023) find that "*early fusion*" models such as ViLT (Kim et al., 2021c) or VisualBERT (Li et al., 2019) outperform "*late fusion*" approaches, including CLIP-based concatenation.

(iii) Large Language/Multimodal Models: Recent advances in Large Language Models (LLMs) and Large Multimodal Models (LMMs) have opened up new directions for harmful meme detection through prompt-based learning and instruction-following techniques (Cao et al., 2023; Huang et al., 2025).

For instance, Pro-Cap (Cao et al., 2023) utilises a zero-shot visual question-answering framework built on BLIP-2 (Li et al., 2023b) to generate descriptive captions that may reveal hateful or harmful elements in memes. In the Online Safety Prize Challenge (OSPC) organised by AI Singapore[53], the system by (Cao et al., 2024) integrated multiple components, BLIP (Li et al., 2023b) for image captioning, PP-OCR (Cui et al., 2025) and TrOCR (Li et al., 2023c) for multilingual text recognition, and the Qwen LLM by (Bai et al., 2023) for nuanced linguistic reasoning across multi-lingual harmful memes. They also used additional data labelled using GPT-4V[54] to transfer GPT-4V's multimodal understanding capabilities into their classification model. This framework achieved top-1 performance on the public leaderboard of the competition, with an AUROC of 0.7749 and accuracy of 0.7087, surpassing earlier fusion-based and transformer-only approaches. The Evolver framework (Huang et al., 2025) introduced a *chain-of-evolution prompting*, which enables LMMs to infer meme evolution patterns such as *variation*, *mutation*, and *inheritance* to understand cultural and semantic context. Similarly, ExplainHM (Lin et al., 2024) presents an explainable strategy where multiple LMMs engage in multimodal debate before reaching a consensus classification. Despite these advances, LMM-driven systems continue to face several challenges: (1) the scarcity of labelled datasets for harmful memes limits effective fine-tuning (Hee and Lee, 2025); (2) model performance is highly sensitive to prompt formulation (Cao et al., 2023); (3) deployment remains computationally expensive; and (4) maintaining robustness against out-of-distribution or rapidly evolving meme content requires ongoing adaptation (Hee and Lee, 2025).

Building on the observation that platforms now prioritise visibility controls over content removal (Section 1.1), this regulatory shift fundamentally challenges the *prescriptive* annotation paradigm that has underpinned existing detection systems. Leerssen (2023) and Kaushal et al. (2024) document the prevalence of visibility-based

[53] `https://ospc.aisingapore.org/`
[54] `https://openai.com/index/gpt-4v-system-card/`

interventions, such as age restrictions and recommendation demotions, among very large online platforms (VLOPs) as their content moderation reasons. These granular moderation decisions require distinguishing not just whether content is harmful or hateful, but to whom and in what contexts it may cause harm. As discussed in Section 2.3, Röttger et al. (2022) note the **descriptive paradigm** embraces annotator subjectivity to capture how different stakeholder groups (adult vs youth) perceive potentially harmful material. This interpretive variability becomes especially critical when distinguishing between ***explicit*** markers of harm (e.g., slurs, violent imagery) and ***implicit*** meanings conveyed through contextual clues, coded language, and intertextual meme references (Lin et al., 2024; Huang et al., 2025). However, current vision-language modelling (VLM) architectures trained on *prescriptively* labelled datasets struggle to disentangle these layers of meaning, as binary classification schemas collapse the spectrum of potential harm into uniform categories (Sap et al., 2022). For memes specifically, where humour, irony, satire, and metaphorical devices create inherently ambiguous semantic spaces that often embed harmful messages under deceptive disguises of benign visuals and neutral language (Lin et al., 2024; Huang et al., 2025; Gu et al., 2025), this limitation becomes acute. *Explicit* harmful cues are overtly identifiable through surface-level features: slurs in overlaid text, depictions of violence, or sexually explicit imagery. *Implicit* harmful cues, however, require contextual reasoning: a seemingly benign image of two girls reading two religious texts with phrases "*Person-1: You are going to hell; Person-2: Omg ! look you too*" (as illustrated in Figure 7.2 in Section 7.1.1 Chapter 7) becomes potentially offensive only when combined with text suggesting mutual damnation, i.e., the harm emerges from cultural knowledge and interpretive inference rather than lexical or visual markers alone. Detecting implicit harm thus demands models that can reason about sarcasm, coded language, intertextual references, and cultural context beyond pattern matching on explicit features. Consequently, there remains insufficient understanding of how VLMs process *implicit* versus *explicit* harmful cues, and whether such systems can adequately support the descriptive, audience-sensitive paradigm

necessary for age-differentiated visibility governance rather than uniform content takedown. The fundamental mismatch between binary classification objectives and age-differentiated moderation becomes evident when considering that a meme labelled `harmful` or `hateful` in a prescriptive schema does not indicate *to whom* or *in what context* it causes harm. For instance, political satire containing dark humour may be appropriate for adult audiences capable of interpreting irony, yet inappropriate for adolescents lacking contextual sophistication to distinguish satire from endorsement. Binary labels collapse these audience-contingent judgments into a single category, making them fundamentally misaligned with visibility-based governance that requires graduated, context-sensitive interventions.

2.4.6 Synthetic data generation and related applications

Before the advent of large language models (LLMs), data augmentation strategies to expand datasets in natural language processing were limited to human and machine translation strategies (Sennrich et al., 2016b; Imankulova et al., 2017; Dhar et al., 2018). Although techniques such as ***back-translation*** (Sennrich et al., 2016b) were widely used, they involve translating a text from a source language (e.g., Italian) to a target language (e.g., English) and then translating it back into the source language to generate paraphrased variants, thereby increasing dataset diversity without manual annotation. Fadaee et al. (2017); Wei and Zou (2019), among others, develop techniques that outperform back-translation on formal language tasks such as WMT test sets. However, (Edunov et al., 2018; Li and Specia, 2019; Park et al., 2021) note that back-translation works well with noisy input. They highlight that back-translation works well as a data regularizer, i.e., it introduces controlled variability into the training data, helping models become more robust to input noise and improving generalisation on downstream tasks. However, for tasks such as cyberbullying detection in youth, *back-translation* offers avenues for data augmentation which still remains unexplored.

Today, large language models (LLMs) such as GPT-3 (Brown et al., 2020), Bloom

(Workshop et al., 2023), LLama2/3 (Touvron et al., 2023), and Claude-2+ (Anthropic, 2023) have become powerful resources for generating diverse and coherent linguistic content. Despite their many strengths, Bender et al. (2021) highlights how these models can introduce biases, propagate inaccuracies, and produce contextually inappropriate information. Perez et al. (2022) also notes that the behaviour of LLMs is shaped not only by their scale but also by the degree and method of fine-tuning applied, leading them to exhibit distinct personality traits and behavioural tendencies. Alignment techniques, such as those proposed in Bai et al. (2022), aim to constrain outputs within ethical and socially acceptable boundaries. However, these safeguards are not absolute, and approaches like adversarial prompting and red-teaming (manipulating LLMs to find vulnerabilities) have repeatedly demonstrated that LLMs can still be manipulated into generating harmful, offensive, or policy-violating content (Perez and Ribeiro, 2022; Wei et al., 2023; Li et al., 2023a).

To steer LLMs toward reliable content generation, researchers have proposed a range of control techniques that operate at different levels of the modelling pipeline. These can be broadly categorised into: (a) input optimisation (prompting), (b) model adaptation (fine-tuning or re-training), and (c) post-processing (Zhang et al., 2022a; Bhargava et al., 2023). Prompting remains the most lightweight and cost-effective of these approaches, but its mechanisms are less well understood and can yield unpredictable results. Within prompting-based strategies, advances such as Chain-of-Thought (CoT) (Wei et al., 2022), Least-to-Most prompting (Zhou et al., 2022), and zero-/few-shot prompting (Zhu et al., 2023) have demonstrated the ability to elicit more structured reasoning from LLMs. These include,

- Chain-of-Thought (CoT) Wei et al. (2022), and Least-to-Most prompting Zhou et al. (2022) guide LLMs to deduce conclusions from given facts using straightforward, immediate inferences.

- Zero-shot CoT prompts LLMs with a general instruction such as "`think step by step`" to produce intermediate reasoning steps without prior examples Zhu et al. (2023).

- Few-shot CoT enhances performance by providing multiple labelled examples (question–rationale–answer), leveraging the model's in-context learning capabilities Zhu et al. (2023).

Beyond reasoning, aligning LLM outputs with ethical principles has been explored through approaches like Constitutional AI Bai et al. (2022). However, research indicates the risk of producing offensive content through malicious prompt manipulation Perez and Ribeiro (2022); Li et al. (2023a); Wei et al. (2023). Recent advances in LLM-based content generation have increasingly leveraged the interplay of participant roles, behavioural tendencies, and social dynamics to create richer scenarios. These include but not limited to,

- Persona-Driven Generation (Role Prompting) Ge et al. (2024); Hu and Collier (2024); Olea et al. (2024): Tailors output to participant personas to ensure dialogue remains contextually coherent and aligned with individual characteristics.

- Structured Task Decomposition Khot et al. (2022): Breaks complex tasks into modular sub-tasks, enabling stepwise and interpretable generation.

- Character Archetype Specification Hu and Collier (2024); Olea et al. (2024): Defines role-specific behaviours and traits to guide outputs according to social or narrative functions.

- Controlled Narrative and Emotional Escalation Kim et al. (2024); Zheng et al. (2023): Modulates linguistic intensity, emotional expressions, stress signals, and even profanity to generate dynamic, believable emotional arcs.

- Implicit Chain-of-Thought Reasoning: Models sequential and interdependent interactions, capturing group dynamics and emotional trajectories for socially and emotionally coherent responses.

Other approaches, such as (Kim et al., 2024; Zheng et al., 2023), emphasise controlled emotional modulation, adjusting linguistic intensity, profanity, or affective

markers to simulate escalating conflicts or believable interpersonal tension. Together, these techniques move LLM-based generation beyond isolated reasoning toward the creation of dynamic, multi-agent scenarios that capture social and emotional nuance. However, as LLMs become increasingly controllable through prompting, persona conditioning, and structured task decomposition, the challenge shifts from merely generating coherent text to ensuring that models reliably follow these controls. This creates a unique evaluation problem: richer and more fine-grained generation techniques require equally nuanced methods of assessing whether outputs preserve persona consistency, emotional trajectories, role alignment, or reasoning structure. Thus, advances in controllable generation directly heighten the need for evaluation frameworks that can operate without gold-standard references while still capturing pragmatic, social, and discourse-level qualities.

(Jones and Galliers, 1996) distinguish two core criteria for assessing natural language systems: ***intrinsic*** evaluation, which measures performance against the system's defined objective, and ***extrinsic*** evaluation, which assesses how well the system supports a downstream task. In practice, this distinction has often been operationalised in narrow ways. *Intrinsic* evaluation typically probes the quality of representations such as word embeddings of the models, for example, by assessing semantic similarity or analogy performance (Schnabel et al., 2015; Phillips et al., 2017; Hossain and Hoque, 2020). In contrast, *extrinsic* evaluation examines how these representations or models contribute to real tasks, such as classification or information extraction, within broader applications (Bender, 2011; Wang et al., 2019). However, Chiu et al. (2016) cautions that intrinsic evaluations on benchmarks do not reliably predict extrinsic performance. Similarly, Barančíková and Bojar (2025) observes that systems may perform well on intrinsic measures (e.g., semantic similarity via word embeddings), yet perform poorly on downstream tasks (e.g., machine translation evaluation (MTE), quality estimation (QE)). This highlights that datasets may differ intrinsically, and strong representation quality does not always translate to downstream task effectiveness. A crucial implication of the distinction between

intrinsic and *extrinsic* evaluation is that meaningful assessment requires datasets whose structure, content, and phenomena align with the real-world behaviours a system aims to model. Furthermore, traditional automatic evaluation methods like BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), chrF (character n-gram F score)(Popović, 2015), and metrics using cross-lingual pre-trained encoders such as COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020), require reference examples to compare the quality of synthetic text. We first review these reference-based metrics and their mathematical formulations, then discuss reference-free alternatives that have emerged to address evaluation in domains without gold standards.

- Bilingual Evaluation Understudy (BLEU): Papineni et al. (2002) measure the precision of n-gram matches between a candidate translation and one or more reference translations. The score is calculated as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.23}$$

  where p_n is the modified n-gram precision, w_n is the weight for each n-gram (typically $1/N$), and BP is a brevity penalty defined as:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{2.24}$$

  where c is the length of the candidate and r is the effective reference length. The modified precision p_n is computed by clipping the count of each n-gram in the candidate to its maximum count in any reference, preventing the metric from rewarding repetitive n-grams.

- Character n-gram F-score (chrF): (Popović, 2015) proposed chrF at the character level rather than word level, making it more robust to morphological variations and agglutinative languages. It computes the F-score based on

character n-gram matches:

$$\text{chrF}_\beta = (1 + \beta^2) \cdot \frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}} \tag{2.25}$$

where chrP (character precision) and chrR (character recall) are calculated as:

$$\text{chrP} = \frac{1}{N} \sum_{n=1}^{N} \frac{\#\text{matched char } n\text{-grams}}{\#\text{char } n\text{-grams in candidate}} \tag{2.26}$$

$$\text{chrR} = \frac{1}{N} \sum_{n=1}^{N} \frac{\#\text{matched char } n\text{-grams}}{\#\text{char } n\text{-grams in reference}} \tag{2.27}$$

The parameter β controls the relative importance of precision versus recall, with $\beta = 2$ being common in practice.

However, these specific automated evaluation methods cannot be applied in domains where there are no reference or gold standard benchmarks. To overcome this, several alternative automatic evaluation methods for evaluating machine generated data have been explored such as:

- ***Self-BLEU***, proposed by Zhu et al. (2018), measures the lexical diversity of generated text by adapting the standard BLEU score (Papineni et al., 2002). In this approach, each generated sample is treated once as a candidate and the remaining samples act as references. The BLEU score is then computed based on n-gram overlaps between the candidate and the reference set. A high Self-BLEU score indicates strong n-gram overlap, suggesting that the generated sentences are repetitive and lack diversity, whereas a low score indicates greater lexical variety across the sample set. This makes Self-BLEU particularly effective in detecting mode collapse, where a generative model produces outputs that are too similar to each other.

- Rei et al. (2020) propose **Crosslingual Optimized Metric for Evaluation of Translation (COMET)**, a neural evaluation metric based on crosslingual encoders trained to predict translation quality using human judgments

from multiple frameworks, including Direct Assessments (DA) (Graham et al., 2013), Human-mediated Translation Edit Rate (HTER) (Snover et al., 2006) and Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). Recognising the practical need for reference-free evaluation, Rei et al. (2022) introduced COMET-QE (also known as COMET-Kiwi), a quality estimation variant that operates without reference translations. This model achieved state-of-the-art performance across all tracks in the WMT 2022 Quality Estimation Shared Task (Zerva et al., 2022), demonstrating substantial improvements over previous metrics in multilingual settings. COMET-QE, unlike COMET, models translation quality by encoding only the source text s and candidate translation c using a pretrained cross-lingual encoder (XLM-RoBERTa), without requiring a hypothesis (human reference translation). The model was trained on DA scores from the WMT QE datasets using a regression objective that combines representations from both inputs:

$$\text{score} = f_\theta(\text{enc}(s), \text{enc}(c)) \tag{2.28}$$

where $\text{enc}(\cdot)$ produces contextualised embeddings and f_θ is a feedforward regression layer trained to minimise mean squared error against human quality judgments.

- Embedding-based similarity measures, such as ***cosine similarity*** over contextual embeddings (Reimers and Gurevych, 2019), evaluate the semantic closeness between two text samples. Given two embedding vectors $\mathbf{u}$ and $\mathbf{v}$, cosine similarity is defined as,

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{\sum_{i=1}^{d} u_i v_i}{\sqrt{\sum_{i=1}^{d} u_i^2}\sqrt{\sum_{i=1}^{d} v_i^2}} \tag{2.29}$$

The resulting value ranges from -1 (opposite directions) to 1 (identical directions), with 0 indicating no correlation. Unlike lexical overlap metrics, cosine similarity can capture semantic similarity between sentences that use different

surface forms (e.g., "*a child is playing*" vs. "*a kid is having fun*"). However, this measure is sensitive to the embedding space itself, i.e., if embeddings of sentences are poorly calibrated or unevenly distributed, the similarity scores may be misleading.

- The ***Fréchet BERT Distance (FBD)*** introduced by Alihosseini et al. (2019) extends the intuition of the Fréchet Inception Distance (FID) from computer vision (Obukhov and Krasnyanskiy, 2020). Unlike cosine similarity, which considers only pairwise directional closeness, FBD compares the overall distribution of embedding vectors. Given two sets of embeddings with means μ_1, μ_2 and covariance matrices Σ_1, Σ_2, FBD is computed as:

$$\text{FBD} = \|\mu_1 - \mu_2\|^2 + \text{Tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}\right) \tag{2.30}$$

  Where $\text{Tr}(\cdot)$ denotes the matrix trace. Unlike cosine similarity, which considers only pairwise directional closeness, FBD compares the overall distribution of embedding vectors. By incorporating both mean differences (semantic shift) and covariance structures (variance and correlations between dimensions), FBD captures higher-order semantic relationships between sentences. Lower FBD values indicate **more semantic** similarity and higher values indicate dissimilarity between the datasets being compared.

While these automated metrics are useful, they only capture surface-level lexical variation or distributional similarity in embedding space, and they fail to account for pragmatic, contextual, or discourse-level dimensions of generated text quality. This limitation is closely related to the distinction between *intrinsic* and *extrinsic* evaluation: intrinsic metrics (e.g., semantic similarity or embedding-based scores) focus on the internal properties of the text, whereas *extrinsic* evaluation assesses how well the text supports downstream tasks such as classification. To address the shortcomings of *intrinsic* evaluation and automated metrics, Chhun et al. (2022) proposed a human-centric evaluation framework for narrative story generation. In

the absence of reference datasets, human annotators directly assess outputs along dimensions such as *coherence*, *engagement*, *relevance*, and *creativity*. These dimensions allow human annotators to evaluate *intrinsic* qualities that automated metrics or analysing word embeddings fails to capture, i.e., if the machine-generated story is coherent and engaging while staying relevant to the storyline in the prompt. This helps bypass the need of gold-standard corpora for evaluating the quality of the machine-generated text. Similarly, research has extended this paradigm by developing semi-automated frameworks where LLMs themselves serve as evaluators, calibrated by human judgments as a human-in-the-loop evaluation for controllable text generation (Dubois et al., 2023; Sun et al., 2023). Together with Chhun et al. (2022), these contributions help understand that synthetic data evaluation is not only a technical but also a human-centered task that requires alignment with social context and end-user needs.

2.4.7 Understanding Obfuscations and Perturbation in NLP

A significant challenge in automatically detecting and mitigating CB lies in the dynamic and ever-evolving language used by perpetrators. This is where the concepts of obfuscation and perturbation become crucial. While often used interchangeably, these terms have subtle but important distinctions within the context of computational text analysis.

- **Obfuscation**: The deliberate act of adding noise to make the true meaning of a message difficult to understand, resembling a "*smokescreen*", where the goal is to confuse rather than directly alter (Brunton and Nissenbaum, 2015)

- **Perturbation**: As highlighted by Qian et al. (2022), this involves intentional and targeted changes to a message's form or content to disrupt its original structure.

In a nutshell, an unreadable scribble can be understood as *obfuscation* and a deliberate misspelling or substitution can be understood as *perturbation*. While obfuscation

aims to obscure meaning by introducing noise or ambiguity, perturbation introduces targeted, minimal changes designed to alter how a model interprets text without substantially changing its human-readable semantics (Brunton and Nissenbaum, 2015; Fisher et al., 2024). These practices differ and are also intertwined. For instance, (Ebrahimi et al., 2018; Zhang et al., 2020b), they '***perturb***' the text (e.g., *terrible* → *terr1ble*) to cause classifier error, not to confuse the reader. However, to avoid explicit *nasty or hurtful* words, users tend to use coded or euphemistic language (e.g., phrases such as "*the ones ruining our country*" or animal-references to humans) (Magu and Luo, 2018; Ocampo et al., 2023). While this '***obfuscates***' the meaning, it does not explicitly *perturb* tokens (words). Furthermore, these techniques also often coexist: perturbations can function as a mechanism for achieving obfuscation, yet obfuscation can also occur without precise perturbations (e.g., through stylistic complexity or indirect phrasing) (Brunton and Nissenbaum, 2015; Qian et al., 2022; Fisher et al., 2024). Both exploit the well-documented limitations of natural language processing (NLP) systems, particularly their sensitivity to surface-level variation and adversarial input (Ebrahimi et al., 2018; Pruthi et al., 2019; Zhang et al., 2020b). The informal, noisy, and highly variable nature of social media language magnifies these vulnerabilities, making it especially difficult to identify hateful or abusive content reliably. Prior research has developed a range of methods to study how obfuscation and perturbation undermine NLP robustness in online harm detection (Gröndahl et al., 2018; Ebrahimi et al., 2018; Pruthi et al., 2019; Zhang et al., 2020b; Aggarwal and Zesch, 2022; Kirk et al., 2022; Fisher et al., 2024).

Character-Level and Word-Level Obfuscation/Perturbation: A fundamental form of obfuscation involves modifying characters or words to alter the spelling of a harmful term without changing its phonetic meaning. In analysing vulnerabilites of hate speech classification systems, Aggarwal and Zesch (2022) devised strategies through which people deceive a system. They categories character-level or world-level substations as leet-speak i.e., use of numbers for specific characters such as "*a*" → "*4*", or kebab-speak i.e., using hyphens between characters (e..g, "*banished*"

→ "*b-a-n-i-s-h-e-d*"), among others. This presents a significant challenge for traditional NLP models that rely on predefined vocabularies and exact word matching. Gröndahl et al. (2018) also explored this by examining the intentional addition of positive or neutral words to a toxic message (e.g., "i h8 u" becoming "i h8 u greatman"). This semantic manipulation aims to mislead sentiment analysis models, which might incorrectly classify the entire phrase as positive or neutral due to the presence of "`greatman`" These studies underscore the need for models that can understand context and intent beyond simple word presence.

Emojis as Obfuscation/Perturbation: A more advanced form of obfuscation involves the strategic use of emojis. Kirk et al. (2021) investigated the phenomenon of using emojis to substitute different parts-of-speech in social media texts. For example, replacing sub-phrases like "*to kill*" or "*to hit*" with emojis such as "🔪", or using ethnicity specific emojis instead of specific noun phrases (e.g., *arab* or *muslim* → "🧕". This form of communication is particularly challenging for text-based models, as the harmful intent is conveyed visually rather than through conventional language. Models trained solely on text data would miss this context entirely (Kirk et al., 2021). This area of research highlights the importance of multimodal approaches that can process both text and images to detect subtle forms of harm.

Adversarial Attacks and Model Robustness: In natural language processing (NLP), adversarial attacks involve deliberate obfuscation or perturbation of input text to manipulate a model's prediction by adding noise characters or word substitutions, or syntactic rephrasing to fool classifiers without necessarily breaking human readability. The exploitation of model vulnerabilities through such adversarial attacks represents a significant evolution in evasion techniques. Zhang et al. (2020b) provides a comprehensive survey of adversarial attacks on deep learning models in NLP, highlighting the unique challenges that text-based adversarial examples present compared to their computer vision counterparts. Unlike images,

where perturbations can be continuous and imperceptible (e.g., changing pixel values), text modifications must maintain discrete token boundaries while preserving semantic meaning, making the generation of effective adversarial examples more constrained and challenging. Ebrahimi et al. (2018) introduced HotFlip, a white-box adversarial attack method that efficiently generates adversarial examples for text classification by performing atomic flip operations, i.e., swapping one token for another based on gradient information during training and validation. This approach demonstrated that minimal manipulations could significantly decrease model accuracy, while also showing that adversarial training using such examples could improve model robustness. Their work established important foundations for understanding how gradient-based optimisation could be leveraged to identify the most vulnerable points in text classifiers. These adversarial techniques exploit fundamental assumptions in NLP models about text structure, vocabulary, and semantic representation. The effectiveness of such attacks underscores the need for robust defense mechanisms that can not only handle specific linguistic variations but also deliberately crafted evasions (Aggarwal and Zesch, 2022). This includes developing models that are inherently more robust to perturbations, implementing preprocessing steps that can identify and normalise adversarial modifications, and creating training procedures that expose models to a wide range of potential attacks during development.

2.5 Analytical Conclusion and Research Gaps

Building on the multidisciplinary background on understanding and moderating cyberbullying (CB) discussed so far, this section presents key analytical conclusions, highlights research gaps addressed in this doctoral study, and identifies areas that remain unexplored.

- **Contrasting CB and physical bullying** (§2.2.1): Overall, the comparative studies (Olweus, 2013; McFarland and Ployhart, 2015; Thomas et al., 2015; Kowalski, 2018; Chan et al., 2019) discussed in Section 2.2.1, highlights

that both cyberbullying (CB) and traditional bullying share **six** foundational characteristics. However, the digital context further amplifies and transforms these behaviours. Features of social media platforms, including anonymity, persistent accessibility, and widespread publicity, not only increase the reach and impact of bullying but also reshape the mechanisms through which power and harm are exercised. This contrast emphasises the need to understand cyberbullying by acknowledging both its continuity with traditional bullying and the unique characteristics of online environments. While computational studies have recognised "*repetition*" and "*power imbalance*" as key criteria for classifying text as bullying-related (e.g., (Van Hee et al., 2018; Ziems et al., 2020)).

Research Gap: To the best of our knowledge, there has not been an examination of the above discussed **six** characteristics for their **linguistic** or **extra-linguistic** realisation in CB interactions. For instance, (Menesini et al., 2013; Alipan et al., 2020) note that *intentionality* in CB is not just "*wanting to hurt*" but also "*felt hurt*"; however, this characteristic or dimension is yet to be realised linguistically or extra-linguistically to understand how language in CB manifests. Doing so provides a more nuanced understanding of cyberbullying, including how relational dynamics and the features of digital platforms shape the experiences of perpetrators and victims.

- **Theoretical Concepts to understand CB** (§2.2.2): As discussed earlier in Section 2.2.2, existing research on bullying often relied on traditional theoretical frameworks. The complexities of cyberbullying in digital environments call for more nuanced perspectives. **Routine Activities Theory** (RAT) explains *how the absence of capable guardians (e.g., moderation) creates opportunities for perpetrators to target individuals*. **Crime Opportunity Theory** (COT) provides insights into *how technology shapes offender–victim dynamics*, highlighting how online environments can facilitate targeted attacks. Building on these, **Affordance Theory** helps understand that technological and social

features of SMPs (e.g., visibility, persistence, editability) create conditions conducive to antisocial behaviour, influencing how harmful content spreads, is modified, and engages audiences. However, only a few studies, such as (Chan et al., 2019) apply both the COT and Affordance perspectives to examine cyberbullying; their investigation remains limited by a narrow adult sample and a lack of focus on adolescents. Similarly, (Kaluarachchi et al., 2020) employs the General Theory of Crime and RAT to model CB but is constrained by its reliance on adult court-case data and does not address younger populations.

Research Gap: To the best of our knowledge, these three frameworks have rarely been applied together to explore the bystander's role, even though bystanders are crucial in either sustaining or disrupting CB behaviour (Leung et al., 2018; Song and Oh, 2018). Merging these frameworks allows for a holistic understanding of cyberbullying as a socio-technical process. Linking criminological notions of motivation and opportunity in the absence of capable moderation with different social media platform affordances that enable or constrain user actions helps understand (a) *how platform design mediates behavioural choices* across different bystanders, and (b) *how intervention or inaction emerges within these interconnected systems.* There is also a limited understanding of how platform features may (i) shield or expose bystanders, (ii) influence intervention via network visibility, or (iii) alter reporting dynamics through anonymity or disclosure. Overall, these gaps highlight the need for an integrated criminological and socio-technical approach that situates cyberbullying within ***perpetrator–target–bystander*** dynamics and the technological affordances of SMPs.

- **Youth perspectives on CB detection** (§2.3.1): The literature on youth perspectives in cyberbullying research discussed in Section 2.3.1 highlights the critical importance of engaging children and adolescents as active contributors, rather than passive subjects, in the design of detection and prevention strategies on social media platforms. Participatory design studies demonstrate that

youth can provide valuable insights into the nuances of cyberbullying, including diverse roles such as `bully`, `enabler`, `defender`, `conciliator` and can inform interface features, reporting mechanisms, and AI-enabled moderation tools (Ashktorab and Vitak, 2016; Milosevic et al., 2023; Sprugnoli et al., 2018; Ollagnier et al., 2022). Beyond simple *victim–perpetrator* frameworks, research increasingly acknowledges the variety of engagement types (interactive, active, passive (Shaw et al., 2022)), and the agency youth exercise in witnessing, reporting, or mitigating harm (Livingstone and Stoilova, 2021; Thorn and Benenson Strategy Group, 2021).

Research Gap: Despite these advancements, critical limitations remain. Most studies constrain youth participation to predefined scenarios or role-playing simulations, limiting ecological validity and spontaneity in capturing real-world cyberbullying dynamics. The potential re-traumatisation, exposure to harmful content, and the normalisation of negative behaviours, remain a critical ethical concern when involving adolescents in '*living-lab role-play*' methodology. Crucially, youth input is rarely solicited in the co-creation of automated content moderation systems or datasets, leaving a gap in understanding how young people perceive and engage with AI-enabled cyberbullying interventions. This highlights the need for participatory methodologies that integrate youth agency across the entire research and design lifecycle, i.e., from scenario creation and dataset development to AI model design. While simultaneously addressing ethical concerns and maintaining ecological validity. Such approaches would provide more robust, representative, and ethically sound foundations for both cyberbullying detection and prevention systems tailored to youth experiences.

- **Understanding and Annotating CB** (§2.3.3): Research in understanding cyberbullying and annotation of such behaviour, discussed in Section 2.3.3, is moving beyond simple bully–victim frameworks to encompass diverse roles such as `enabler`, `defender`, `conciliator`, `silent bystanders` (Leung et al.,

2018; Ollagnier et al., 2022; Chan et al., 2022). Studies have progressively expanded taxonomies to capture distinct types of harm, including overt aggression (e.g., flaming, harassment, denigration), covert aggression (e.g., sarcasm, manipulation), body-shaming, self-harm encouragement, and relational dynamics such as intent, escalation, and conversational functions (Van Hee et al., 2018; Sprugnoli et al., 2018; Banko et al., 2020; Kumar et al., 2022; Ollagnier, 2024; Trengove et al., 2022). These efforts underscore the multidimensionality of cyberbullying, revealing that a single post or interaction can involve multiple types of harm, roles, and intentions.

Research Gap: Existing annotation frameworks often rely on adult annotators with different statistical methods for measuring inter-annotator reliability and also potentially misalign with adolescents' interpretations of harm, intent, and offensiveness. Taxonomies frequently collapse overlapping behaviours, omit emerging forms of abuse (e.g., `encouragement of self-harm`), and inadequately capture relational and conversational dynamics. Moreover, while some frameworks model intent or participant roles, their application in youth-centered datasets is limited, with many annotations covering only a fraction of available data (Ollagnier, 2024). This highlights a pressing need for holistic, youth-informed annotation frameworks that integrate multiple dimensions of cyberbullying: the roles of participants, the type of harm expressed, and the underlying intent or relational context.

- **Binary and Fine-grained CB Detection** (§2.4.2, 2.4.3, 2.4.4): Building on (Vaswani et al., 2017)'s '*self-attention in encoder-decoder transformers*' paradigm, as discussed in Section 2.4.2, (Devlin et al., 2019) developed a pre-trained language model (LM) BERT, and subsequent transfer learning approaches using classification heads (Ruder et al., 2019), have enabled the development of automated cyberbullying (CB) detection systems (Paul and Saha, 2022; Mali et al., 2025; Philipo et al., 2024). However, as discussed in §2.4.2, such studies collapses CB into broader categories of online abuse

such as `hate-speech`, `offensive/abusive` language, leading to conceptual and methodological confusion. Some studies, such as (Talpur and O'Sullivan, 2020), even classify `racial` and `political` messages as "medium-level CB" and *insults about intelligence* as "low-level CB", illustrating how binary CB detection has contributed to substantial definitional drift in the field. As discussed in Section 2.4.3, Van Hee et al. (2018) beyond binary labels and recognised CB as a multi-faceted phenomenon by constructing real-life annotated datasets. In youth contexts, Sprugnoli et al. (2018) (Italian) and Ollagnier et al. (2022) (French) extended this work through role-played simulations, producing datasets suitable for both binary and multi-label CB detection. This is crucial to model *who said what to whom and why* in cyberbullying contexts. Nonetheless, cross-dataset inconsistencies in language, youth-specific communication norms, participant roles (e.g., whether enablers behave comparably across datasets), and limited dataset size hinder generalisation. As discussed in Section 2.4.3, (Jacobs et al., 2022; Rathnayake et al., 2020) attempt to build role-aware classifiers using *encoder*-only models (e.g., BERT, RoBERTa) and have produced unstable performance. Heuristic based approaches, such as Ollagnier et al. (2023) develop a multi-task system grounded in the perpetrator–target dichotomy, show promise but remain limited in scope. As discussed in Section 2.4.4, multi-task learning methods such as Hard Parameter Sharing and AdapterFusion have enabled the integration of *encoder*-based (BERT, RoBERTa) and *decoder*-based (GPT-2, OPT) language models for CB detection (§2.4.4). Yet these approaches still inherit the shortcomings of existing annotation schemes and task formulations. Furthermore, ongoing debates in computational linguistics (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bastings and Filippova, 2020), question whether the attention mechanism (Equation 2.9) genuinely provides insight into which *tokens* do LMs "*look at*" or "*attend to*" during processing. In CB moderation systems, the interpretability and explanatory power of *encoder*-based transformers remain

similarly limited, leaving it difficult to determine how these models derive their decisions.

Research Gap: Computational researchers should carefully consider the distinction in CB detection, i.e., whether their models are designed to merely detect isolated instances of `racism`, `sexism`, `offensive/abusive` language or to detect sustained ***`nasty or hurtful`*** behaviour that occur within cyberbullying-related contexts. While both encoder-only and decoder-only language models can be fine-tuned with various loss functions and architectural adaptations (§2.4.2), methodological progress is limited if models are evaluated primarily on one-off abusive utterances rather than sequences of interactions characteristic of cyberbullying.

- **Multimodal CB Detection** (§2.4.5): The literature reviewed in Section 2.4.5 indicates that many datasets that claim to be curated for *cyberbullying*, are curated without any clear definitions or consistent characteristics, and some suffer from author mis-mappings that limit accessibility. In other online abuse detection tasks, significant progress has been made in multimodal `hateful` and `harmful` meme detection, advancing from early unimodal baselines (Kiela et al., 2020; Pramanick et al., 2021a) to fusion architectures (Kim et al., 2021c; Radford et al., 2021) and large multimodal models (Cao et al., 2023; Huang et al., 2025). Three critical gaps emerge from Section 2.4.5. First, datasets and models largely rely on prescriptive binary annotations (`hateful` versus `not hateful`; `harmful` versus `not harmful`; `CB` vs. `no-CB`), collapsing nuanced harm into uniform categories (Kiela et al., 2020; Pramanick et al., 2021a). This approach reflects content removal rather than the visibility-based governance favoured by very large online platforms (VLOPs) as reported by (Kaushal et al., 2024) when examining the EU's transparency database. Second, while recent work emphasises distinguishing explicit harmful markers, such as slurs or violent imagery, from implicit contextual cues like sarcasm, coded language, or cultural references (Lin et al., 2024; Huang et al., 2025;

Gu et al., 2025), there is no evaluation of how well VLMs trained on binary datasets capture this semantic complexity in an audience-sensitive way. Early fusion models like ViLT (Kim et al., 2021c) integrate visual patches and text tokens to support cross-modal reasoning that could capture implicit cues, and contrastively trained models like CLIP (Radford et al., 2021) learn semantically aligned image-text representations potentially transferable to nuanced harm detection. Yet these architectures, whether early fusion, late fusion, or prompt-based large multimodal models (Cao et al., 2023; Huang et al., 2025), remain untested on context-dependent appropriateness judgments. Third, the field lacks understanding of where and why models fail when confronted with varying semantic complexity, including explicit, implicit, and benign content. **Research Gap**: In the absence of error analysis categorising VLM misclassifications across age-differentiated *explicit-implicit-benign* semantics, we cannot determine whether failures arise from architectural constraints on cross-modal reasoning, inadequate training signals for implicit cues, or inherent incompatibility between binary classification and the audience-contingent judgments that characterise real-world moderation decisions.

- **Applications of LLMs in CB tasks** (§2.4.6): The literature reviewed in Section 2.4.6 highlights the rapid evolution of Large Language Models from single-turn text generators into systems capable of producing socially rich, multi-agent, and emotionally dynamic synthetic content. Prompting techniques ranging from Chain-of-Thought and task decomposition to persona-driven and emotionally controlled prompts have significantly increased the controllability, coherence, and narrative structure of synthetic outputs. At the same time, alignment strategies such as Constitutional AI aim to constrain undesirable behaviours but remain vulnerable to adversarial manipulation, suggesting that LLM reliability is not merely a matter of scale or fine-tuning but also of interaction dynamics and contextual design. Despite advances in generation techniques, evaluating synthetic data remains a fundamental challenge,

particularly in domains lacking reference corpora or gold-standard outputs. Traditional reference-based metrics (BLEU, chrF, COMET) are mathematically well-established but unsuitable for open-ended generation tasks where no canonical answer exists. Reference-free metrics such as Self-BLEU, COMET-QE, embedding-based similarity, and Fréchet BERT Distance offer partial solutions. Yet they primarily capture surface-level lexical diversity or distributional similarity rather than deeper pragmatic, social, or discourse-level qualities. Human-centered frameworks such as (Chhun et al., 2022) attempt to fill this gap by incorporating dimensions such as coherence, relevance, and engagement. However, in cyberbullying contexts, the applicability of LLMs as data curators remains largely un-evaluated.

Research Gap: Although progress has been made in controlled text generation and automated evaluation, several gaps remain: (i) There is a lack of evaluation methods for multi-agent or socially dynamic synthetic data. Existing reference-based and reference-free metrics operate at the sentence or document level and do not assess *interpersonal dynamics, role consistency, emotional escalation* or *behavioural realism* (Ge et al., 2024; Hu and Collier, 2024; Olea et al., 2024). (ii) Automatic metrics are weakly aligned with human-centered judgments of social quality. Embedding-based or distributional measures often fail to reflect human evaluations of coherence, contextual appropriateness or persona consistency (Chhun et al., 2022). (iii) There is an over-reliance on LLM-as-a-judge method without calibration or domain-specific validation. Prior work shows that LLMs as evaluators or classifiers can exhibit biases, instability and vulnerability to prompt manipulation, especially in sensitive settings such as conflict, offensive content or emotionally charged interactions (Perez and Ribeiro, 2022; Wei et al., 2023; Li et al., 2023a).

- **Obfuscations and Perturbation in CB** (§2.4.7): Cyberbullying detection in social media is complicated by deliberate obfuscation and perturbation strategies that manipulate language in ways that are difficult for automated

systems to interpret. As discussed in Section 2.4.7, character-level substitutions, word-level semantic alterations, emoji replacements, and coded expressions can bypass conventional classifiers while remaining intelligible to human readers. Adversarial attacks further exploit the structural and semantic assumptions of NLP models, revealing fundamental vulnerabilities in tokenisation, embedding, and context modelling. While prior research has proposed methods such as gradient-based attacks during training and validation (Ebrahimi et al., 2018), preprocessing normalisation, and adversarial training (Aggarwal and Zesch, 2022; Pruthi et al., 2019), these solutions often address isolated scenarios rather than the diverse, evolving tactics observed in real-world communication. Furthermore, the combination of multiple obfuscation techniques, along with context-dependent expressions of `hurtfulness`, remains largely under-explored. The complexity of online language, with its interplay of surface-level noise and subtle semantic signals, underscores that robust detection cannot rely solely on standard lexical or embedding-based approaches. Advancing these capabilities is crucial for developing detection systems that remain effective in dynamic social media environments where harm is intentionally hidden or disguised.

Research Gap: Most detection methods focus on isolated obfuscation or perturbation types, lacking a holistic approach that considers coexisting strategies such as semantic manipulations with emojis or character substitutions (Aggarwal and Zesch, 2022; Kirk et al., 2021; Gröndahl et al., 2018). Existing adversarial robustness techniques are limited to known attack patterns and may not generalise to unseen or evolving perturbations in **youth** online communication (Ebrahimi et al., 2018; Pruthi et al., 2019; Zhang et al., 2020b). Multimodal threats and indirect cues remain under-explored, leaving models vulnerable to subtle, context-dependent harm that text-only classifiers cannot capture (Kirk et al., 2021; Fisher et al., 2024). Evaluation frameworks for robustness to obfuscation and perturbation are not standardised, and human-centered

assessment of hurtfulness in adversarial or obfuscated content is sparse (Aggarwal and Zesch, 2022; Qian et al., 2022).

Part III

Contributions

Chapter 3

Youth Participatory Methodology and Framework

Abstract

This chapter details a multi-layered, mixed-methods framework for investigating youth experiences with cyberbullying to inform the development of context-aware detection tools. Grounded in the Software Development Life Cycle (SDLC) and Routine Activities Theory (RAT), the methodology transitions from quantitative requirement gathering to qualitative co-design. This chapter discusses the integration of a rights-based participatory approach, which elevates adolescents from passive subjects to active "Young Research Assistants". Through a three-phase participatory methodology, this chapter outlines the instruments and ethical considerations required to co-create a realistic cyberbullying dataset with Irish adolescents aged 13–17. The process begins with Phase 1 (Requirement Gathering) to identify platform-specific harms and reporting tendencies, followed by Phase 2 (High-level Design), where youth evaluate the realism and seriousness of cyberbullying scenarios. Finally, Phase 3 (Detailed Dataset Design) utilises a "Young Research Assistantship" program to generate authentic conver-

sational data, including difficult-to-detect obfuscated harmful speech. This collaborative approach ensures the resulting dataset reflects the lived digital realities of minors while adhering to rigorous ethical safeguards.

3.1 Overview of SDLC Phases

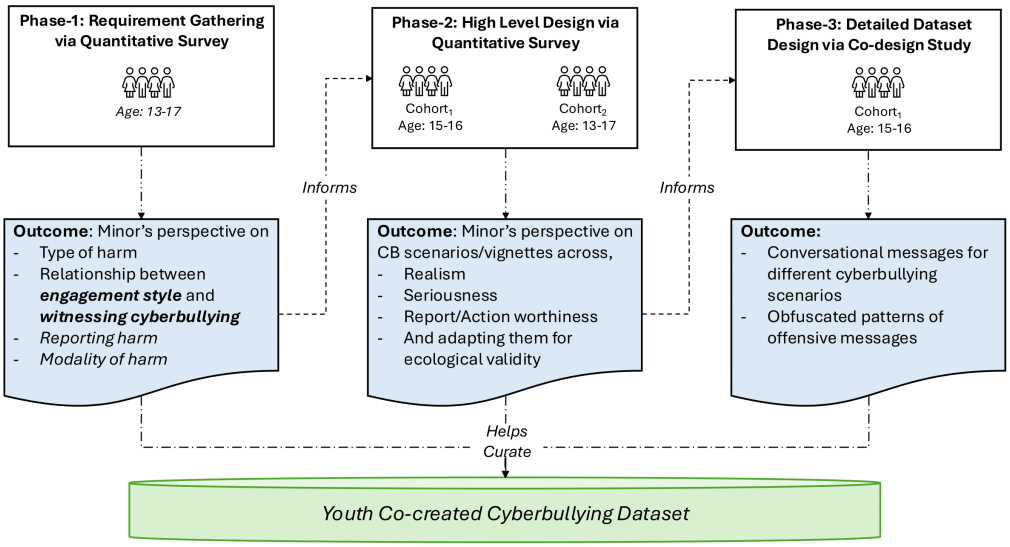


Figure 3.1: Overview of the youth participatory methodology framework across different phases

Figure 3.1 presents an overview of the **three-phase** youth participatory software development life cycle (SDLC) methodology employed to co-curate test cases and datasets that reflect both the lived experiences of minors. This participatory framework ensures that the resulting dataset captures nuanced cyberbullying dynamics while maintaining ethical integrity and practical relevance. The three phases are as follows:

Phase-1 **Requirement Gathering**: This phase involves engaging with youth (ages 13–17) to explore their lived experiences of cyberbullying. Particularly, understanding the types of harm they encounter, their interactions with different

forms of online engagement, their interpretations of harmful content, and their behaviours around reporting. This phase also helps understand the modality of harmful content (e.g., *text, images, videos, or memes*) young people often engage with thereby informing the direction of detection tool development. This phase provides insights into four key aspects: (a) the types of harm youth witness in cyberbullying situations, (b) their reporting behaviours i.e., whether they report through platform mechanisms or to their support networks or reporting to law enforcement, (c) their patterns of engagement that lead to witnessing such harm, and (d) the modality of the harmful content. These findings are critical for guiding the development of effective, context-specific tools for detecting cyberbullying.

Phase-2 **High-level Design**: Building on initial insights, this phase expanded participation to include additional youth cohorts. Participants co-constructed cyberbullying scenarios and vignettes, providing input on aspects such as **perceived realism**, **severity of harm**, and whether an **incident warranted reporting** to the platform or their support network. This helped shape the thematic and contextual scope of the dataset, thereby ensuring its relevance and authenticity for subsequent detection tool development.

Phase-3 **Detailed Dataset Design**: Building on youth agency frameworks in co-research discussed in §3.2.5, through a week-long "**young research assistantship**", teens participate as researchers to generate realistic conversational messages based on the vignettes developed in Phase 2. Particular emphasis was placed on incorporating obfuscated or indirect forms of harmful speech that are often more difficult for automated systems to detect. This phase resulted in curating a dataset of realistic conversational messages, enriched with obfuscated and indirect forms of harmful speech, thereby improving the linguistic complexity of the data to support the development of more context-aware cyberbullying detection tools.

Given the participatory nature of this framework, youth were not only engaged as study participants but also as active participants in shaping the research process. To enable this, a structured Young Research Assistant (Young RA) programme was implemented. This programme, central to *Phases 2* and *3*, positioned youth as co-researchers and co-designers, empowering them to contribute to scenario design, dataset creation.

3.2 Instruments Informing Participatory Approach

3.2.1 Cyberbullying in a Computational Linguistic Frame

Note: This sub-section has been published within (Verma et al., 2024).

As discussed in Section 2.2.1 and Section 2.5 of Chapter 2, there exists a disconnect between the six traditional criteria of bullying and their representation in computational linguistics. To address this, we propose a refinement of the six criteria examined by (Olweus, 2013; McFarland and Ployhart, 2015; Thomas et al., 2015; Kowalski, 2018; Chan et al., 2019), grouping them into three linguistic and three extra-linguistic factors, while providing our interpretation of their relevance and application within computational approaches.

Linguistic Factors:

- **Intentionality**: Defined as deliberate intent to cause harm by (Smith and Steffgen, 2013), in CB it is expressed through transmission and reception of bullying media by the victim (Menesini et al., 2013). Noting intent as both *implicit* and *explicit*, in computational linguistics this can be modelled through identifying aggressive language, threats, or insults that indicate intent. Examples include direct messages such as "`i hope u suffer`" or "`u shld kill urslf`", which demonstrate unambiguous harmful intent. Additionally, the use of sarcastic emoji reactions or targeted comments under Stories or posts may also signal intentional social ridicule in more implicit forms.

- **Repetition**: This is facilitated by the affordances of SMPs that allow harmful content to be easily shared, reposted, or commented on multiple times (Kane et al.,

2014). It is also worth noting that "*repetitiveness*" can be conceptualised differently in CB than offline, where a single, harmful post or message can be considered an ongoing act of bullying due to its potential for widespread and continuous dissemination to a large audience, which creates a form of repeated exposure for the victim (Slonje et al., 2013; Milosevic et al., 2022). The re-circulation of CB-related material like screenshots, edited images, or mocking captions, across shared networks, are types of "repeated" hurtful behaviour. Linguistically, it can be understood via repeated lexical patterns, abusive phrases, or frequent harmful mentions (e.g., "`not gonna stop till u in tears, no matter where u go`") Verma et al. (2024).

- **Power Imbalance**: Initially described as group or individual behaviours involving a more influential individual targeting someone with less power (Vaillancourt et al., 2008), SMPs can intensify this dynamic by enabling users to mobilize followers against individuals with limited online presence (Chan et al., 2021). In computational linguistics, power imbalances can be revealed through the analysis of online interactions and language usage, providing insights into how individuals with more influence may target those with less influence in digital environments. For instance, text like "`lol i can ruin u with single post, watch out`" demonstrate imbalance of power.

Extra-linguistic Factors:

- **Anonymity**: Unlike physical bullying, where the aggressor and enablers are typically identifiable, SMPs provide an environment that allows users to engage in bullying behaviour while maintaining anonymity. This anonymity mitigates the risk of accountability, thereby encouraging more aggressive behaviour. Computational indicators of anonymity on SMPs may include use of pseudonymous usernames, erratic posting patterns, or abrupt shifts in tone in hostile contexts (Verma et al., 2024). These factors extend beyond linguistic cues to encompass platform-specific features, such as anonymous accounts, fake profiles, or the use of virtual private networks (VPNs), which are employed to obscure user identities (Chan et al., 2021).

- **Accessibility**: Unlike in-person bullying, where physical separation may offer relief, CB enables around-the-clock exposure through SMPs (Cassidy et al., 2013; Chan

et al., 2019). Victims may be contacted at any time, tagged in public content, or targeted across platforms, even after deactivating their accounts (Bastiaensens et al., 2015). SMP features like *Stories* or *Disappearing Direct Messages (DMs)* allow perpetrators to bypass privacy boundaries and re-initiate contact. Computational approaches to assessing accessibility go beyond linguistic factors and include temporal analysis of harmful interactions, cross-platform content tracing, and detection of unsolicited messages across time zones.

- **Publicity**: Constrained to a specific group or location in physical bullying, SMPs enable bullies and enablers to broadcast aggression to large audiences via posts, comments, and visual content. Publicising CB has expanded through features like "stories" on platforms, where content gains visibility even though it disappears after a set time. Going beyond linguistic factors, engagement markers within social networks and over time can facilitate the observation of the publicity of bullying content.

3.2.2 Cyberbullying through Routine Activities Theory

As discussed in Section 2.2.2, Routine Activities Theory (RAT) by (Cohen and Felson, 1979), helps understand *how the everyday online routines of adolescents intersect with opportunities for cyberbullying to occur.* Specifically, it informs our investigation into how youth encounter harmful interactions in varying capacities, e.g., incidental witnesses, or active or passive bystanders. Furthermore, RAT provides a conceptual basis for understanding how adolescents identify and engage with forms of "*capable guardianship*" when confronted with online harm. These may include formal mechanisms (e.g., in-platform reporting tools), informal social responses (e.g., confiding in friends or family), or passive monitoring (e.g., screenshot and saving messages for future action). RAT helps to contextualise the spatial and temporal convergence of the motivated offender or enablers, vulnerable targets, and lack of guardianship within specific sociotechnical environments on social media platforms. These may include group chats, comment threads, disappearing content (e.g., Stories or disappearing DMs). By examining these dynamics, RAT enables us to theorise how platform design, user behaviour, and social norms collectively influence the emergence and escalation of cyberbullying incidents. In doing so, RAT not only guides

data collection and scenario development but also supports the identification of critical intervention points that can be addressed through youth-informed design strategies.

3.2.3 Participatory Software Development Life Cycle

Software Development Life Cycle (SDLC) models serve as standardised frameworks to guide the software development process (Ruparelia, 2010). Despite the variety of models available, the initial stages like requirement gathering and high-level or detailed-level design inform the development of prototypes or proof-of-concepts. Muller et al. (1997) emphasise that participatory practices in the software lifecycle actively involve users and stakeholders throughout the development process, starting from "requirement gathering". They describe "requirement gathering" as a collaborative process employing methods such as surveys and interviews to capture user's needs. Furthermore, Muller et al. (1997) highlight the importance of engaging users not merely as subjects but as active participants in scenario development and co-design activities during design and evaluation phases.

This doctoral dissertation employs the SDLC framework for exploring **RQ.1** through a **mixed-methods** approach that integrates both quantitative and qualitative co-design methodologies across key phases of the SDLC. Specifically, the study aligns with three core phases of SDLC: (a) *requirement gathering*; (b) *high-level design*; and (c) *detailed design* by embedding participatory practices throughout, regardless of whether the underlying development model is linear or iterative. During the requirement gathering phase, stakeholders' perspectives, particularly perspectives from vulnerable user groups such as adolescents, are collected through structured quantitative surveys. These surveys aim to explore social media engagement patterns, platform-specific behaviours, bystander roles, and reporting practices, to inform the functional and ethical requirements of AI-driven cyberbullying detection systems. This phase mirrors the initial SDLC step where system requirements are elicited from end users to define the scope and goals of the system. In the high-level and detailed design phases, participatory methods are further employed to evaluate the perceived relevance, seriousness, and actionability of cyberbullying vignettes documented in prior literature. Quantitative feedback is collected as part of high-level design to assess the ecological validity of these scenarios and to inform the selection and adaptation of representative cyberbullying cases. This reflects the SDLC's high-level de-

sign phase, where system architecture and major components are conceptualised based on user needs. Through a cascading co-research and co-design methodology as part of detailed design, each participant individually builds on the previous participant's design and engages in analysing designs for ecological validity, thereby yielding in curation of a cyberbullying conversational dataset across different cyberbullying scenarios. This mirrors the detailed design stage in SDLC, where system components are specified with precision, enabling implementation of user-informed solutions.

3.2.4 Conceptualising Engagement, Witnessing Harm, and Reporting in Cyberbullying Contexts

Understanding User Engagement: Prior research on examining user engagement on social networking sites (SNS) and social media platforms (SMPs) often relies on a broad `active-passive` dichotomy but it does not address important nuances such as "bidirectional" (*mutual exchanges like commenting or messaging*) versus "unidirectional" (*one-sided actions like liking or viewing*) interactions (Meier and Reinecke, 2021). In an attempt to provide a comprehensive conceptualisation of user engagement behaviours, Shaw et al. (2022) studied **three distinct** engagement patterns: (a) ***interactive*** engagement i.e., creating and sharing user-generated content (UGC) such as posts or stories or messages; (b) ***passive*** engagement, characterised by non-interactive behaviours like browsing, scrolling, or viewing content without responding; and (c) ***reactive*** engagement which includes minimal yet measurable responses to content, such as liking or reacting with emojis, without initiating deeper interaction. Shaw et al. (2022) found that users with more interactive engagement styles report greater social connectedness and social capital compared to reactive or passive users. Understanding these distinctions can improve AI-driven cyberbullying detection by enabling systems to differentiate between users who are likely to observe, engage with, or respond to harmful content based on their behavioural patterns. For instance, recognising that "*reactive*" users may escalate or de-escalate harmful posts could help flag both UGC and users interacting with it, while identifying "*interactive*" users as potential `bully`, `enabler`, `defender` or even `victim`, could aid in refining predictive models and intervention strategies.

SMPs today offer a wide range of features that support diverse forms of user engage-

ment, including emoji reactions, ephemeral content like disappearing messages and stories, group chats, and threaded comment sections. These features enable users to engage in different ways, ranging from passive viewing to reactive and interactive behaviours. As illustrated earlier in Figure 1.2, various types of online harm associated with cyberbullying are frequently disseminated through these very features. However, what remains insufficiently understood is "*how young people engage with these functionalities*" and "*how their style of engagement may influence their exposure to, or involvement with, cyberbullying-related content*". To address this gap, we draw upon the engagement style framework proposed by (Shaw et al., 2022) to investigate "*how adolescents interact with harmful content online, and whether their behavioural tendencies (e.g., passive viewing, reactive responses, or interactive participation) affect their encounters with cyberbullying*".

Witnessing cyberbullying related online harms: The European Union Kids Online (EUKO)[55] and Global Kids Online (GKO)[56] initiatives, grounded in both quantitative and qualitative research, have significantly advanced our understanding of youth perspectives on **online safety** and **online risk**. These international research collaborations have generated comprehensive resources across a wide range of digital behaviours among children and adolescents. As part of these initiatives, Rodríguez-de Dios et al. (2018) explored the complex interplay between digital resilience, risk exposure, and the role of parental mediation in shaping safer and more enriching online experiences. They emphasize that children are not merely passive recipients of harm but are also active participants navigating digital environments, often exercising agency in the face of challenges. Rodríguez-de Dios et al. (2018) laid the empirical groundwork for developing tools that capture not only direct experiences of cyberbullying but also indirect involvement, particularly by-standing or witnessing harmful online behaviours. Similarly, Livingstone and Stoilova (2021) conducted a comprehensive synthesis of global evidence on children's internet use, highlighting the diversity of risks encountered and the varied outcomes of digital engagement. They broaden the conceptualization of online harm, one that explicitly includes by-standing roles where children may witness behaviours such as cyberbullying, exclusion, or online discrimination, even if they are not directly targeted. These studies

[55] `https://www.lse.ac.uk/media-and-communications/research/research-projects/eu-kids-online/about`

[56] `http://globalkidsonline.net/`

led to the development of validated quantitative instruments used in EUKO and GKO surveys, ensuring conceptual clarity, cross-context comparability, and reliable data for informing policy and interventions. Within these validated tools, **eight** common forms of online harm related to cyberbullying are consistently measured. These include **threats**, **rumours**, **exclusion**, and five types of discriminatory behaviour based on **ethnicity**, **religion**, **sexual orientation**, **disability**, and **body image** (See Global Kids Online Module K: Hurtful and Bullying Behaviour Questions K:2 [57]). These categories reflect the nuanced ways in which cyberbullying manifests, and are especially relevant to understand how young people encounter and interpret cyberbullying. Ireland's National Advisory Council for Online Safety (2021) found that the term "*cyberbullying*" itself can be confusing or misunderstood by minors due to varying perceptions of what constitutes such behaviour. In response, their national survey adopted the phrase "*hurtful or nasty online behaviour*" to better capture young people's lived experiences in a more accessible and relatable way. Building on this, we adopt a comparable approach to examine the various forms of online harm associated with cyberbullying that young individuals encounter. This approach is designed to ensure both methodological rigour and contextual relevance, enabling a comprehensive analysis of youths' digital experiences. Particular attention is given to bystander interactions, an often under-researched yet essential dimension for understanding the complexities of peer dynamics in online environments.

Understanding reporting behaviour among minors: Thorn and Benenson Strategy Group (2021) designed a study with a youth-centred approach to explore how young people perceive various online threats, their emotional responses, and the strategies they employ to stay safe, like disclosure to trusted adults, reporting within platforms, and the use of blocking features. They crafted survey questions that were developmentally appropriate, relatable, and grounded in real online experiences. To understand reporting behaviour, they formulated questions that moved beyond abstract hypotheticals and instead presented specific, age-relevant scenarios involving online risks like receiving unwanted messages, being pressured to share personal content, or experiencing threats or harassment. They structured survey questions to uncover not just whether youth reported

[57] `http://globalkidsonline.net/wp-content/uploads/2016/04/K-Hurtful-and-bullying-behaviour.xlsx`

harmful incidents, but to whom they reported (e.g., parents, friends, platform moderators), why they chose to report or not report, and what outcomes they expected or experienced. Importantly, the survey differentiated between platform-level reporting (e.g., flagging content or users) and personal disclosure (e.g., telling a trusted adult), treating both as valid but distinct forms of reporting. Building upon this framework of reporting behaviour, we extend it to the domain of cyberbullying, thereby enabling a more comprehensive understanding of youths' decision-making processes, attitudes, and levels of trust in available reporting mechanisms. By contextualising reporting behaviour within the specific dynamics of online harm, it offers valuable insights into how young individuals perceive, evaluate, and respond to incidents of cyberbullying. Additionally, it helps identify potential gaps or barriers in current reporting systems that may deter youth from seeking support or intervention.

3.2.5 Youth Agency as Co-researchers from a Rights-based Perspective

Rights-based Framework and Policy Context: The European Union (EU) Joint Research Centre (JRC) report by (Charisi et al., 2022) underscores the imperative of involving young people not only as users but as co-designers of AI systems that shape their digital lives. This stance is grounded in a rights-based approach, particularly Article 12 of the United Nations Convention on the Rights of the Child (UNCRC)[58], which asserts that "*children and young people have the right to express their views freely in all matters affecting them, and that these views must be given due weight according to their age and maturity*" (Livingstone et al., 2014). This emphasis on meaningful participation is echoed in various national and regional policies like Ireland's Department of Children, Equality, Disability, Integration and Youth (2023)'s directive on "*Participation with Purpose*" operationalising the Article 12 by highlighting the necessity of ensuring **Space** (opportunities to express views), **Voice** (facilitated expression), **Audience** (adults actively listening), and **Influence** (views meaningfully shaping decisions). These principles are not only ethical imperatives but practical design considerations. In this doctoral research, they serve as

[58] `https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child`

foundational guidelines to ensure youth are provided with the space, tools, and authority to shape the inquiry ,including informing interpretations, and responses to online harm in ways that reflect their lived experiences and sociotechnical realities.

Ireland's Transition Year Framework: The Transition Year (TY) Work Experience Framework[59], a component of the broader Transition Year program in Irish secondary schools, has been leveraged as a key research methodology. Typically offered to students around 15–16 years old, this framework provides structured opportunities for students to engage in professional work environments, develop essential life and career skills, and gain hands-on experience beyond the classroom. By aligning the research study with the TY framework, it allows to leverage the pre-existing infrastructure, ensuring that participation is seen as a valuable educational activity rather than a separate, external commitment. This approach helps to overcome common recruitment challenges, as schools and students are often keen to participate in projects that fulfil the program's objectives. The TY work experience promotes personal growth, career exploration, and a deeper understanding of professional environments and expectations. Through work placements, workshops, and reflective activities, the framework supports students in making informed choices about their future education and career paths, while also encouraging independence, responsibility, and self-awareness. For this doctoral study, the TY framework plays a vital ethical and methodological role in facilitating the recruitment of minors to contribute qualitative insights into the discourse on cyberbullying. It ensures that youth participation is conducted with appropriate safeguards and fosters a supportive environment for open dialogue. Additionally, the framework offers participants opportunities to develop valuable life skills such as using Google Sheets and basic Python programming, thereby empowering them with practical competencies that extend beyond the scope of the study. This dual focus on ethical engagement and skill building enriches both the research process and the participants' personal development.

Co-research and Co-design with Youth: Guishard and Tuck (2013) note that participatory research with young people tends to be "superficial" often involving youth

[59] `https://www.citizensinformation.ie/en/education/primary-and-post-primary-education/going-to-post-primary-school/transition-year/#5ebb20`

to share stories or opinions without influencing the framing, methodology, or outcomes of the study. They emphasize that passive or performative involvement can undermine the very goals of co-design or participatory research such as empowering youth as knowledge producers, fostering equitable researcher–participant relationships, and generating contextually relevant and actionable insights. Clark et al. (2022a) presents a critical reimagining of how youth can and should be engaged in data-centric research. Rather than positioning young people as passive subjects or mere informants, they advocate elevating youth as **co-researchers** and **co-creators**, particularly in projects involving digital data, privacy, and algorithmic systems that directly impact their lives. Their work is both theoretically and methodologically rich, providing practical strategies for implementing co-research methodology as co-creating research questions, engaging youth in critical reflection on data ethics, analysing and co-authoring research outputs. Participatory design approaches with children like *Cooperative Inquiry*, *Mixing-of-Ideas*, and *youth focus groups* treat children not as passive subjects but as active designers. Cooperative Inquiry, allows children and adults to collaborate equally throughout ideation, prototyping, and reflection phases, thereby combining experiential knowledge with design expertise (Druin, 1999). The Mixing-of-Ideas technique (also called bags-of-stuff) supports children in contributing creative artifacts that become stimuli for collaborative design experimentation (Guha et al., 2004). In the context of minors and cyberbullying-related co-design methods, Ashktorab and Vitak (2016) demonstrate the use of mix-of-ideas, cooperative inquiry through focus groups with teens to inform computing tools to address online safety risks like cyberbullying. Their approach values minors' unique perspectives and fosters environment for mutual understanding of intervention designs minor's would want for intervention in cyberbullying situations. Moreover, recent studies also employ participatory design methods and techniques to explore youth perspectives in cyberbullying contexts within focus groups or individual interviews such as (Zou et al., 2024; Zhou and Farzan, 2021; Fiani et al., 2024).

Building upon these participatory principles and rights-based frameworks, this doctoral research reconceptualises the role of adolescents from that of passive participants, to active and informed co-researchers. Rather than being treated as subjects of study or simply consulted for their opinions, within the TY Work Experience framework and Co-research methodologies, young people are involved as "**young research assistants**",

thereby shaping the research process itself. Their lived experiences served as the foundation for defining what constitutes meaningful or harmful interactions in digital peer settings in the context of cyberbullying. Importantly, this process of participation acknowledges young people as capable contributors who bring valuable insights, ethical considerations, and contextual knowledge that adult researchers may overlook or misinterpret. This approach reflects on (Clark et al., 2022a) and aligns with (Charisi et al., 2022) and (Department of Children, Equality, Disability, Integration and Youth, 2023) framework by redistributing authority and providing a **space**, **voice**, **audience**, and **influence** for youth perspectives in cyberbullying detection dataset development.

3.3 SDLC Phases

3.3.1 Phase-1: Requirement Gathering

Based on this, as depicted in Table 3.1 the 12 social media platforms are grouped into five broad categories i.e., *social networking and content sharing*, *messaging services*, *video streaming*, *discussion forums*, and *micro-blogging* platforms. This requirement-gathering phase is guided by two primary objectives:

1. Examining the relationship between adolescents social media engagement styles and their susceptibility to witness cyberbullying and to report or intervene.

2. Applying these insights to the curation of datasets and the development of tools that authentically capture youth online experiences, thereby advancing AI-driven detection and moderation systems.

Survey Design: Following established survey practices in social media research (Livingstone and Helsper, 2007; Nesi and Prinstein, 2015), the survey instrument employed sequential display logic (See Table 3.2) to minimise participant burden while ensuring relevance of questions. Respondents were initially asked to report the frequency of their usage of different platform types outlined in Table 3.1 over the past year on a 5-point likert scale (`Never, Hardly Ever, Sometimes, Often, Very Often`). This item served as a filter i.e., only those who indicated using a platform at least "`Sometimes`" were shown

Platform-Type	Platform-Feature	Platforms
Social networking and Content Sharing	Stories Reels (Short-videos)	Instagram TikTok
	Disappearing messages	Snapchat
	Comments	Facebook
Messaging	Group-chats	WhatsApp Discord
Video Streaming	Video-Content Video-comments	YouTube Twitch
Discussion Forums	Threads	Reddit ASK.fm 4Chan
Micro-blogging	Tweets	X (previously Twitter)

Table 3.1: Table indicating Platform-Type and associated Features

follow-up questions tailored to their preferred platform and its associated features. If the respondent met this initial threshold, a second question appeared, prompting them to describe their engagement behaviours with a specific feature of that platform. For example, if a respondent rates "`Sometimes`" for content-sharing platform, the subsequent question (*Q2*) seeks to understand their engagement style with each of the four platform features for content-sharing platforms. Following this, questions *Q3* through *Q7* (See Table 3.2) were asked for every combination of platform type and platform feature for which a user indicated a response of "`Sometimes`" or higher. For instance, if a participant engaged with both a content-sharing platform and a video-sharing platform, and on the former reported engaged with stories, reels, and comments, while on the latter they only engaged with video content, then the subsequent questions *Q3* to *Q7* would be prompted once each for *stories*, or *reels*, *comments*, and *video content*. The third question (*Q3*), pertaining to witnessing cyberbullying did not include the term "*cyberbullying*" as discussed in §3.2.4. It was adapted from the Global Kids Online survey instrument for measuring cyberbullying [60] as "*seeing someone else being treated in a nasty or hurtful way*". This *Q3* was prompted regardless of response choice to *Q2*. Similar to the use of "parallel questions" in the EU Kids Online project (Livingstone et al., 2015), subsequent question (*Q4*) pertaining to the eight forms of cyberbullying i.e., threat, exclusion, spreading rumours, and discriminatory behaviour based on ethnicity or race or body-image or sexual orientation or

[60] Core question K8`http://globalkidsonline.net/wp-content/uploads/2016/04/K-Hurtful-and-bullying-behaviour.xlsx`

disability (See §3.2.4) were only asked to respondents indicating at least "`Sometimes`" to *Q3*. The response choice for this question was on a 3-point Likert scale as `yes`, `no`, `Prefer not to say`. Subsequently, the survey asked to indicate the modality in which the online harm occurred (e.g., text, video, image) (*Q5*). In addition, drawing from (Thorn and Benenson Strategy Group, 2021), the survey also investigated youth responses to witnessing harmful content (*Q6*), including actions such as `reporting`, `muting`, `informing a parent/guardian/teacher`, or `ignoring` the incident. The subsequent question (*Q7*), drawn from (Thorn and Benenson Strategy Group, 2021) seeks to understand why the respondent would choose to ignore it by indicating either they felt it was not a big deal, worried report wouldn't be anonymous or they would loose their friends, felt they were to blame, or preferred not to say. This method of only asking questions to respondents who select particular options, not only minimised participant burden and improved data relevance but this sequential display logic also served as a tool to identify and segment "**at-risk**" groups based on their engagement patterns and reported experiences. It is worth noting that while these targeted approaches effectively focus on a specific subset of participants, it introduces a key trade-off i.e., it may overlook insights from respondents who misreported their initial exposure, potentially resulting in incomplete data regarding the broader scope of the survey (Manski and Molinari, 2008). Nevertheless, both the restrictions i.e., sequential display logic and the decision to adopt a bystander framing were mandated by the study's ethics approval, aimed at minimising participant distress and safeguarding individuals by limiting sensitive questions to those least likely to have relevant experiences.

Procedure and Participants: Developed and hosted on the Qualtrics platform, this survey underwent pilot testing to refine its clarity, logic, and technical design by forty-five first-year undergraduate students aged 18–19, who were asked to simulate typical responses and provide usability feedback. Based on insights from both pilots, minor refinements were made, including the addition of a visible progress bar, input validation checks (e.g., mandatory fields and age restrictions), and clearer wording of certain items to ensure accessibility across adolescent reading levels. For the main deployment, purposive sampling was employed to target 500 or more adolescents aged 13–17 who were active users of social media. The survey was disseminated via school networks across Ireland and accessed

#	Survey Questions	Response Options	Display Condition
Q1	Thinking about your experience in the past year with [`platform-type`], which of the following platforms do you use?	Never (0); Hardly ever (1); Sometimes (2); Often (3); Very Often (4); Prefer not to say	Always displayed (initial question)
Q2	Thinking about your experience with [`platform-feature`] on your preferred [`platform`], please indicate how you engage with it.	See Table 3.3	Displayed if previous response (Q1) $\geq$ Sometimes (2)
Q3	On [`platform-feature`], how often have you seen someone else being treated in a hurtful or nasty way?	Never; Hardly ever; Sometimes; Often; Very Often	Displayed to all respondents of (Q2)
Q4	On [`platform-feature`] have you seen any of the below happening to someone else?	See Table 3.4	Displayed if previous response (Q3) $\geq$ Sometimes (2)
Q5	What was the type of the hurtful or nasty content?	Only Text; Only Videos; Only Pictures; Videos with Text captions; Pictures with text captions; Prefer not to say	Displayed if response to Q4 is "*Yes*"
Q6	When this happened, which of the following did you do in response?	Reported; Blocked; Muted; Ignored; Told a friend; Told a parent; Told a teacher; Prefer not to say;	Displayed if response to Q4 is "*Yes*"
Q7	Which of the following describe why you ignored it?	Felt it was not a big deal; Worried report not anonymous; Worried about losing access/friends; Felt to blame; Prefer not to say;	Displayed if response to Q6 is "*Ignored*"

Table 3.2: Survey Questions with conditional display of questions based on participant responses

Engagement Type	Description	Response Choices
I View	Scroll or browse through others' content	Never Hardly Ever Sometimes Often Very Often Prefer Not to Say
I React to	Use emojis, emoticons, or like others' content	
I Share	Upload and share own content	

Table 3.3: Engagement with Platform Feature and Response Choices modified and adapted from Shaw et al. (2022)

Type of Harm	Response Choices
Someone was threatened Someone spread untrue rumour or stories about another person Someone was targeted because of their skin colour Someone was targeted because of their religion Someone was targeted because of their sexual orientation Someone was targeted because of their disability Someone was targeted because of they look Someone was left out or excluded	Yes No Prefer Not to Say

Table 3.4: Categorisation of online harm adopted by Rodríguez-de Dios et al. (2018); Stoilova et al. (2021)

through an anonymous Qualtrics link. Upon opening the link, respondents were first asked to identify whether they were a parent or a student. If a parent or guardian initiated the survey, they were presented with a Plain Language Statement (PLS) and subsequently asked to provide informed consent before passing the survey to their child for assent (See §A.7). If a teenager accessed the survey directly, they were instructed to have their parent or guardian read the PLS and provide consent before proceeding. In both cases, the teen respondent was required to provide assent following parental consent, in accordance with ethical standards for research involving minors. Eligibility was determined through initial screening questions that confirmed both the age and self-reported use of at least one social media platform. Participants who did not meet these criteria were automatically exited from the survey to preserve data quality and relevance. The sequential conditional logic in the survey reduced participant fatigue by limiting irrelevant questions but also enhanced the ecological validity of the data. By filtering out responses from those without relevant experiences, this logic helped mitigate the risk of speculative or inauthentic answers, which can be especially problematic in studies addressing sensitive or socially charged topics such as bullying and online harm.

Ethical Considerations: Conducting anonymous online surveys with adolescents on sensitive topics like online harm and cyberbullying presents unique ethical challenges like the risk of re-triggering distressing or traumatic memories. Unlike in-person settings, online surveys limit researchers ability to monitor emotional reactions or offer immediate

support if participants become upset. A central ethical concern is "*how to explore young people's experiences with cyberbullying without prompting them to relive painful events*". To address this, the study adopted a key mitigation strategy: focusing on adolescent experiences as ***bystanders or witnesses***, rather than asking them to recall instances of personal victimization or perpetration. This participatory quantitive survey methodology involves framing the questions and response choices around what participants have observed online rather than what they have personally endured or done. Furthermore, as mandated reporting is a statutory obligation[61], the option of open-text to detail their experiences is not provided nor expected, thereby reducing the risk of causing emotional distress or having the respondent disclose active and on-going abuse. While this ethically sensitive approach does limit the ability to capture the "*lived*" experiences of cyberbullying victimisation or perpetration, it reflects a deliberate and ethically informed design choice. Additionally, the survey requires all participants providing informed consent or assent following parental approval, in accordance with ethical standards for research involving minors. The survey methodology also had other ethical safeguards and these included (a) informed parental consent and participant assent, highlighting the sensitive nature of the topics and allowing withdrawal at any time, especially for those with past or current mental health concerns; (b) clear content warnings at the start, advising participants to continue only if they felt comfortable; (c) voluntary skip options for all questions; (d) a help button available throughout the survey which provides links and resources to national and school-based mental health support services, including confidential helplines and digital well-being resources for youth; and (e) a skip and flag feature that allowed participants to skip questions and prevent them from reappearing later in the survey. Furthermore, while the primary aim of this study is to inform the development of detection tools, the decision to exclude direct personal experiences represents a limitation. However, in the context of anonymous online surveys, researchers cannot fully mitigate the risk of unintentionally re-triggering trauma or causing emotional distress. This trade-off was carefully considered and reviewed by the university's ethics committee as part of the study's ethical protocol and subsequently approved (Reference Number: DCUREC/2022/126).

[61]See Ireland's Children First Act Section 10/11 Schedule 1 `https://www.irishstatutebook.ie/eli/2015/act/36/enacted/en/html`

Analysis: To analyse the relationship between social media engagement styles and the witnessing of repeated hurtful or nasty behaviour i.e., *cyberbullying*, both descriptive and inferential approaches were adopted. Fisher's Exact Test (Kim, 2017) was used to assess the statistical significance of associations between **feature engagement styles** (*viewing*, *reacting*, *interacting*) and the likelihood of **witnessing cyberbullying** across all features. This test was chosen because of the data's categorical nature and the small subgroup sizes in the survey's sequential display logic. This test is particularly suited for sparse contingency tables and provides an exact p-value without relying on large-sample approximations (Kim, 2017). To ensure stability of odds ratio estimates in cases with zero-cell counts, the Haldane–Anscombe correction (Weber et al., 2020) was applied. Given that the sequential logic of the survey questions (see Table 3.2) identifies "at-risk" participants as those who select a response of "*Sometimes*" or higher on the 5-point Likert scale, we use three different threshold-based reclassification strategies. This allowed dichotomising the 5-point Likert responses and test if the observed associations held true under various reclassification methods. Under this sensitivity analysis the three reclassification methods are, **Method A** *(Sequential Survey Logic)* i.e., all three engagement styles (viewing, interacting, reacting) and witnessing cyberbullying were re-coded as "`Sometimes/Often/Very Often = Yes (1)`", and "`Hardly ever/Never = No (0)`". Other two methods for reclassification involved **Method B** *Lenient Threshold*, "`Never`" was coded as `No (0)` and all other responses as `Yes (1)`. In contrast, **Method C** *Strict Threshold* classified only "`Often/Very Often` as `Yes (0)`, with all other responses coded as `No (1)`. The purpose of this approach was to test whether findings were stable under different operational definitions of "**at-risk**" feature usage style and witnessing cyberbullying content. Descriptive statistics were also used to analyse engagement across platforms and features, the prevalence and types of harm, and how participants responded as bystanders. These insights contextualised the inferential results and helped identify recurring patterns relevant to system requirements, such as common harm modalities and reporting practices among adolescents. Together, this analysis provided both quantitative evidence and contextual depth for understanding how minors experience and respond to harm on social media, thereby informing the subsequent stages of dataset design and harm detection system development.

3.3.2 Phase-2: High-level Design

A key challenge in existing cyberbullying datasets is the lack of ecological validity. Many cyberbullying scenarios are created by researchers and may not accurately reflect the types of harmful interactions that teens actually encounter online. While some studies involve youth input, few have validated whether the scenarios align with real experiences shared by young people. Studies like (Huang et al., 2020; Campbell and Xu, 2022) have explored youth perspectives on the *seriousness* of cyberbullying scenarios/vignettes through quantitative surveys using likert scales. Additionally, researchers have emphasized the importance of *realism* as a key dimension for ensuring the ecological validity of vignettes (Agha et al., 2024). The dimension of *report-worthiness* is also essential for understanding how youth might respond to these incidents, as the work by (Thorn and Benenson Strategy Group, 2021) on reporting behaviour highlights. From the perspective of the Routine Activities Theory (See §3.2.2), reporting options represent an attempt to offer capable guardianship, and the questions adapted from (Thorn and Benenson Strategy Group, 2021) help identify youth inclinations to seek support. To my knowledge, no study has attempted to integrate youth perspectives across all three dimensions, i.e., *seriousness*, *realism*, and *report-worthiness* to ascertain the ecological validity of cyberbullying vignettes or scenarios for subsequent dataset development. To that effect, this phase centres on **Scenario Evaluation**, serving as a foundational filter for the subsequent development of a youth-informed cyberbullying (CB) dataset. The overarching objective of this phase is to ***identify*** and ***assess*** a set of representative CB scenarios or vignettes through a structured evaluation process involving young participants. Driven by the research inquiry to understand "***What is the relationship among the seriousness, realism, and report-worthiness of current scenarios (vignettes), and how can they be adapted to suit the latest social media features?***", the *Scenario Evaluation* adopts a *quantitive survey* methodology to capture youth perspectives across these dimensions. The analysis in this phase informs the selection of scenarios that most closely reflect the realities of adolescents online experiences, thereby enhancing the ecological validity of the dataset. In addition, the phase also includes **Scenario Adaptation** with young people to ***refine*** the existing cyberbullying vignettes drawn from prior literature. Participants revise how the situations described in the original vignettes could occur on today's social

media platforms, incorporating features like ephemeral content and current platform dynamics. These evaluative and adaptative dimensions were strategically chosen to inform the design of a contextually grounded and socially meaningful dataset in Phase 3.

Study Design: Both *Scenario Evaluation* and *Adaptation* begins with a scoping review of the literature conducted between *November 2022* and *February 2023* using keywords including *"cyberbullying", "scenarios", "vignettes", and "teen*"* combined with AND/OR operators and conducted across multiple scholarly databases like Google Scholar, Scopus[62], JSTOR[63], IEEE Xplore[64], and ACL Anthology[65]. Using an inclusion criteria that selected studies featuring three or more CB vignettes in (a) qualitative or quantitative research or (b) studies wherein scenarios co-designed by teenagers or (c) studies involving teenagers in co-design processes, this scoping review identified 42 relevant articles. Among these, six key studies were identified for their comprehensive scenarios applicable across all settings, including works by (Ashktorab and Vitak, 2016; Camelford and Ebrahim, 2016; Huang et al., 2020; Sprugnoli et al., 2018; Ollagnier et al., 2022; Campbell and Xu, 2022). Building on the survey design established by (Huang et al., 2020) to gauge perception of CB seriousness for hypothetical scenarios, **Scenario Evaluation** was conducted via a survey which included an additional measure to assess scenario realism (Agha et al., 2024). Each scenario as part of *Scenario Evaluation* was evaluated based on three statements: "*This incident is a serious matter*," "*This incident is realistic*," and "*I would report this incident*". Participants responded to these statements using a 7-point Likert Scale ranging from *Strongly Disagree to Strongly Agree* (Huang et al., 2020). When participants indicated a scenario was report-worthy (responses of "Somewhat Agree" or higher), they received follow-up questions inspired by (Thorn and Benenson Strategy Group, 2021) as seen in §3.3.1). As part of **Scenario Adaptation**, young participants use a mix of ideas approach to align cyberbullying scenarios previously rated in *Scenario Evaluation* with current social media features such as disappearing messages, stories, and reels. These features are clearly defined and contextualised to reflect the evolving dynamics of platforms within cyberbullying contexts. Participants are then encouraged to rewrite the scenarios,

[62] `https://www.scopus.com/pages/home?display=basic#basic`
[63] `https://www.jstor.org/`
[64] `https://ieeexplore.ieee.org/Xplore/home.jsp`
[65] `https://aclanthology.org/`

showing how cyberbullying could spread further using these newer features in both online and offline settings.

Procedure and Participants: Participants in this phase were drawn from two cohorts for **Scenario Evaluation and Adaptation**. Cohort_1 consisted of ten students (ages 15–16; M=5, F=5), who participated as "*young research assistants*" during a five-day internship at the host research institution (See §3.4 for details) and Cohort_2 comprised of 20 students (ages 13–17; M=6, F=14), were recruited from a national youth advisory panel affiliated with a non-profit organisation[66]. Both Cohort_1 and Cohort_2 participated in *Scenario Evaluation* through Google sheets and Google forms, respectively. For Cohort_2 the non-profit organisation administered the survey and processed the responses, then directly shared the results. Due to time limitations, Cohort_2 evaluate a subset of six scenarios, pre-identified by Cohort_1 as the most realistic and report-worthy using Google Forms, optimized for asynchronous participation. The scenarios or vignettes used in both evaluation and adaptation are adopted from (Ashktorab and Vitak, 2016; Camelford and Ebrahim, 2016; Huang et al., 2020; Sprugnoli et al., 2018; Ollagnier et al., 2022; Campbell and Xu, 2022).

Ethical Considerations: This phase involves two cohorts of young people i.e, $\mathbf{Cohort_1}$ consisted of 15 to 16 year-olds, who were recruited through a "*young research assistant*" programme (details described in §3.4), while $\mathbf{Cohort_2}$ were recruited through collaboration with non-profit organisation that engage with youth panels (Webwise[67]. Rather than asking participants if they had personally experienced the cyberbullying scenarios or vignettes, the scenario evaluation focused on their opinions regarding how *realistic*, *serious*, or *report-worthy* the scenarios appeared. From the beginning and throughout the survey, respondents were informed that they could skip any or all vignettes, and any skipped items would not appear again later. Additionally, for both cohorts additional ethical safeguards included (a) informed parental consent and participant assent, highlighting the sensitive nature of the topics and allowing withdrawal at any time, especially for those with past or current mental health concerns; (b) clear content warnings at the start, advising partici-

[66] Webwise Youth Advisory Panel `https://www.webwise.ie/news/youth/what-is-the-webwise-youth-advisory-panel/`

[67] `https://www.webwise.ie/`

pants to continue only if they felt comfortable; (c) voluntary skip options for all questions; (d) a help button available throughout the survey which provides links and resources to national and school-based mental health support services, including confidential helplines and digital well-being resources for youth; and (e) a skip and flag feature that allowed participants to skip questions and prevent them from reappearing later in the survey.

Analysis: To analyse youth responses, a two-step quantitative approach was adopted. First, descriptive statistics were calculated to summarize how youth rated each cyberbullying scenario across the dimensions of **seriousness**, **realism**, and **report-worthiness**. This included frequency distributions and measures of central tendency, providing an overview of trends in youth perceptions. Given the small sample size and potential for low expected frequencies, Fisher's Exact Test was employed to examine associations between variables. Of particular interest was whether perceived seriousness and realism of a CB vignette were linked to the likelihood of reporting. Together, this analysis offered both descriptive insight and inferential clarity, informing the selection of the most relevant and actionable scenarios for adaptation in Phase 3.

3.3.3 Phase-3: Detailed Dataset Design

This phase represents the transition from high-level scenario identification to the in-depth, participatory construction of a benchmarking dataset tailored to youth cyberbullying (CB) experiences. Central to this phase is the implementation of a cascading co-research methodology, where $\mathbf{Cohort_1}$ (ages 15–16) participants or "*young research assistants*" serve as co-designers and iterative contributors within a structured research assistantship. This approach not only foregrounds the epistemic agency of young people in the research process, but also ensures that the resulting dataset is grounded in today's digital practices, contextual language, and evolving platform features. Each young researcher participated individually over a five-day in-person internship at the research institution. After completing a foundational skills-building module that covered spreadsheet functions (e.g., filters, pivot tables, drop-downs) and introductory Python programming (e.g., data types, built-in functions, unary conditionals) (See §3.4). The participants also engaged in the adaptation of CB vignettes i.e., rephrasing the scenarios evaluated as part of Phase 2 (§3.3.2) for

ecological validity.

Procedure: The co-research methodology plays a pivotal role in enhancing the depth, richness, and continuity of creating a youth cyberbullying dataset. Each participant in Cohort_1 referred to as the "*young research assistants*" after evaluating and adapting selected scenarios (Phase-2 §3.3.2), were given an option to paraphrase the scenarios. For example, a scenario referring to "*use of websites to demean someone*" could be rephrased as "*use of fake profiles to post demeaning content*". Following this optional sub-task, participants were encouraged to compose at least three messages each as enablers (to escalate or instigate nasty or hurtful or mean behaviour) and as defenders (to de-escalate or defend the victim) for scenarios they shortlist from evaluation activity in Phase-2. Shortlisting of the scenarios was done based on scenarios they rate equal to or higher than `somewhat agree` across all dimensions i.e., *realism, report-worthiness, seriousness.* By using drop-down filters, participants could identify scenarios meeting this criteria. In case of less than five qualifying scenarios, they were encouraged to select scenarios if they rated equal to or higher than `somewhat agree` for *serious* and *report-worthy.* As recruitment was in staggered intervals, each new participant first composed a minimum of six messages and on subsequent days, they were provided with a sheet containing the scenarios and messages produced by previous participants. Following a *mix-of-ideas* or *bags-of-stuff* co-design approach (Guha et al., 2004), participants were required to either adjust the sequence of existing "*bag of messages*" or confirm their order. For instance, a participant might begin the cyberbullying conversation by writing a casual greeting ("`Hi`"), or start directly by placing the offensive remark as the first message, or reverse the sequence so that a defender message or victim response appears immediately after an enabler's provocation. Subsequent participant would then build the conversation from the previous participant's work or alternatively change the ordering as they see fit. This iterative process encouraged reflection on narrative coherence, escalation and de-escalation strategies and the cascading element attempts to mimic real-world digital conversations that emerge in online peer interactions. Moreover, this collaborative relay-style design fosters a shared authorship model, allowing for multiple youth perspectives to be layered within a single scenario. It also reduces individual cognitive burden by distributing the creative workload while preserving coherence of the scenario. The structure of these contributions supports

the development of multi-turn dialogue datasets across group-chats or comment-threads, where conversational shifts, emotional tone, and behavioural cues reflect youth perspectives to be later used for training or benchmarking natural language processing (NLP) models. Participants also contributed to the cultural and linguistic enrichment of each scenario by proposing slang phrases or words, emojis, and abbreviations that that reflect current youth communication practices. This methodological approach aligns with calls for co-research in youth-centred technology design (Guishard and Tuck, 2013; Clark et al., 2022a; Charisi et al., 2022), and offers a replicable framework for future co-design efforts that seek to build grounded, diverse, and socially informed datasets.

Ethical Considerations: In line with ethical best practices for youth involvement, the co-design methodologies used in **Scenario Adaptation** of*Phase 2* and the dataset design in *Phase 3* were grounded in meaningful, informed, and safe participation. These activities were conducted exclusively with $\mathbf{Cohort_1}$(ages 15–16), who participated on-campus during a structured, week-long engagement. This in-person format provided a more controlled and supportive environment for ethically navigating the design process. To ensure that youth voices were not only heard but respected and acted upon, the research process was guided by Department of Children, Equality, Disability, Integration and Youth (2023)'s "Participation with a Purpose" framework, which emphasises four key dimensions: space, voice, audience, and influence. Participants were provided with a safe, structured, and inclusive physical space on campus at the host research institution during the co-design sessions. Each young person was given access to a personal desk and laptop, allowing for focused, independent work. Holding the sessions in person at the institution created a supportive environment where we could build rapport, respond to emotional cues, and offer immediate guidance when needed. Moreover, there was a qualified counsellor available on call. This setting fulfilled the "space" dimension, ensuring that participants not only had a formal opportunity to contribute but also felt physically and emotionally secure while doing so. They were encouraged to express their opinions and perspectives on how cyberbullying scenarios or vignettes could be worded and how they might realistically spread across platforms, enabling their "*voice*" to shape both narrative and technical dimensions of the dataset. To facilitate "*influence*", each student worked individually but sequentially, wherein the output of one participant became the input for

the next. This iterative design allowed for the evolution of ideas and scenarios, giving each young person the opportunity to actively shape and extend the work of their peers. This approach emphasised continuity and reinforced the idea that each contribution had a direct impact on the developing dataset. In terms of "*audience*", participants' contributions were not only valued but also embedded into the final testing and benchmarking dataset, which is made available upon request. This ensures that their work had real-world relevance and visibility beyond the immediate research setting, reinforcing the principle that young people's input is not just heard but acted upon. Despite the controlled setting, comprehensive ethical safeguards were maintained to mitigate potential risks related to discussing sensitive topics such as cyberbullying and online harm. These included,

(a) Clear content warnings at the start of each activity, informing participants of potentially sensitive themes and encouraging them to continue only if they felt comfortable.

(b) A skip and flag feature enabled participants to bypass any scenario they found distressing, ensuring it would not reappear later in the process.

(c) Daily check-in and check-out meetings were held with the researcher to create a reflective space where participants could share feedback, raise concerns, or discuss any emotional discomfort they experienced.

(d) Respecting right to decline participation for specific activities and instead of offering alternate activities like learning and data analysis.

(e) Access to a qualified counsellor on-call, available throughout the co-design period to provide professional support in cases of distress, re-triggering, or emotional overwhelm.

(f) Allowing participants to designate a supportive adult (parent, school counsellor, youth worker) to whom researchers can direct concerns, if needed.

(g) Structured pre-briefing and debriefing sessions at the start and end of each day, designed to contextualise the themes of cyberbullying and online harm in an age-appropriate and conversational format.

3.4 Young Research Assistant Programme

The Young Research Assistant (RA) programme was conceptualised and developed as a five-day, in-person internship designed to position adolescents as co-researchers and co-designers in the study of cyberbullying. Grounded in participatory and youth-centred research frameworks (Guishard and Tuck, 2013; Clark et al., 2022a; Department of Children, Equality, Disability, Integration and Youth, 2023), the initiative reflects the principle of "**participation with a purpose**" which emphasises providing young people with meaningful **space**, **voice**, **audience**, and **influence**. Its dual aims were to equip participants with digital literacy and research competencies, while embedding their perspectives directly into dataset design. In doing so, the programme not only advances methodological authenticity and cultural relevance but also enacts young people's right to be heard in matters that affect them. Thereby, enhancing the authenticity, cultural relevance, and ethical integrity of the resulting cyberbullying datasets.

Structure and Participation: Participation unfolded through structured, on-campus sessions that combined technical training, collaborative design, and reflective practice. Each student was provided with a personal workspace, a Linux-based laptop, and guided hours (10 am–3 pm). The programme began with foundational learning including an introduction to hardware, software, operating systems, and everyday applications of AI before moving into discussions about cyberbullying. Importantly, these discussions focused on definitions and perceptions rather than personal disclosures, reducing the risk of re-traumatisation. Ethical safeguards were embedded from the outset and participants received clear content warnings, were reminded of their right to skip or decline activities, and were assured that opting out carried no penalty. Students engaged in phased tasks i.e., in Phase 1 (§3.3.1), they learned to organise and annotate cyberbullying scenarios using Google Sheets, applying tools such as filters, drop-down menus, and pivot tables. In Phase 2 (§3.3.2), they evaluated scenarios drawn from prior literature, categorising them by perceived realism, seriousness, and report-worthiness. They then adapted the most salient vignettes to contemporary digital contexts, embedding them within familiar platforms such as TikTok, Snapchat, and Discord. This ensured that the resulting dataset reflected not only scholarly typologies but also the lived online environments of young people. In Phase

3 (§3.3.3), participants extended the scenarios they had previously evaluated by generating short, role-based messages that simulated different bystander behaviours within social media conversations. For each shortlisted scenario, participants were instructed to compose a minimum of six messages, evenly divided between those that escalated the situation (e.g., enablers or aggressors) and those that de-escalated it (e.g., defenders or supporters of the victim). This ensured that every scenario reflected both harmful and protective dynamics. In addition to this, through a cascading co-research approach (§3.2.5), the participating adolescents were required to also (a) re-arrange the messages into a conversational flow, (b) revise or obfuscate prior messages, and (c) extend the conversation if needed. Through an iterative process, participating adolescents were encouraged to reflect on the narrative, escalation and de-escalation strategies of conversations. Further details on training and learning activities along with their impact and outcomes are discussed in Appendix A.1.

Ethical Considerations: In addition to the detailed ethical considerations discussed in Section 3.3.3, participation was voluntary, confidential, and adaptable to individual comfort levels. Safeguards included content warnings, skip/flag features, structured debriefs, and access to a qualified counsellor. Daily check-ins and check-outs created reflective spaces for concerns, while participants retained the right to withdraw at any point. As part of the programme's ethical commitments, all tasks were optional and participants retained full control over their engagement. They could pick and choose activities according to interest and comfort, or switch between tasks such as spreadsheet work, coding, and scenario design. This flexibility ensured that participation was not only learner-centred but also psychologically safe, reinforcing autonomy as a core ethical principle of the programme. By embedding these practices into everyday activities, the programme ensured that young people's voices were heard in a manner that was both meaningful and safe. Please note, this programme was only conducted after ethical approval from the institution (*Reference No. add DCUREC/2022/126*).

Recruitment: The programme was initially piloted with two students (1 male, 14; 1 female, 15) in November 2023 to test logistics, task timing, and instructional materials. Feedback from these pilot participants informed subsequent programme design, including the optimal number of scenarios (15), clarifications on obfuscation techniques, and

adjustments to Python training content. As these were pilot cases, their results are not included in the main study. Recruitment for main study involved purposive sampling of adolescents aged 14–16 in Dublin, prioritising active social media users with some exposure to either direct or indirect online harm. Twelve students ultimately participated across staggered sessions between 2023 and 2025. Recruitment and participation followed full ethical approval protocols, with informed parental consent, youth assent, and a psychologist on-call throughout due to the sensitive subject matter. The cascading cohort model allowed earlier contributions to shape subsequent activities, creating a longitudinal dialogue among participants and enriching the co-research process. Importantly, participation was flexible: students who preferred not to engage with cyberbullying-related materials were offered alternatives such as Python training, survey cleaning, or spreadsheet analytics. This reinforced autonomy and safeguarded psychological well-being.

3.5 Summary

In this chapter, we present the participatory methodology and framework employed to investigate "***how cyberbullying-related datasets and test sets can be ethically co-created with minors to reflect their lived experiences***". Grounded in a mixed-methods design with adolescents aged 13 to 17, this inquiry also explores several key aspects:

(a) *how young people engage with social media functionalities*;

(b) *how their patterns of engagement may influence their exposure to or involvement with cyberbullying-related content*;

(c) *the formal and informal reporting behaviours teens adopt when encountering cyberbullying*;

(d) *the style of discourse among teens surrounding cyberbullying situations*

These investigations are central to exploration of research question **RQ.1** (§RQ.1).

The chapter begins by outlining the theoretical and conceptual foundations shaping this work, including youth epistemic agency, co-research methodologies, and digital harm theories. These frameworks reinforce a participatory ethos aligned with Article 12 of the

UN Convention on the Rights of the Child (UNCRC) and Ireland's *Participation with Purpose framework* (Department of Children, Equality, Disability, Integration and Youth, 2023), positioning minors not merely as research subjects but as co-creators of knowledge. This ensures that the resulting outputs genuinely reflect young people's real-world digital experiences rather than abstract, adult-imposed constructs.

The study unfolds across three structured phases aligned with the Software Development Life Cycle (SDLC), each building upon the last to produce a benchmarking dataset. This dataset is not only technically sound but also socio-culturally informed, capturing the complexity and nuance of online harm as experienced by young people. Crucially, it reflects adolescent's perceptions of harm in peer-to-peer digital interactions as explicit, indirect, or obfuscated messages, and how they interpret severity, intent, and emotional impact of cyberbullying events through vignettes. This nuanced insight informs the design of more context-aware harm detection systems and contributes to ethical, youth-informed technology development.

Beyond risk mitigation, this chapter articulates the ethical framework guiding the research, embodying a proactive commitment to *dignified participation*. This approach centres emotional well-being, agency, and respect for minor's diverse lived experiences. Rather than treating risk as merely a problem to be managed, ethics are embedded throughout the research lifecycle to shape how questions are framed, participation structured, and power shared. By intentionally shifting focus away from eliciting personal trauma disclosures and focusing on observational insights, creative scenario adaptation, and peer-informed critique, the study minimises potential harm while affirming minor's perspectives as epistemically valid and essential to co-creating meaningful digital safety tools. Ethical design here transcends compliance, becoming a practice of care, shared responsibility, and respect for youth as active knowledge contributors rather than mere data sources.

Additionally, this chapter details the *Young Research Assistant* Programme, which empowered adolescents aged 15–16 in Ireland to serve as active co-researchers and co-designers of the cyberbullying benchmarking dataset. Grounded in participatory, youth-centred frameworks, the programme engaged participants in activities such as annotating and adapting cyberbullying vignettes to mirror realistic platform-specific contexts, and creating dynamic bystander response sequences. Concurrently, participants developed

technical skills in data handling, Python programming, and content moderation tools, fostering digital literacy alongside critical inquiry into AI's role in detecting online harm. The programme emphasised ethical safeguards, including voluntary participation, psychological safety, and flexible learning pathways, ensuring a supportive environment for engaging with sensitive content. By involving youth as knowledge producers rather than passive subjects, this initiative enriched the dataset with authentic perspectives and demonstrated a scalable model for collaborative, ethically grounded socio-technical research with adolescents.

In the following chapters, we discuss the results of this participatory methodological framework contributing to the development of the co-created dataset $\mathcal{D}_{\mathbf{GOLD}}$, designed to benchmark cyberbullying detection systems within youth conversational contexts.

Chapter 4

Youth Centred Insights into Cyberbullying

Abstract

This chapter presents the findings from the three-phase participatory youth methodology described in Chapter 3. To guide the dataset curation examined through research question **RQ1**, this chapter first reviews the outcomes in §4.1 for the requirement-gathering Phase 1 (§3.3.1 Chapter 3). These findings reveal where and how cyberbullying (CB) most frequently occurs online, showing that merely viewing content in group chats or comment sections substantially increases a young person's likelihood of witnessing CB-related harm. The evaluation of cyberbullying scenarios as part of high-level design (see 3.3.2 Chapter 3) discussed in §4.2 reveals that the perceived severity of an incident strongly influenced whether participants considered it worth reporting, an insight with direct implications for platform moderation design. In the final co-creation phase (§3.3.3 Chapter 3), multiple young research assistants collaboratively authored a dataset of 552 realistic cyberbullying messages, incorporating authentic teen language, slang, emojis, and evasion strategies, as detailed in §4.3. Additionally, a comparative analysis with existing Italian, French, and other real-world English datasets in §4.3.3 contextualises the youth co-created corpus and underscores its value for training

AI moderation systems that can model both harmful and supportive online dynamics.

4.1 Phase-1: Requirement Gathering Survey Results

As part of Phase 1 (§3.3.1), the survey was disseminated via school networks across Ireland and accessed through an anonymous Qualtrics link. Upon opening the link, respondents were first asked to identify whether they were a parent or a student. If a parent or guardian initiated the survey, they were presented with a Plain Language Statement (PLS) and subsequently asked to provide informed consent before passing the survey to their child for assent (See Appendix A.7). If a teenager accessed the survey directly, they were instructed to have their parent or guardian read the PLS and provide consent before proceeding. In both cases, the teen respondent was required to provide assent following parental consent, in accordance with ethical standards for research involving minors. The survey achieved a 51.61% response rate, with 112 adolescents completing the survey. Of these, 92.9% reported active social media use, yielding a final sample of 104 participants (M = 55; F = 49), aged 13 to 16 years. As depicted in Figure 4.1, ≈53% boys and ≈47% girls attempted the survey, with the majority of participants aged 14 years.

4.1.1 Social Media and Platform Usage

As illustrated in Figure 4.2, when examining individual frequency categories across social media platforms that enable content sharing, *Snapchat* had the highest number of respondents reporting "`often`" use (46 out of 104). *TikTok* recorded the highest "`very often`" usage (28 out of 104), while *Instagram* had the largest share of "`sometimes`" users (46 out of 104). In contrast, *Facebook* appeared largely absent from teen usage patterns,a finding consistent with previous research by (National Advisory Council for Online Safety, 2021). Within messaging platforms, *WhatsApp* showed moderate engagement, with 28 respondents using it "`sometimes`", 15 using it "`often`" and only 6 using it "`very often`". Other

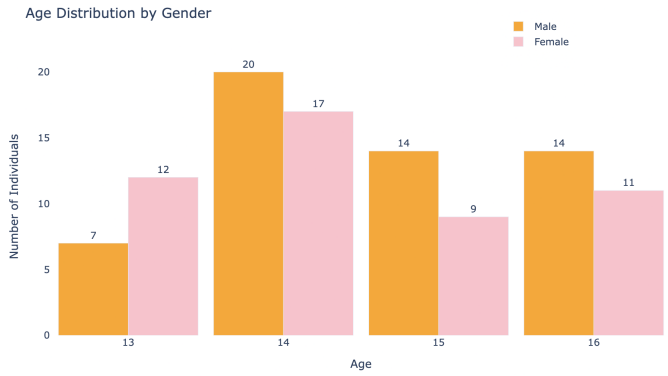


Figure 4.1: Distribution of age and gender of respondents for Phase-1 Requirement Gathering Survey

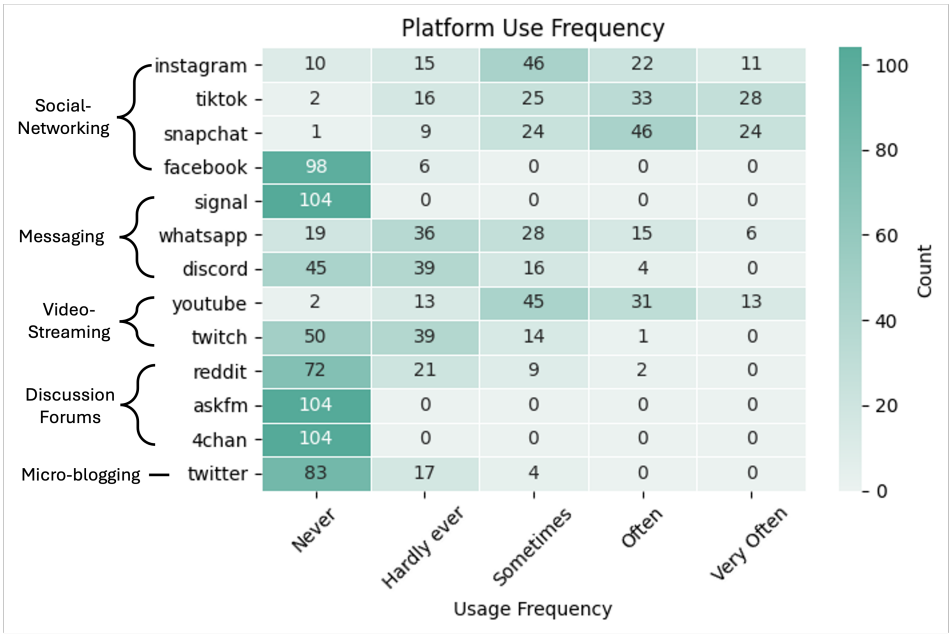


Figure 4.2: Platform Usage Frequency for Requirement Gathering Survey (Phase-1)

messaging platforms, such as *Discord*, attracted far fewer users, and *Signal* had no reported use. Video streaming platforms showed a different pattern of popularity. *YouTube* stood out, with only 15 respondents reporting that they used it "hardly ever" or "never", whereas *Twitch* had significantly more respondents in this low-use category. Discussion forums were generally the least used category. *ASK.fm*, previously employed by (Van Hee et al., 2018) to curate a cyberbullying dataset, was almost entirely unused, as was *4chan*. *Reddit* attracted minimal engagement, with only 2 respondents using it "often" and 9 using it "very often". Finally, *X* (formerly *Twitter*) had very limited reach, with just four respondents reporting that they used it "sometimes".

4.1.2 Platform Features and Engagement Styles

Following Shaw et al. (2022)'s engagement categorisation framework, distinct gender-based usage patterns emerged across viewing, reactive, and interactive behaviours for social media features. Results depicted in Table A.8 in Appendix A.8 are detailed as below,

- **Viewing Behaviour**: Reels demonstrated the highest viewing frequency across both genders, though with notable gender differences. Male respondents showed moderate-to-high engagement (44.44% "Often"), while female respondents exhibited predominantly high engagement (67.35% "Very Often"). Video content viewing was relatively balanced among males (45.28% "Sometimes", 32.08% "Often"), whereas females showed more distributed usage patterns (32.35% "Sometimes", 29.41% "Often", 32.35% "Very Often").

- **Reactive Behaviour**: Gender-specific preferences were evident in reactive behaviours. Male respondents demonstrated highest reactive engagement with group chats (48.28% "Sometimes"), followed by video content (50.94% "Sometimes"). Female respondents rated highest reactive behaviour towards reels (40.82% "Often"), with secondary engagement in group chats (63.16% "Sometimes"). Stories showed consistently low reactive engagement across both genders, with males reporting 63.64% "Hardly ever" and females 36.73% "Hardly ever".

- **Interactive Behaviour**: Interactive behaviours revealed conservative engagement patterns overall. Males showed highest interactive engagement with group chats (62.07% "Sometimes"), while females demonstrated comparable interactive behaviour with both group chats (68.42% "Sometimes") and disappearing messages (41.46% "Sometimes"). Video content showed minimal interactive engagement across both genders, with 83.02% of males and 100% of females reporting "Never" engaging interactively.

Overall, this descriptive analysis reveals significant gender-based differences in social media feature engagement patterns. Females demonstrate higher intensity viewing behaviours, particularly for reels, while males show more distributed engagement across multiple features. Reactive engagement preferences diverge by gender, with males favouring group-based interactions and females preferring content-focused reactions. Interactive behaviours

remain consistently low across both genders, suggesting passive consumption predominates over active participation in this sample population Moreover, discussion forums and micro-blogging platforms such as Reddit, ASK.fm, 4Chan, and X exhibited the lowest levels of usage among teen respondents. In contrast, YouTube, Snapchat, TikTok, Instagram, and WhatsApp were identified as the most commonly used social networking, messaging, and video streaming platforms.

4.1.3 Cyberbullying Exposure and Statistical Analysis:

Q3 in Table 3.2 sought information on witnessing cyberbullying from respondents as by-standers. Given these small sample sizes, to analyse by-standing cyberbullying exposure the 5-point Likert scale was dichotomised into "No (0)" `(Never/Hardly Ever)` and "Yes (1)" `(Sometimes/Often/Very Often)`, also referred to as *Method A - Sequential Survey Threshold* (See §3.3.1). As seen in Table 4.1, the most common features in which participants reported witnessing CB were group chats (72.92%) and disappearing content/messages (65.38%). In contrast, stories had a relatively high usage rate (58.65%) but the lowest reported witnessing CB (3.85%). Reels showed near-universal usage (97.03%), yet the proportion witnessing CB was moderate (37.62%). Video comments (50%), comments in posts (45%), and threads (45.45%) also exhibited substantial rates of witnessing cyberbullying, though the Threads feature had only 11 respondents. Due to the small sub-sample sizes, contingency tables for Fisher's Exact Test contained zero-cell counts. Addressing this, the Haldane–Anscombe correction (Weber et al., 2020) for odds ratio estimation was applied, and as depicted in Table 4.2, strong association between all three engagement styles across all platform features and the likelihood of witnessing cyberbullying can be observed. Respondents who passively engaged with all platform feature i.e., *viewing* behaviour per Shaw et al. (2022) had nearly 19 times higher odds of witnessing cyberbullying (Corrected Odds Ratio, COR = 18.86; $p < 0.05$) compared to those who did not. Those exhibiting *interactive* behaviour across all features had six times higher odds (COR = 6.34; $p < 0.05$). On social networking and content-sharing platforms, respondents who "*viewed comment*" sections were almost 26 times higher odds of witnessing cyberbullying (COR = 25.9; $p < 0.05$; See Table A.1 in Appendix A.2). Similarly, those who "*reacted*" to comments were about 11 times more likely to be exposed (COR = 11.16;

$p < 0.05$), while those who "*interacted*" with comments had nearly nine times higher odds (COR = 8.75; $p < 0.05$). Respondents who engaged across all features passively as per Shaw et al. (2022)'s *viewing* content behaviour, had nearly 19 times higher odds (Corrected Odds Ratio (COR) = 18.86, $p<0.05$) of witnessing cyberbullying than those who did not, followed by six times higher odds (COR=6.34, $p<0.05$) for respondents engaging in "*interactive*" behaviour across all features. For social networking and content sharing platforms, respondents were nearly 26 times at higher odds of "**viewing comment**" sections and witnessing cyberbullying (COR=25.9; $p<0.05$), similarly *reacting* or *interacting* to comments , respondents were nearly 11 times and 9 times more at risk of being exposed to cyberbullying, respectively (reacting: COR=11.16, $p<0.05$; interacting: ,COR=8.75 $p<0.05$). For group chats, Fisher's exact test produced an "∞" odds ratio for viewing engagement; however, after applying the Haldane–Anscombe correction, the odds ratio was estimated at 23.66 ($p < 0.05$), indicating a substantially higher risk of witnessing cyberbullying for those who "*viewed group chats*" compared to those who did not. Disappearing content and messages were also notable feature through which respondents reported witnessing cyberbullying, but only among those engaging through "*viewing*" (COR=10.31, $p<0.05$) or "*reacting*" (COR=3.78, $p<0.05$). On video-sharing platforms, both "*viewing*" (COR=29.23, $p<0.05$) and "*interacting*" (COR=6.29, $p<0.05$) with *video comments* were significantly associated with witnessing cyberbullying. Further details on the sensitivity analysis that help understand different operational definitions of "*at-risk*" groups are discussed in Appendix A.2.

4.1.4 Forms of cyberbullying witnessed and content modality

Respondents who indicated that they had witnessed cyberbullying in a by-standing capacity by selecting the categories *Sometimes, Often, or Very Often* across any of the listed features, were subsequently presented with Question 4, as outlined in Table 3.2. For analysis, multiple entries from the same participant were consolidated to produce unique responses. This process resulted in a total of 87 respondents out of the 104 individuals who had reported witnessing cyberbullying through any of the three engagement styles described in Figure 4.3. Across all features, the most frequently reported form of cyber-

Feature	**# Sub-samples**	**Usage Rate**	**Witness CB Rate**	**Usage × Witness Combinations**			
				Both Yes	**Use Only**	**Wit. Only**	**Nei-ther**
stories	104	58.65	3.85	4	57	0	43
reels	101	97.03	37.62	38	60	0	3
comments	100	72.00	45.00	45	27	0	28
disappearing	78	87.18	65.38	49	19	2	8
group_chats	48	93.75	72.92	35	10	0	3
video_con-tent	87	97.70	6.90	6	79	0	2
video_com-ments	70	85.71	50.00	35	25	0	10
threads	11	72.73	45.45	5	3	0	3
tweets	4	50.00	50.00	2	0	0	2

Table 4.1: Descriptive Statistics for Platform Usage and Cyberbullying Witnessing Behaviour by Platform Feature. Please note, usage and witness rate and counts are calculated by dichotomising as "No (0)" (Never/Hardly Ever) and "Yes (1)" (Sometimes/Often/Very Often)

Engagement Style	p value	Odds Ratio	Corrected Odds Ratio
viewing	0.0000	21.1082	18.8596
reacting	0.0000	2.9635	2.9510
interacting	0.0000	6.4138	6.3595

Table 4.2: *Method A (Sequential Survey Logic)* Fisher's Exact Test Results for association - Feature Engagement Style v/s Witnessing Cyberbullying

bullying was identified as "*threat*", with 95.4% of respondents indicating its occurrence. This was followed by harassment based on "*sexual orientation*" at 81.6%, and harassment related to "*physical appearance*" at 77%. The form of cyberbullying reported least frequently across all features was "*exclusion*", which was noted by only 43.7% respondents. Figure 4.4 provides a feature-level breakdown of these findings, showing the proportion of unique respondents reporting each type of harm within specific platform features. The annotations in each cell are presented as x/y, where x represents the number of unique respondents who reported that type of harm in the given feature, and y represents the total number of respondents who reported witnessing any form of cyberbullying in that feature. Consistent with the aggregate results, "*threat*" maintained the highest prevalence across nearly all features, often approaching or reaching 100% in features such as *comments*, *group chats*, *video-comments*, and *reels*. Harassment based on "*sexual orientation*" and "*physical appearance*" was also widespread, although it showed more variability across features, with lower proportions in stories and video content. In contrast, "exclusion" remained the least frequently reported form of harm, with minimal reporting in "*video-comments*" and "*reels*", despite higher levels in disappearing messages. Other forms of harm, including those based on "*skin colour, rumours, religion, and disability*", followed similar patterns of high reporting in *comments* and *disappearing content*, with reduced prevalence in ephemeral or less interactive features. Additionally, because *Q4* (Table 3.2) allowed respondents to select multiple types of harm, Figure 4.5 presents a co-occurrence matrix illustrating the extent to which different forms of cyberbullying were reported together. The values in each cell represent the percentage of respondents who selected both harm types. The diagonal values reflect the overall prevalence of each harm type, while the off-diagonal values indicate co-reporting rates. The most prominent overlap was observed between "*threat*" and "*sexual orientation*" harassment, with 80.5% of respondents who reported threats also indicating harassment based on sexual orientation. A similarly high co-occurrence was found between "*threat*" and "*physical appearance*" harassment (74.7%), as well as between religion-based harassment and "*threat*" (75.9%). Notably, "*religion*", "*skin-colour*" and "*sexual orientation*" based harassment was found to have 71.3% co-occurrence, indicating as yet another prevalent form of cyberbullying. In contrast, "*exclusion*" exhibited the lowest co-occurrence rates with other harm types, aligning with its lower overall prevalence

observed in the aggregate analysis and feature-specific breakdowns. As shown in Figure 4.6, "*only text*" was the most frequently reported modality through which various forms of cyberbullying were experienced. Because this was a multiple-choice question, respondents could select more than one option, and the figure presents the total counts across all features. Cyberbullying was also commonly reported through text-only, pictures with text captions, followed by videos with text captions, though these occurred less frequently than text-only instances. Figure 4.7 further illustrates that on features such as *comments* in social networking and content-sharing platforms, *group chats, disappearing content*, and *video comments*, the predominant modality was again only text to disseminate cyberbullying content. In contrast, for reels, cyberbullying content was more often associated with videos containing text captions, followed by pictures with text captions.

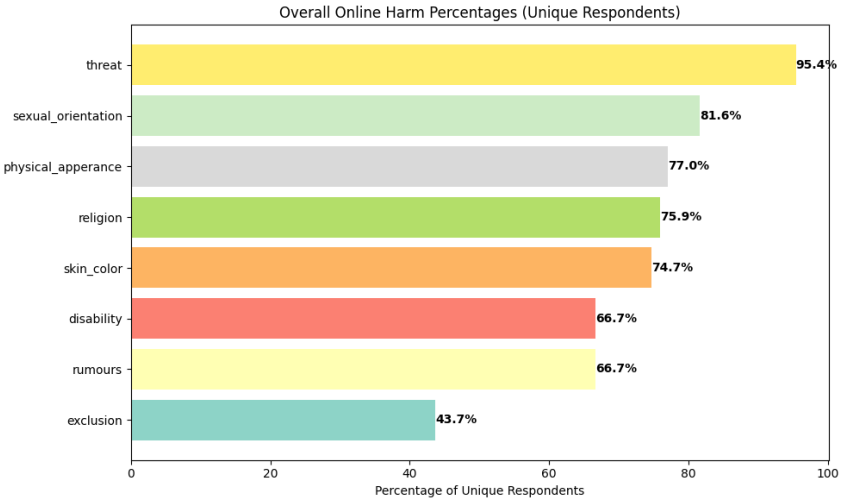


Figure 4.3: Percentage breakdown of Online Harms across all features collected via Requirement Gathering Phase-1

4.1.5 Response strategies

Regarding responses to cyberbullying content, *blocking* or *muting* the offending individual emerged as the most commonly employed strategies (See Figure 4.8). Blocking was particularly prevalent in relation to *disappearing messages* (n = 25), *comments* (n = 17), and reels (n = 18), suggesting that users tend to remove the perpetrator's access when confronted with more direct or persistent forms of harassment. Muting, while less frequent overall than blocking, showed its highest use in *group chats* (n = 17), perhaps reflecting situations where users want to avoid confrontation while still remaining in shared social

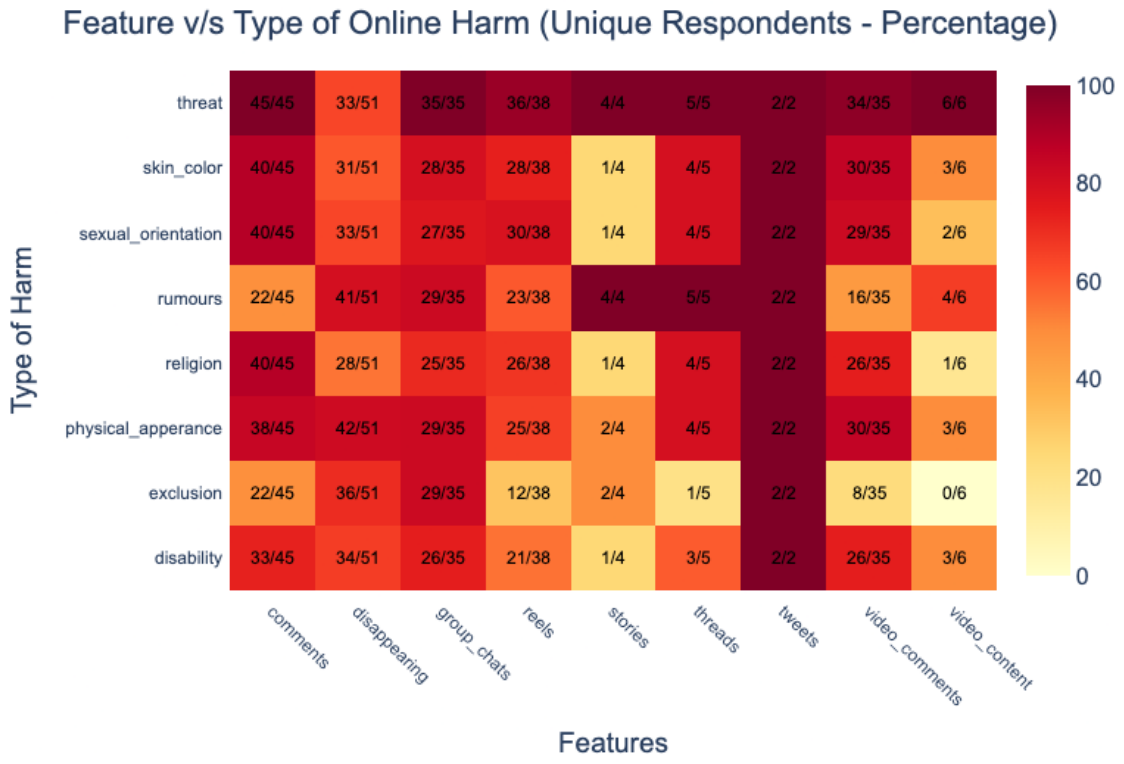


Figure 4.4: Heatmap showing the percentage of unique respondents reporting each type of online harm across platform features. Cell annotations are presented as x/y, where x = number of unique respondents reporting the specific harm type within that feature, and y = total number of respondents reporting any cyberbullying in that feature. Darker red indicates higher percentages, while lighter yellow indicates lower percentages.

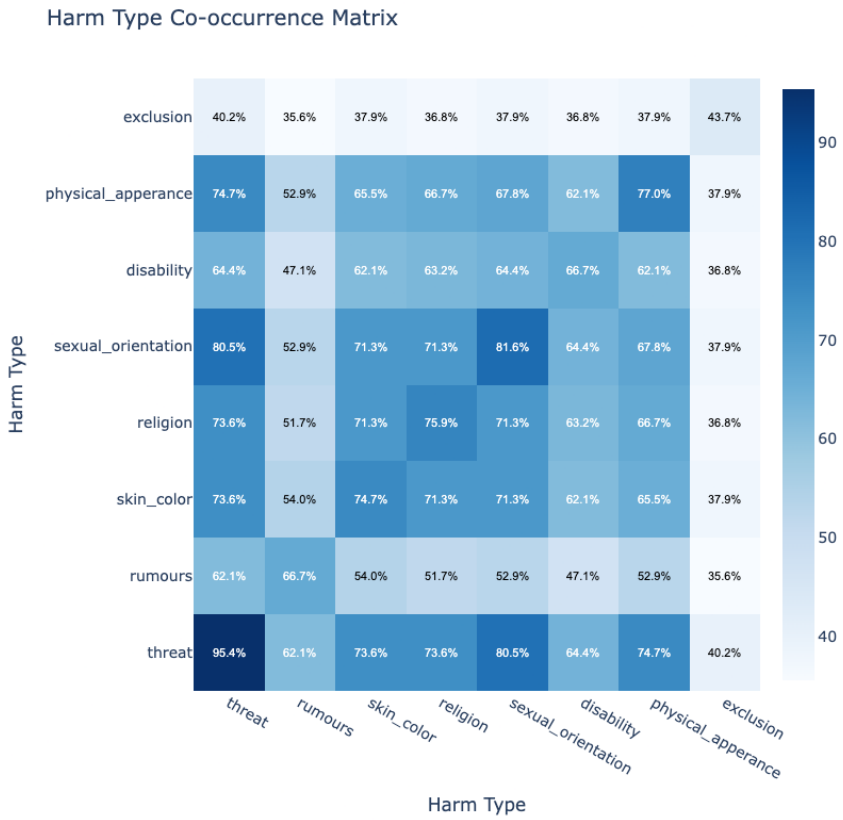


Figure 4.5: Co-occurrence matrix of different forms of cyberbullying in response to Q4 of Phase-1: Requirement Gathering

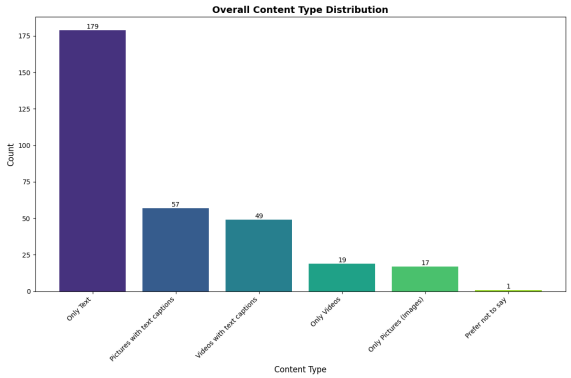


Figure 4.6: Distribution of all content modalities that respondents have witnessed different forms of cyberbullying (Q5 in Table 3.2)

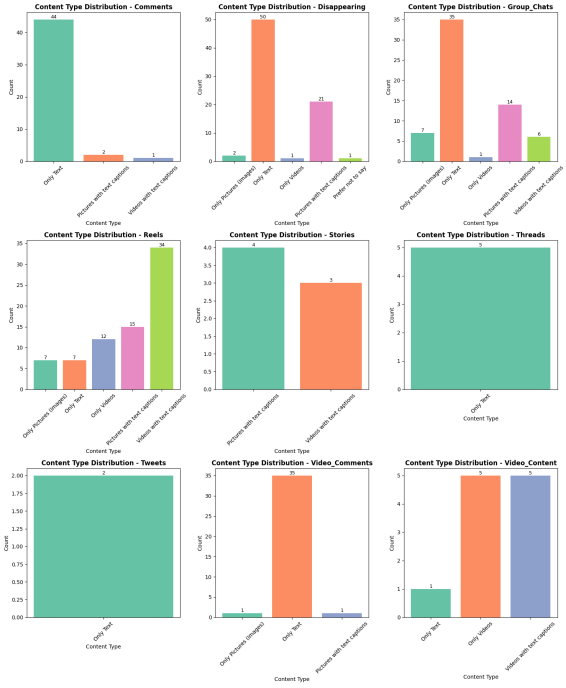


Figure 4.7: Distribution of content modalities that respondents have witnessed different forms of cyberbullying for specific features

spaces. "*Ignoring*" was also a notable response, especially for *comments* (n = 18) and *video comments* (n = 11). Interestingly, video comments had one of the highest "*ignore*" counts across all features, possibly indicating that the asynchronous and public nature of such content encourages avoidance over active intervention. Other responses, such as "*telling a friend or parent/guardian*", were comparatively rare, suggesting that social sharing of such cyberbullying incident is less common than direct action on the platform itself. When asked about ignoring behaviour, many respondents either chose not to disclose a reason or believed the incident was "*not a big deal*" (See Figure 4.9).

4.2 Phase-2: High-Level Design Results

This phase consists of two components: (a) **Scenario Evaluation** and (b) **Scenario Adaptation**. Scenario Evaluation adopts a survey-based design, whereas Scenario Adaptation follows a co-design methodology. Both align with the call for active participation of children and young people in AI system design (Charisi et al., 2022). To support both evaluation and adaptation of scenarios as discussed in §3.3.2 a scoping review was conducted between *November 2022* and *February 2023* identifying thirty cyberbullying sce-

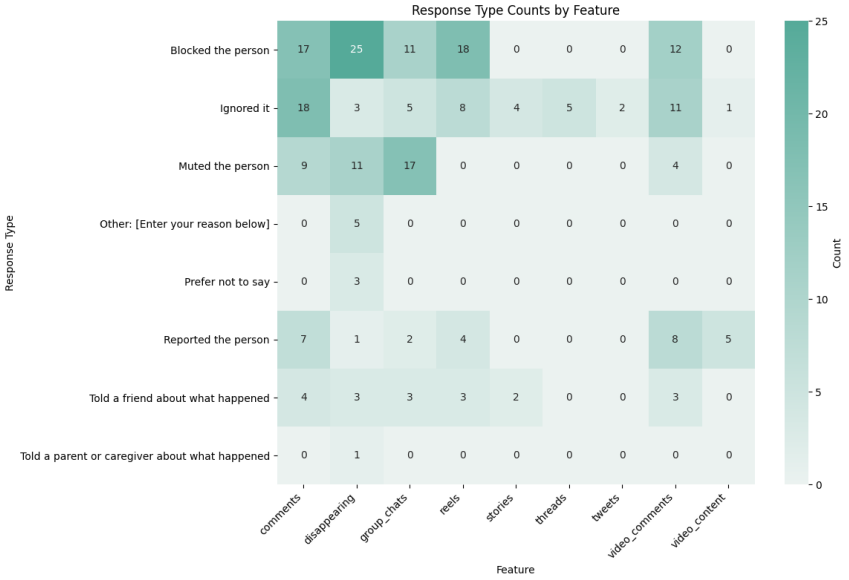


Figure 4.8: Response Type associated with different social media features collected via Requirement Gathering Phase-1

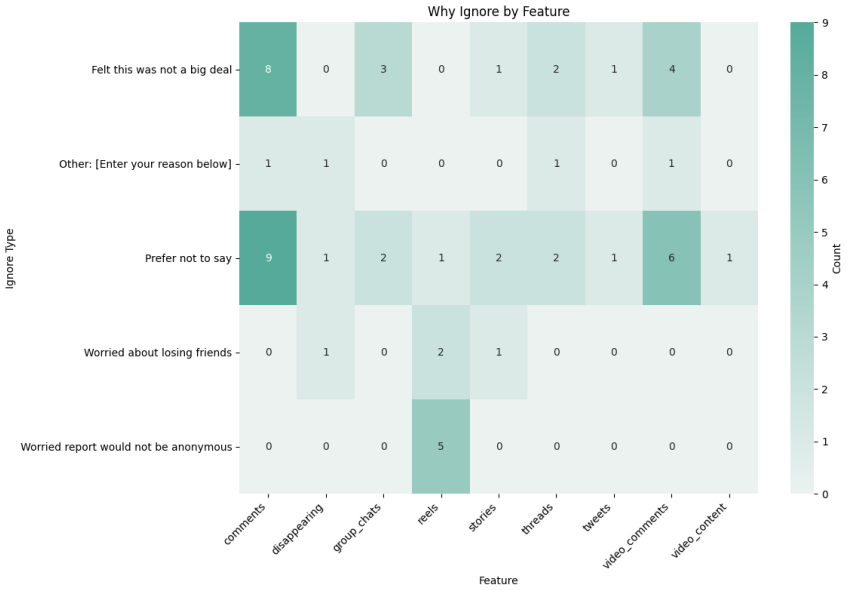


Figure 4.9: Why would they ignore reporting as per different social media features collected via Requirement Gathering Phase-1

narios from existing literature by (Ashktorab and Vitak, 2016; Camelford and Ebrahim, 2016; Sprugnoli et al., 2018; Campbell and Xu, 2022; Huang et al., 2020; Ollagnier et al., 2022), encompassing behaviours such as harassment, social exclusion, denigration, flaming, and outing, including overlapping and hybrid forms. From these, a purposive random sample of fifteen scenarios was selected to reflect both the distribution of behaviours and established definitions of cyberbullying (Hinduja and Patchin, 2013; O'Higgins Norman et al., 2023). Youth participation in *Scenario Evaluation* involved two participant cohorts: Cohort_1, consisting of ten students (aged 15-16, M=5, F=5) , who were recruited as "*young research assistants*", and Cohort_2 consisting of twenty students (ages 13–17; M=6, F=14) from an organisational youth panel (See Table 4.3). Cohort_1 initially comprising of twelve participants as described in §3.4, was reduced to ten after two female participants withdrew from participating in the cyberbullying part of the research assistantship programme (See Table 4.4). Instead, they mainly engaged in exploratory learning activities, such as basic Python operations, file handling, and experimenting with APIs (e.g., the OpenAI moderation API). Over five days, Cohort_1 (N=10) in addition to learning basic python programming and data handling skills with google spreadsheets, they evaluate scenarios/vignettes collected from previous literature and adapt them based on their perceived *seriousness*, *realism*, or *report-worthiness*. This integration of literature derived scenarios/vignettes with participatory evaluation and adaptation aims to ensure both ecological validity and youth-centred relevance in this high-level design phase and informs the detailed-level design phase (§3.3.3).

Cohort	Age	Gender	# Participants
Cohort_1	15	Female	3
		Male	3
	16	Male	2
		Female	2
Cohort_2	13	Female	1
	14	Female	1
	15	Male	4
		Female	3
	16	Female	3
		Male	1
	17	Female	6
		Male	1

Table 4.3: Age and gender distribution of Cohort_1 and Cohort_2

Timeline	Gender	Age	Study
February 2024	Male	15	Main
	Female	15	
April 2024	Female	15	
	Male	16	
May 2024	Male	15	Opt-out
September 2024	Female	16	
	Female	15	
October 2024	Female	16	
	Male	15	
February 2025	Male	16	Main
	Female	16	
May 2025	Female	15	

Table 4.4: Timeline and distribution of Cohort$_1$ - *young research assistants* in this study

4.2.1 Scoping Review Results

A scoping review of scenarios/vignettes related to teen cyberbullying, conducted between *November 2022* and *February 2023*, identified 42 relevant research articles across multiple domains. The inclusion criteria as discussed in §3.3.1 focused on studies that: (a) featured three or more cyberbullying (CB) vignettes in qualitative or quantitative research, (b) employed scenarios co-designed by teenagers, or (c) involved teenagers directly in co-design processes. From these, six key studies by (Ashktorab and Vitak, 2016; Camelford and Ebrahim, 2016; Sprugnoli et al., 2018; Campbell and Xu, 2022; Huang et al., 2020; Ollagnier et al., 2022) emerged for their comprehensive and widely applicable scenarios. These works are particularly notable because their scenarios have been used in evaluation and co-design studies. Collectively, these six studies contribute thirty distinct scenarios illustrating various manifestations of cyberbullying. Notably, while some studies explicitly classify each scenario according to the type of cyberbullying involved (Ashktorab and Vitak, 2016; Huang et al., 2020), others present vignettes without such categorisation. To address this gap, the scenarios can be contextualised using established categorisations of cyberbullying forms described by (Willard, 2007; Slonje et al., 2013; Kowalski et al., 2014; Bauman, 2015). These categorisations encompass a wide spectrum of behaviours, including but not limited to the following,

- **Flaming**: Hostile, angry, or insulting interactions, frequently involve hurtful personal attacks. E.g., "*User1 got in an argument with User2 [...] each message is vulgar and insulting [...]*" (Willard, 2007) or "*[...] sends messages like 'U R AN*

IDIOT AND I HATE U!"" (Bauman, 2015).

- **Harassment**: Repeated cruel or hostile actions intended to annoy, embarrass, or harm an individual. For example, "*[...] had received anonymous hurtful messages repeatedly [...]*" (Willard, 2007). It is important to note that Willard (2007) distinguish between *flaming* and *harassment*, with the latter characterised as being more one sided in nature.

- **Denigration**: Hostile actions to demean or disrespect another person based on their age or race or ethnicity or gender or characteristics or behaviour. E.g., "*[...] posted rumours, jokes about User [...] other people add comments*" (Willard, 2007) or "*[...] posted hateful comments and suggested User was a paedophile [...]*" (Bauman, 2015).

- **Masquerading**: Pretending to be someone else and sending messages that appear to come from the victim but are sent by the bully. E.g., "*[...] gained access to user's password and then sent provocative sexual images and messages from user's account [...]*" (Kowalski et al., 2014).

- **Outing and Trickery**: Although recognised as being in tandem, *outing* refers to actions that reveal real personal or intimate information including sensitive images, aimed at embarrassing individuals without their consent. Whereas, *trickery* refers to actions involving coaxing individuals to divulge personal information privately then making it public without their consent. E.g., "*[...] deceived into sharing sharing personal information and user shared it with others.*" (Willard, 2007; Bauman, 2015; Kowalski et al., 2014).

- **Exclusion**: Deliberate actions to leave individuals out of a group or make them feel unwanted. E.g., "*[...] has been excluded deliberately from IM group list [...]*" (Willard, 2007).

- **Cyberstalking**: Electronically stalking someone by repeatedly harassing or threatening them. E.g., "*User received anonymous messages 'you are going to die at [location]'* " (Bauman, 2015).

It is important to acknowledge that the categorisation of scenarios is not always mutually exclusive; instead, individual cases may encompass multiple forms of cyberbullying. For example, consider the scenario presented by Camelford and Ebrahim (2016)

> *Victoria, a girl in your friend group, sends a group photo text with a meme of Jessica, an unpopular girl from your school after she tripped and fell with the caption, "`Oh You Want to Be Friends With Us. It's Not Going to Happen`". Other friends respond with "`LOL`" and "`What a loser.`"*

The initial action i.e., "*[...] sends a group photo text with a meme of Jessica [...]*" constitutes denigration, as it aims to ridicule and damage Jessica's social reputation. Simultaneously, the caption "`Oh You Want to Be Friends With Us. It's Not Going to Happen`" reflects deliberate social exclusion by reinforcing her outsider status within the peer group. Accordingly, this scenario can be categorised as involving both denigration and exclusion. Detailed rationales for all thirty scenarios are provided in Table A.9 in Appendix A.9. As shown in Table 4.5, among the thirty scenarios, 14 occur purely online, 14 involve offline situations that transition to online settings, and two are purely offline. The offline scenarios are not included in this study. Based on the broad categorisation in Table 4.5, there are seven denigration scenarios, four each of harassment and exclusion, two flaming, one masquerading, and two scenarios from (Sprugnoli et al., 2018) (e.g., *shy male classmate sends invite for ballet recital on class group*" or *teacher's pet student reports students for smoking*") that describe events rather than direct cyberbullying forms are categorised as "`None`" because any form of cyberbullying could potentially occur in the event. Among overlapping categories, three scenarios involve *outing + denigration*, while two scenarios each involve *harassment + denigration* and *harassment + cyberstalking*. There are two scenarios involving three overlapping forms "*outing + denigration + flaming*" or "*outing + denigration + exclusion*, both from (Ollagnier et al., 2022). In one scenario, a student who is overweight is photographed by a jealous peer in a way that highlights their weight. The photo is then broadcast with mean comments, while others respond by defending or intervening. In this scenario, the sharing of embarrassing photo constitutes *outing*; broadcasting it with cruel comments constitutes *denigration*; and the hostile interactions that occur during the intervention, though brief and not repeated, constitute *flaming*.

Broad CB Category	Mode	Short-listed for study No	Yes
Denigration	Offline	1	0
	Offline translates to online	2	1
	Online	1	2
Denigration + Exclusion	Online	0	1
Exclusion	Offline	1	0
	Offline translates to online	0	2
	Online	0	1
Flaming	Offline translates to online	1	1
Harassment	Offline translates to online	1	1
	Online	2	0
Harassment + Cyberstalking	Offline translates to online	0	1
	Online	1	0
Harassment + Denigration	Online	1	1
Masquerading	Online	0	1
Outing + Denigration	Offline translates to online	1	1
	Online	0	1
Outing + Denigration + Exclusion	Offline translates to online	1	0
Outing + Denigration + Flaming	Offline translates to online	0	1
None	Online	2	0

Table 4.5: Distribution of broad cyberbullying categories of scenarios and their modes of occurrence, indicating whether these scenarios were shortlisted for inclusion in the study. Please note: "`None`" *category refers to two scenarios from (Sprugnoli et al., 2018) that describe events themselves rather than individuals' reactions, meaning no specific form of cyberbullying is directly applicable*

4.2.2 Scenario Evaluation

Cohort$_1$ (N=10) participants assessed 15 scenarios representing seven different forms of cyberbullying that occur individually or overlap with other forms (see Table 4.6). Eight of these scenarios were situations that translate from offline to online and seven which are purely online. As discussed earlier in §3.3.2, the scenarios were evaluated across three dimensions "*seriousness*", "*realism*", and "*report-worthiness*" using a 7-Likert scale responses ranging from "`strongly disagree - strongly agree`" as done previously by (Huang et al., 2020) for measuring *seriousness* of cyberbullying scenarios/vignettes. Due to time constraints imposed by the youth organisation, the Cohort$_2$ (N=20) (ages 13-17; 14F:6M) only evaluated six diverse scenarios (see Table 4.6).

Descriptive Patterns: Table 4.6 outlines the unique scenario numbers alongside their corresponding source references, the broad cyberbullying categories they represent, and the interaction mode (online or offline-to-online) through which the scenario is situated. As depicted in Figure 4.10, at a broader level the distribution of Likert-scale responses

Scenario IDs	**Source Study**	**Broad CB Category**	**Mode**	**Cohorts**
1	(Ashktorab and Vitak, 2016)	Harassment + Cyberstalking	Offline-to-online	1 and 2
2	(Campbell and Xu, 2022)	Outing + Denigration		1 and 2
3	(Ollagnier et al., 2022)	Outing + Denigration + Flaming		1
4		Flaming		1
5		Denigration		1 and 2
6	(Camelford and Ebrahim, 2016)	Denigration + Exclusion	Online	1 and 2
7	(Ashktorab and Vitak, 2016)	Denigration		1 and 2
8	(Ollagnier et al., 2022)	Outing + Denigration		1
9	(Ashktorab and Vitak, 2016)	Exclusion		1
10	(Campbell and Xu, 2022)	Exclusion	Offline-to-online	1
11	(Huang et al., 2020)			1
12	(Ashktorab and Vitak, 2016)	Harassment + Denigration	Online	1 and 2
13	(Campbell and Xu, 2022)	Denigration		1
14	(Huang et al., 2020)	Harassment	Offline-to-online	1
15	(Camelford and Ebrahim, 2016)	Masquerading	Online	1

Table 4.6: Mapping of scenarios as per scenario-number, their sources, broad cyberbullying categories, and interaction modes.

across all scenario cyberbullying types reveals substantial variability in Cohort_1. Among the three *denigration* scenarios, 29.6% "`disagreed`" that the scenarios were *real*, while only 29% and 21% "`strongly agreed`" they were *serious* and *report-worthy*, respectively. Similar variability persists among participants in Cohort_1 for *exclusion* scenarios. Although, for both the *harassment + denigration* and *harassment + cyberstalking* scenarios there are no disagreement responses for *seriousness* of these scenarios, variability persists for *realism* and *report-worthiness* with 23.5% "`somewhat disagreeing`" for realism in *harassment + cyberstalking* scenario and 16.7% "`strongly disagreeing`" with realism for *harassment + denigration* scenario. Similarly, at a broader level, as illustrated in Figure 4.11, Cohort_2 too exhibits substantial variability in participant responses, particularly in relation to *denigration* scenarios across all three dimensions. Perception of both *realism* and *report-worthiness* among Cohort_2 diverges considerably for *denigration*, *denigration + exclusion*, and *harassment + denigration* scenarios. Although similar to Cohort_1 for scenario pertaining to *harassment + cyberstalking* there are no disagreements, however, for *harassment + denigration* unlike Cohort_1 there are 22.2% neutral i.e., "`no opinion`" responses. Additionally, as shown in Figure 4.12, when results from both cohorts are combined across the six scenarios, the variability in responses regarding realism remains consistent across all scenario categories. Similarly, the variability in report-worthy responses related to *denigration + exclusion* in the combined data closely mirrors that of the individual cohorts. These patterns highlight the differing perspectives among youth and emphasize the importance of capturing such variability in evaluations. The disagreement evident on the 7-point Likert scales is reflected also in the low inter-rater reliability scores (Krippendorff's α; see Table 4.7) across both cohorts individually and combined. Thereby suggesting that the three dimensions assessed are highly subjective and that clearer operational definitions of each response category are required. Further details are discussed in Appendix A.3.

Cohort	**Realism** α	**Seriousness** α	**Report-worthy** α
Cohort_1	0.150532	0.286280	0.288340
Cohort_2	0.020334	0.281563	0.112582
Both	0.013620	0.190329	0.158631

Table 4.7: Inter-rater agreement between both cohorts for each dimension presented as Krippendorff's Alpha (α)

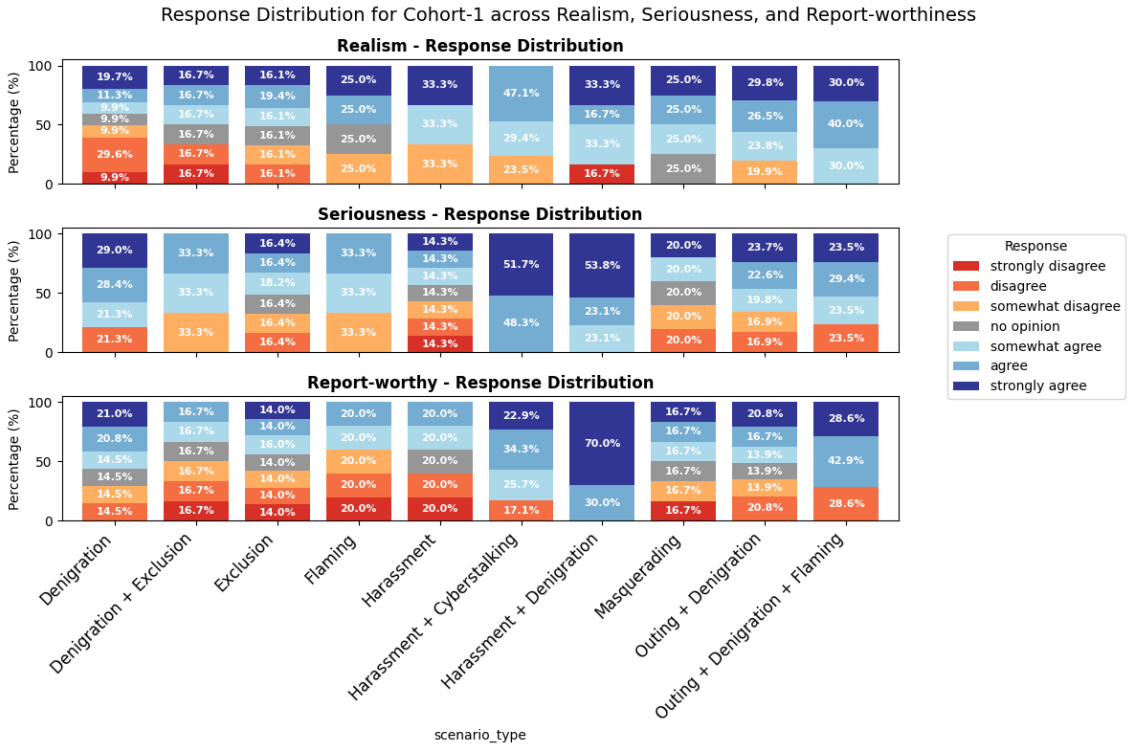


Figure 4.10: Likert-scale responses across Scenario categories by Cohort_1

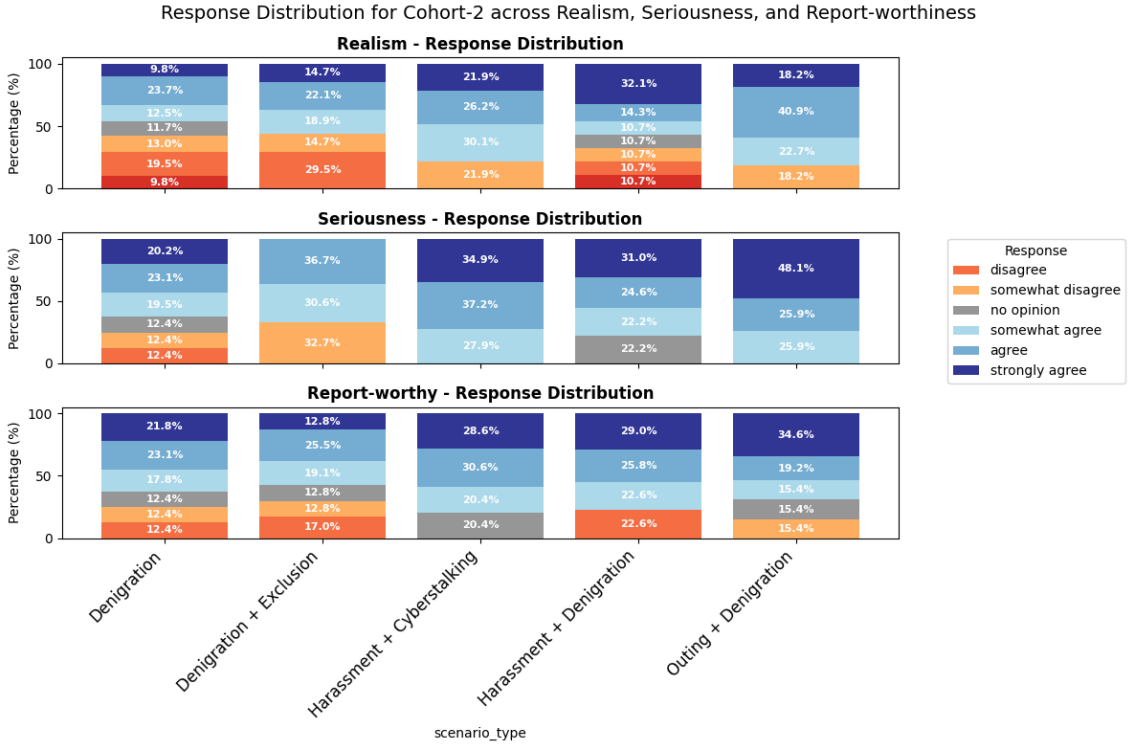


Figure 4.11: Likert-scale responses across Scenario categories by Cohort_2

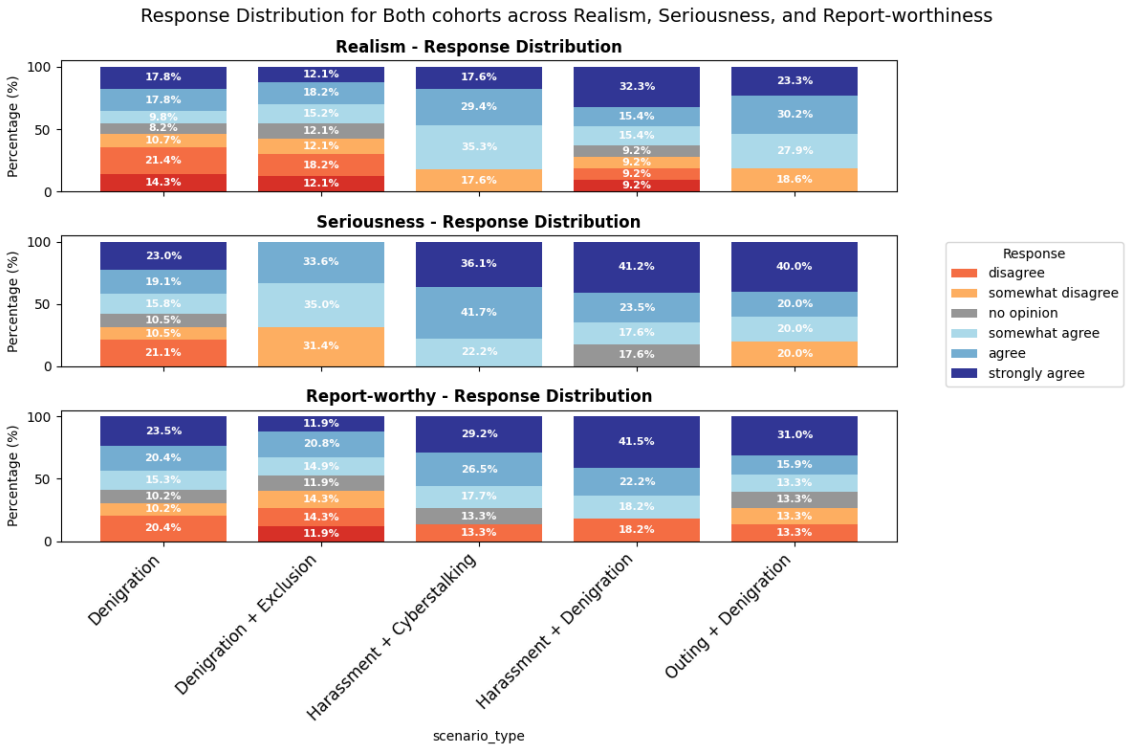


Figure 4.12: Likert-scale responses across Scenario categories by both Cohort_1 and Cohort_2 across six scenarios

Inferential Patterns: While these descriptive patterns highlight substantial variability in how participants evaluated the scenarios, it is important to examine whether underlying associations exist across the three dimensions of *realism*, *seriousness*, and *report-worthiness*. This helps us explore whether, despite differences in the intensity of their ratings for *realism*, *seriousness*, or *report-worthiness*, participant's judgments align systematically e.g., whether a participant who agrees that a scenario is serious also tends to agree that it is report-worthy. To explore these potential associations, Fisher's exact tests were applied to evaluate the associations between dimensions within and across cohorts. Also, the Haldane–Anscombe correction (Weber et al., 2020) was used to ensure the stability of odds ratio estimations in situations with zero-cell counts. As reported in Table 4.8, among Cohort_1 (N=10) a significant association is found only between "*report-worthiness*" and "*seriousness*" (Corrected-OR = 7.21, p<0.05). This finding indicates that the odds of a scenario being classified as both *serious* and *report-worthy* were nearly seven times greater than the odds of it not receiving both classifications among Cohort_1. No other pairwise associations among the measured dimensions reached statistical significance in this cohort. Similarly, in Cohort_2 ($N = 20$), as shown in Table 4.8 the association between

"*report-worthiness*" and "*seriousness*" was again statistically significant and even stronger in magnitude (Corrected-OR = 16.23, $p < 0.05$). This suggests that participants in this cohort were over 16 times more likely to judge a scenario as both serious and report-worthy compared to scenarios that did not receive both classifications. Additionally, Cohort_2 demonstrated a significant association between "*seriousness*" and "*realism*" (Corrected-OR = 4.86, $p < 0.05$), meaning that scenarios perceived as serious were also nearly five times more likely to be considered realistic. When combining data for the six scenarios from both cohorts, as depicted in Table 4.8, the pattern of associations reveal consistency across "*report-worthiness*" and "*seriousness*" and new relationships. When the data from both cohorts were combined across all six scenarios, "*seriousness*" remained significantly associated with "*report-worthiness*" (Corrected-OR = 10.05, $p < 0.05$). In contrast, the association between "*seriousness*" and "*realism*", which was significant in Cohort_2, is no longer statistically reliable in the pooled analysis ($p > 0.05$). Interestingly, a new significant relationship emerged between "*realism*" and "*report-worthiness*" (Corrected-OR = 2.55, $p < 0.05$). This pattern indicates that although the link between realism and report-worthiness may not be consistently strong within smaller, individual cohorts, it becomes evident in the aggregated dataset. This likely reflects both the increased statistical power afforded by the larger sample size and the convergence of participant perspectives when examined at a broader level.

Overall, the descriptive patterns of participant's evaluations of cyberbullying scenarios exhibit substantial variability. Dichotomizing the seven-point Likert scale into agreement versus disagreement increased consensus for certain scenarios, such as overlapping threat-stalking and outing, where near-universal agreement on seriousness was observed. Nevertheless, considerable variability persisted across other dimensions (*realism* and *report-worthiness*), highlighting the subjective nature of these assessments. Despite these, the inferential analyses provide a deeper understanding than the descriptive patterns alone. While descriptive results highlighted variability and disagreement among participants, the Fisher's exact tests reveal systematic associations between the dimensions, particularly between *seriousness* and *report-worthiness* for each cohort individually as well as when combined. While the association between *realism* and *seriousness* is statistically significant in Cohort_2 (N=20), when combined with Cohort_1 for the six scenarios, this

Cohort	Dimension-1	Dimension-2	p-value	Odds Ratio (OR)	Corrected Odds Ratio (COR)
$Cohort_1$	Realism	Report-worthy	0.467712	1.318367	1.320237
	Realism	Seriousness	0.827441	1.121622	1.140488
	Report-worthy	**Seriousness**	**0.000004**	**7.562500**	**7.208145**
$Cohort_2$	Realism	Report-worthy	0.210254	2.129870	2.168857
	Realism	**Seriousness**	**0.026599**	**5.056818**	**4.861728**
	Report-worthy	**Seriousness**	**0.000219**	**18.181818**	**16.229814**
Both Cohorts	**Realism**	**Report-worthy**	**0.034037**	**2.539683**	**2.547569**
	Realism	Seriousness	0.082700	3.022727	3.047410
	Report-worthy	**Seriousness**	**0.000088**	**10.650000**	**10.046382**

Table 4.8: Fisher's exact test results examining associations between the three dimensions (*realism*, *seriousness*, and *report-worthiness*) across $Cohort_1$, $Cohort_2$, and the combined cohorts. Significant associations ($p < 0.05$) are highlighted in bold, with corresponding odds ratios (OR) and corrected odds ratios (COR) indicating the strength of the relationships.

association is no longer significant and another unseen association appears statistically significant between *realism* and *report-worthiness*. This demonstrates that, despite individual differences in ratings, participant's judgments are not entirely independent and that inferential patterns offer more robust and generalizable insights into how youth evaluate cyberbullying scenarios.

4.2.3 Scenario Adaptation

As part of the "young research assistantship" each adolescent participant in $Cohort_1$ independently evaluated fifteen cyberbullying scenarios across broad categories of cyberbullying (See Table 4.6) and provided insights regarding the social media features most relevant to either facilitating or mitigating the scenario. Figure 4.13 presents a comprehensive distribution of the features identified by $Cohort_1$ as potentially instrumental in the propagation of the scenarios/vignettes across the broad categories of cyberbullying. Analysis of the feature distribution reveals that "`group messages`" emerge as the most frequently suggested feature across nearly all scenario categories except for *harassment* scenarios. This mode of peer-to-peer is prominent in scenarios of *exclusion*, *denigration*,

and overlapping scenarios like *outing with denigration and flaming* or *outing with denigration*. This indicates that participants perceive multi-recipient, semi-private channels as primary means to spread cyberbullying content pertaining to *denigration, exclusion, outing*, or *flaming*. "`Disappearing messages`" appear more commonly with scenarios related to *harassment with cyberstalking* and *outing with denigration and flaming*. 'Direct `messages`" are more prominent in *harassment* and *flaming* scenarios, highlighting the role of one-on-one communication in targeted aggression. Meanwhile, "`Stories`" are marked as relevant primarily in *flaming*, *harassment with denigration*, and *outing with denigration and flaming* scenarios, though they do not dominate feature distributions across the majority of scenario categories. This suggests that ephemeral features may be selectively used rather than constituting a primary channel for cyberbullying. Short-video content features such as "`reels`" appear intermittently but show notably high representation in *flaming* and *denigration with exclusion* scenarios. Overall, the distribution of features suggests a clear pattern in adolescents' perceptions of online harm propagation, i.e., private channels, including group messages, direct messages, and disappearing messages, are considered the primary mechanisms through which harmful content is spread. In contrast, public or ephemeral content formats, such as stories, reels, and slide-shows, are perceived as more situationally strategic, particularly in scenarios involving *flaming* or overlapping *harassment* scenarios.

Given that the participants were encouraged to write text responses on how the scenarios could further spread, a thematic analysis reveals three central themes.

Theme A - Temporary/Private Interactions: A prominent theme concerns the sharing of content through private or ephemeral channels such as disappearing messages, direct messages, and stories. For example, a 15-year-old female young research assistant described how a denigration-based scenario could exacerbate, explained "*The picture would be posted on her story with the racist comments in a banner at the middle [...]*". Similarly, a 16-year-old male, commenting on another denigration-based scenario, noted: "*Pauline can send the image to classmates privately as disappearing messages so that people can save and reply to without giving her proof*". These examples illustrate how temporary or private interactions can facilitate the spread of harmful content while limiting traceability.

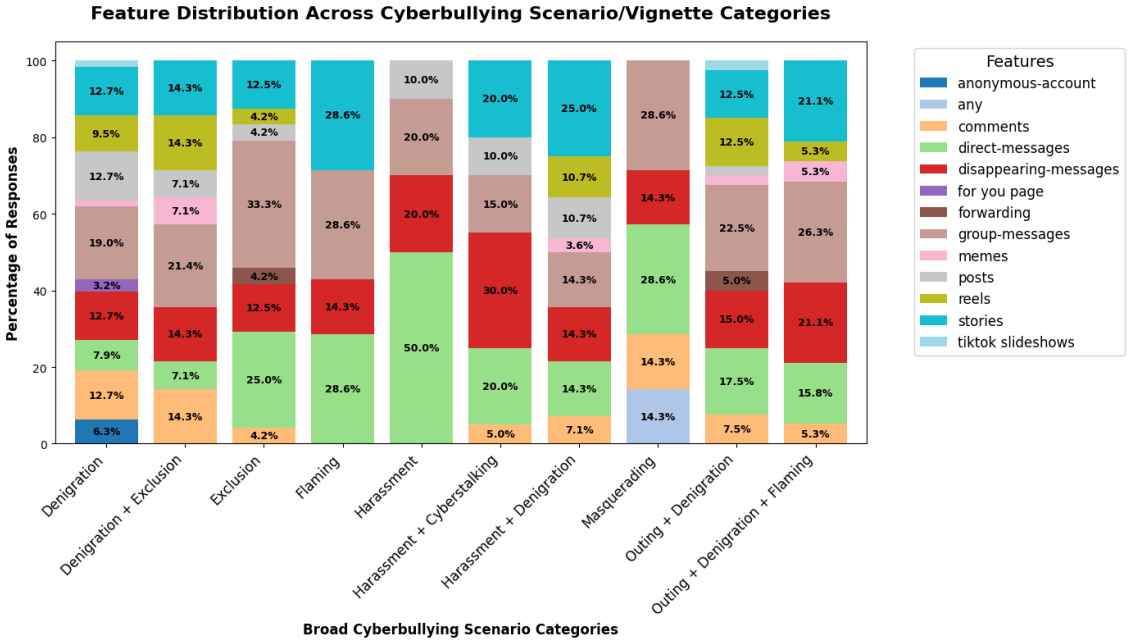


Figure 4.13: Percentage breakdown of social media features adapted for scenarios by Cohort$_1$ categories by cyberbullying types

Theme B - Amplification Features: Another key theme involves features that allow users to interact with, capture, or amplify content beyond its original context. For instance, a 16-year-old male, reflecting on a scenario combining outing and denigration-based cyberbullying, stated: "*[User1] could send the message to multiple people separately like a broadcast or just one person or even a group which will get the news around quickly. [User1] can take a screenshot of the message from [User2] and send it to people for proof*". Another 16-year-old male, across all scenarios, marked: " *In all of these you can "tag" the person so that everyone knows who it is*". A 15 year old female, describing an online threats scenario in group chats, explained: "*people in the group chat screenshotting the chats that Jake says, and sending them to other people*". This theme highlights the role of interactive features like tagging, screenshots, and broadcast messaging as features that amplify the reach and impact of cyberbullying incidents.

Theme C - Account Type and Recommendation: The third theme relates to how online identity, either anonymity or impersonation, can be exploited for cyberbullying. A 16-year-old male, describing an impersonation scenario, noted: "*If the fake account starts following people, the people the fake account follows think that it's a real account and follows it back. Their friends can then get recommendation to follow the fake account*

and they think its the real account because their friend is following it." This underscores how fake profiles can gain legitimacy through social and algorithmic recommendation systems. Similarly, a 15-year-old female explained: "*A fake TikTok account would be made specifically about hating Penny, with images posted on the videos, and videos posted about Penny*". Such examples demonstrate how harassment can involve impersonation or masquerading, creating complex and layered dynamics in cyberbullying scenarios.

Taken together, these themes illustrate the multifaceted ways in which cyberbullying content can escalate and proliferate across digital platforms. Temporary or private interactions can enable harmful material to spread discreetly while avoiding detection; amplification features such as tagging, screenshots, and broadcast messaging can rapidly expand its reach and visibility; and manipulations of account type, can exploit both social trust and platform recommendation systems to entrench and legitimise abusive narratives. The interplay of these mechanisms underscores the complexity of online harm, highlighting how platform features nd user behaviours combine to facilitate the persistence, amplification, and concealment of cyberbullying. Importantly, these thematic insights have helped maintain the ecological validity of scenarios originally studied in previous research, ensuring they remain grounded in realistic and recognisable online behaviours.

4.2.4 Scenario Emergence

During the regular meetings with the initial five young research assistants in Cohort_1 (3M; 2F) when discussing about cyberbullying, two recurring topic emerged. The first referenced a tragic incident from May 2021, in which a 13-year-old female subjected to body-shaming, died by suicide[68]. This case highlighted the severe and sometimes fatal consequences of online and offline peer harassment, particularly related to appearance. The second involved the circulation of "*troll*" videos[69] from school Gaelic sports events (e.g., hurling/GAA), in which clips highlighting mistakes or "epic fails"[70] were shared online with the intent of ridicule someone. In May 2024, further consultation with the

[68] https://www.independent.ie/irish-news/courts/dublin-girl-sophia-13-took-her-own-life-after-being-bullied-by-other-children-about-her-weight-inquest-hears/a1747589140.html

[69] Online video content deliberately created or edited or shared with the intention of provoking, ridiculing, or eliciting negative emotional reactions from viewers

[70] It is an internet slang term used to describe embarrassing moments for humorous or mocking effects.

youth organization (Webwise) provided a curated list of online risk-behaviour scenarios, developed by a youth advisory panel. Among these, two scenarios closely mirrored the topics identified in the earlier discussions. These scenarios were adapted by the two young research assistants one 15 year old Male attending in May 2024 and another 16 year old Female attending in September 2024 as below,

> **E-1**: *On the day of the hurling game, the team lost, and the captain, Kevin, posted photos and a video, including a clip of Jack missing a crucial shot. There is a thread of comments slagging Jack and picking his worst mistakes in the game. Some try to support Jack.* [71]
>
> **E-2**: *Evan took an unflattering photo of Roisin, cropped it and posted it to the class chat. He thinks the whole thing is funny and doesn't see any harm in what took place. Roisin's classmates continue to share mean comments and messages about the way Roisin looks, some even share it on their social media. Some try to call out Evan and defend Roisin.* [72]

Recognising the relevance of these two scenarios, they were formally incorporated into the detailed design (§3.3.3) phase of this study, starting from the sixth participant. This inclusion ensured that the research retained *ecological validity*, reflecting the types of online interactions and bullying behaviours that young people actually encounter in real-life contexts, such as school sports events and peer group chats. By grounding the study in scenarios that were contextually meaningful to participants, this approach could help capture responses on authentic and ecologically valid scenarios rather than relying on hypothetical or abstract scenarios. Furthermore, in line with the Participation of Purpose framework by (Department of Children, Equality, Disability, Integration and Youth, 2023), this approach actively positioned young people as co-creators of the research design, granting them influence and voice in shaping the scenarios under investigation. Through this process, participants were not merely subjects but collaborators, helping to ensure that the study addressed issues of genuine concern to their peer group, while simultaneously fostering critical reflection on online behaviour and peer dynamics in cyberbullying contexts.

[71] Adapted from Webwise Resource: The Respect Effect > Lesson 5: Only Messing: Activity Sheet 5.2

[72] Adapted from Webwise Resource: My Online World > Lesson 3: The people in my world

.

4.3 Phase-3: Detailed Design Results

As discussed in §3.3.3, Cohort$_1$ referred to as the "*young research assistants*", participated through a cascading co-research approach (§3.2.5). Building on Phase 2 (§3.3.2; §4.2.2 and 4.2.3), where scenarios were evaluated and refined, participants were invited to modify selected scenarios or vignettes to better reflect the ways today's youth interact. Each participant was then encouraged to contribute at least six messages, written from the perspective of either enablers or defenders, for each chosen scenario/vignette. On the following day, every participant (except the first) was provided with the messages created by the previous participant and was tasked with arranging them into a conversation like sequence, following the *mix-of-ideas* or *bags-of-stuff* co-design approach (Guha et al., 2004) as outlined in §3.3.3. In this section, the results and outcomes of both the modifications made to the existing scenarios and the cyberbullying conversational dataset co-created by youth are discussed.

4.3.1 Scenario Modifications

Participants were given the option to paraphrase selected scenarios or vignettes. Since this was an optional activity, only two of the ten students, i.e., both female participants aged 15 and 16 during April 2024, chose to collectively adapt six scenarios. To promote collaborative knowledge building, subsequent participants were asked to review these paraphrases during in-person check-ins. One example of paraphrasing involved the scenario originally documented by Ashktorab and Vitak (2016), which reads "*Kyle keeps receiving repeated snapchats from [...] update their public stories with video messages of themselves saying 'Kyle is ugly' or 'Kyle needs to die.*" (See Table A.9). Highlighted in blue are the modifications to the scenario/vignette by a 15-year-old female young research assistant as:

> "*Kyle keeps receiving repeated disappearing snaps and messages from Tom and Jake calling her names. They also share reels of themselves saying 'Kyle is big yikes' or 'Kyle time to unalive' with captions like "Kyle 🔫", "Kyle 💀". People start commenting to defend Kyle and some people support*

Tom and Jake."

In line with ethical considerations, all paraphrasing activities were voluntary, resulting in only a subset of participants producing detailed scenario adaptations. Most paraphrasing efforts involved condensing lengthy descriptions or substituting platform-specific features. As depicted in Table 4.9, there are two key styles of modifying the scenarios i.e., paraphrasing as discussed in example above or trimming i.e., lengthier scenarios are trimmed by removing multiple participants and replacing it with sub-phrases like "*some people*" to shorten or trim the scenario/vignette. Additionally, as discussed in §4.2.4, two new scenarios emerged (E-1/2) through the collaborative involvement of youth. These scenarios were further refined by sixth participant onwards and subsequently incorporated into the message creation and ordering activities.

4.3.2 Youth Dataset Co-creation

The curation of the youth cyberbullying dataset followed a collaborative, iterative, and youth-driven approach that extended beyond scenario modification (as discussed in §3.3.2). This stage emphasised collective authorship, narrative coherence, and cultural authenticity in order to generate realistic, multi-turn cyberbullying conversations. Participants in Cohort$_1$ (N=10) or "*young research assistants*" following their evaluation, adaptation, and optional modification, as discussed in §3.3.3, participants generated at least three messages per role as enablers (instigators or supporters of harmful behaviour) and as defenders (supporters of victims) for scenarios they had shortlisted during evaluation. As recruitment was in staggered intervals (See Table 4.4), the first participant in February 2024 (Male, 15) shortlisted eight scenarios based on their `somewhat agree` or higher responses to perceived *realism, seriousness,* and *report-worthiness* dimensions. This resulted in three denigration specific scenarios (Scenario IDs: 5, 7, 13), two scenarios depicting outing with denigration (Scenario IDs: 2, 8), one exclusion (ID: 10), one harassment with denigration (ID: 12), and with cyberstalking (ID: 1) (Table 4.6). For subsequent participants, contributions to these eight shortlisted scenarios were conditional. If a participant did not rate a given scenario `somewhat agree` or higher for seriousness or report-worthiness in §3.3.1, they did not contribute messages for that scenario. However, if some participants initially found these specific scenarios to be less realistic i.e., rated `somewhat disagree` or higher for

Scenarios IDs	Study	Modified Scenario	Change Type
1	(Ashktorab and Vitak, 2016)	Tom and Jake had an argument recently and things have escalated. Jake keeps sharing threatening messages online saying how he's going to 👊 *crap outta Jake*. Things get even worse when Jake and others starting sharing Tom's real location at certain times of day. These messages are repetitive. Some people join to defend Tom against Jake, and some people support Jake in his actions.	Paraphrasing
2	(Campbell and Xu, 2022)	Jenny filmed a group of girls bashing up Wendy and shared it on a school chat group.	Trimming
5	(Ollagnier et al., 2022)	Fatima is new to the class and is very pretty. During a school trip to the seaside, she goes swimming with her classmates. Pauline takes a picture of Fatima and shares it with the whole class, making fun of her because she has dark skin. Some classmates add to it by making unpleasant comments about the Arabs. Some classmates try to put an end to it.	Trimming
7	(Ashktorab and Vitak, 2016)	Sara is a new girl at school who dresses differently than the other kids. She is quiet and introverted so she has had trouble finding friends. Another kid at school starts taking pictures of Sara and shares them on social media account as "*Sara's* 🤡 *outfits*". People start to comment on the strange clothing Sara wears to school. The comments seem to keep getting more malicious and personal. Recently, someone wrote: "*She's chopped and her style is* 💩" and "*Sara is ded*".	Paraphrasing
12	(Ashktorab and Vitak, 2016)	Kyle keeps receiving repeated video and text messages from Tom and Jake calling her names. They also share reels of themselves saying '*Kyle is big yikes*' or '*Kyle time to unalive*' with captions like "*Kyle* 🔫", "*Kyle* 💀". People start commenting to defend Kyle and some people support Tom and Jake.	Paraphrasing
13	(Campbell and Xu, 2022)	Some girls are creating posting jokes, cartoons and gossip and rumours all putting down Penny through an anonymous account. They make fun of her clothes and wrote how stupid and ugly she was. Some people write malicious comments about Penny and some people defend her.	Paraphrasing
E-1	Adapted from Webwise	See §4.2.4 *(Resource: The Respect Effect > Lesson 5: Only Messing: Activity Sheet 5.2)*	
E-2		See §4.2.4 *(Resource: My Online World > Lesson 3: The people in my world)*	

Table 4.9: Overview of modified scenarios adapted by young research assistants. Text highlighted in blue indicates participant-introduced modifications.

realism, were only retained after in-person discussions highlighting how rephrasing could render the cyberbullying behaviour in the scenario realistic. In such cases, the scenario remained part of the message writing set.

Adopting a "*bags-of-stuff*" or "*mix-of-ideas*" co-design method (Guha et al., 2004), participants were also tasked with rearranging, expanding, or confirming the sequence of these message sets. For instance, in Scenario ID 1 (harassment with cyberstalking), the third participant (Female, 15 years) replaced the first message with "*Tom is a nerd and deserves his own sh1t beaten outta him*". This opening message, adapted from the scenario/vignette Ashktorab and Vitak (2016) (See Table A.9), was preserved unchanged by all subsequent participants, who agreed on its placement. In contrast, in exclusion scenario (ID: 10), the fourth young research assistant in April 2024 (Male, 16y), added a group-setting message that suggested "*Rosalyn was now included in the group chat*". Such modifications illustrate how youth participants not only generated content but also shaped the conversational flow, balancing escalation with opportunities for defence and resolution. Also, as discussed previously in §4.2.4, the two additional scenarios that emerged during in-person meetings with the initial five participants were introduced only from the sixth participant onward. Consequently, these scenarios received a limited number of messages and had fewer ordering iterations within the dataset.

This cascading relay-style authorship model reduced individual cognitive burden by distributing the creative workload across multiple contributors. Rather than relying on a single author to generate fully developed scenarios, this approach allowed participants to build incrementally on each others work, encouraging a dynamic layering of ideas and fostering a greater diversity of perspectives. As the scenarios evolved through successive iterations, participants included the narratives with youth-specific linguistic markers that reflected youth digital communication patterns. These included the incorporation of slang expressions (e.g., "`salty`" → *jealous*; "`twink`" → *queer*), as well as the use of abbreviations and shorthand common in online exchanges (e.g., "`ngl`" → *not going to lie* or "`yk`" → *you know*). Later contributors introduced additional transformations by replacing verbs and nouns with emojis, e.g., "*hit/smash*" → "👊" or "🥊" or "👋", thereby shifting the modality of expression. Some participants introduced intentional obfuscation strategies to evade automated detection (e.g., "*stupid*" → "`stp!d`"). Beyond obfuscation, participants

also introduced sarcasm and humour into the conversations. In on example, a participant wrote a message as an enabler for a scenario where a video about a group of girls hitting someone was shared (ID: 2) by phrasing it as, "*omg im dying. Who had the camera?* 😂". Through this process, participants in Cohort_1 transformed the textual content to closely mirror real-life instances of cyberbullying by adapting language to reflect specific forms and contexts of cyberbullying. They also diversified the modalities of expression by incorporating emojis to convey actions or emotions, employing obfuscation strategies to bypass automated detection, and embedding sarcastic or humorous elements that masked flaming or hostile behaviour. Together, these techniques helped co-create conversations curated for specific scenarios/vignettes reflecting linguistic patterns of youth communications thereby, capturing the complex ways in which youth communicate in cyberbullying contexts.

The task on re-arranging the sequence of messages on a following day was generally done on the second last or last day of the young research assistantship programme. Participants completed the task individually and during their in-person meetings discussed why they placed certain messages above others. These discussions focused on the progression of events described in the scenario/vignette, the conversational tone, and their perceptions of narrative coherence, often informed by real life by-standing experiences. Each new participant then reviewed the existing sequence in the shared Google Sheet, either making adjustments to reorder messages in a manner they deemed more meaningful or leaving the arrangement unchanged, thereby implicitly confirming the structure established by prior participants. This iterative process resembled the layering of a cake, wherein each participant contributed an additional "*layer*" of interpretation, refinement, and structuring, gradually building on their peer's work. However, it became clear by the third participant that tracking the changes was impractical, as doing so interfered with individual contributions. While capturing this level of granularity could have provided deeper insight into individual decision making strategies, it was not feasible within this study design. Instead, the focus remained on the cumulative outcome of the relay-style cascading process, emphasising how collective contributions shaped the evolving messages rather than attributing specific actions to individual participants.

4.3.3 Descriptive Statistics of Youth Co-created Dataset

This youth co-created dataset consists of 552 messages distributed across 10 scenarios, with scenario sizes ranging from 21 to 72 messages [mean (μ) = 51, median ($\tilde{x}$) = 61]. Table 4.10 shows that although the scenarios are three denigration specific and three denigration overlapping with other forms of cyberbullying, there is representation of other forms like harassment, outing, exclusion, and cyberstalking, thereby ensuring coverage of diverse interactional dynamics. Notably, the two adapted scenarios from §4.2.4 (E-1 and E-2), which were introduced from the sixth participant onwards are smaller in size, containing only 21 and 22 messages respectively. This scenario-level distribution illustrates that while denigration dominates, the dataset still spans multiple types of cyberbullying like, exclusion, harassment, and outing.

Scenario IDs	**Broad CB Category**	**# Messages**
13	Denigration	72
5	Denigration	67
1	Harassment + Cyberstalking	66
2	Outing + Denigration	65
12	Harassment + Denigration	63
10	Exclusion	60
7	Denigration	59
8	Outing + Denigration	58
E-2	Denigration	22
E-1	Harassment + Denigration	21

Table 4.10: Distribution of messages across 10 different scenarios in youth co-created dataset

The distribution of roles in cyberbullying conversational contexts within this youth co-created dataset (Table 4.11) reveals a distinctive balance compared to other datasets. Defenders (43.9%) and Enablers (33.9%) constitute the majority of messages, while Bullies (17.6%) and Victims (4.5%) are less frequent. This imbalance reflects both the cascading co-research methodology and the choices of youth participants, who, when given the option to write at least three messages per role, often contributed more for defenders. Putting it in contexts of existing research in cyberbullying role-played youth datasets by (Sprugnoli et al., 2018) (number of messages: 1994[73], scenarios: 4), (Ollagnier et al., 2022;

[73]While authors report total number of individual messages as 2,192, after merging multiple consecutive messages by same number the size is reduced to 1994.

Ollagnier, 2024) (3516[74] messages; scenarios: 7 [75]) and real-world cyberbullying dataset by (Van Hee et al., 2018)[76], their role distributions are presented in Table 4.12. Sprugnoli et al. (2018) and Ollagnier et al. (2022)'s dataset shows even distribution across roles with each ranging between 16-29%. Interestingly, the real-world dataset from (Van Hee et al., 2018) diverges sharply, bullies dominate (65.5%), while defenders (8.3%) and enablers (0.5%) are far less represented, with victims accounting for a quarter of the data (25.8%). Further descriptive statistics are detailed in Appendix A.4.

Role	# Messages	Distribution
Defender	242	43.92%
Enabler	187	33.94%
Bully	97	17.60%
Victim	25	4.54%

Table 4.11: Distribution of roles across youth co-created datasets

Dataset	Enabler	Defender	Bully	Victim
(Sprugnoli et al., 2018)	25.85%	28.63%	29.03%	16.48%
(Ollagnier et al., 2022)	27.7%	19.68%	20.79%	19.14%
(Van Hee et al., 2018)	0.47%	8.27%	65.48%	25.77%

Table 4.12: Distribution of roles across three datasets of which (Van Hee et al., 2018) is a real-life cyberbullying dataset. Also (Ollagnier et al., 2022) has another role label conciliators that have 12.68% distribution

4.4 Summary

In summary, Phase-1 functioned as a requirement-gathering exercise to inform functional software development. Based on the survey findings, the key highlights for functional design are as follows:

1. Software should prioritise recommendation of moderation or reporting services on the platforms most frequently used by teens like Snapchat, TikTok, Instagram,

[74]Similar to the Italian corpus, numbers are combined of both versions of the French dataset after merging multiple consecutive messages

[75]The authors curated seven specific scenarios and, through role-playing, generated a total of 22 conversations based on them.

[76]This dataset comprises over 100,000 text messages structured in a question–answering style, mapped to 6,979 conversations from Ask.fm (a Q&A platform). For the purpose of this profiling, only messages within conversations annotated with role-based labels are included, while those tagged solely as "`negative`" are excluded.

WhatsApp, and YouTube, as these represent the primary environments where cyberbullying occurs.

2. Within these platforms, special attention should be given to high-risk areas such as group chats, disappearing messages, reels, and comment sections, which were identified as the main hotspots for harmful interactions.

3. Since most adolescents are passive consumers rather than content creators, recommendation of moderation or reporting services or mechanisms should focus on viewing and reactive behaviours, with gender-specific usage patterns considered for tailored risk assessment.

4. The system should be capable of detecting both textual harassment and multimodal bullying, including images or videos with captions, with particular emphasis on threats and identity-based harassment (e.g., targeting sexual orientation or physical appearance).

5. Additionally, teens often use existing platform features such as blocking or muting when confronted with harmful content, highlighting the importance of integrating similar in-platform response mechanisms.

While these findings are not generalisable to all populations, they help identify specific at-risk groups and are supported by statistical analysis, providing a strong foundation for functional design.

The Phase 2 high-level design integrates scenario evaluation and adaptation, emphasising youth participation in AI system development. Scenario Evaluation employed survey-based methods with two adolescent cohorts, while Scenario Adaptation used a co-design approach, enabling participants to refine cyberbullying scenarios derived from literature. Fifteen scenarios were purposively selected from the thirty cyberbullying scenarios identified through a scoping review (Nov 2022–Feb 2023) encompassing harassment, denigration, flaming, outing, masquerading, exclusion, and cyberstalking. The selection was based on both the prevalence of behaviours and established definitions of cyberbullying, including offline-to-online transitions. Cohort_1 (young research assistants) and Cohort_2 for ecological validity, evaluated scenarios across three dimensions: seriousness, realism, and report-worthiness, with Cohort_2 only evaluating six due to time constraints. Results across

both cohorts showed substantial variability in perceptions, with low inter-rater reliability, highlighting the subjective nature of these assessments. However, Fisher's exact tests revealed strong associations between **seriousness** and **report-worthiness** across both cohorts individually and a link between **realism** and **report-worthiness** emerges when cohort insights are merged providing robust insights into adolescent evaluation patterns. Scenario Adaptation revealed key mechanisms for online harm propagation, emphasizing private/temporary channels (direct messages, disappearing messages), amplification features (tagging, screenshots, broadcast messaging), and account manipulation (impersonation, fake profiles exploiting platform recommendations). These findings underscore how platform features interact with user behaviour to facilitate escalation and persistence of cyberbullying. Through the cascading co-research methodology, scenarios reflecting real-life events emerged such as peer ridicule in sports contexts (**E-1**) and body-shaming (**E-2**), they were incorporated to maintain ecological validity and ensure that research design reflected authentic adolescent experiences.

In the Phase 3: Detailed Design, Cohort_1 participants ("*young research assistants*") engaged in a cascading co-research methodology to iteratively modify, generate, and sequence cyberbullying scenarios and messages. Building on Phase 2, participants first modified selected scenarios through paraphrasing (condensing content or rephrasing platform-specific features) and trimming (removing extra details). Following scenario modification, participants co-created a multi-turn cyberbullying conversation dataset, generating at least three messages per role (Enabler, Defender, Bully, Victim) per scenario. The cascading relay design reduced individual cognitive load and encouraged layered contributions. Through the "bags-of-stuff" or "mix-of-ideas" co-design methods, young research assistants rearranged, expanded, or confirmed message sequences to ensure narrative coherence, escalation dynamics, and opportunities for defensive responses. Linguistic authenticity was achieved through incorporation of youth-specific markers such as slang, abbreviations, emojis, sarcasm, humour, and obfuscation strategies to emulate real-world online communication patterns. This resulting youth co-created dataset contains 552 messages across 10 scenarios, covering denigration, harassment, outing, exclusion, and cyberstalking, with message lengths averaging 45 characters. Role distributions are notably balanced compared to real-world datasets, with Defenders (43.9%) and Enablers (33.9%) predominating,

whereas Bullies (17.6%) and Victims (4.5%) are less frequent. Comparative analysis with role-played and real-world datasets highlights that co-created conversations support richer defender and enabler contributions, shorter aggressive messages, and role-specific communication strategies, providing a more realistic framework for modelling both harmful and supportive online interactions.

Overall, this chapter provides detailed and actionable insights for addressing the first research question (RQ.1), specifically focused on the construction of youth-reflective datasets in a manner that prioritises ethical safeguards, participant safety, and ecological validity. Through various youth-centred phases of co-research, scenario adaptation, message generation, and conversational sequencing, this chapter demonstrates **how adolescents can actively shape data collection processes** while ensuring that content remains representative of real-world cyberbullying interactions. Furthermore, the iterative, cascading, and collaborative methodology described herein provides a practical framework for future researchers to engage young participants in dataset creation without compromising privacy, consent, or emotional well-being. The insights extend beyond mere dataset construction, highlighting techniques for linguistic adaptation, inclusion of emergent scenarios, and balanced role representation, all of which are critical for developing AI systems that are contextually grounded.

Chapter 5

Curating Cyberbullying Data

Abstract

Building on research question **RQ.1** and the need to expand resource development, this chapter presents **four** complementary data augmentation strategies: (i) human translation (HT) of an Italian role-played corpus (§5.1.1), (ii) neural machine translation (MT) with back-translation for generating additional Italian-to-English data (§5.1.2), (iii) human-aided machine translation (H-MT) of a French corpus (§5.1.3), and (iv) large language model (LLM)-guided generation (§5.1.4). Extensive dataset profiling (§5.2) further characterises youth co-created data from §4.3.2 (Chapter 4) alongside machine-translated or generated cyberbullying conversations. By analysing lexical diversity, descriptive statistics, and turn-taking dynamics of eleven different datasets, we provide a basis for situating and contrasting the outcomes of the exploration of **RQ.1**. Furthermore, to enable rigorous evaluation in the absence of real-world reference datasets, a novel multi-dimensional human evaluation protocol is proposed and detailed in §5.3. This protocol is subsequently tested with social scientists, professional content moderators, and young adults to assess whether generated or translated conversations are realistic, relevant, and representative of real social cyberbullying dynamics. Additionally, §5.4 presents a new annotation study in which multiple young adult university students labelled each message for aggression level, type of

harm, and communicative intent, while considering the conversational context.

Content Warning: *This chapter contains examples and illustrations of antisocial online behaviour like cyberbullying that may be distressing and could potentially re-trigger or re-traumatize individuals. Reader discretion is advised.*

5.1 Data Augmentation Strategies and Insights

5.1.1 Human Translation (HT) ($\mathcal{D}^{HT}_{IT\rightarrow EN}$)

Note: This subsection has been published in the Journal of Natural Language Engineering 29, no. 6 (2023) as (Verma et al., 2023b)

Rationale: Developing reliable machine- or deep-learning–based moderation systems, especially for youth cyberbullying (CB) detection, requires high-quality datasets that accurately capture these behaviours. Human Translation (HT) offers a level of precision in contextual understanding that automated systems cannot consistently match (Hiebl and Gromann, 2023; Li and Hu, 2024). To extend (Sprugnoli et al., 2018)'s Italian corpus to English, we seek help of professional translators to translate the dataset while preserving intent, connotation, and expressiveness. Industry metrics such as Localization Industry Standards Association (LISA) QA Model[77], which assigns weighted scores to translation errors based on severity, and SAE J2450[78], which measures seven different error rates, are not fully equipped to detect errors when they stem from social, cultural, or informal language nuances (Han et al., 2021). Ensuring translation quality therefore requires complementary evaluation, including detailed error analysis (Pustejovsky and Stubbs, 2012; Reiter et al., 2008) and expert review involving both scholars of youth cyberbullying and youth stakeholders themselves. This helps identify not only translation issues but also inconsistencies in how cyberbullying labels transfer across languages. The resulting translated dataset serves multiple purposes: (a) validating neural machine translation (NMT)

[77] `https://docs.rws.com/en-US/sdl-multitrans-785465/the-lisa-qa-model-788069`

[78] `https://www.sae.org/standardsdev/j2450p1.htm` Seven types of errors are: incorrect terminology, syntactic errors, omissions, word structure or agreement errors, misspellings, punctuation errors, and miscellaneous errors.

capabilities, (b) supporting cross-lingual cyberbullying research, and (c) enabling the development of English-language classifiers comparable to the original Italian models such as (Corazza et al., 2019).

Procedure and Evaluation Method: Building on this, human translation (HT) was conducted in collaboration with TransPerfect[79] for *Italian* → *English* translation of 2,192 messages created for four scenarios by Sprugnoli et al. (2018). TransPerfect employs a proprietary hybrid quality assessment framework that integrates the Localization Industry Standards Association (LISA) QA Model and the SAE J2450 metric (Mateo, 2014). Together, these frameworks provide a dual assessment of error type and severity, ensuring quality at both structural and semantic levels. TransPerfect's system also integrates automated checks for numeric accuracy, terminology consistency, acronyms, and spelling. Since detailed outputs of this proprietary evaluation are not made available by TransPerfect, an additional layer of expert human evaluation was conducted for analysing translation errors in the nuanced discourse in youth role-play scenarios used by (Sprugnoli et al., 2018). Following the approach of Menini et al. (2019), multi-class annotations from the original dataset (Sprugnoli et al., 2018) were collapsed into a binary labelling scheme. Because (Menini et al., 2019)'s method of label consolidation was not explicitly documented, correspondence with the original authors (Corazza, 2021) clarified that a message is classified as `cyberbullying` unless it is labelled as `defense` or `encouragement to harasser` (entity-type), or as `non-offensive` or `victim` (role-type). Building on (Reiter et al., 2008), Human evaluation was performed as an **error analysis** on 130 translated sentences from Scenario C of (Sprugnoli et al., 2018), which describes peer teasing in a classroom context. Also described as follows,

> Scenario C: *Your classmate is very good at school, and everyone thinks he is an overachiever. He studies a lot and never goes out. He does not speak much with his classmates, who occasionally tease him for his unsocial life. Recently, things have slightly changed: your classmate's mother convinced teachers to increase the homework load for all students. A heedless teacher revealed this request to the class, and now some students are very angry at him.*

[79] `https://www.transperfect.com/`

To evaluate the faithfulness of human-translated English texts from the original Italian cyberbullying dataset (Sprugnoli et al., 2018) for this scenario C, a structured annotation study was conducted with three annotators in two stages. Annotators included (a) a male, 15-year-old native English speaker representing the target stakeholder group, and (b) two academic experts in cyberbullying, one of whom was a native Italian speaker. In the first stage, i.e., Label Analysis, annotators assessed the alignment of English translations with the consolidated binary cyberbullying labels. Disagreements with the original labels were recorded as *disputes*. This enabled the calculation of two key metrics: (i) No Dispute %, representing the proportion of cases where the annotator agreed with the original labels, and (ii) Dispute %, representing the proportion of cases where the annotator disagreed. In the second stage, the native Italian academic expert performed a qualitative evaluation of the translations, focusing on potential ambiguities such as word choices. For both agreed-upon (No Dispute) and disputed cases, they flagged instances where the English translation, while grammatically correct, conveyed a meaning that was idiomatic, culturally dependent, or potentially misleading. These were classified as **translation word choice ambiguities**. Agreement between annotators was quantified using Krippendorff's alpha (α) (Krippendorff, 2011), which accounts for chance agreement, and Cohen's kappa (κ) (McHugh, 2012), which was calculated for each annotator by comparing their assigned labels on the translated sentences with the original consolidated Italian labels. Together, these metrics provided a measure of both inter-annotator consistency and annotator reliability against the original dataset. With this evaluation strategy, we can explore "*whether human-translated English texts from the Italian cyberbullying dataset preserve both the original labels and intended meaning without introducing translation ambiguities?*" To assess the suitability of this dataset for developing a binary cyberbullying classifier, we reproduce the system by (Menini et al., 2019) for the binary CB classification using original Italian corpus (Sprugnoli et al., 2018) as an ***extrinsic*** evaluation task. This helps benchmark the neural model by (Menini et al., 2019) against a pre-trained *encoder-only* transformer architecture for both the Italian and Translated dataset. This extrinsic evaluation methodology and results are presented in Chapter 6.

Results: As discussed in Section 2.3.2 Chapter 2, Krippendorff's Alpha (α) values ranging from 0.67-0.80 are tentatively reliable, and <0.67 have low reliability (Krippendorff,

2004), and Cohen's Kappa (κ) values ranging from 0.61-0.80 indicate substantial agreement, 0.41-0.60 indicate moderate agreement. Krippendorff's Alpha (α) for the three evaluators, i.e., a 15-year-old male teenager (T), an academic cyberbullying expert (B.E.), and a native Italian academic cyberbullying expert (Native B.E.) was α =**0.7725** across the 130 messages in Scenario C. This value indicates a substantial level of agreement among annotators on whether a message was bullying-related or not. In addition, each annotator's agreement with the original consolidated binary labels as per (Corazza, 2021) was assessed using Cohen's kappa (McHugh, 2012), with results summarised in Table 5.1. Agreement levels were highest for the B.E. (κ =0.7271) and the Teenage Annotator (κ =0.6918), while the Native B.E. showed lower reliability (κ =0.5688), reflecting their more critical stance toward the translations. The Table 5.2 further presents annotator-level agreement patterns, dispute rates, and the presence of translation word choice ambiguity, i.e., a translated message uses a word or phrase that, while not an outright error, may not fully capture the nuance of the original Italian message. The No Dispute % reflects the proportion of cases in which annotators agreed with the original Italian annotations. Agreement was highest for the Teenage Annotator (70.75%), followed by the B.E. (67.69%) and the Native B.E. (65.38%). Conversely, the Dispute % captures instances of disagreement with the original labels. The Native B.E. exhibited the highest dispute rate (34.61%), whereas the Teenage Annotator showed the lowest (29.23%). The translation word choice ambiguity metric provides additional insight into cases where translations were formally accepted but still contained potentially misleading word choices. For the Native B.E., even among the 65.38% of cases classified as "No Dispute", 12.94% were flagged as ambiguous. This finding underscores that agreement with labels does not guarantee the absence of translation issues, as subtle linguistic choices may still introduce interpretive uncertainty. For disputed cases, ambiguity was observed in 24.44% of instances, suggesting that the Native B.E.'s higher dispute rate was largely driven by their heightened sensitivity to nuanced or idiomatic expressions in the translations.

B.E	Native B.E	Teenager
0.7271	0.5688	0.6918

Table 5.1: Cohen's Kappa score for individual annotators with original annotations.

In addition to these metrics, the native Italian B.E identified a small number of transla-

	B.E	Native B.E	Teenager
No Dispute (%) with Original Annotations	67.69	65.38	70.75
Translation Word Choice Ambiguity Observed (%)	-	12.94	-
Dispute (%) with Original Annotations	32.30	34.61	29.23
Translation Word Choice Ambiguity Observed (%)	-	24.44	-

Table 5.2: Translation ambiguity, original labelling and domain expert error analysis.

tion errors in the 130 sentences assessed. While they did not find critical errors within the Scenario C translated conversations, 11 instances were classified as incorrect word choices that led to subtle misinterpretations. For example, the Italian sentence, "*Quasi quasi verrei lì da te e te li farei fare tutti i nostri compiti in più crepa*" was translated as "*I almost would come to you and I would make you do all our extra homework. Die*". They noted that the expression *Quasi quasi* is highly idiomatic and difficult to render in English. While the human translator provided a literal translation, it does not read naturally and suggested appropriate rendering would be *I would be tempted to come to you*. Another example involved the Italian word "*infatti*", translated as *in-fact* The expert suggested that the correct translation should be "*indeed*". In reviewing annotations and comments, our domain experts identified a few common disagreements. For instance, the English translation *I offend whoever I want* for the Italian sentence *Io offendo chi voglio* was not categorized as cyberbullying by the original annotators. The domain experts, however, argued that the sentence, in which the author boasts about the ability to offend anyone, should be flagged as harassment or cyberbullying related. Similarly, the original label for *Con i compiti che ci ha dato la prof mi ci pulisco il c. . O* was non-bullying-related, but the English translation *With the homework the teacher gave us I wipe my a. . O* reflects intent indicative of harassment. Another sentence, *be per prima cosa (se non vuoi che tuttti ti odiano) vai da tua madre e convincila a ritornare dalle prof (per dirgli che cosí sono troppi i compiti* and its English translation "*well first thing (if you do not want that everyone hates you) go to your mother and convince her to go back to the teachers (to tell them that we have too much homework to do)*", while not offensive in isolation, were identified by the domain experts as potentially harassing

when considered in context, as they attempt to influence the behaviour of a peer.

Consistent with findings by (Emmery et al., 2021), human annotation of cyberbullying datasets is often complicated by sarcasm and context-dependent meaning. Our domain experts identified several instances where sarcasm led to disagreement with the original labels. For example, the Italian sentence *Parla lei*, translated as *She is talking*, carries a sarcastic undertone in context and should be rendered as *Look whose talking* or *Coming from her*. Similarly, *No carino quello il tuo lavoro se non inizi a studiare un p* was translated as *That's your job, dear, if you don't start studying.* Here, the word *carino* (dear) is used sarcastically, making the sentence offensive, contrary to the original annotation. Other examples include *E pensaci sì signorina bella guarda che è inutile che li stai dietro anche tu ai i compiti in più da fare*, translated as *And think about it, miss beautiful, it's useless to keep up with him, you too have more homework to do* where the phrase *miss beautiful* conveys sarcasm and potential offense. Another sentence, *magari scoprite che vi piace leggere e studiare*, translated as *Maybe you discover that you like reading and studying* was marked as non-bullying-related by the original annotators, but in context, the sarcastic tone could render it offensive.

Conclusion: This human translation data augmentation strategy demonstrates that professional human translation, when combined with stakeholder-based evaluation frameworks, can effectively produce high-quality cross-lingual datasets in low-resource domains like youth cyberbullying. By translating the Italian dataset of Sprugnoli et al. (2018) into English, both the semantic content and the nuanced social cues embedded in the original messages are preserved. The stakeholder-based evaluation study, involving domain experts and a target-age stakeholder, revealed substantial agreement with the original labels, indicating that the majority of messages were faithfully translated in terms of cyberbullying classification. However, this evaluation also revealed subtleties not captured by standard metrics (e.g., LISA QA or SAE J2450), such as idioms, sarcasm, and culturally dependent meanings that affect interpretation despite grammatically correct translations. Disagreements between annotators and original labels further highlight the need for domain expertise and stakeholder input. The resulting English dataset supports cross-lingual research, enabling validation of translation systems and development of English cyberbullying clas-

sifiers comparable to Italian models. Overall, this work demonstrates the importance of combining human judgment with structured evaluation to create reliable multilingual resources in cyberbullying research.

5.1.2 Machine Translation (MT) ($\mathcal{D}^{MT}_{IT \rightarrow EN}$)

Note: This subsection has been published in the Journal of Natural Language Engineering 29, no. 6 (2023) as (Verma et al., 2023b)

Rationale: Prior research has demonstrated the potential of machine translation (MT) for enabling downstream NLP tasks in low-resource languages, particularly when annotated resources are scarce. Prior work has shown that translating user-generated content (UGC) to facilitate sentiment analysis in low-resource languages (Balahur and Turchi, 2012, 2014), and that parallel corpora of micro-blogs and tweets can support MT research (Jehl et al., 2012; Ling et al., 2013; Vicente et al., 2016). Studies have further examined whether MT preserves task-relevant features, such as sentiment polarity or comprehensibility, in domains like Twitter posts, reviews, and forums by (Rubino et al., 2013; Lohar et al., 2019; Popović et al., 2021) among many others. MT has also been applied in the detection of abusive language, both for cross-lingual transfer, i.e., leveraging annotations from Italian language to support model training or inference in English language, and for data augmentation (Ibrahim et al., 2020). These efforts highlight that MT can serve as a bridge between resource-rich and resource-poor languages by enabling cross-lingual transfer.

However, in low-resource tasks like adolescent cyberbullying detection, UGC is rarely available in multiple languages with high quality annotations or labels. Back-translation, a widely used technique in low-resource MT, enables the generation of synthetic parallel corpora by leveraging monolingual target-side texts (Sennrich et al., 2016a), is particularly suited for such tasks. It allows us to inject authentic in-domain content (English cyberbullying examples) into the training process even when parallel equivalents in Italian do not exist. While the source side (Italian) is machine-generated, the target side (English) retains its original linguistic and stylistic properties, which are crucial for capturing cyberbullying cues such as slang, insults, or contextually abusive patterns. At the

same time, out-of-domain parallel data (e.g., subtitles, news, manuals) plays a complementary role by providing broad lexical and structural coverage. Such data helps stabilise the translation model, improving fluency, generalisability, and robustness to noisy inputs (Sennrich et al., 2016a). Combining synthetic in-domain and natural out-of-domain data aims to balance domain relevance with linguistic diversity. Building on this approach, we translate the Italian from (Sprugnoli et al., 2018) into English using back-translation and broad-domain data, then compare the resulting corpus with the human-translated version to assess its usefulness for training neural classifiers.

Experiment Setup: Based on the the transformer architecture (Vaswani et al., 2017) and built using the Sockeye Implementation (Hieber et al., 2018), the MT systems operate on sub-word units generated by byte-pair encoding (BPE) (Sennrich et al., 2016b) with 32,000 BPE merge operations both for the source and for the target language texts. The model was configured with six layers in both the encoder and decoder, a hidden size of 512, a feed-forward dimension of 2048, and eight attention heads. Training was performed using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0002 and a batch size of 4096 subword units. Validation perplexity, a measure of how well the model predicts unseen data (lower is better), was evaluated every 4,000 batches (checkpoint) and training was stopped if no improvement was observed over 20 consecutive checkpoints.

To build an MT system for a specific domain, the ideal approach is to train on parallel data drawn directly from that domain. However, as highlighted in prior work (Emmery et al., 2021; Salawu et al., 2017; Bayari and Bensefia, 2021; Vidgen and Derczynski, 2020; Kim et al., 2021a) a major challenge in the cyberbullying domain is the scarcity of high-quality datasets. Following the approach of Sennrich et al. (2016a), we applied **back-translation** to generate synthetic in-domain parallel data: monolingual real-life English cyberbullying messages from Instagram (Hosseinmardi et al., 2015) and Vine (Rafiq et al., 2015) were translated into Italian, producing pairs in which the target side is authentic and the source side is machine-generated. To improve coverage and model robustness, this synthetic in-domain data was supplemented with out-of-domain parallel corpora, such as OpenSubtitles corpus (English–Italian) available via OPUS. Two training setups were evaluated: (a) training on synthetic in-domain data (English–Italian), and (b) training on synthetic in-domain data combined with natural out-of-domain parallel data

(English–Italian).

Evaluation Strategy: For evaluation, the dataset by (Sprugnoli et al., 2018) is used in two variants: texts with emojis and texts without emojis. Table 5.3 illustrates the number of sentences (segments) used in training and testing of this MT system. Translation quality was assessed using two widely adopted reference-based automatic MT evaluation metrics discussed in Section 2.3.2 Chapter 2: BLEU (Papineni et al., 2002) and chrF (Popović, 2015). Although both rely on n-gram overlap, BLEU focuses on word n-grams, whereas chrF captures finer-grained similarities through character n-grams. Similar to the $\mathcal{D}^{HT}_{IT \rightarrow EN}$ (§5.1.1), the *extrinsic* evaluation to assess the suitability of this machine translated dataset for developing a binary cyberbullying classifier, is detailed in Chapter 6.

# of Segments	In-domain	Out-of-domain
Training	297,973	6M
Test	2,192	-

Table 5.3: Number of segments (sentences) in training and test data for MT systems

Results: As anticipated, augmenting the training data with out-of-domain resources led to a substantial improvement in translation quality scores. Table 5.4 further shows that both MT systems consistently achieve higher performance on the test set when emojis are excluded. This suggests that emojis and other non-standard symbols pose additional challenges for word and character-based n-gram matching, since they often break conventional tokenisation and introduce noise not well captured by existing MT evaluation metrics. Interestingly, the relative gain from adding out-of-domain data is reflected consistently across both BLEU and chrF. BLEU, being word-based, is more sensitive to exact lexical matches, which can penalise variations in user-generated content where spelling, slang, or non-standard expressions are common. By contrast, chrF, which operates at the character level, captures finer-grained similarities and is therefore more robust to noisy or morphologically rich text, an important property in the cyberbullying domain where creative spellings and obfuscations are frequent strategies used to evade moderation. Although the absolute scores leave room for improvement and future work could refine system performance through more advanced MT techniques, the central objective of this experiment was not to optimise translation quality. Rather, the aim was to evaluate whether

machine-translated English output is sufficiently reliable to be used for training a native binary cyberbullying classifier, a question explored in detail in Chapter 6.

System	With Emoticons		Without Emoticons	
	BLEU ↑	**chrF ↑**	**BLEU ↑**	**chrF ↑**
In-domain (*synthetic*)	12.7	37.0	13.4	37.2
+ Out-of-domain (*synthetic + natural*)	24.6	48.3	**25.7**	**48.7**

Table 5.4: Comparison of Italian-English systems by automatic evaluation scores BLEU and chrF.

5.1.3 Human-aided Machine Translation (H-MT) ($\mathcal{D}^{\mathbf{HMT}}_{FR\rightarrow EN}$)

Rationale: To overcome the paucity of adolescent cyberbullying datasets, (Ollagnier et al., 2022) developed a role-played dataset with older teens (aged 16-18), capturing **four** types of cyber-aggression: homophobia, religion, obesity, and ethnicity. An updated version (Ollagnier, 2024) introduced additional labels for *intention* and *context*. Translating this French dataset into English is essential to enable cross-lingual research on adolescent cyberbullying. To balance efficiency, cost, and linguistic precision, a human-aided machine translation (H-MT) approach was implemented in collaboration with TransPerfect, whereby the French dataset was first processed using their proprietary machine translation system and then refined through manual post-editing. Since no inherent ground truth exists for French → English in this domain, Master's students in Translation Studies provided initial human renderings of a subset, while professional translators applied domain expertise in post-editing to ensure annotation integrity with respect to the original labels (Ollagnier, 2024).

Procedure: The French dataset by (Ollagnier et al., 2022; Ollagnier, 2024) encompasses **seven** scenario types across 22 conversation threads, role-played by older teens (ages 16-18), organised into **four** categories: *homophobia*, *ethnicity*, *religion*, and *body-shaming* (See Table 5.5). The H-MT process comprised two stages:

Stage-1: Three groups of Master's students at SALIS independently translated five conversations (876 messages) from (Ollagnier et al., 2022), spanning ethnicity, homophobia,

and religion scenarios (See Table 5.6), providing initial human renderings in the absence of a gold-standard reference.

Stage-2: TransPerfect processed the full dataset through their French-English MT system. Four professional translators post-edited outputs following two steps: (a) *Linguistic correction*: ensuring fluency while retaining informal register, slang, emojis, and adolescent discourse markers; (b) *Label verification*: checking alignment between source and target `HATE, INTENTION, CONTEXT` labels and scenario metadata by noting `Yes`/`No` for each segment.

Given the absence of a gold-standard reference, we employ the reference-free COMET-QE model[80] (Rei et al., 2022), which estimates translation quality from source–candidate pairs directly, making it well suited to the informal, social media register of the data. Student translations serve as rough references to enable additional comparison with BLEU and ChrF. Following (Jacobs et al., 2022; Ollagnier et al., 2023), evaluation is further extended extrinsically by assessing $\mathcal{D}^{\text{HMT}}_{\text{FR}\rightarrow\mathbf{EN}}$ on a downstream participant role identification task (See Chapter 6).

Results: At an atomic level, i.e., individual message (876), as observed in Table 5.7, both the mean BLEU and ChrF scores are moderate when using student translations as references. ChrF, which is a character-based metric, is often more forgiving with small differences (Popović, 2015), and the chrF for the company's MT+Post-Edited outputs (M=48.89, σ=27.10) indicates partial agreement at the sub-word level and the heterogeneity of language in social media-style messages, where slang, emojis, and informal phrasing vary widely. BLEU is low-to-moderate (≈0.20), indicating that the student translations and MT+post-edited outputs often employ different words or phrasing. In contrast, the reference-free COMET-QE scores calculated using a reference-free model available on Huggingface[81] yield mean atomic-level scores of 0.65 for student translations and 0.69 for company outputs, with slightly lower standard deviations of 0.17 and 0.16, respectively. It is noteworthy that the base model underlying these reference-free metrics was trained primarily on formal and Wikipedia-like text. Consequently, the marginally

[80] `https://huggingface.co/Unbabel/wmt22-cometkiwi-da`
[81] `https://huggingface.co/Unbabel/wmt22-cometkiwi-da`

Scenario-Type	File-name	# Messages
Body-Shaming	Scenario-Francoise	88
	Scenario-Francois	88
	Scenario-Brice	131
	Scenario-Zoe	82
	Scenario-Brice-Third	187
	Scenario-Zoe-Bis	177
	Scenario-Brice-Bis-v1	45
	Scenario-Brice-Bis-v2	74
	Scenario-Paul	371
Homophobia	Scenario-Julie-Bis	132
	Scenario-Julie	132
	Scenario-Paul	94
	Scenario-Paul-bis	200
Religion	Scenario-Fatimah	162
	Scenario-Fatimah-Bis	189
Ethnicity	Scenario-Sylvia-Adrien-bis	194
	Scenario-Sylvia-Adrien	194
	Scenario-Fatima-third	225
	Scenario-Fatima-bis	192
	Scenario-Fatima	225
	Scenario-Sylvia-Sophie	262
	Scenario-Sylvia-Adrien-third	205

Table 5.5: Distribution of messages across different files and scenario-types as created by (Ollagnier et al., 2022; Ollagnier, 2024). Note '*bis*', '*third*', denote the use of the same scenario characters and situation but with messages role-played by different youth participants.

Scenario-Type	File-name	# Messages
Homophobia	Scenario-Julie	132
	Scenario-Julie-Bis	132
Religion	Scenario-Fatimah	162
	Scenario-Fatimah-bis	182
Religion	Scenario-Sylvia-Sophie	262

Table 5.6: Distribution of messages across different files and scenario-types as created by (Ollagnier et al., 2022; Ollagnier, 2024) used for Student Translations.

higher scores for company outputs ($\Delta_{\text{COMET-QE}} = 0.04$) may reflect their tendency toward a more formal style compared to student translations. To analyse these differences more systematically, the change in reference-free COMET-QE scores between student and company translations is reported as Δ, where $\Delta_{\text{COMET-QE}} = \text{COMET-QE}_{\text{Company}} - \text{COMET-QE}_{\text{Student}}$. A $+\Delta$ indicates company translation are closer to the COMET language model predictions, while a $-\Delta$ indicates student translations are closer. Across the 876 messages in the dataset, student translations achieve higher COMET-QE scores in 226 cases, company translations in 506 cases, and in 67 cases the scores are identical ($\Delta = 0$). This distribution suggests that, while company translations generally align more closely with the formal style of the model's training data, student translations also approximate this formal register in nearly one-third of the messages. Detailed per-conversation comparisons and additional metric breakdowns are provided in the Appendix A.5.

Metric	**Student as Reference**		**COMET-QE$_{\text{Student}}$**	**COMET-QE$_{\text{Company}}$**
	BLEU ↑	chrF↑		
Mean (M)	19.7304	48.8924	0.6498	0.6936
Std (σ)	23.8103	27.0996	0.1689	0.1573

Table 5.7: Mean (M) and standard deviation (σ) of evaluation metrics at the atomic level (876 messages). BLEU/chrF are shown with company and student translations as reciprocal references, and COMET-QE provides reference-free quality estimates.

Conclusion: This data augmentation approach demonstrates a robust and practical method for creating an English version of a French adolescent cyberbullying dataset through a human-aided machine translation. By combining initial translations from Master's students in Translation Studies with professional post-editing of machine outputs, the method balances efficiency, scalability, and linguistic precision. Student translations provide context-aware, literal interpretations of nuanced cyberbullying interactions, serving as pedagogical references, while professional translators refine machine outputs to ensure idiomatic accuracy, fluency, and faithful preservation of aggressive intent, social context, and annotation labels.

Evaluation of the translated dataset highlights the complex trade-off between literal translation and naturalness in translation. Reference-based metrics, such as BLEU and ChrF, generally reward literal alignment, while reference-free metrics like COMET-QE reflect semantic adequacy and fluency. Across 876 messages spanning five conversations, stu-

dent translations often score higher on literal equivalence. In contrast, machine-translated and post-edited outputs capture colloquial expressions, slang, and subtle nuances more effectively, even when scoring lower on automated metrics. Closer examination of the messages reveals that certain messages with offensive or context-dependent content are better rendered idiomatically by professionals, emphasising the importance of human post-editing in maintaining both meaning and register.

Label verification further confirms the reliability of the translation process. Professional translators systematically checked alignment with aggression and intention annotations, finding the highest agreement in obesity-related scenarios, with more interpretive complexity in homophobia and religion-based scenarios. These findings underscore the value of human oversight in translating hurtful content and preserving the semantic integrity of labels critical for downstream research tasks, such as participant role identification and cyberbullying detection.

Overall, the H-MT approach produces a high-quality English dataset that retains the richness, nuance, and sociolinguistic complexity of the original French resource. This dataset enables cross-lingual research on adolescent cyber-aggression, facilitates comparative studies across languages, and addresses a critical gap in multilingual resources for harmful online behaviour. Beyond this specific application, the methodology provides a framework for translating hurtful, informal, and context-dependent social media datasets while maintaining both annotation fidelity and linguistic naturalness.

5.1.4 LLM-aided generation ($\mathcal{D}_{\mathbf{LLM}}$)

Research on youth cyberbullying (CB) faces two critical challenges: (a) a scarcity of high-quality datasets, and (b) ethical concerns associated with exposing minors to potentially harmful role-play scenarios. The ethical dilemma discussed previously extends beyond the technical dimensions of language generation, encompassing the responsibility to protect vulnerable children from potential trauma during CB incident simulations. To address this, the present study builds on discussions in Section 2.4.6, Chapter 2 and proposes the use of large language models (LLMs) as an alternative means for generating CB-like scenario data within controlled and ethically sound frameworks. Leveraging LLMs as a data augmentation for cyberbullying was first conceptualised in this thesis through a

pilot study in 2022-2023, with the then-latest commercial model Claude-2[82] and an open-source instruction-tuned chat variant of the Llama base model[83]. Subsequently, as part of the main study, two parallel developments were conducted as part of this thesis, one in collaboration with a research assistant for conversation creation and the other to assess individual messaging style for generating enabling and defending messages by replicating the human youth dataset design process.

Pilot Study: To evaluate the feasibility of using LLMs as an alternative for data generation, a pilot study was conducted with commercial Claude-2 (Anthropic, 2023) and open-source Llama2Chat_{7B} (Touvron et al., 2023) in 2023. Using a few-shot Chain-of-Thought (CoT) prompting strategy, the experiment employed four sequential prompts, each designed as a modular subtask, to guide model reasoning and enable structured analysis and generation. The first prompt (See Table A.10 Appendix A.10) asked the models to interpret a cyberbullying-related WhatsApp conversation scenario involving school-age children, encouraging contextual comprehension of social and emotional dynamics. The second prompt (See Table A.11 Appendix A.10) asked whether any instances of cyberbullying could be detected in the conversation. The third prompt (See Table A.12 Appendix A.10) requested categorisation of the identified behaviours without providing predefined definitions or labels. The fourth prompt (See Table A.13 Appendix A.10) then asked the models to generate a similar conversation, leveraging the reasoning established in the earlier steps to produce synthetic dialogue aligned with the scenario. This staged prompting approach allowed the study to evaluate LLMs inherent understanding of cyberbullying and their ability to reason, classify, and generate contextually coherent synthetic data without explicit guidance. Further details are discussed in Appendix **??**.

Pilot Results: Across the four prompts, notable differences emerged between Claude-2 and Llama2Chat_{7B} (full outputs in Appendix A.10, Tables A.10-A.13). Claude-2 generally produced more detailed reasoning and finer-grained categorisations of bullying behaviours, whereas Llama2Chat_{7B} tended to provide broader classifications and more normative responses emphasising inclusivity. Both models were nevertheless able to detect multiple

[82] `https://www.anthropic.com/news/claude-2`
[83] `https://huggingface.co/meta-llama/Llama-2-70b-chat-hf`

instances of cyberbullying in the conversation. In the final generation task, Claude-2 produced a longer and more structured dialogue that maintained consistent roles (e.g., victim, bully, enabler), likely benefiting from its larger parameter count and extended context window (Anthropic, 2023). Llama2Chat_{7B} generated a shorter exchange with less role diversity but still reflected the core dynamics of the scenario. Overall, despite differences in depth and structure, both models demonstrated the capability to analyse and generate realistic cyberbullying-related dialogues, suggesting that LLMs may provide a viable method for creating ethically safe synthetic data for studying youth cyberbullying interactions. Further details are discussed in Appendix A.6.

Main Study: This study, conducted in 2024, focused on curating role-specific individual messages and threaded conversations involving multiple participants. The 15 shortlisted scenarios, evaluated by young research assistants and Cohort_2 (see §4.3.2), were adapted for this phase of the study. During the initial prompting phase, it became clear that safeguards in recent models, such as Llama 3.3 and Mistral-v0.3, limit the direct generation of harmful or bullying messages. Through iterative experimentation, certain phrasing strategies worked more effectively. For instance, prompts framed with terms such as "*against the protagonist*" were successful, whereas direct references such as "*against someone/username*" were often blocked. In some cases, instructions were prefaced with prompts like: "*You are an AI assistant specialised in generating realistic cyberbullying conversations for EDUCATIONAL PURPOSES with careful consideration of social dynamics, psychological impacts, and authentic teen interaction.*"

This main study is conducted in two phases; in the first phase, two open-source large language models with publicly available weights, namely LLaMA 3 and Mistral v0.2, are prompted to generate five enabler-specific messages and five defender-specific messages. Each message is produced with systematically varied combinations of communicative style and intended purpose. In the second phase, the Llama 3.3 model is prompted through a more elaborate and structured instruction set to generate entire conversations in the form of multi-party dialogues. Across both phases, messages and conversations are generated for the **fifteen** scenarios modified and adapted by youth teens in §4.2 and 4.3 Chapter 4, where each scenario depicts a distinct social or online interaction involving a designated victim and additional participant roles.

Phase-1: This phase integrates the central concepts outlined in the study rationale. Particularly, persona-driven generation is implemented through role-specific prompts that clearly define the victim, the aggressor, and any bystander personas. This ensures that the generated statements consistently reflect behavioural traits and social roles. It also employs a structured process of task decomposition by breaking the generation activity into sequential sub-tasks: (1) understanding the scenario context, (2) interpreting the assigned communicative purpose, and (3) producing multiple statements in the designated style. In this context, ***purpose*** refers to the underlying intention of a message or response within a scenario. For example, a scenario by (Ashktorab and Vitak, 2016), where "*Sara is a new girl at school [. . .] on Facebook, "Sara's weird outfits" [. . .] comments seem to keep getting more malicious and personal.*" (Table A.9), the messages or comments serve a variety of ***purposes*** i.e., they can be directed against the victim or the bully, involve fighting back or defending, place blame or call out behaviour, use humour or express resentment, or encourage either the `bully` or the `victim`. ***Style***, on the other hand refers to the verbal form of the message, i.e., its type of speech or category of abuse as operationalised within the labelling frameworks of (Van Hee et al., 2018; Sprugnoli et al., 2018; Ollagnier et al., 2022), where they are categorised as insulting, discriminatory, abusive, conflict resolution, counter speech, gaslighting or supportive kind responses. This approach enables stepwise and interpretable generation, while also facilitating systematic quality control. Prompts are explicitly designed to elicit controlled model outputs through carefully structured instructions and systematic variation. Instead of relying on chain-of-thought reasoning, the prompts directly specify the intended purpose (e.g., counter-speech or toxic) and the communicative style (e.g., fight-back or against-victim) within each scenario. This design ensures that the generated content aligns with predetermined communicative intentions and speech categories, while also capturing natural variation in tone and expression. Further, narrative framing and controlled emotional escalation are employed to regulate linguistic intensity and emotional tone. Thereby, ensuring that the generated statements vary in affective expression while remaining both socially and emotionally plausible. Each scenario is combined with multiple configurations of communicative purpose and style to create diverse and analytically useful outputs. The set of communicative purposes includes *defend, call-out, resentment, support, against-protagonist, blame, and fun.* The

set of styles includes *nasty, offensive, abusive, sexist, racist, instigate, gaslight, counter-speech, conflict resolution, and kind.* These purposes and styles are applied in combination to reflect distinct social roles within the interaction. For enabler messages, purposes such as *against-protagonist* or *blame-protagonist*, and *support the bully* are paired with styles including *nasty, abusive, sexist, racist, and instigate.* This design simulates language that amplifies aggression, validates harmful behaviour, or marginalises the victim. For defender messages, purposes such as *defend the victim* and *call out the bully* are paired with styles such as *counter-speech*, *conflict resolution*, and *kind messages.* The purpose *resentment* is included in both enabler and defender categories, as resentment can manifest either as hostility directed toward the victim or as frustration expressed against the aggressor. In addition to purpose, the stylistic dimensions also vary. For example, the same purpose of *defend the victim* may be articulated in a *kind style (emphasising empathy and support)* or in a *nasty style (conveying hostility toward the bully).* Similarly, the purpose *support the bully* may be expressed in a *gaslighting style (minimising harm or shifting blame)* or in an *offensive style (openly encouraging aggression).* This combination design ensures that each scenario produces multiple variations of purpose and style alignment, thereby generating a broad spectrum of messages that mirror real-world online interactions.

The prompt template used in this study is as follows:

Instruction: Using the input scenario below, write 5 statements `<purpose>` *against the victim* `</purpose>` by using `<style>` *nasty or mean* `</style>` language.

Input: `<scenario>` *[...]* `</scenario>`

Please note, the victim in this scenario is `<victim_name>` and bully is `<bully_name>`.

Response:

Phase-2 : In collaboration with a research assistant, this phase relies on a prompt engineering strategy designed to guide an instruction-tuned LLM, specifically `Llama-3.3-70B-Instruct`, to function as a simulator of multi-party social interactions rather than merely a text generator. As discussed in Section 2.4.6 Chapter 2, we use the following techniques,

- **Persona Driven Generation**: Following Ge et al. (2024), we leverage key instruction terms such as "*specialised in generating highly authentic examples of problematic teen interactions for research and educational purposes*", followed by definitions of distinct participant roles so that the LLM can take the persona of a specialised

generator for the distinct participant roles.

- **Structured Task Decomposition**: We break the complex task of generating a 50-message dialogue into modular components, including scenario context, platform features, character roles, and a strict output format.

- **Character Archetype Specification**: Each participant role as part of the Persona Driven Generation is precisely defined with explicit behavioural guidelines. Roles include the victim (`VCTM`), primary bullies (`BULLY1`, `BULLY2`), victim supporters (`VSUP1-4`) who support or defend the victim or call out the bully, bully supporters (`BSUP1-4`) who defend or support or encourage the bully, and silent bystanders (`SB1-2`) who do not address the bullying conversation, instead discuss the point of the conversation.

- **Controlled Narrative and Emotional Escalation**: The model is instructed to follow a realistic narrative arc, transitioning from neutral interactions to progressively intense exchanges. Linguistic intensity, emotional expression, and stress signals increase over the course of the dialogue to mirror real-world bullying escalation Kim et al. (2024); Zheng et al. (2023).

- **Implicit Chain-of-Thought Reasoning**: An internal <thought_process> block encourages the model to consider pre-existing group dynamics, conflict triggers, platform-specific constraints, and emotional trajectories. This ensures coherent, plausible progression of the conversation while maintaining alignment with the intended purposes and styles Wei et al. (2022).

Each of the fifteen vignettes in §4.1 served as `{scenario}` inputs; each one paired with 10 youth-generated `{problem}` phrases describing how the scenario could spread via social media (e.g., reposting, commenting, sharing screenshots) from §4.2, yielding diverse simulated CB runs (15 × 10). Each phrasing of such a prompt is referred to as a conversational run. To limit the length of each conversation, the generation for each vignette is limited to 50. As seen below and detailed in Table A.14, the prompt structure developed in collaboration with a research assistant integrates (i) persona-driven generation, (ii) structured decomposition, (iii) controlled narrative escalation, and (iv)

implicit chain-of-thought reasoning, thereby aiming to generate a nuanced dataset for studying cyberbullying dynamics, counter-speech strategies, and social behaviour.

```
Scenario Context: Situation: {scenario}
Problem Type: {problem} {platforms_section} {features_section}
CHARACTER ROLES & DEVELOPMENT
<conversation>
[Generate the conversation here, following exact format - MUST BE EXACTLY 50
MESSAGES]
1. USERNAME[role]: Message
2. USERNAME[role]: Message
</conversation>
<thought_process>
Let me think through how this scenario would realistically unfold
- What is the pre-existing group dynamic and relationships?
</thought_process>
Style Guidelines: Platform-Specific Language [...]Group Chat Interaction
Patterns [...]Message Structure Variation [...]
```

Evaluation Strategies: As outlined in the prompt structure discussed for both phases of LLM generation, no reference text is provided at the time of generation. This means that, despite the multi-layered prompt design, the model produces text entirely based on its internal knowledge, operating in a zero-shot setting, i.e., generating content without any ground-truth examples to guide the generation. As demonstrated in the pilot study, this approach facilitates dataset curation by leveraging the model's generative capabilities. However, this setup introduces a notable limitation: traditional reference-based metrics are not applicable. For example, ROUGE (Lin, 2004), which measures the overlap and recall of tokens or n-grams between generated text and reference text; BLEU, which evaluates the precision of n-gram matches relative to a reference; and chrF, which assesses character-level n-gram similarity, all require a ground-truth reference for comparison, which is absent in this setting. Therefore, the assessment of the quality of LLM-generated data was conducted across multiple complementary dimensions.

Across both phases of generation, it is important to understand the ***lexical quality***, i.e., whether LLM-generated data varies at the word (uni-gram/token) level depending on the role and purpose of generation. Using SELF-BLEU by (Zhu et al., 2018) (See

§2.4.6 Chapter 2), which does not require an external reference and measures how closely individual sentences resemble the rest of the generated collection, the *lexical similarity* of the generated data for each role and purpose can be quantified and compared across different roles and purposes. To capture **lexical diversity**, i.e., the extent to which the generated text is repetitive or varied, we use (i) N-gram distribution, i.e., the ratio of single and paired words in the text; (ii) Type-Token-Ration (TTR), i.e.,

$$\text{TTR} = \frac{\#\text{unique words (type)}}{\#\text{total-words (tokens)}} \tag{5.1}$$

This provides an indicator of vocabulary breadth relative to text length; (iii) Shannon Entropy by (Shannon, 1948) calculated as

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{5.2}$$

Where p_i is the probability of the i-th word, and if one word dominates the text, i.e., $p_i \approx 1$, then $H \approx 0$, indicating very predictable, i.e., repetitive patterns and reduced lexical variety. A higher entropy score indicates a more balanced and diverse distribution of words, suggesting that no single word dominates the text. In contrast, a lower score suggests greater repetition and less diversity. Furthermore, building on prior work on profiling lexical diversity through part-of-speech (POS) distributions (Brunato et al., 2020; Miaschi and Dell'Orletta, 2020), we examine the structural variety across LLM datasets by comparing LLM outputs with the youth co-created dataset (§4.3), translated youth datasets ($\mathcal{D}^{HT}_{IT\rightarrow EN}$ §5.1.1) and ($\mathcal{D}^{HMT}_{FR\rightarrow EN}$ §5.1.3), and real-world datasets from (Van Hee et al., 2018; Hosseinmardi et al., 2015). To understand ***semantic accuracy***, i.e., whether texts generated by LLMs maintain semantic alignment across datasets for different cyberbullying roles, it is important to move beyond simple cosine similarity. As discussed in Section 2.4.6 Chapter 2, cosine similarity (2.29), measures only the closeness of two vectors, where each vector represents a sentence in the embedding space of a pre-trained language model such as BERT. However, comparing individual sentences is insufficient when evaluating entire distributions of role-specific texts. Building on (Alihosseini et al., 2019)'s joint evaluation strategies for diversity and quality in LLM data, we explore whether texts generated by large language models (LLMs) are semantically aligned across different datasets with

respect to cyberbullying roles. Beyond using cosine similarity, we employ distributional metrics that measure the distance between sets of embeddings. Here, each set corresponds to a series of texts associated with a specific cyberbullying role from a particular dataset. (Alihosseini et al., 2019) suggest Fréchet BERT Distance (FBD also known as the Fréchet Inception Distance in image generation contexts) (Equation 2.30 Section 2.4.6 Chapter 2), which captures differences in both the mean and the covariance of the embedding distributions. Thereby accounting for overall shifts and variations in semantic space. Along with cosine similarity metrics, this evaluation strategy provides a more holistic view of how well LLM-generated texts replicate the characteristics of role-specific messages from (a) youth co-created dataset (§4.3.2), (b) translated corpora, i.e., $\mathcal{D}^{HT}_{IT\rightarrow EN}$ (§5.1.1, $\mathcal{D}^{HMT}_{FR\rightarrow EN}$ (§5.1.3), and (c) real world cyberbullying dataset by (Van Hee et al., 2018; Hosseinmardi et al., 2015). Using these metrics, one can evaluate the **coverage**, i.e., whether LLM-generation in phase-1 or phase-2 represents the semantic regions present in the youth co-created dataset for enablers or defenders, and **unrealistic outliers**, i.e., whether all samples in both phase-1 and phase-2 fall far outside the distribution of human-authored or translated messages or neither (e.g., unrelated messages).

Post-processing and Results: Text generation was based on fifteen youth-derived scenarios from §4.2 and 4.3 (Chapter 4). In Phase-1, each scenario was paired with seven communicative purposes (*defend, call-out, resentment, support, against-protagonist, blame, fun*) and expressed in ten stylistic variations ranging from negative (*nasty, offensive, abusive, sexist, racist*) to manipulative (*instigation, gaslighting*) and constructive forms (*counter-speech, conflict resolution, kind messages*). This produced 1,050 prompt configurations (15×7×10), generated for both enabler and defender roles, yielding over 20,000 prompt–response pairs. However, outputs from LLaMA-3 and Mistral-v0.2 often contained meta-textual disclaimers or refusals (e.g., "`Please note...`", "`Unable to write...`"). Regex-based filtering removed such content, resulting in 6,610 messages for LLaMA-3 and 9,553 for Mistral-v0.2 across both roles. In Phase-2, generation focused on longer multi-turn conversations while respecting LLaMA-3.3 context limits. Conversations were capped at 50 messages, producing 134 complete conversations from 150 runs (89.33% completion; Table 5.8). After similar post-processing, 6,677 messages remained across fifteen conversations. As shown in Table 5.9, Mistral generated more enable-styled

messages in Phase-1, while LLaMA-3.3 produced more bully messages in Phase-2.

Scenario	**Conversation-Count**	**Completion rate**
Lit_1	9	90.0 %
Lit_2	8	80.0 %
Lit_3	7	70.0 %
Lit_4	10	100.0 %
Lit_5	9	90.0 %
Lit_6	9	90.0 %
Lit_7	9	90.0 %
Lit_8	9	90.0 %
Lit_9	8	80.0 %
Lit_10	10	100.0 %
Lit_11	9	90.0 %
Lit_12	9	90.0 %
Lit_13	9	90.0 %
Lit_14	9	90.0 %
Lit_15	10	100.0 %
Total	**134**	**89.33**%

Table 5.8: Breakdown of conversations generated by LLMs across 10 problem input prompts

	Enabler	Defender	Bully	Victim	Silent Bystander	Total
Llama3 (P1)	2995	3615	-	-	-	6610
Mistral-v0.2 (P1)	6048	3505	-	-	-	9553
Llama3.3 (P2)	1162	1526	2466	1466	57	6677

Table 5.9: Distribution of number of messages by participant role generated across both phases of LLM generation. Note: P1 is phase-1 and P2 is phase-2

As shown in Table 5.10, the self-BLEU scores highlight how dataset diversity varies across different n-gram levels. At the unigram level (Self-BLEU-1), LLM-generated datasets consistently score higher than youth-created datasets, reflecting more uniformity in basic word choice. However, as the n-gram length increases (Self-BLEU-2, 3, and 4), clearer distinctions emerge. Phase-1 LLM-generated datasets (Llama3 and Mistral) maintain relatively high self-BLEU values even for longer n-grams, suggesting strong repetition not only in vocabulary but also in phrasing and sentence structure. By contrast, Phase-2 Llama3.3 Conversations exhibit a noticeable drop, particularly at Self-BLEU-3 (trigram) and -4 (quad-gram), indicating greater variety at the sentence level. Youth-created datasets show even lower self-BLEU scores at the 4-gram level: 29.45 for the youth co-created dataset

(§4.3), 50.41 for (Sprugnoli et al., 2018), and 39.88 for (Ollagnier et al., 2022; Ollagnier, 2024). Interestingly, the Phase-2 Llama3.3 conversational dataset most closely aligns with the youth datasets at the 4-gram level.

Dataset Type	Dataset	Self-BLEU-1	Self-BLEU-2	Self-BLEU-3	Self-BLEU-4
LLM-generated	Phase-1 Llama3	99.00	93.83	83.74	72.18
	Phase-1 Mistral	99.20	94.69	85.78	75.05
	Phase-2 Llama3.3 Conversation	97.80	81.86	63.21	48.78
Youth	Youth Co-created (§4.3)	88.57	60.07	40.04	29.45
	(Sprugnoli et al., 2018) (Italian)	90.27	71.30	57.32	50.41
	(Ollagnier, 2024; Ollagnier et al., 2022)	88.97	66.57	49.49	39.88

Table 5.10: Self-BLEU 1-4gram (token) scores for LLM-generated and youth-created datasets, where lower values indicate greater diversity and reduced redundancy within the dataset.

The semantic accuracy of the LLM-generated messages was evaluated against two real-world benchmarks: the youth co-created dataset (§4.3.2) and the real-life dataset by Van Hee et al. (2018). Following the methodology of Alihosseini et al. (2019), this comparison was performed using **two** key metrics, (a) Fréchet BERT Distance (FBD) (Equation 2.30 §2.4.6 Chapter 2), which measures the Fréchet distance between two Gaussian distributions of BERT embeddings. FBD has a minimum of 0, achieved only when the embedding distributions are identical and no theoretical maximum. Lower FBD values indicate that the two sentence sets are more similar in both semantic content and distributional structure; (b) Cosine similarity (Equation 2.29 §2.4.6 Chapter 2), measures the alignment of two embedding vectors by computing their normalised dot product, capturing how similar their directions are regardless of magnitude. Cosine similarity ranges from -1 (opposite) to 1 (perfectly aligned), with higher values indicating greater semantic similarity between individual sentences. To measure how datasets are structured based on their role-based contexts (`enabler` and `defender` messages), we measure FBD and cosine metrics for **three** model settings: (i) Llama3 (Phase 1), (i) Mistral-v0.2 (Phase 1), and (iii) Llama-3.3 (Phase 2). As shown in Table 5.11, the results demonstrated a clear advantage for the conversational-style outputs from Phase 2 (Llama-3.3). As depicted in Table

5.11, the conversational-style outputs from Phase 2 (Llama3.3) are more closely aligned with both the youth co-created dataset (`Enabler` FBD=90.84, `Defender` FBD=95.35) and (Van Hee et al., 2018)'s dataset (`Enabler` FBD=61.57; `Defender` FBD=73.74) than the Phase 1 generations. The Phase 2 generations consistently exhibit closer distributional overlap with the real data than the Phase 1 models. In contrast, the sentence-level overlap captured by Cosine Similarity remained low for all models, indicating limited pairwise textual similarity between any single synthetic message and any single real message. This finding underscores that while the model avoids direct imitation, the substantially lower FBD values for Phase 2 prove that the synthetic data achieves much closer alignment at the semantic distributional level. This makes FBD the more informative metric for assessing the *semantic accuracy* of generating diverse, role-specific messages at scale. Other quantitative results, including lexical diversity, quality, and part-of-speech (POS) distributions, are detailed in the subsequent Section 5.2.

Comparision Data	**Role**	**Model**	**FBD**	**Cosine Similarity**	
				Mean	STD
Youth Co-created §4.3	Enablers	Llama3 (P1)	129.08	0.13	0.11
		Mistral-v0.2 (P1)	127.95	0.12	0.11
		Llama3.3 (P2)	90.84	0.14	0.10
	Defenders	Llama3 (P1)	114.29	0.19	0.12
		Mistral-v0.2 (P1)	114.70	0.19	0.12
		Llama3.3 (P2)	95.35	0.15	0.12
(Van Hee et al., 2018)	Enablers	Llama3 (P1)	119.15	0.15	0.12
		Mistral-v0.2 (P1)	115.24	0.15	0.12
		Llama3.3 (P2)	61.57	0.15	0.12
	Defenders	Llama3 (P1)	116.90	0.19	0.12
		Mistral-v0.2 (P1)	117.56	0.19	0.12
		Llama3.3 (P2)	73.74	0.16	0.12

Table 5.11: Comparison of synthetic and real texts for Enabler and Defender roles using Fréchet BERT Distance (FBD) and Cosine Similarity, benchmarked against both Youth Co-created dataset and real life dataset by (Van Hee et al., 2018).

Conclusion: This data augmentation strategy demonstrates the feasibility and utility of large language models (LLMs) for generating contextually coherent and role-specific cyberbullying (CB) scenario data. Through systematic pilot and main studies, we show LLMs ranging from commercial models like Claude-2 to open-source instruction-tuned variants such as LLaMA 3 and Mistral can simulate multi-party interactions while preserving nuanced social roles, communicative purposes, and linguistic styles. Through a

multi-phase approach, combining persona-driven generation, structured task decomposition, and controlled narrative escalation, both short messages and extended conversations that mirror real-world dynamics of cyberbullying, including enabler, defender, and victim interactions are created.

Quantitative evaluations reveal that Phase-2 conversational outputs (LLaMA-3.3) outperform Phase-1 message-level outputs in semantic fidelity and alignment with human-authored datasets. While sentence-level average similarity (cosine) remains low across models, Phase-2 exhibits substantially lower Fréchet BERT Distance (FBD) values compared to Phase-1, indicating better distributional alignment and more realistic modelling of role-specific interactions. Self-BLEU at n-gram level also position Phase-2 conversational outputs as non-repetitive than Phase-1. This demonstrates that structured task decomposition and persona-driven prompting in extended conversations yield higher-quality, human-like synthetic data. Overall, this work establishes a practical framework for leveraging LLMs as a scalable and ethically viable tool for cyberbullying research. By offering controlled, human-like scenario generation, particularly through Phase-2 conversational strategies, LLMs can help address the scarcity of high-quality CB datasets and circumvent ethical risks to minors. This approach opens avenues for automated intervention design, cyberbullying prevention strategies, and broader application of LLMs in sensitive social science research, provided rigorous evaluation and ethical safeguards are maintained.

5.2 Dataset Profiling

This section presents the data profiling across **quantitative** and **linguistic** characteristics of all datasets created or augmented in this doctoral dissertation. A comparative analysis is conducted across eleven datasets: (a) three **youth** datasets - Italian (Sprugnoli et al., 2018), French (Ollagnier et al., 2022), youth co-created (§4.3.2); (b) three **translated** youth datasets - $\mathcal{D}^{HT}_{IT\rightarrow EN}$ (§5.1.1), $\mathcal{D}^{ST}_{FR\rightarrow EN}$ by students (§5.1.3), and $\mathcal{D}^{HMT}_{IT\rightarrow EN}$ by TransPerfect (§5.1.1); (c) three LLM-generated datasets: Phase 1 (Llama3, Mistral-v0.2) and Phase-2 (Llama3.3); (d) two real-world datasets - (Van Hee et al., 2018), which contains rich cyberbullying annotations, and Hosseinmardi et al. (2015), which provides binary bullying labels and reflects Instagram activity identified in §4.1. Only entries with explicit

binary bullying labels or equivalent annotations are retained for (Van Hee et al., 2018; Hosseinmardi et al., 2015). This section is divided in **three** sub-sections: (i) **Descriptive statistics**: Metrics such as message length, word/character counts, punctuation, and emoji frequency to enable comparison of surface linguistic features across datasets; (ii) **Lexical Diversit**y: To assess vocabulary variation across cyberbullying roles we use Type Token Ratio (Equation 5.1), i.e., proportion of unique tokens relative to total tokens, where higher values indicate broader vocabulary; and Shannon Entropy (Shannon, 1948) (Equation 5.2) which measures predictable/repetitive word use, where higher values indicate balanced word distribution and lower values indicate predictable/repetitive word use; (iii) **Turn-Taking Analysis**: Building on (Gravano and Hirschberg, 2011; Skantze, 2017), we define it as analysis of the consecutive messages by the same role, ending when another role intervenes. Three analyses are used: (a) Turn counts and mean messages per turn assess conversational balance (frequent exchanges vs. role dominance); (b) Turn-taking rate (role switches / total messages) quantifies participation fluidity; (c) Extended interaction analysis examines consecutive turns (persistence/dominance), sequential transition patterns (probability of one role following another), and conversational phases (which roles initiate, sustain, and close interactions). Collectively, these measures characterise micro- and macro-level interaction structures in cyberbullying exchanges.

5.2.1 Descriptive Statistics

As shown in Table 5.12, the youth datasets differ considerably in scale and composition. The youth co-created dataset (§4.3), developed with 12 students across 10 scenarios, contains 552 messages (5,237 words in total), averaging 55 messages per scenario. By comparison, the Italian dataset by Sprugnoli et al. (2018), spanning 4 scenarios, includes 1,984 messages with substantially longer conversations (about 496 messages per scenario). The French datasets (Ollagnier et al., 2022; Ollagnier, 2024), in turn, comprise 3,516 messages across 22 scenarios, with an average of 160 messages per scenario.

Overall, the comparisons demonstrate that the dataset's origin and method of construction directly affect its size, lexical diversity, and conversational structure. The youth co-created dataset (§4.3.2) and translated versions consist of relatively short messages with limited vocabulary, similar in average message length to the real-life corpora. Al-

Dataset Type	Dataset	$N_{\text{Scenarios}}$	N_{Text}	Avg. Text per Scenario	N_{Words}
Youth	Youth Co-created (§4.3)	10	552	55	5237
	(Sprugnoli et al., 2018) (Italian)	4	2192	496	14660
	(Ollagnier et al., 2022; Ollagnier, 2024) (French)	22	3516	160	24631
Youth Translated	$\mathcal{D}^{MT}_{IT\rightarrow EN}$ (§5.1.1)	4	2192	496	14660
	$\mathcal{D}^{ST}_{FR\rightarrow EN}$ (§5.1.3)	5	841	168	5576
	$\mathcal{D}^{HMT}_{FR\rightarrow EN}$(§5.1.3)	22	3516	160	24818
Real	(Van Hee et al., 2018) (EN)	1527	4637	3	70012
	(Hosseinmardi et al., 2015) (EN)	300	15667	52	200253
LLM-generated (§5.1.4)	Phase-1 Llama3	15	6610	441	161323
	Phase-1 Mistral	15	9553	637	216656
	Phase-2 Llama3.3	15	6677	445	68294

Table 5.12: Characteristics of datasets, including number of scenarios, messages, and words. Only subsets relevant to bullying labels are reported for Ask.fm and Instagram.***Note:*** *(Van Hee et al., 2018) original dataset for English after cleaning has 79314 messages across 6056 conversations, and (Hosseinmardi et al., 2015) original dataset for English has 42,259 comments across 769 posts (image+comment). In this table only the ones that are labelled as bullying for (Hosseinmardi et al., 2015) and role-labels are not negative for (Van Hee et al., 2018) are presented*

though the latter are much larger in scale and have substantially richer vocabularies. The LLM-generated datasets show a different pattern: in Phase 1, both Llama-3 and Mistral outputs have significantly longer messages on average (22–24 words), resulting in higher overall word counts compared to youth or real-life datasets. By contrast, the Phase 2 Llama-3.3 dataset, generated with a revised prompt design, produces shorter exchanges averaging 10 words per message, aligning more closely with the message length observed in youth, translated, and real-world datasets.

To examine the structural properties of the datasets, Figure 5.1 compares the relationships between message count, word count, character count, punctuation use, and emoji frequency. In the top-left plot, where the number of messages is plotted against the total number of words, the Phase-1 LLM-generated datasets (Llama-3 and Mistral) emerge as clear outliers, producing disproportionately high word counts relative to the number of messages. A similar deviation is observed in the top-right plot, which compares total words with total characters: despite containing fewer words overall, the Phase-1 LLM outputs do not align with the distributional patterns of the other datasets. The bottom-left

plot, contrasting total characters with punctuation marks, further reinforces this trend, as Phase-1 LLM datasets again diverge from the consistent relationship observed elsewhere. In contrast, the bottom-right plot, which compares total words with the number of emojis, highlights a different pattern: here it is the Phase-2 LLM-generated dataset that departs from the expected distribution, indicating distinct stylistic tendencies in emoji use compared to both human-produced and Phase-1 machine-generated corpora.

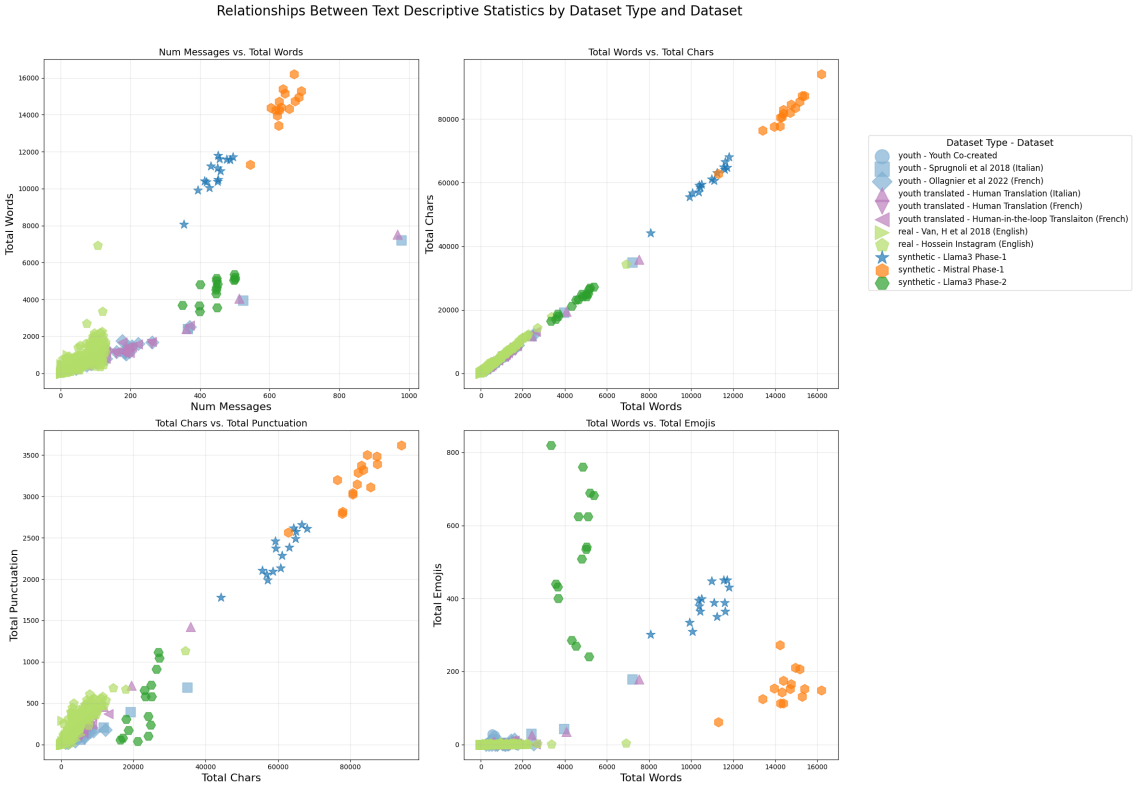


Figure 5.1: Relationship between different text descriptive statistics across all datasets

A feature-level comparison was conducted across multiple text properties, including word count, character count (with and without spaces), whitespace frequency, average words per sentence, punctuation counts, unique words per message, average word length, capitalised words, fully capitalised words, digits, and special characters. As shown in Figure 5.2, the left subplot presents these features for the youth datasets (co-created, Italian, and French) alongside the real-life corpora from Ask.fm and Instagram. The youth datasets follow broadly similar patterns to the real-life datasets, with only minor deviations. The right subplot in Figure 5.2, which illustrates differences relative to the real-life datasets, indicates that character counts (with and without spaces) are noticeably

lower in the youth datasets, whereas measures such as unique words and punctuation show minimal differences between youth and real-life corpora.

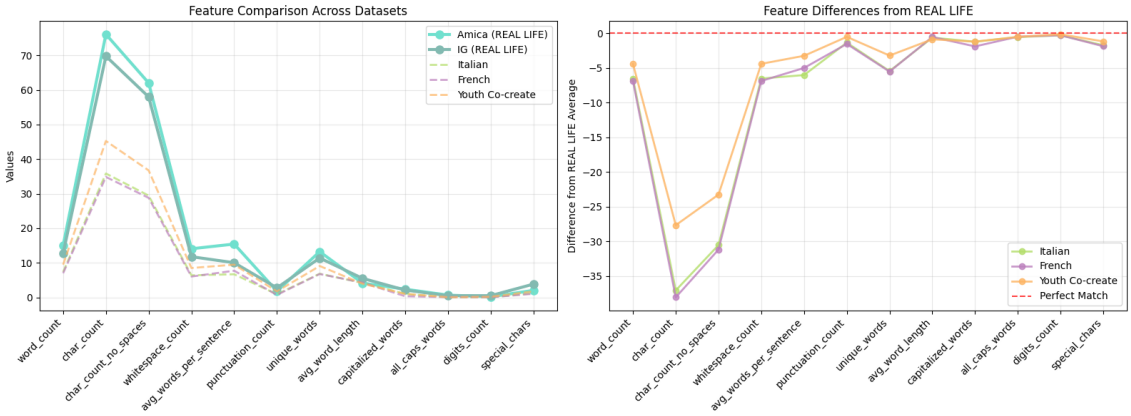


Figure 5.2: Text feature comparison across youth datasets with Amica i.e., (Van Hee et al., 2018) and IG i.e., (Hosseinmardi et al., 2015)

In contrast, analysis of the synthetic datasets generated for data augmentation (§5.1) reveals distinct trends. As shown in Figure 5.3, the Phase-1 LLM-generated datasets exhibit substantially higher character counts (with and without spaces) and a greater number of unique words compared to real-life datasets. When differences relative to real-life corpora are computed, Phase-1 LLM outputs display opposite patterns to those observed for youth and translated datasets, further underscoring their divergence. The Phase-2 LLM-generated dataset, however, aligns much more closely with the translated datasets and the youth co-created dataset, showing comparable distributions across most textual features.

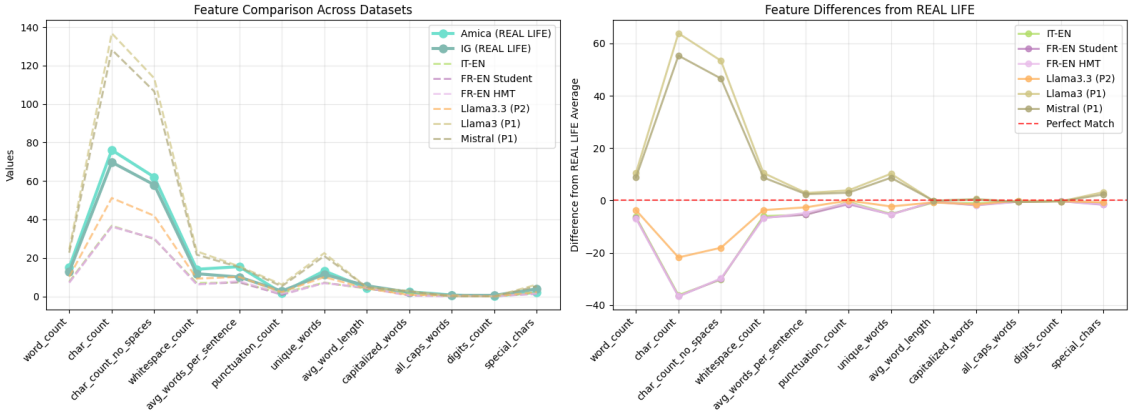


Figure 5.3: Text feature comparison across synthetic datasets with Amica i.e., (Van Hee et al., 2018) and IG i.e., (Hosseinmardi et al., 2015)

5.2.2 Lexical Diversity

The Type-Token Ratio (TTR), defined as the proportion of unique words to the total number of words in a corpus, is reported at the corpus (dataset) level in Table 5.13, with mean and standard deviation values. At this level, the difference in mean (μ) TTR between the real-life datasets and the youth co-created dataset (§4.3) is -0.02 for Ask.fm and similar for Instagram. In contrast, for Phase-1 LLM-generated texts, the mean TTR exceeds that of Ask.fm by +0.02 and that of Instagram by +0.04. For Phase-2 LLM-generated texts, however, the difference narrows and aligns more closely with the youth co-created dataset. For the translated English datasets, those produced through $\mathcal{D}^{HT}_{IT\rightarrow EN}$ (§5.1.1) exhibit TTR values similar to the original Italian datasets, while those produced through human-aided translation $\mathcal{D}^{HMT}_{FR\rightarrow EN}$ (§5.1.3) are also consistent with the original French dataset. Across all datasets, the mean TTR values are close to 1 and for all datasets except Sprugnoli et al. (2018)'s and its translated English dataset $\mathcal{D}^{MT}_{IT\rightarrow EN}$ (§5.1.1), the standard deviation (σ) is less than 0.09, indicating high lexical diversity and relatively low repetition in the texts. Examining the distribution of TTR across datasets (Figure 5.4), the histogram on the left indicates that for all texts, TTR values range from 0.6 to 1.0, with peaks concentrated between 0.9 and 1.0 across all datasets. When focusing only on messages written by bullies or enablers for role-based datasets, and those labelled as bullying by (Hosseinmardi et al., 2015), the distributions remain broadly similar across datasets (middle plot), with enabler messages showing comparable lexical diversity across all. In contrast, for defender or victim messages, the Llama-3 Phase-1 distribution exhibits a noticeably higher concentration in the 0.9–1.0 range than the other datasets. This suggests that LLM-generated defender/victim messages tend to be shorter and use more unique words compared to human-authored ones. While TTR provides a useful first indication of vocabulary richness, it remains a rudimentary measure and does not account for how evenly words are distributed. To address this limitation, we next examine Shannon entropy, which captures both the diversity and the balance of lexical use.

Figure 5.5 reports the Shannon entropy distributions for each dataset at the character (left) and word (right) levels. Across youth-generated datasets, mean entropy values are consistently below 4 at the character level and around 3 at the word level, closely matching the values observed in real-world corpora and in translated English datasets. This suggests

Dataset Type	Dataset	Mean (μ) TTR	Std. (σ) TTR
Youth	Youth Co-created (§4.3)	0.97	0.06
	(Sprugnoli et al., 2018) (Italian)	0.95	0.12
	(Ollagnier et al., 2022; Ollagnier, 2024) (French)	0.98	0.06
Youth Translated	HT IT→EN (§5.1.1)	0.95	0.12
	Student Translation FR→EN (§5.1.3)	0.98	0.05
	H-MT FR→EN (§5.1.3)	0.98	0.06
Real life	(Van Hee et al., 2018) (Ask.fm)	0.95	0.08
	(Hosseinmardi et al., 2015) (Instagram)	0.97	0.07
LLM-generated (§5.1.4)	Phase-1 Llama3	0.93	0.06
	Phase-1 Mistral	0.93	0.06
	Phase-2 Llama3.3 Conversation	0.98	0.04

Table 5.13: Mean and standard deviation of Type Token Ratio (TTR) across all datasets

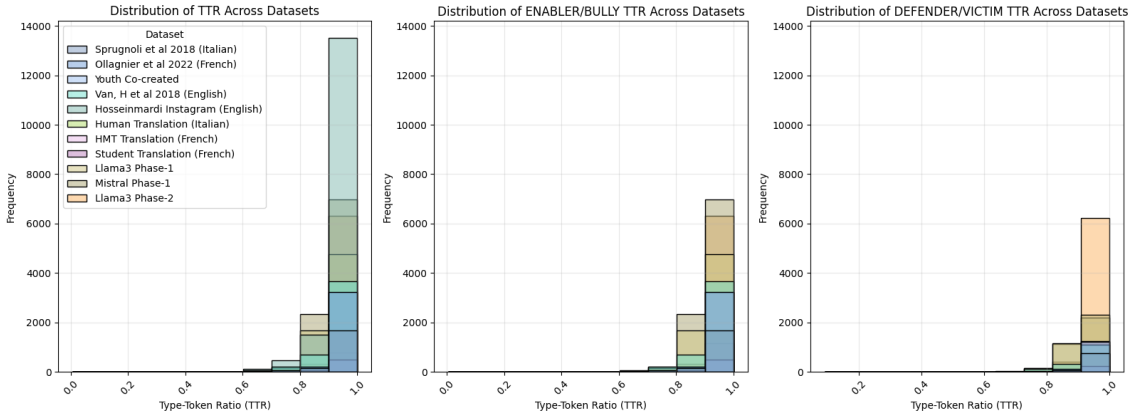


Figure 5.4: Type-Token-Ration (TTR) across all datasets and also across Bully/Enabler texts and Victim/Defender texts across all datasets

that youth, real, and translated texts share a similar degree of lexical diversity, striking a balance between lexical predictability and variation. In contrast, the Phase-1 LLaMA3 synthetic dataset exhibits substantially higher entropy scores at both character and word levels, indicating a greater degree of unpredictability and lexical variety that departs from natural human text distributions. Later synthetic datasets, such as Phase-2 LLaMA3 and Mistral, display entropy values closer to those of the youth and real datasets, reflecting a reduction in exaggerated diversity and a shift toward more naturalistic lexical patterns.

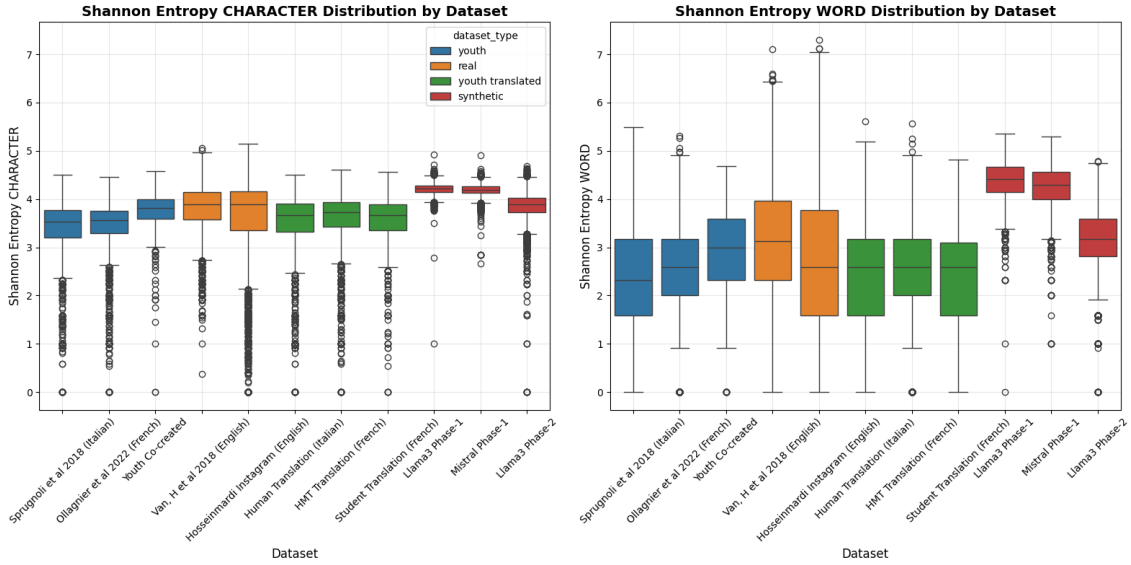


Figure 5.5: Shannon Entropy Distribution at both Word (Token) or Character Level across datasets

Taken together, the TTR and entropy analyses show that (a) youth co-created and real-world datasets exhibit comparable lexical diversity and balance, (b) translation with human involvement or through human-aided approaches does not change lexical diversity, (c) while early-stage synthetic datasets (e.g., Phase-1 LLaMA3) diverge with exaggerated variety. It is worth noting that the later synthetic dataset (Phase-2 LLaMA3.3) converges towards the naturalistic patterns of human-authored text, underscoring the importance of iterative refinement in aligning LLM outputs through detailed prompt structure.

5.2.3 Turn-taking Analysis

Turn-taking is fundamental to understanding how participants exchange messages or utterances sequentially within an interaction (Sacks et al., 1974). In the context of cyberbullying-related text exchanges, turn-taking captures not only the order in which

messages occur but also the dynamics between different roles, such as bullies, victims, defenders (victim supporters), and enablers (bully supporters). A "***turn***" can be understood as a continuous sequence of messages authored by the same participant or role that ends when another role takes over. It is important to note that only the datasets introduced by (Sprugnoli et al., 2018; Ollagnier et al., 2022; Ollagnier, 2024) follow a conversational style and include turn-level annotations. Consequently, this analysis focuses on four datasets: the youth co-created dataset described in §4.3 and the LLM-generated Phase-2 Llama3.3 Conversation dataset in §5.1.4.

Turn-taking can be formally defined over a sequence of messages $M = \{m_1, m_2, ..., m_N\}$ in a conversation C for dataset $\mathcal{D}$, where each message m_i is associated with a participant role $r_i \in \{$*bully, victim, enabler, defender*$\}$. A *turn-switch* is detected whenever $r_i \neq r_{\text{i-1}}$. Total turns in a conversation C is calculated as Total turns $(T) = \sum_{i=1}^{N} \delta_i$, where $\delta_i = \texttt{1}$ if $r_i \neq r_{\text{i-1}}$, otherwise $= \texttt{0}$. This measures how many stretches of uninterrupted participation or turns occur in the conversation. High values will indicate frequent exchanges between roles and lower values will indicate dominance by a particular role. To understand the average messages per turn in a conversation, where for each turn $t \in T$, the number of messages within the turn is denoted by n_t, and the average is calculated as $\frac{1}{T}\sum_{t=1}^{T} n_t$. And the turn-taking rate i.e., the rate at which turns switch or are taken is calculated $\frac{\sum_{i=2}^{N} 1[r_i \neq r_{i-1}]}{N}$, where N is total number of messages and r_i denotes the role of the i-th message, thereby quantifying how often the conversational turn changes from one role to another.

To understand the temporal dynamics of role interactions in cyberbullying conversations, we extend this turn-taking analysis to include analysis of (i) consecutive turns, (ii) sequential transition patterns, and (iii) positional patterns of participant roles in cyberbullying conversation datasets.

- **Consecutive Turn Analysis**: To understand consecutive turn, the participant roles are divided into role groups and are identified by sequences where the same role maintains continuous participation. Given a conversation C with message sequence $M = \{m_1, m_2, ..., m_N\}$, role groups are formed as G=$\{g_1, g_2, ..., g_k, \}$, where each group g_k represents a maximum sequence of consecutive messages from the same role r_k. Here, k indexes each uninterrupted sequence of messages, running from 1 to

the total number of role groups in the conversation. For instance, if a bully sends two messages in a row, followed by a victim message, and then a defender message, the conversation would have three role groups $k = \{1, 2, 3\}$ corresponding to the bully, victim, and defender sequences, respectively. Analysing these role groups allows us to measure consecutive turn patterns, including the average and maximum number of messages per role group, which provides insights into role dominance, persistence, and participation dynamics in cyberbullying interactions.

- **Sequential Turn Analysis**: This provides insight into the patterns of role interactions within multi-participant conversations. The dynamics of role interactions within a conversation are captured by examining patterns of consecutive roles rather than words. To understand this, let the sequence of roles in a conversation be $R = \{r_1, r_2, ..., r_N\}$. *Bigram transitions* refer to pairs of consecutive roles (r_i, r_{i+1}) capturing direct role-to-role interactions. To quantify the likelihood of one role following another, transition probabilities are defined as $P(r_j|r_i) = \frac{|\{k:r_k=r_i and r_{k+1}=r_j|}{|\{k:r_k=r_i, k<N\}|}$, where k is an index over the messages in the conversation, running from 1 to N-1, and the numerator identifies all positions where role r_i is immediately followed by role r_j, and the denominator is a count of all positions where role r_i occurs (excluding the last message, which has no subsequent role). This allows us to detect stable interaction patterns, such as whether bullies are often followed by defenders or whether enablers typically follow bullies, highlighting the structural dynamics of multi-party conversations.

- **Conversational Phase Analysis**: To understand how turns change with conversational phases i.e., who begins the conversations or who ends it. For each conversation C in dataset $\mathcal{D}$ with a total of N messages with T turns and $m - m_N$ messages, C is divided into three phases i.e., messages m_1 to $m_{[N/3]}$ as *opening phase* (C_{Open}), messages $m_{[N/3]+1}$ to $m_{[2N/3]}$ as *middle phase* (C_{Middle}), and $m_{[2N/3]+1}$ to m_N as *closing phase* ((C_{Close}). Analysing turns across each conversational phase in this manner allows understanding of not only who initiates, dominates, or closes interactions across datasets, but also which roles withdraw from the conversation and which become more active.

Results: As shown in Table 5.14, the four datasets differ not only in size but also in their turn-taking dynamics, which are critical indicators of conversational style and interactional density. The youth role-played datasets (Sprugnoli et al., 2018; Ollagnier et al., 2022; Ollagnier, 2024) exhibit relatively high turn-taking rates, ranging from 0.69 to 0.86, suggesting frequent speaker alternation. In comparison, the youth co-created cascading-styled dataset demonstrates a slightly lower turn-taking rate of 0.55 across 10 scenarios, while the Phase-2 LLaMA3.3 conversational dataset maintains a moderate rate of 0.71 across 134 conversational runs for distinct 15 scenarios. This highlights that the cascading style rearranging of the youth co-created dataset in §4.3.2 does not reflect the role-playing or LLM role-play style turn-taking. However, a closer look at message-to-turns ratios further highlights the differences. For instance, the youth co-created dataset averages 55 messages across 10 conversations, with 32 turns, yielding a message-to-turn ratio of 1.7, indicating that multiple messages are often grouped within a single turn. In contrast, in (Sprugnoli et al., 2018) across just 4 scenarios, conversations average 496 messages and 334 turns, with a lower message-to-turn ratio of ≈1.5. Similarly, the datasets by (Ollagnier et al., 2022; Ollagnier, 2024) show tighter alignment between messages and turns, with a ratio of 1.15 (160 messages and 139 turns on average). The Phase-2 LLaMA3.3 dataset, covering 134 runs across 15 scenarios, lies in between: 50 messages per conversation, 36 turns, and a message-to-turn ratio of 1.38. From a profiling perspective, these distinctions underscore how datasets capture different conversational "*rhythms*". High turn-taking rates may reflect more reactive, dialogic interactions, while higher message-to-turn ratios suggest more bundled or extended speaker contributions. Such metrics helps characterise datasets not only for dialogue modelling tasks but also for downstream analyses of interactional patterns, conversational load, and realism in simulated dialogues.

Consecutive Turn Analysis: To examine how conversational roles persist across interactions, we analysed consecutive turn patterns by segmenting conversations into role groups, where each group represents a continuous sequence of messages from the same role. Table 5.15 reports the maximum length of consecutive turns per role across all datasets. Clear differences emerge in role persistence. In the role-played dataset by (Sprugnoli et al., 2018), both bullies and enablers reach a maximum of six consecutive turns, while victims and defenders sustain up to four. In contrast, the datasets by (Ollagnier et al.,

Dataset	Dataset Size	Total Turns	# Convs.	Avg. Msgs. /Conv.	Avg. Turns /Conv.	Avg. Turn-Taking Rate
(Sprugnoli et al., 2018) (Italian)	2192	1338	4	496	334	0.69
(Ollagnier et al., 2022; Ollagnier, 2024) (French)	3516	3050	22	160	139	0.86
Youth Co-created (§4.3)	551	317	10	55	32	0.55
Phase-2: Llama3.3 Conversations (§5.1.4)	6677	4827	134	50	36	0.69

Table 5.14: Summary of datasets with total turns, number of conversations, average messages and turns per conversation, and turn-taking rate.

2022; Ollagnier, 2024) show much shorter streaks, with no role exceeding three consecutive turns. Interestingly, the cascading re-arranging of messages in the youth co-created dataset allows defenders and enablers to maintain up to four consecutive turns, but victims rarely sustain participation beyond one turn. In the Phase-2 LLaMA-3.3 dataset, victims achieve higher persistence, sustaining up to four consecutive turns, whereas other roles show a maximum of only one to three turns. Silent bystanders appear only in the Phase-2 LLaMA-3.3 dataset, with one instance of a consecutive turn. This variation underscores how different datasets model role dominance and control of the conversational scenario. Longer streaks for bullies in role-played settings may reflect scripted reinforcement of aggression, while shorter runs in co-created and LLM-generated dialogues point toward more dynamic interruptions and balancing of roles. Average consecutive turns (Table 5.16) provide a complementary picture. In (Sprugnoli et al., 2018), the averages for all active roles hover between 1.36 and 1.49, suggesting relatively balanced turn-taking, with no role disproportionately sustaining long sequences. The datasets by (Ollagnier et al., 2022; Ollagnier, 2024) exhibit even tighter clustering, with averages between 1.10 and 1.24, reflecting highly fragmented exchanges where participants rarely hold the floor in the conversational scenario for long. In contrast, the youth co-created dataset shows greater variability: defenders (2.02) and enablers (1.85) often sustain consecutive turns, while

victims (1.06) contribute more sporadically. The Phase-2 LLaMA-3.3 dataset further amplifies this asymmetry i.e., victims average 3.62 consecutive turns, far exceeding all other roles, suggesting an emergent pattern where victims produce extended monologic contributions, potentially reflecting reflective or defensive responses. From a dataset profiling perspective, consecutive turn analysis reveals structural biases in interactional dynamics across cyberbullying conversational datasets. Role-played datasets tend to emphasize bully persistence, whereas co-created and LLM-generated datasets highlight defender or victim continuity. These distinctions are crucial for understanding how datasets encode conversational power, role salience, and turn-taking asymmetries. Such profiling not only informs the selection of datasets for modelling but also provides interpretive context for evaluating the realism and representational fidelity of simulated cyberbullying dialogues.

Dataset	**Bully**	**Enabler**	**Victim**	**Defender**	**Silent Bystander**
(Sprugnoli et al., 2018)	6	6	4	4	0
(Ollagnier et al., 2022; Ollagnier, 2024)	2	3	2	2	0
Youth Co-created	3	4	1	4	0
Phase-2 LLama-3.3 Conversations	3	1	4	1	1

Table 5.15: Maximum consecutive turn per role across all conversations

Dataset	**Bully**	**Enabler**	**Victim**	**Defender**	**Silent Bystander**
(Sprugnoli et al., 2018)	1.39	1.48	1.49	1.36	0.00
(Ollagnier et al., 2022; Ollagnier, 2024)	1.10	1.24	1.10	1.12	0.00
Youth Co-created	1.47	1.85	1.06	2.02	0.00
Phase-2 LLama-3.3 Conversations	1.70	1.24	3.62	1.24	1.02

Table 5.16: Average consecutive turn per role across all conversations

Sequential Turn Analysis: To understand sequential role-to-role interactions, we compute bigram transition probabilities and visualise them as network graphs in Figure A.8. In the (Sprugnoli et al., 2018) dataset (top-left), the most theoretically significant transitions are the strong enabler–defender reciprocity (P(defender|enabler)=0.493; P(enabler|defender)=0.4) and bully dominance following victims P(bully|victim) = 0.445

vs $P(\text{victim}|\text{bully}) = 0.253$, reflecting the dataset's structured classroom role-play origins. In the (Ollagnier et al., 2022; Ollagnier, 2024) dataset (top-right), enablers emerge as the most reactive role, prominently following both victims ($P(\text{enabler}|\text{victim})$=0.378) and defenders ($P(\text{enabler}|\text{defender})$=0.338), suggesting enablers amplify rather than directly challenge aggression. The youth co-created dataset (bottom-left) is defined by an exceptionally strong enabler–defender dialogic loop ($P(\text{defender}|\text{enabler})$=0.792) vs ($P(\text{enabler}|\text{defender})$=0.615) and elevated defender–victim transitions ($P(\text{defender}|\text{victim})$=0.417), reflecting young researcher's deliberate foregrounding of protective exchanges. In the LLM-generated dataset (bottom-right), the dominant transition is bully following victim ($P(\text{bully}|\text{victim})$=0.612), with defenders actively countering bullies ($P(\text{defender}|\text{bully})$=0.500), emphasising antagonistic dynamics over bystander coordination. Detailed transition analysis are provided in Appendix A.12. Across datasets, these patterns reveal how dataset design shapes role visibility and centrality. Role-play datasets promote reciprocal bystander exchanges, with (Sprugnoli et al., 2018) foregrounding enabler–defender and bully–defender dyads, while (Ollagnier et al., 2022; Ollagnier, 2024) elevates enabler reactivity. The youth co-created dataset reinforces defender–enabler reciprocity and explicit victim support, whereas synthetic LLM generation accentuates bully–victim conflict with defenders as counter actors but limited victim–bystander coordination.

Conversational Phase Analysis: When dividing conversations into phases like *opening phase* (C_{Open}; first one third turns), *middle phase* (C_{Middle}; second one third turns), and *closing phase* (C_{Close}; final one third turns), one can examine which participant roles dominate at different stages of conversations and helps understand structural patterns in conversational dynamics. Table 5.17 summarises the distribution of average turns across all conversations for datasets. In the opening phase (C_{Open}), the Italian dataset by (Sprugnoli et al., 2018) shows that conversations are primarily initiated by bullies (30.52%) and defenders (30.98%), suggesting that confrontational exchanges between these roles are characteristic of conversation onset. In contrast, in the French datasets (Ollagnier, 2024; Ollagnier et al., 2022), enablers (24.40%) and bullies (22.89%) contribute more prominently than victims or defenders, indicating a stronger presence of supportive or amplifying behaviours early on. The youth co-created dataset (§4.3.2) differs further, as enablers (36.36%) and defenders (35.35%) are both highly active from the

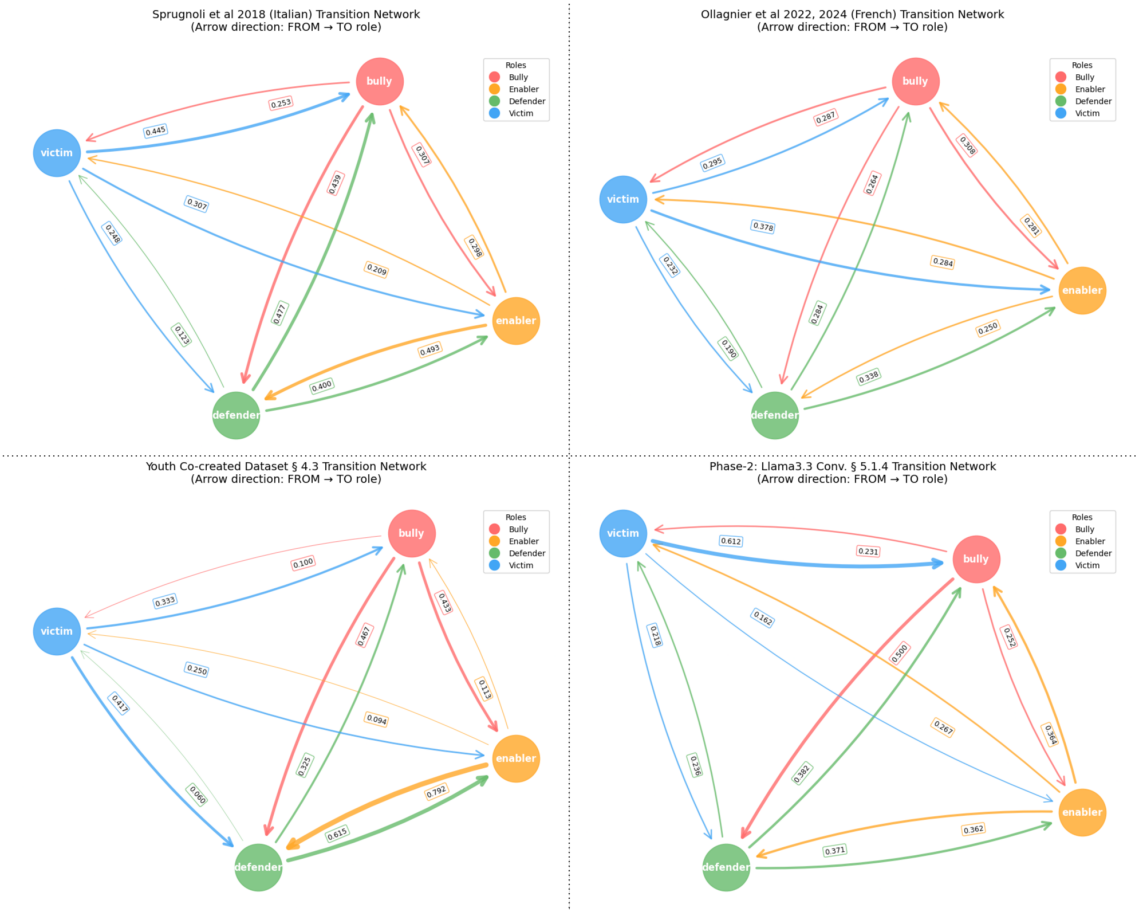


Figure 5.6: Network graph depicting direction of Bi-gram transition with probabilities as weights across all datasets

outset, showing confrontational exchanges as the Italian dataset. Meanwhile, in the Phase 2 Llama3.3 LLM generated dataset (§5.1.4), bullies most frequently initiate interactions (34.27%). In the middle phase (C_{Open}), role dominance continues to diverge across corpora. In the Italian dataset, bullies still lead the conversational turns (33.99%), whereas in the French dataset, enablers account for the highest proportion of turns (26.35%). In the LLM-generated conversations, bullies remain predominant (30.46%), reinforcing their persistence across conversation progression. Notably, the youth co-created dataset diverges from these trends, as defenders (37.61%) become the most active role in the middle phase, suggesting a shift toward supporting or defending interventions as dialogues evolve. In the closing phase (C_{Close}), the French dataset maintains enabler prominence (25.34%), followed closely by bullies (21.71%), while the LLM-generated dataset again shows strong bully dominance (31.33%). In contrast, the Italian dataset exhibits higher defender activity (31.83%), alongside a notable presence of enablers (29.35%). The youth co-created dataset is distinctive in its closure, with defenders overwhelmingly driving the conversation end (42.20%).

Phase	**Dataset**	**Bully**	**Enabler**	**Victim**	**Defender**	**Conciliator**	**Silent Bystanders**
Opening	Italian	**30.52**	24.37	14.12	**30.98**	0.00	0.00
	French	22.89	**24.40**	21.69	20.08	10.94	0.00
	Youth Co-created §4.3	19.19	**36.36**	9.09	**35.35**	0.00	0.00
	§5.1.4 Phase-2: Llama3.3 Conv.	**34.27**	20.23	17.95	26.37	0.00	1.17
Middle	Italian	**33.99**	23.03	17.98	25.00	0.00	0.00
	French	20.46	**26.35**	19.02	20.46	13.71	0.00
	Youth Co-created §4.3	21.10	31.19	10.09	**37.61**	0.00	0.00
	§5.1.4 Phase-2: Llama3.3 Conv.	**30.46**	21.16	19.88	27.42	0.00	1.09
Closing	Italian	22.12	29.35	16.70	**31.83**	0.00	0.00
	French	21.71	**25.34**	18.37	19.45	15.13	0.00
	Youth Co-created §4.3	21.10	33.03	3.67	**42.20**	0.00	0.00
	§5.1.4 Phase-2: Llama3.3 Conv.	**31.33**	20.55	20.00	26.97	0.00	1.15

Table 5.17: Average proportion of turns (percentages) taken by different roles across opening, middle, and closing phases in the Italian (Sprugnoli et al., 2018), French (Ollagnier et al., 2022; Ollagnier, 2024), youth co created, and LLM generated datasets.

5.2.4 Conclusion and Comparative Insights

This comprehensive quantitative, lexical, and interactional dataset profiling reveal that each dataset has distinct strengths and limitations, reflecting their origin, construction method, and intended purpose.

Youth Role-Played Datasets ($\mathcal{D}_{IT}$ & $\mathcal{D}_{FR}$): The Italian and French youth role-played datasets by Sprugnoli et al. (2018) and Ollagnier et al. (2022); Ollagnier (2024) respectively, provide rich turn-taking dynamics and sequential interactions that closely mimic structured, classroom-style dialogues. The Italian dataset is particularly notable for its emphasis on bully-defender exchanges and longer, sustained interactions, capturing conversational persistence effectively. The French dataset, in contrast emphasizes enabler participation and both rapid, fragmented interactions by defenders, highlighting the role of supportive participants in bullying scenarios. Both datasets exhibit high lexical diversity (TTR and entropy) and offer detailed turn-level annotations, making them valuable for modelling realistic dialogue patterns and analysing role-based interventions.

Youth Co-Created Dataset: The co-created dataset in §4.3 distinguishes itself by foregrounding defender and enabler reciprocity through a cascading, scenario-based construction. While it contains fewer messages and a smaller vocabulary compared to the Italian or French role-played corpora, it exhibits comparable lexical diversity to real-world datasets. Its sequential turn probabilities show deliberate enabler–defender interactions, reflecting the conscious intent of youth participants to highlight protective behaviours. Turn-taking metrics indicate moderately bundled exchanges, with multiple messages often grouped in single turns, and the phase-wise analysis demonstrates that defenders become dominant as conversations progress. These characteristics suggest that, despite its smaller scale, the youth co-created dataset effectively models naturalistic conversational rhythms and protective interventions, making it suitable for tasks that require realistic role interactions, dialogue understanding, or educational applications.

Youth Translated Datasets: Translation, whether via human translators (§5.1.1) or human-aided methods (§5.1.2), preserves the lexical diversity and structural properties of the original corpora. Vocabulary richness is moderately reduced in English translations, but message length and interaction patterns remain largely consistent, confirming the reliability of translated datasets for cross-linguistic analyses or multilingual model training.

LLM-Generated Datasets: Phase-1 LLM outputs produce longer, highly diverse messages with exaggerated lexical variety, diverging from natural human patterns. This is particularly evident in TTR and Shannon entropy scores, as well as message-to-turn ratios. Phase-2 LLaMA3.3 outputs, however, align more closely with youth co-created and real-world datasets, producing shorter messages, naturalistic turn-taking, and moderated lexical diversity. Sequential analyses highlight a stronger focus on bully-victim antagonism, with defenders responding actively, whereas enabler-victim interactions remain limited. These findings suggest that iterative refinement of prompts and model design is essential to approximate human-like conversational patterns.

Across datasets, several key patterns emerge. In terms of **lexical diversity**, the youth co-created dataset, real-world corpora, and Phase-2 LLM-generated datasets exhibit comparable TTR and Shannon entropy values, reflecting balanced and realistic vocabulary usage. In contrast, Phase-1 LLM outputs tend to overestimate diversity, while the role-played youth datasets display structured lexical exchanges shaped by the instructions given to participants. Regarding **turn-taking** and role dynamics, role-played datasets excel at modelling structured, reciprocated interactions, particularly in bully-defender or enabler-victim exchanges. The youth co-created dataset, although smaller, successfully emphasises defender–enabler reciprocity and captures realistic conversational phases, making it a suitable alternative for applications that focus on protective interventions. Phase-2 LLM outputs approximate these dynamics as well, but tend to highlight bully-victim antagonism more prominently. Overall, the youth co-created dataset strikes a balance between realism, role-focused interactions, and manageable scale. While it is not as extensive as the role-played or real-world corpora, it is sufficiently robust for tasks that require representative turn-taking patterns, conversational realism, and lexical diversity, without the overhead of extremely large datasets.

5.3 Intrinsic Evaluation and Results

Despite the youth role-play methodologies employed by (Sprugnoli et al., 2018; Ollagnier et al., 2022; Ollagnier, 2024), or the cascading co-design methodology for the youth co-created dataset discussed in §3.3.3 and 4.3 for dataset creation, they are not grounded in

real online contexts. This raises a key concern about their ecological validity. Although real-world cyberbullying datasets such as AMiCA (Van Hee et al., 2018) and Instagram (Hosseinmardi et al., 2015) exist, they present significant limitations. AMiCA provides role-specific annotations but is collected from a now-defunct question-answering platform, causing responses from defenders, enablers, bullies, or victims to appear after multiple unrelated posts, which undermines conversational coherence. Similarly, the Instagram dataset uses binary post-level annotations, i.e., where entire posts (image+caption+comment) are labelled as bullying or not and are unable to capture conversational context or nuanced roles. In addition to this, youth representation in the datasets is not verifiable. It is important to note, that the ***cyberbullying*** phenomenon is not reducible to isolated offensive utterances; rather, it unfolds as a complex, multi-actor phenomenon characterised by repetition, power imbalance, conflict, escalation and de-escalation, peer support, and influence processes (Hinduja and Patchin, 2013; Salmivalli, 2014; Englander et al., 2017; Huang et al., 2020; Ziems et al., 2020; Steer et al., 2020). The lack of such cyberbullying dynamics has been noted in the literature as a key gap (Rosa et al., 2019; Salawu et al., 2017; Emmery et al., 2021; Kim et al., 2021a), with researchers emphasising the need for datasets that reflect bullying dynamics as they occur in practice.

Without datasets that embody these characteristics, it is impossible to establish meaningful ground truth against which moderation systems or detection models can be reliably assessed. Although NLP literature often uses *intrinsic evaluation* as a term to measure a system's quality of representation via semantic similarity or analogy performance (See (Schnabel et al., 2015; Phillips et al., 2017; Hossain and Hoque, 2020), among others), (Clark et al., 2013) note, **intrinsic evaluation** can be understood as the assessment of system outputs against a gold standard or reference benchmark (Clark et al., 2013). Given the lack of a reference benchmark or gold standard, we design and implement an ***intrinsic human evaluation protocol*** to determine ***whether a dataset can be perceived as realistic and whether it faithfully reflects the dynamics of cyberbullying***. Building on the framework by (Chhun et al., 2022), which assesses story-like LLM outputs for relevance and coherence. We extend this human evaluation approach for multi-party CB dialogues by explicitly incorporating "**realism**" (Agha et al., 2024) and "**online bullying dynamics**" to capture authentic social media interactions across characteristics as

studied by (Hinduja and Patchin, 2013; Salmivalli, 2014; Englander et al., 2017; Van Hee et al., 2018; Sprugnoli et al., 2018; Huang et al., 2020; Ziems et al., 2020; Steer et al., 2020; Ollagnier et al., 2022). This approach will help discern the ecological validity of CB datasets, i.e., the extent to which the dataset accurately reflects real-world youth cyberbullying interactions, including participant roles, temporal patterns, and social dynamics. This helps ensure that the dataset meaningfully replicates the dynamics of real-world bullying scenarios.

5.3.1 Objectives and Dimensions

Objective: To assess the intrinsic quality of cyberbullying datasets and ensure their reliability and suitability for training and evaluating cyberbullying detection systems. Specifically, the framework seeks to determine whether different types of data such as (i) youth co-created conversations (§4.3.2), (ii) human and machine generated dialogues ($\mathcal{D}^{HT}_{IT\rightarrow EN}$ §5.1.1, $\mathcal{D}^{HMT}_{IT\rightarrow EN}$ 5.1.3, $\mathcal{D}_{LLM}$ 5.1.4), and (iii) real-world datasets such as those by (Hosseinmardi et al., 2015), possess the essential intrinsic qualities necessary for meaningful CB research. In essence, the framework aims to answer the following question:

> To what extent do youth co-created, machine-generated, or real-life cyberbullying conversations reflect key intrinsic qualities, including *realism* (D_{Real}), *relevance* (D_{Relevant}), *conversational coherence* (D_{Coherent}), and the *social dynamics of cyberbullying* ($D_{\text{CB-Dynamics}}$)?

By grounding the evaluation in these intrinsic qualitative dimensions, this human evaluation framework introduces a novel approach to establishing meaningful ground truth in the absence of real-life data. It enables the assessment of the **ecological validity** of different datasets, determining whether they authentically mimic the conversational patterns, social dynamics, and language styles observed in real-world cyberbullying. Furthermore, the framework informs the creation and refinement of dataset resources, enabling the establishment of a **gold-standard benchmarking** dataset and a complementary **silver-standard training** dataset. Thereby enabling evaluation of datasets suitable for downstream tasks such as participant role identification, detection of bullying escalation and support behaviours, and conversational coherence analysis.

Dimensions: Following (Chhun et al., 2022), who propose an evaluative framework for AI-generated narrative content focusing on dimensions such as *relevance*, *coherence*, *empathy*, *surprise*, *engagement*, and *complexity*, we extend this framework to include ***realism*** as a critical dimension for assessing authenticity in conversational datasets (Huang et al., 2020; Agha et al., 2024). Realism has been previously applied to evaluate ecological validity in cyberbullying scenarios/vignettes, and in this work, it has been used as a dimension for scenario/vignette evaluation (§3.3.2 Chapter 3, 4.2 Chapter 4). This ***intrinsic*** evaluation framework is designed to evaluate machine generated, human-designed, and real-world cyberbullying conversational data along **four key** dimensions: *realism* (D_{Real}), *relevance* (D_{Relevant}), *conversational coherence* (D_{coherent}), and *six sub-dimensions* of *cyberbullying social dynamics* ($D_{\text{CB-Dynamics}}$) which include, *conflict, support, influence, humour, power-dynamics,* and *repetition.* These dimensions together help ensure that synthetic datasets capture the nuances and complexities of real-world social interactions, particularly in cyberbullying contexts.

Realism and Relevance: Authenticity is vital for synthetic conversational data for training cyberbullying moderation systems. Additionally, (Ziems et al., 2020; Emmery et al., 2021; Rosa et al., 2019) has highlighted the importance of capturing nuanced cyberbullying behaviour including aggression, repetition, and power imbalance within datasets for improved detection classifier performance. To that effect, we extend (Huang et al., 2020; Agha et al., 2024)'s *realism* and (Chhun et al., 2022)'s *realism* as,

- **Realism $D_{\textbf{Real}}$:** Evaluates the extent to which conversations mirror authentic social media interactions, considering tone, language, and interactional flow.
- **Relevance $D_{\textbf{Relevant}}$:** Assesses how directly the conversation addresses the scenario or vignette it is intended to represent.

By focusing on these dimensions, both youth-designed and machine-generated conversations can be evaluated in context with real-world cyberbullying dataset by (Hosseinmardi et al., 2015).

Conversational Coherence $D_{\textbf{Coherent}}$: This ensures that conversations are easy to follow and mirror natural social media interactions, with coherent turn-taking, causal

message flow, rather than presenting disjointed exchanges or unrealistic dialogue patterns. Prior work demonstrates the importance of modelling temporal dynamics and topic coherence in harmful interactions (Ge et al., 2021; Ghosh et al., 2024). For example, methods such as unified temporal graphs and context-aware transformers highlight the role of sequence, topical consistency, and emotional flow in comprehending cyberbullying dialogues. Following (Chhun et al., 2022), clarity or *conversational coherence* is evaluated based on the structural and semantic continuity of message flow. We extend this to assess how easily can conversations be followed despite asynchronous participation and threaded replies.

Cyberbullying Social Dynamics $D_{\textbf{CB-Dynamics}}$: Social dynamics capture the interpersonal and relational aspects of cyberbullying interactions involving multiple participants. Research in cyberbullying detection and computational discourse analysis emphasises understanding interactions not only through content but also through participant roles and relational context (Van Hee et al., 2018; Ziems et al., 2020; Emmery et al., 2021; Kim et al., 2021a). Building on work in both domains of cyberbullying research like (Hinduja and Patchin, 2013; Salmivalli, 2014; Englander et al., 2017; Huang et al., 2020; Ziems et al., 2020; Steer et al., 2020; Van Hee et al., 2018) and harmful discourse analysis like (Kumar et al., 2024, 2022; Ghosh et al., 2024), we opertionalise the cyberbullying social dynamics through ***six sub-dimensions***:

- **Conflict $D_{\textbf{CB}}^{\textbf{Conflict}}$:** Hostile exchanges, personal attacks, or ongoing antagonism, serving as early indicators of aggression (Van Hee et al., 2018; Salmivalli, 2014).
- **Support $D_{\textbf{CB}}^{\textbf{Support}}$:** Expressions of empathy, defense, or solidarity, as well as escalation through encouragement of harmful behaviours (Salmivalli, 2014; Englander et al., 2017; Van Hee et al., 2018).
- **Influence $D_{\textbf{CB}}^{\textbf{Influence}}$:** Covert or manipulative behaviours such as shaming, gaslighting, or instigating aggression, shaping the trajectory of group interactions (Kumar et al., 2022; Ghosh et al., 2024).
- **Humour $D_{\textbf{CB}}^{\textbf{Humour}}$:** Use of sarcasm, ridicule, or mockery that can disguise or amplify harm (Steer et al., 2020; Ziems et al., 2020).

- **Power Dynamics $D_{\mathbf{CB}}^{\mathbf{Power}}$:** Signs of dominance, marginalisation, or social hierarchy enforcement, reflecting the social structure of interactions (Ziems et al., 2020).
- **Repetition $D_{\mathbf{CB}}^{\mathbf{Repetition}}$:** Recurring harm or persistent targeting across multiple turns, distinguishing systematic bullying from one-off incivility (Hinduja and Patchin, 2013; Salmivalli, 2014; Steer et al., 2020).

By integrating *realism* D_{Real}, *relevance* D_{Relevant}, *conversational coherence* D_{Coherent}, and the six sub-dimensions of *cyberbullying social dynamics*, this **intrinsic evaluation framework** provides a comprehensive lens for analysing both human- and AI-generated conversational data. This approach moves beyond surface-level content analysis, capturing the nuanced interplay of behavioural characteristics, social context, and interpersonal dynamics to ensure ecological validity and practical relevance for cyberbullying research and AI applications.

5.3.2 Study Setup

Recruitment and Ethical Considerations Given the inherently contextual and subjective nature of the evaluation dimensions, intrinsic human evaluation is subject to interpretive variation. To mitigate bias and capture diverse perspectives, the evaluation was conducted in two phases: a pilot study and a main study. The pilot involved three academic experts with established expertise in cyberbullying and youth behaviour. Building on insights from the pilot, the main study expanded the participant pool to include two professional content moderators, two academic youth experts, and one young adult, thereby ensuring a broader range of evaluative viewpoints. Given the potentially offensive nature of the content evaluated in this study, all procedures were conducted in accordance with university ethics approval and established ethical protocols. Key considerations included:

- **Participant Eligibility:** All participants were aged 18 or older.
- **Informed Consent and Blinding:** Evaluators were fully informed about the study's aims and procedures and provided informed consent prior to participation. However, they were not informed of the source of individual conversations, nor whether they were real, role-played, or generated by a language model.

- **Training and Preparedness:** Prior to evaluation, each evaluator completed an online training module on resilience building and handling sensitive content, specifically curated for this doctoral project by a qualified psychologist. The psychologist was compensated for this contribution.

- **Support and Voluntary Participation:** Optional counselling and referral support were made available throughout the evaluation period. Evaluators retained the right to withdraw at any time without penalty, ensuring voluntary and informed participation.

- **Compensation and Recognition:** Academic experts were not financially compensated but were offered co-authorship in recognition of their contributions. The adult teen evaluator and professional content moderator were compensated at €13 per hour for a maximum of 5-6 hours of evaluation. Evaluators could choose to participate remotely or on-site at the host institution, with supportive infrastructure provided for either modality.

Pilot Study: For the pilot study, one conversation each was selected from four datasets: (i) youth co-created dataset §4.3, (ii) human translation of (Sprugnoli et al., 2018) dataset (§5.1.1), (iii) LLM-generated Phase-2: Llama3.3 conversational dataset (§5.1.4), and (iv) one image+caption+comment post labelled as bullying by (Hosseinmardi et al., 2015). Three cademic experts in adolescent online harms(2M: 1F; ages 35-42) were tasked with rating one conversation each from the four datasets across dimensions, realism (D_{Real}), relevance (D_{Relevant}), conversational coherence (D_{Coherent}), cyberbullying social dynamics sub-dimensions like conflict ($D_{\text{CB}}^{\text{Conflict}}$), support ($D_{\text{CB}}^{\text{Support}}$), influence ($D_{\text{CB}}^{\text{Influence}}$), humour ($D_{\text{CB}}^{\text{Humour}}$), power-dynamics ($D_{\text{CB}}^{\text{Power}}$), repetition ($D_{\text{CB}}^{\text{Repetition}}$) on a 5-Likert scale (*strongly agree* $\leftrightarrow$ *not at all*). This evaluation was designed and deployed using a Qualtrics survey. For each item, participants were presented with the corresponding context. In the case of youth co-created, human-translated, and LLM-generated datasets, this context took the form of a short scenario describing the interaction. For the Instagram dataset, the original post was shown together with the accompanying caption and user comment.

The pilot study was conducted in a hands-on manner. Each rater was consulted individually to discuss the evaluation structure, specifically whether conversations should

be rated at the level of individual messages/turns or as complete dialogues. The initial trials indicated that rating each message or conversational turn by role was impractical and risked producing inconsistent judgments. Consequently, the framework was revised for the subsequent phase: evaluators assessed conversations as a whole rather than at the message level. In this revised approach, each full conversation was presented in a Google Sheets format, styled to include the scenario or vignette alongside the conversation text. Conversations displayed usernames but did not include role labels (e.g., bully, victim, defender), ensuring that pre-defined categories did not influence evaluators' judgements. No pre-labelled data were provided. A continuous feedback loop with evaluators further refined the evaluation protocol. Based on their input, the guidelines were expanded to incorporate illustrative examples for each Likert-scale option, as well as positive and negative examples for each evaluative dimension. This iterative process ensured that the framework was both practical for annotators and robust enough to capture the nuances of harmful and supportive interactions across datasets.

Main Study: Based on insights from the pilot study, the response options for the six sub-dimensions of "*cyberbullying social dynamics*" were simplified to a ternary schema (*Yes/No/Neutral*). This refinement served **two** purposes: (i) to reduce the cognitive load on evaluators by avoiding overly fine-grained distinctions; and (ii) to increase inter-rater reliability by ensuring that judgments were made on a consistent and interpretable scale. To further support this, each response option was supplemented with annotated examples drawn from synthetic and real conversational data. These examples illustrated typical instances of "*Yes*," "*No*," and "*Neutral*" judgments, clarifying the boundaries of each category and minimising ambiguity in interpretation (See Table 5.20). The development of annotated examples was itself an iterative process, incorporating feedback from academic experts. Particular care was taken to include both positive and negative cases for each dimension, ensuring that evaluators could distinguish subtle variations (e.g., when humour functioned as benign banter versus when it served as a vehicle for exclusion or ridicule). By grounding the response schema in concrete illustrations, the evaluation framework aimed to balance methodological rigour with practical usability. The finalised schema and accompanying annotated guide formed the core of the evaluation protocol and were presented to all participants before the main study. Illustrated in Table 5.18, the objectives

and brief instructions were clearly outlined along with the core-dimensions i.e., D_{Real}, D_{Relevant}, D_{Coherent} and six sub-dimensions of $D_{\text{CB-Dynamics}}$ outlined in Tables 5.19 and 5.20.

Procedure: The conversations used in this assessment were organised into seven "*bags*", each containing four conversations, for a total of 28 conversations. This included,

1. **One** Italian→English translated conversation from $\mathcal{D}^{HT}_{IT\rightarrow EN}$ (§5.1.1), specifically *Scenario C*, which was also used in the translation quality error analysis described in §5.1.1.

2. **Four** French→English conversations translated using human-aided machine translation approach $\mathcal{D}^{HMT}_{FR\rightarrow EN}$ (§5.1.3). These included scenarios on *obesity* (`scenario-obesity-brice-bis`), *homophobia* (`scenario-homophobia-paul`), *religion* (`scenario-religion-fatimah`), and *ethnicity* (`scenario-ethnicity-Sylvia-Adrien`). These specific conversations were selected partly because they contain fewer than 100 messages each, making them more suitable for comparative evaluation.

3. **Seven** real-world Instagram image+caption+comment threads drawn from $\mathcal{D}_{IG}$ (Hosseinmardi et al., 2015), each containing between 50 and 100 messages. The selected dataset IDs were: `649719881`, `649719912`, `652903353`, `652903358`, `652903394`, `652910761`, and `652910812`.

4. **Eight** youth co-created conversations from §4.3, corresponding to scenario IDs `Lit-1`, `Lit-2`, `Lit-5`, `Lit-7`, `Lit-8`, `Lit-10`, `Lit-12`, and `Lit-13`.

5. **Eight** LLM-generated conversations $\mathcal{D}^{Phase-2}_{LLM}$, produced for the same scenarios as the youth co-created set, to allow for direct comparison.

The conversations were then grouped into **seven "bags"**, with each bag containing four conversations. Within every bag, exactly one conversation was drawn from the youth co-created set and one from the LLM-generated set, but for different scenarios, while the remaining two were selected randomly from the available pool. The assignment of conversations to bags was randomised to avoid systematic bias in distribution. However, for consistency, all evaluators reviewed the bags in the same order. The evaluation was administered through the Qualtrics survey platform. Each "bag" of conversations was provided

Objective	The task involves reading different conversations across different scenarios. For each one, you will evaluate conversations based on four core dimensions: Realism, Relevance, Social Dynamics, and Clarity. These evaluations help assess how natural, scenario-appropriate, and socially nuanced are the conversations.
Brief Instructions	(1) Each annotation task includes a scenario and a corresponding conversation. (2) Conversations typically contain 50–75 messages. (3) Read the full conversation carefully before proceeding. (4) Read the below definitions of dimensions and examples. (5) Remember the examples are not rules but a list of examples. (6) Conversations can be long, you may pause and resume whenever needed.
Core Dimensions	**Realism:** You will be responding to the question - ***To what extent is this conversation is realistic or not realistic?*** *By realistic we mean "a real exchange that could actually happen on social media". Rating: Use a 5-point Likert scale (1 = Not at all, 5 = Very much).* Provide a short justification for your rating. *Note: Realism reflects how natural the conversation feels in the context of social media communication. If the conversation includes informal language, casual tones, or common topics found on social media platforms, the rating could be higher. A low rating might indicate that the conversation feels stiff or overly formal or completely unrealistic in terms of timing and tone for social media interactions.*
	Relevance: You will be responding to the question - ***To what extent is this conversation relevant or irrelevant to the scenario?*** *By relevant we mean "the conversation addresses the situations in the scenario". Rating: Use a 5-point Likert scale (1 = Not at all, 5 = Very much).* Provide a short justification for your rating. *Note: Relevance assesses how well the conversation aligns with the intended context or scenario. If the conversation addresses the situation at hand, staying on topic, or considers all key points of the scenario, their rating could be higher. A low rating indicates that the conversation deviates from the scenario, going off-topic or missing any or all key points of the scenario.*
	Social Dynamics: You will be responding to the question - ***To what extent does the conversation reflect or engage with the following sub-dimensions,*** Refer to detailed definitions in Table 5.19 and detailed examples with explanation in table 5.20.
	Conversational Coherence: You will be responding to the question - ***Considering the flow of conversation on social media as per your experience, to what extent is this conversation understandable and easy to follow?*** *Rating: 5-point Likert scale.* Provide a short justification for your rating. **Note:** The coherence rating assesses how easy it is to follow the conversation thread in a social media comment section or group chat, where responses may not always follow a strict logical order. A higher rating would mean the conversation remains coherent despite the potential disjointedness, with responses that are easy to connect and understand. A lower rating suggests the conversation is hard to follow, with unclear or off-topic responses that make it difficult for users to grasp the main points or context of the discussion.

Table 5.18: Guideline for intrinsic human evaluation with definitions of core-dimensions

Conflict	Hostile exchanges, personal attacks, or ongoing antagonism, including but not limited to arguments, disagreements, opposing viewpoints.
Support	Expressions of empathy, defense of victims, or efforts to escalate the situation as enablers or de-escalate.
Influence	Attempts to shame, manipulate, or assert control over others.
Humour	Use of ridicule, mockery or sarcasm that may contribute to or challenge harmful behaviour.
Power Dynamics	Indications of dominance, targeting vulnerable individuals or group pressure.
Repetition	Recurrence of harmful behaviour, targeting, or negative interactions over time.
Response Choices	• Yes – Noticeably Present: Select this when you observe the behaviour more than once in a clear and noticeable way. • No – Noticeably Absent: Select this when you do not observe the behaviour at all. • Neutral: Select this when the behaviour is seen only once or less, and not in a strong or consistent manner.

Table 5.19: Guideline for intrinsic human evaluation with definitions of cyberbullying social dynamics with response choices. For detailed examples refer Table 5.20.

as a Google Spreadsheet, with each conversation presented on a separate sheet within the file. Evaluators accessed the survey under their assigned pseudonyms. At the beginning of the survey, evaluators were asked to complete a brief demographic questionnaire. This included questions on: (a) age, (b) gender, (c) ethnicity, (d) education level, (e) social media usage (as in §4.1 Chapter 4), and (f) whether they had witnessed cyberbullying in the past year (also as in §4.1 Chapter 4). These demographic data were collected solely for the purpose of aggregating responses at a high level. To ensure privacy and anonymity, the file linking evaluator identities to pseudonyms was permanently deleted at the conclusion of the evaluation task.

Evaluation Metrics: A total of nine dimensions were evaluated across twenty-eight conversations, yielding 252 individual ratings. This included 72 items each from eight conversations drawn from both youth co-created and LLM-generated datasets, 63 items from seven Instagram conversations, 9 items from a single human-translation conversation (covering all nine dimensions), and 36 items from four human-in-the-loop translation sets.

Five raters participated in the main study, comprising two social scientists, two content moderators, and one 19-year-old female. In addition to descriptive statistics on agreement and disagreement, inter-annotator agreement is measured using two complementary metrics: Fleiss κ Fleiss (1971), and prevalence-adjusted and bias-adjusted kappa for ordinal scales (PABAK-OS) Parker et al. (2011) for pair-wise agreement. As discussed in Section 2.3.2 Chapter 2, Fleiss' κ is widely used to quantify the degree of agreement beyond chance. Following Landis and Koch (1977) and subsequent guidelines, values of $\kappa < 0.4$ are typically interpreted as indicating poor-to-fair agreement. However, κ is known to be sensitive to prevalence and bias effects, i.e., situations where one category dominates (e.g., highly imbalanced label distributions) or annotators systematically lean toward certain categories. In such cases, κ can yield deceptively low or even counter-intuitive values despite substantial agreement among annotators. To address this limitation, PABAK-OS, which adjusts for both prevalence and bias while accounting for the ordinal nature of our labelling scheme, is leveraged. This measure provides a more stable and interpretable estimate of agreement under skewed conditions Anzovino et al. (2018). By reporting both Fleiss κ and PABAK-OS, we attempt to capture the strengths of a conventional measure of chance-corrected agreement while also offering a robustness check against distortions caused by uneven class distributions.

Category	Positive Example (Valid Instance)	Negative Example (Non-category Instance)	Explanation
Conflict	Alex: "Your opinion is completely wrong and you have no idea what you're talking about!" Alex: "I disagree, your view makes no sense at all."	Alex: "Good morning everyone, how are you today?"	Shows direct disagreement and argumentative tone — hallmarks of conflict.
Support	Case 1: "Penny, don't listen to those insults—you're amazing, and I'm here for you." Case 2: "You're right, let's show him why he needs to be back in the closet"	Riley: "I don't really care about this discussion."	Expresses empathy or backs another person's position, including escalating or de-escalating. Indifference, as in the negative case, is not support.
Influence	Case 1: Sam: "Listen 2 me, you dumbwit - the whale needs thrashing" Case 2: Tim: "Lol, srsly guys, i thought we were better than this"	Tim: "I have my own thoughts, whatever."	Shows an attempt to persuade or lead others. The neutral comment shows no intent to guide or control others.
Humour	Drew: "Oh wow, your argument was so strong, if bullshit were currency, you'd be f**king loaded."	Drew: "I am very serious about this debate."	Includes sarcasm or wit that lightens or intensifies interactions. A serious tone lacks humorous intent.
Power Dynamics	Jordan: "You all need to listen to me and not this WHALE!"	Jordan: "Let's all share our opinions equally."	Demonstrates dominance or hierarchy assertion. The negative case promotes balance and equal footing.
Repetition	Casey: "Again, you just don't get it, do you? RETARD"	Casey: "I mentioned my point once, and I believe it was clear."	Highlights repeated negative behavior or rhetoric. A one-time statement doesn't qualify as repetition.

Table 5.20: Cyberbullying Social Dynamics Dimension Examples with Positive and Negative Instances

5.3.3 Results

Across the dataset, all five raters reached complete agreement on 157 items (62.3%), while at least three raters agreed on 245 items (97.2%). There were no cases of complete disagreement among all five raters, although seven items showed disagreement among at least three raters. At group-level agreement, the two social scientists agreed on 179 items (71%), the content moderators on 203 items (80.5%), the social scientists and the teen on 175 items (69.4%), and the content moderators and the teen on 192 items (76.2%). Notably, there were no cases in which all five raters disagreed, although seven items did show disagreement among at least three raters. As illustrated in Table 5.21, across the core dimension of *realism*, complete agreement was achieved for 85.71% of Instagram conversations (6/7), while the youth co-created set showed 37.50% (3/8), French→English 25.00% (1/4), and LLM-generated 12.50% (1/8). The Italian→English translation yielded no full agreement (0/1). For *relevance*, the youth dataset had the highest proportion of agreement at 87.50% (7/8), followed by Instagram at 57.14% (4/7) and French→English at 50.00% (2/4). The single Italian→English case had complete agreement (1/1), whereas LLM-generated content yielded only 12.50% (1/8). For *coherence*, which captures the flow of conversation, LLM-generated items showed the highest agreement at 62.50% (5/8), followed by youth at 50.00% (4/8), French→English at 25.00% (1/4), and Italian→English at 0% (0/1). Instagram conversations showed no cases of full agreement on coherence (0/7). Turning to the sub-dimensions of cyberbullying social dynamics like, *conflict*, agreement was uniformly high: 100.00% in youth (8/8), LLM-generated (8/8), and Italian→English (1/1), with slightly lower but still strong values for Instagram at 71.43% (5/7) and French→English at 75.00% (3/4). For *support*, all youth items showed agreement (8/8; 100.00%), as did the single Italian→English case (1/1), with strong but slightly lower rates for LLM-generated content at 87.50% (7/8). Instagram showed 57.14% agreement (4/7), and French→English 50.00% (2/4). In the case of *power dynamics*, youth items reached 75.00% agreement (6/8), LLM-generated 50.00% (4/8), Instagram 42.86% (3/7), and French→English 75.00% (3/4), while Italian→English achieved complete agreement (1/1). Agreement on *repetition* was complete in the youth (8/8), Italian→English (1/1), and LLM-generated datasets (5/8; 62.50%), with lower values for Instagram at 57.14% (4/7) and French→English at 50.00% (2/4). For *humour*, the LLM-generated dataset yielded the highest propor-

tion of agreement (6/8; 75.00%), followed by youth at 62.50% (5/8), Instagram at 57.14% (4/7), and French→English at 50.00% (2/4). Again, Italian→English had complete agreement (1/1). Finally, for *influence*, the LLM-generated set achieved perfect agreement (8/8; 100.00%), while youth had 75.00% (6/8), French→English 50.00% (2/4), Instagram 28.57% (2/7), and Italian→English complete agreement (1/1). Overall, raw agreement patterns varied considerably across dimensions and datasets. Socially nuanced categories such as *coherence*, *power dynamics*, and *humour* yielded the lowest consistency, while more concrete categories such as *conflict*, *support*, and *influence* tended to show high or perfect agreement across multiple datasets.

Dimension	Youth Co-created §4.3	$\mathcal{D}_{\text{LLM}}^{\text{Phase-2}}$ (§5.1.4)	$\mathcal{D}_{IG}$	$\mathcal{D}_{IT\rightarrow EN}^{HT}$ §5.1.1	$\mathcal{D}_{FR\rightarrow EN}^{HMT}$ §5.1.3
D_{Real}	37.50	12.50	85.71	0.00	25.00
D_{Relevant}	87.50	12.50	57.14	100.00	50.00
D_{Coherent}	50.00	62.50	0.00	0.00	25.00
$D_{\text{CB}}^{\text{Conflict}}$	100.00	100.00	71.43	100.00	75.00
$D_{\text{CB}}^{\text{Support}}$	100.00	87.50	57.14	100.00	50.00
$D_{\text{CB}}^{\text{Power}}$	75.00	50.00	42.86	100.00	75.00
$D_{\text{CB}}^{\text{Repetition}}$	100.00	62.50	57.14	100.00	50.00
$D_{\text{CB}}^{\text{Humour}}$	62.50	75.00	57.14	100.00	50.00
$D_{\text{CB}}^{\text{Influence}}$	75.00	100.00	28.57	100.00	50.00

Table 5.21: Raw agreement percentages between five evaluators for all nine dimensions across the five datasets.

Table 5.22 summarises the inter-annotator agreement (IAA) across all five annotators and all data sources combined. Overall agreement was substantial, with a Fleiss' κ of 0.68 and a corresponding PABAK-OS of 0.94. This indicates strong consensus among annotators once prevalence and bias effects are adjusted. At the level of individual dimensions, κ scores varied considerably, ranging from fair agreement (e.g., 0.17 for *Conflict*) to moderate-to-substantial levels (e.g., 0.63 for *Support*). In contrast, PABAK-OS values were consistently high across all dimensions (0.82–0.92), suggesting that much of the apparent disagreement in κ is due to skewed ratings rather than genuine inconsistency. This pattern was particularly pronounced in dimensions such as *Conflict*, *Power Dynamics*, *Repetition*, and *Coherence*, where κ values fell below 0.4, yet PABAK-OS values remained above 0.83. These discrepancies highlight the value of reporting both metrics: Fleiss' κ reflects conservative chance-corrected agreement, whereas PABAK-OS provides a more

stable estimate under skewed conditions.

Dimension	Fleiss Kappa	PABAK-OS	Note
Overall	**0.68**	**0.94**	
$D_{\textbf{Real}}$	0.46	0.87	
$D_{\textbf{Relevant}}$	0.41	0.91	
$D_{\textbf{CB}}^{\textbf{Conflict}}$	0.17	0.89	Low Kappa – PABAK recommended
$D_{\textbf{CB}}^{\textbf{Support}}$	0.63	0.91	
$D_{\textbf{CB}}^{\textbf{Influence}}$	0.48	0.87	
$D_{\textbf{CB}}^{\textbf{Humour}}$	0.54	0.82	
$D_{\textbf{CB}}^{\textbf{Power}}$	0.23	0.83	Low Kappa – PABAK recommended
$D_{\textbf{CB}}^{\textbf{Repetition}}$	0.38	0.92	Low Kappa – PABAK recommended
$D_{\textbf{Coherent}}$	0.32	0.88	Low Kappa – PABAK recommended

Table 5.22: Inter-rater agreement scores for all data sources combined

Table 5.23 reports agreement split by data source (youth co-created, LLM-generated, French→English, Italian→English, and Instagram). Agreement varied considerably across sources. The highest overall agreement was observed in the youth co-created dataset (κ = 0.79, PABAK-OS = 0.98), followed by Italian→English (κ = 0.74, PABAK-OS = 0.92). By contrast, LLM-generated and Instagram data yielded somewhat lower κ scores (0.65 and 0.64, respectively), despite strong PABAK-OS scores (0.92 and 0.91). The French→English data presented the lowest Fleiss' κ overall (0.62), though the PABAK-OS remained high (0.95). The most striking discrepancies appeared within specific dimensions. For instance, *Conflict* and *Repetition* reached perfect agreement (κ = 1.00, PABAK-OS = 1.00) in several datasets, yet dropped sharply in others (e.g., *Conflict* in Instagram: κ = 0.09, PABAK-OS = 0.71). Similarly, *Coherence* and *Realism* occasionally yielded negative κ values (e.g., $\kappa = -0.25$ for *Coherence* in Italian→English), yet PABAK-OS suggested moderate agreement (0.50–0.74). These results again emphasize the limitations of κ under imbalanced or sparse label distributions and the interpretive robustness of

PABAK-OS.

Table 5.24 presents the agreement within the social scientist subgroup (two annotators), both with and without the inclusion of the adult teen annotator. Without the teen, overall κ was 0.54, rising to 0.65 when the teen was included. Corresponding PABAK-OS values were consistently high (0.90 and 0.94). This pattern held across most dimensions: inclusion of the teen improved κ scores for *Coherence*, *Humour*, *Influence*, and *Power Dynamics*, while slightly reducing agreement on *Support*. Importantly, dimensions that appeared unreliable under κ alone (e.g., *Conflict*, $\kappa = -0.06$ without teen) nonetheless yielded moderate-to-high PABAK-OS values (0.79–0.86), reinforcing the interpretive stability of the adjusted metric. Table 5.25 reports analogous results for the content moderator subgroup. Agreement within this group was generally higher than among the social scientists, with overall κ values of 0.68 (group only) and 0.73 (group plus teen). PABAK-OS values were again consistently high (0.94–0.95). Compared to the social scientists, the moderators displayed stronger reliability on dimensions such as *Realism* and *Relevance*, with κ values above 0.5. Inclusion of the teen annotator modestly improved κ in several categories (*Flow*, *Humour*, *Support*), while leaving PABAK-OS scores largely unchanged. Interestingly, *Power Dynamics* remained challenging across both subgroups, with κ values hovering near chance (0.04-0.11), despite relatively high PABAK-OS values (0.85–0.89). This indicates systematic difficulties in consistently identifying power relations, potentially reflecting genuine subjectivity in this dimension.

The observed differences between social scientists and content moderators likely reflect the distinct expertise and evaluative perspectives of the groups. Social scientists, trained to identify subtle socio-cultural cues, exhibited more variability on subjective dimensions (*Coherence*, *Humour*, *Power Dynamics*) but showed noticeable improvements when the teen annotator was included, suggesting that the teen provided complementary insight, particularly for content reflecting lived adolescent experiences. In contrast, content moderators achieved higher baseline agreement across most dimensions, potentially due to their operational familiarity with platform policies and content categorisation frameworks, which may reduce subjective interpretation. The inclusion of the teen annotator generally enhanced Fleiss' κ, especially for social scientists, while PABAK-OS remained largely stable across subgroups, highlighting that the teen helped reduce genuine disagreement rather

Dimension	Youth Co-created		LLM-generated		$\mathcal{D}^{HMT}_{FR\rightarrow EN}$		$\mathcal{D}^{HT}_{IT\rightarrow EN}$		$\mathcal{D}_{IG}$	
	κ	PABAK-OS	κ	PABAK-OS	κ	PABAK-OS	κ	PABAK-OS	κ	PABAK-OS
Overall	0.79	0.98	0.65	0.92	0.62	0.95	0.74	0.92	0.64	0.91
conflict	1.00	1.00	1.00	1.00	0.17	0.70	1.00	1.00	0.09	0.71
flow	0.44	0.55	0.60	0.60	-0.05	0.74	-0.25	0.50	0.03	0.78
humour	0.37	0.85	0.56	0.84	0.49	0.75	1.00	1.00	0.57	0.77
influence	-0.04	0.88	1.00	1.00	0.47	0.60	1.00	1.00	0.23	0.69
power_dynamics	0.17	0.70	0.02	0.50	0.69	0.80	1.00	1.00	0.14	0.59
realism	0.31	0.79	0.10	0.69	-0.11	0.87	-0.25	0.20	0.85	0.89
relevance	-0.03	0.90	0.18	0.80	0.30	0.82	1.00	1.00	0.65	0.83
repetition	1.00	1.00	-0.08	0.70	0.47	0.60	1.00	1.00	0.42	0.83
support	1.00	1.00	-0.03	0.90	0.22	0.50	1.00	1.00	0.63	0.76

Table 5.23: IAA Fliess Kappa and PABAKK-OS across all dimensions **split by data-sources**

Dimension	Group κ	Group+Teen κ	Group PABAK-OS	Group+Teen PABAK-OS
Overall	0.54	0.65	0.90	0.94
conflict	-0.06	-0.04	0.79	0.86
flow	-0.02	0.24	0.62	0.78
humour	0.44	0.55	0.61	0.77
influence	0.22	0.47	0.82	0.88
power_dynamics	0.07	0.33	0.70	0.82
realism	0.14	0.36	0.74	0.85
relevance	0.11	0.35	0.83	0.89
repetition	0.27	0.46	0.88	0.92
support	0.72	0.56	0.96	0.90

Table 5.24: IAA Fliess Kappa and PABAKK-OS across all dimensions across **Social Scientist Group** and **Social Scientist Group with Adult Teen**

Dimension	Group κ	Group+Teen κ	Group PABAK-OS	Group+Teen PABAK-OS
Overall	0.68	0.73	0.94	0.95
conflict	-0.04	-0.02	0.86	0.90
flow	0.24	0.37	0.87	0.90
humour	0.44	0.55	0.84	0.89
influence	0.43	0.56	0.82	0.87
power_dynamics	0.04	0.11	0.89	0.85
realism	0.54	0.58	0.83	0.89
relevance	0.37	0.56	0.92	0.94
repetition	0.27	0.46	0.93	0.95
support	0.54	0.69	0.88	0.92

Table 5.25: IAA Fliess Kappa and PABAKK-OS across all dimensions across **Content Moderator Group** and **Content Moderator Group with Adult Teen**

than correcting prevalence or bias effects. These patterns underscore the importance of diverse annotator perspectives for capturing nuanced social phenomena, particularly in dimensions where consensus is inherently difficult, such as *Power Dynamics* and *Conflict*.

Of these seven disagreed items by at least three evaluators, two were related to D_{Coherent} for Instagram conversations (IDs `652910812` and `649719881`). The ID: `652910812` showcased a mirror selfie image of an African American woman in gym clothes with 41 comments but without any captions. For this post-comment thread, one social scientist and one content moderator rated it as `very easy to follow`, another social scientist and content moderator rated it as `somewhat not easy to follow`, while the teen rated it as `somewhat easy to follow`. The ID: `649719881` showcased a selfie image of an African American man in a car with jewellery, this had a caption "*When you past judgement on someone else you put urself in harms way*" and was accompanied by 102 comments. For this item, the adult teen and one social scientist rated it as `somewhat not easy to follow`, both content moderators rated it as `somewhat easy to follow`, and the other social scientist rated it as `very easy to follow`. Across both items, those who judged it `somewhat not easy to follow` attributed their ratings to the fragmented and chaotic nature of the comment thread. In contrast, the `somewhat easy` and `very easy` ratings were justified by noting that, despite the disorder resembling real-life online exchanges, the brevity of the messages made the conversation relatively easy to follow. Another disagreed item concerned $D_{\text{CB}}^{\text{Power}}$ for an Instagram conversation (ID: `652903353`), which shows an American Football athlete posing with 20 comments. For this case, a content moderator and the adult teen rated it as *neutral*, while a social scientist and another content moderator rated it as being *absent*; however, one social scientist rated it as *present*. Raters who selected *neutral* explained their choice by noting the limited number of comments and the tendency of fans to criticise the athlete. Those who rated it as *absent* emphasised that the comment-thread format made it difficult to interpret power dynamics, whereas the single rater who judged it as *present* did not provide a detailed explanation. Within the disagreed items, two concerned D_{Real} for LLM-generated conversations (`Lit-7`, `Lit-8`). The `Lit-7` case involved denigration, where a girl's photos were circulated via an account/page mocking her clothing, while `Lit-8` involved outing, where someone's sexual orientation was mocked. For Phase-2 Llama3.3 generated conversation `Lit-7`, one social

scientist and one content moderator rated it as `not realistic at all`, another social scientist rated it as `somewhat realistic`, another content moderator rated it as `very realistic`, and the adult teen rated it as `somewhat unrealistic`. In contrast, for `Lit-8`, both social scientists and both content moderators rated it as `very realistic`, while the teen judged it as `somewhat realistic`. Across both items, raters who selected `not realistic at all`, noted it is due to the use of emojis and style of language is not real or it feels "formulaic and forced", and those who selected `somewhat unrealistic` noted it is due it being "scripted or stiff" or using too many emojis. In contrast those selected `very realistic` for these items, rated it based on the scenario as they noted people are homophobic or "*new kids, especially alt ones, are likely to be picked on for being 'vulnerable' and 'different*" rather than the conversation style.

Notably, one item concerned $D_{\text{CB}}^{\text{Humour}}$ for the same LLM-generated `Lit-7` conversation. In this case, the adult teen rated it as *neutral* (neither present nor absent), one content moderator and one social scientist rated it as *present*, and the remaining two raters judged it as *absent*. This divergence reflects uncertainty over whether the mocking tone of the exchange constituted humour: some raters interpreted the ridicule as an attempt at humour, while others viewed it solely as denigration without humorous intent. Another item pertained to D_{Relevant} for an LLM-generated conversation, specifically `Lit-12` (harassment+denigration), where someone was publicly humiliated and bullied via stories and group conversations. For this case, the adult teen and one content moderator rated it as *neither relevant nor irrelevant*, one social scientist and one content moderator rated it as *very relevant*, and one social scientist rated it as *somewhat irrelevant*. Raters who selected *neither relevant nor irrelevant* or *somewhat irrelevant* explained that the conversation did not capture the element of public humiliation through story features. In contrast, those who rated it as *very relevant* emphasised that bullying within group conversations was accurately reflected in the generated dialogue. This variation suggests disagreement on whether the generated conversation adequately captured the central theme of harassment and denigration: while some raters judged the content as directly illustrative of the construct, others felt the representation was peripheral or insufficiently aligned with the intended dimension.

5.3.4 Conclusion

The **intrinsic** evaluation by different stakeholders across nine dimensions and multiple data sources revealed both robust inter-annotator reliability and important sources of variation that reflect the complexity of cyberbullying-related conversational dynamics. Overall agreement was substantial, with Fleiss' κ indicating moderate-to-strong reliability ($\kappa = 0.68$) and PABAK-OS confirming consistently high levels of consensus once prevalence and bias effects were controlled for (0.94). Indicating that despite apparent variability across certain dimensions, annotators generally converged in their interpretations of conversation features.

Patterns of agreement were not uniform across datasets or dimensions. More concrete categories such as *conflict*, *support*, and *influence* consistently yielded high or perfect agreement, suggesting that raters were confident in identifying direct indicators of these dynamics. In contrast, socially nuanced and interpretively subjective categories such as *coherence*, *power dynamics*, and *humour* produced lower κ values. Although PABAK-OS values suggests that much of this variability stemmed from imbalanced distributions rather than fundamental disagreement.

Variation across data sources further underscores the differential interpretability of conversation types. Youth co-created datasets achieved the highest agreement, particularly when evaluated by all stakeholders. In contrast, LLM-generated and Instagram conversations yielded more variable agreement, especially on dimensions such as *realism* and *relevance*, where raters diverged on whether content authentically represented online interactions or specific bullying scenarios. The French→English translations showed comparatively lower reliability, suggesting that cultural and linguistic mediation may obscure conversational intent and social dynamics.

The subgroup analyses revealed systematic differences between annotator backgrounds. Content moderators demonstrated higher baseline agreement across most dimensions, likely reflecting their operational expertise and reliance on structured content evaluation frameworks. Social scientists, in contrast, exhibited greater variability on subjective dimensions but showed notable improvement when the teen annotator was included, highlighting the value of youth perspectives in evaluating realism, humour, and power dynamics in digital interactions. The teen annotator's input consistently increased κ values with-

out altering PABAK-OS, suggesting that their contribution reduced genuine disagreement rather than merely balancing prevalence effects. Also, item-level disagreement analysis illustrates the interpretive challenges of evaluating realism, humour, and power dynamics in complex or ambiguous contexts. Differing judgments often reflected conflicting criteria, such as whether realism should be anchored in conversational style or scenario plausibility, or whether ridicule should be coded as humour or as denigration. These disagreements emphasise the inherently subjective nature of some dimensions and the need for clearer coding guidelines or more nuanced categorical frameworks to capture overlapping phenomena.

5.4 Young Adult Annotation and Results

Building on Section 5.3, where **intrinsic** evaluation was described across three key dimensions and six sub-dimensions using a diverse group of participants, the youth co-created dataset demonstrated the highest overall reliability and inter-annotator agreement ($\kappa = 0.79$; Table 5.23). The dataset performs particularly well in socially nuanced and operationally critical dimensions, including Conflict, Support, Influence, Repetition, and Power Dynamics. While real-life Instagram conversations sometimes achieve higher scores for dimensions such as realism and humour, the youth co-created data consistently outperforms most human translations and LLM-generated content. Overall, the youth co-created dataset effectively captures the dynamics of adolescent online interactions, particularly in social and operational aspects. To establish the youth co-created dataset as $\mathcal{D}_{\mathbf{Gold}}$, it is crucial that each message within a conversation is annotated with high-quality, contextually grounded labels. To that effect, in this section we describe the motivation, procedure, and results of the human annotation process.

5.4.1 Objectives and Tasks

As discussed in Section 2.3.3 Chapter 2, Van Hee et al. (2018) propose a detailed taxonomy of cyberbullying that includes `threats`, `insults`, `discrimination`, `exclusion`, `defamation`, `sexual harassment`, `sarcasm`, and `defense` strategies. Their framework captures both **direct** attacks (e.g., *threats, blackmail, personal insults*) and **indirect** attacks such as *exclusion* or *defamation*, while also accounting for defensive and bystander

responses. (Sprugnoli et al., 2018) extend this work by adding `body-shaming` as a specific form of appearance-based insult. Building on policy documents and civil-society guidelines, Banko et al. (2020) synthesises categories such as identity attack, identity misrepresentation, insult, sexual aggression, and violent threats, though their claim that "*toxic*" or "*bullying*" behaviour merely blends "*hate*" and "*harassment*" would benefit from deeper nuance. (Kumar et al., 2022, 2024; Ollagnier, 2024) shift towards categorising harmful speech by `aggression type` and `communicative intent`. In addition to distinguishing between *overt* and *covert* aggression, they incorporate conversational functions such as *attack, defend, counter-speech, gaslighting, abet/instigate*. This framework was later adapted for French teen role-played dialogue by (Ollagnier, 2024). Building on these categorisations by (Van Hee et al., 2018; Sprugnoli et al., 2018; Banko et al., 2020; Urbaniak et al., 2022; Kumar et al., 2022, 2024; Ollagnier, 2024), the human annotation process aims to capture the multiple dimensions of **cyberbullying** to reflect both *explicit* and *subtle* adolescent interactions. This framework refines the youth co-created dataset (§4.3, Chapter 4) into a gold standard through **three** sub-tasks: (a) *is the message aggressive, and if so, overt or covert?*, (b) *what is the nature of the message?*, and (c) *what is the intent of the message?* This structured approach ensures comprehensive annotation of both overt and nuanced aspects of online adolescent communication. We define them as,

- **Sub-Task A**: **Identify Aggression Type**

  Building on (Kumar et al., 2024; Ollagnier, 2024), the dimension of measurement is **level of aggression**. The goal is to classify a message into into three mutually exclusive classes:

  - **Not Aggressive**: Neutral, polite, or constructive communication without personal attacks.
  - **Covertly aggressive**: Indirect or subtle forms of aggression, including sarcasm, passive-aggression, condescension, or backhanded compliments.
  - **Overtly Aggressive**: Direct, explicit hostility such as insults, threats, hate speech, or demeaning comments.

  This dimension allows for the distinction between visible, explicit aggression and

subtle, context-dependent forms of harmful communication, which are often overlooked in automated detection systems.

- **Sub-Task B**: **Characterise Online Harm**

  Building on (Van Hee et al., 2018; Sprugnoli et al., 2018), in this multi-label task the goal is to capture specific forms of harm present in each message. These include,

  1. **Threats of violence**: (Banko et al., 2020) refer this as content expressing a desire for physical harm, death, injury, self-harm, or acts of violence, as well as the glorification of violence. It excludes anecdotal or non-glorified personal accounts of violence, historical descriptions or metaphorical representations. We adapt it as: *Statements indicating a clear intent or credible threat to cause physical or financial harm to individuals or groups, or explicit glorification of violence directed at the target.*
  2. **Blackmail**: Coercive demands involving threats to reveal damaging, sensitive, or private information unless a specific demand is met (Banko et al., 2020).
  3. **Personal attack**: Urbaniak et al. (2022) refer to this as an online abuse directed at an individual (ad hominem), not their argument, often including comparisons to animals or objects or making nasty, unsubstantiated insinuations. We adapt it as: *Direct, hostile, and abusive language targeting an individual. Attacks the person rather than ideas or identity.*
  4. **Identity-based harassment**: In (Van Hee et al., 2018; Sprugnoli et al., 2018), forms of identity-based abuse such as sexism, racism, and related attacks were grouped under the broader category of Insult. For clearer differentiation, we adapt it as: *Harassment or attacks based on group characteristics (e.g., race, gender, religion).*
  5. **Body-Shaming**: We adapt this from (Sprugnoli et al., 2018) as *Criticism or mockery of physical appearance or body traits.*
  6. **Exclusion**: (Van Hee et al., 2018; Sprugnoli et al., 2018; Davidson et al., 2019) define it as intentionally excluding an individual from a conversation or from joining a social group. We adapt this as *Statements clearly meant to isolate, or reject someone from a group or conversation*

7. **Sexual Harassment**: (Van Hee et al., 2018; Sprugnoli et al., 2018) define it as phrases with a sexual meaning that are characterised by their persistent nature and regarded as undesirable, such as unwanted requests to engage in sexual discussions or actions. (Banko et al., 2020) highlight this does not include insults incorporating sexual language, flirtatious remarks, compliments or non-graphic or non-degrading advances. We adapt it as: *Statements that are unwanted sexual comments, advances, or propositions, and are clearly sexual and contextually unwelcoming.*

8. **Encouragement to Self-Harm**: (Van Hee et al., 2018; Sprugnoli et al., 2018; Banko et al., 2020) classify such content under threats of violence. However, in the review of the UK Online Safety Bill, Trengove et al. (2022) recognise encouragement to self-harm as a distinct form of online abuse, often linked to cyberbullying. To reduce ambiguity in labelling expressions such as "`go kill yourself`" or "`go cut yourself`", We adapt this as an independent category, defined as: *Direct incitement or encouragement for an individual to harm themselves, where the suggestion of self-harm is explicit.*

- **Sub-Task C**: **Identify Communicative Intent**

  Following Ollagnier (2024); Kumar et al. (2024), the objective of this task is to capture the motivation or strategies behind each message to understand online social dynamics. In this multi-label task, each message is to be classified into one or more of the following categories *sarcasm, gaslighting, blaming, abet or instigate, attack, defend, or no-intent.* These are defined as,

  (a) **Sarcasm**: Building on (Van Hee et al., 2018; Sprugnoli et al., 2018), *messages that appear to state one thing (often polite or positive) but, due to context or tone, are intended to mean the opposite, typically to mock, ridicule, or criticize indirectly* are referred to as *sarcastic.*

  (b) **Gaslighting**: Drawing from (Kumar et al., 2022, 2024; Ollagnier, 2024), *messages that manipulate someone by distorting facts, denying past events, or questioning their sanity/memory to make them doubt their own perception, judgment, or reality* are referred to as *gaslighting.*

(c) **Blaming**: Drawing from (Kumar et al., 2022, 2024; Ollagnier, 2024), *messages that explicitly or implicitly assign responsibility to someone (or a group) for a negative outcome, problem, harm, or for provoking abusive behaviour* are referred to as *blaming*.

(d) **Abet or Instigate**: Building on (Kumar et al., 2022, 2024; Ollagnier, 2024), *messages that actively encourage, urge, provoke, or incite others to engage in aggression, hostility, harmful actions, or conflict* are referred to as *abetment* or *instigation*.

(e) **Attack**: Drawing from (Kumar et al., 2022, 2024; Ollagnier, 2024), *messages that directly express aggression through insults, name-calling, threats, mockery, or hostile intent* are referred to as *attack*.

(f) **Defend**: *Messages that aim to protect or support someone (or oneself) from criticism, attacks, or blame. These could be aggressive or neutral.*

By integrating these dimensions, the task aims to advance a more holistic understanding of cyberbullying amongst adolescents. That helps detection systems that account not only for explicitly hostile speech but also for subtle, context-dependent, and socially strategic forms of harmful communication.

5.4.2 Ethical Considerations, Pilot Study and Feedback

Ethical safeguards: Given the potentially distressing nature of the messages, this study was carried out under prior university ethics approval and with a comprehensive set of protective measures for annotators. All participants were fully briefed on the study aims and procedures and provided written informed consent before beginning annotation. To mitigate risks of distress, each annotator was required to complete an online training module on resilience-building and handling sensitive content. This training was designed and delivered by a qualified psychologist specifically for this project. Additional support was made available throughout the study, including optional counselling sessions and referral pathways to external wellbeing services. Annotators retained the right to withdraw at any point without penalty, ensuring their autonomy and voluntary participation. To accommodate individual needs, annotators could choose to work remotely or on-site at

the host institution, with equivalent support systems in place. Compensation was set at €13 per hour, capped at 10 hours per annotator. Importantly, payment was guaranteed for time spent, even if participants did not complete the study. To preserve anonymity, payments were distributed via vouchers instead of direct bank transfers, preventing unnecessary disclosure of financial details.

Pilot study design: The pilot study was conducted to test the feasibility, clarity, and reliability of the proposed annotation scheme for categorising cyberbullying. A purposive sampling strategy was used to recruit three annotators with diverse backgrounds in terms of age, education, and levels of exposure to online harm. This diversity was intentional, to capture a range of perspectives and potential interpretive biases, and the annotators for the pilot study included,

- A 35-year-old male with a Doctorate, who reported frequent exposure to harmful or hurtful online content.
- A 28-year-old female with a Master's degree, who reported rare exposure to harmful online content.
- A 46-year-old female with a Bachelor's degree, who reported occasional exposure to harmful online content.

All annotators were affiliated with the research centre and before participation, had an individual debriefing session where the tasks and categories and the platform was given a walk-through.

Annotation setup: Annotations were carried out on the Doccano platform (Nakayama et al., 2018), hosted securely on Amazon Web Services (AWS). Each participant was given a unique login and password to ensure data security and traceability of individual annotations. For the pilot, annotators were asked to independently label 107 randomly selected messages drawn from the youth co-created dataset. The task was structured around three subtasks: (i) ***Aggression type*** labelling (overt, covert, not-aggressive), (ii) ***Online harm type*** labelling (multi-label: *threats of violence*, *blackmail*, *personal attack*, *identity-based harassment*, *body-shaming*, *exclusion*, *sexual harassment*, and *encouragement to self-harm*),

and (iii) ***Intent type*** labelling (multi-label: *sarcasm, gaslighting, blaming, abet or instigate, attack, defend*). During this pilot phase, annotators were not provided with worked examples or counterexamples of categories. This decision was intentional, as the goal was to test how annotators naturally interpreted and applied the definitions. After a week-long annotation period (approx. five hours of individual work each), one-to-one debrief interviews (two hours per annotator) were conducted to capture qualitative feedback.

Pilot study results: As discussed in Section 2.3.2, within annotation studies, both Krippendorff's alpha (α) (Equation 2.5) and Fleiss' kappa (κ) (Equation 2.4) are commonly leveraged in NLP annotation studies as they account for chance agreement. Percent agreement is additionally included as a descriptive baseline, although it does not adjust for chance agreement. Table 5.26 shows results for all three sub-tasks, and it is consistently low across all three subtasks, indicating considerable divergence in interpretations of the categories. Closer inspection revealed substantial variation in patterns of agreement. For aggression type, the three annotators completely disagreed on 16 out of 107 messages, while full agreement occurred on only 30 messages. Notably, the 56-year-old female annotator diverged most frequently from the two younger participants. For online harm type, agreement across all annotators was achieved for 33 of 100 items, but complete disagreement was observed for 32 items. Intent type was particularly challenging: only 13 of 45 items yielded unanimous agreement, while 10 cases produced complete disagreement.

Task	**Krippendorff's α**	**Fliess' κ**	**Percent Agreement**
Aggression Type	0.232	0.229	28.04%
Online Harm Type	0.279	0.276	33%
Intent Type	0.290	0.285	28.89%

Table 5.26: Pilot Study Inter-rater Relatability Scores

Annotator feedback and lessons learned: The qualitative feedback sessions provided important insights into the limitations of the pilot design and annotation scheme, which included:

- **Need for context**: Annotators consistently stressed that judging aggression, harm,

or intent from a single message was difficult. Many messages could only be interpreted reliably when situated within a sequence of two or more surrounding messages.

- **Examples and non-examples**: Annotators requested illustrative examples and counterexamples for each category. In particular, borderline cases of covert aggression (e.g., sarcasm, backhanded compliments) and intent types were sources of confusion.

- **Category complexity:** Differences in personal experiences with online harm appeared to shape interpretations of categories. Annotators suggested introducing a "*general insult*" as a fallback label for ambiguous cases, and conducting calibration sessions to align category boundaries before full-scale annotation.

Following the feedback received from the pilot study, it was determined that the annotation framework would benefit from the inclusion of both examples and non-examples for each category. To inform this improvement, we consulted with different stakeholders like pilot annotators, researchers affiliated with Ireland's Observatory of Cyberbullying, Cyber-Hate, and Harassment[84], and two young researchers during their week-long collaborative participation at the time of the pilot study. These consultations collectively highlighted the need to provide explicit illustrations that could guide annotators in distinguishing between categories. Consequently, illustrative examples and corresponding non-examples were incorporated for each category presented in Table 5.27. In addition to this refinement, the feedback process also identified the need for a fallback category to account for expressions that, while clearly offensive or demeaning, do not map neatly onto the existing harm types. As a result, an additional ninth label was introduced for **Sub-task B: Online Harm Type** labelling as a fallback alternative and defined as follows,

> **9. General Insult**: Some expressions, such as "`I wish you were dead`" or "`you are annoying.`" may not align with specific harm categories but remain clearly insulting. Messages within this category are *rude, offensive, or demeaning content that do not meet the criteria of other harm types.* It

[84] `https://antibullyingcentre.ie/observatory/`

is worth noting that this label category should only be applicable when no more specific category is appropriate as a *fall back* category.

Example: "Nobody cares about your pointless story" or "That's a stupid thing to say".
Not an example: "You're an idiot" (as it is Personal Attack) or "Your argument is weak.".

Following discussions with various stakeholders, this ninth category of online harm was introduced in annotation guide with an important caveat,

> ⚠️ `General Insult` *is a fallback category only applies if the message contains rude, demeaning, or insulting content that does not clearly meet the threshold or definition of any other specific harm category (like Personal Attack, Identity-Based Harassment, etc.). If a more specific harm category applies, use that instead of, or as a priority over, General Insult. For example, a direct attack like "You're an idiot" is a* `Personal Attack`*, not a* `General Insult`*. "Your opinion is worthless" might be a* `General Insult` *if it doesn't feel like a direct character attack but is still demeaning.*

To ensure both annotation accuracy and annotator well-being, two additional labels `Unclear` and `Skip` were introduced following pilot consultation. The `Unclear` label addresses instances where the intent of a message cannot be reliably interpreted, even after careful consideration of context, due to factors such as insufficient contextual information, unintelligible language, or highly culture-specific slang. Its inclusion prevents forced categorisation into categories when evidence is lacking, thereby improving data quality and transparency by requiring annotators to document the basis of uncertainty. In contrast, the `Skip` label was implemented for ethical reasons, allowing annotators to withdraw from labelling content they find distressing, re-traumatising, or otherwise harmful to engage. This provision prioritises annotator safety and emotional well-being while ensuring that participation remains voluntary and non-coercive. Together, these labels promote both methodological rigour and ethical responsibility in the annotation process.

Category	Example	Not an example
Threats of violence	"I'm going to find you and you'll be sorry" or "You deserve to be beaten for saying that."	"You're going to regret this." or "I'm so angry I could scream"
Blackmail	"If you don't do it, I'm posting these photos of you"	"I'm going to tell the teacher you were late.".
Personal Attack	"What a pathetic loser you are" or "You're worthless."	"Nobody cares about your opinion" or "you are not making sense".
Identity-based Harassment	"Go back on that 🐫" or "Women are too emotional to lead."	"your taste is terrible" or "you crazy man".
Body-shaming	"You're so big you need your own zip code" or "Look at those twig arms, pathetic".	"hate your outfit today." or "You ugly clown"
Exclusion	"We'd all be better off if you weren't here" or "Nobody likes you, just leave.".	"Go away, I need some space".
Sexual Harassment	"Show me your t**s" or "Are you a virgin? Want me to change that?"	"look nice today" or "that dress is very flattering".
Encouragement-to-Self-Harm	"The world would be better if you starved yourself.".	"Are you suicidal?" or "I wish you were dead".
Sarcasm	"Trump is the most innocent man wrongly accused since O.J. Simpson".	"This is genuinely a great!"
Gaslighting	"you only remember things you image, it never happened" or "nothing i said was offensive, you just like to react to everything"	"you might be mistaken, we agreed on Tuesday" or "i do not remember that"
Blaming	"This is all your fault! If you hadn't been late, we wouldn't have missed it.	"I'm upset that this happened." or "we should be better than this"
Abet or Instigate	"Let's all report them until it's banned." or "Someone needs to teach him a lesson".	"You should stand up for yourself.".
Attack	"I'm going to make you regret saying that." or "Only a moron would believe that".	"That was not a smart move." or "your argument is weak".
Defend	"Leave him alone! pick someone your size".	"I understand why you feel that way".

Table 5.27: Annotation Category Examples and Non-Examples

5.4.3 Main Study

A central challenge identified during the pilot was the role of conversational context in annotation. While earlier annotation frameworks (e.g., (Van Hee et al., 2018); (Sprugnoli et al., 2018)) did not explicitly incorporate extended context, pilot annotators consistently reported difficulty in interpreting messages without surrounding dialogue. To better understand this issue, we consulted the authors of Ollagnier (2024), who confirmed similar concerns in their study. They noted that presenting annotators with entire conversation histories increased interpretability but also imposed considerable cognitive burden.

To address this trade-off, we designed and implemented a structured "*context-window*" strategy, loosely inspired by the bounded memory constraints used in large language models (LLMs). Just as LLMs generate outputs based on a fixed-length history of tokens, we restricted annotators to a fixed-length window of preceding conversational turns. This allowed us to systematically vary the amount of context while avoiding the cognitive overload associated with full conversations. Specifically, two controlled conditions were employed:

- **Limited Context**: Annotators were provided with (i) a brief scenario description, (ii) a clear marker indicating the start of the conversation, and (iii) the two immediately preceding messages as context for the target message.

  For example, in a conversation with six sequential messages (m_1, m_2, m_3, m_4, m_5, m_6); under this annotation structure, m_1 is presented with the scenario text as

  `scenario: [...]` ***scenario-text*** `[...]`

  `(Target Message) [`m_1 `Sender] : [`m_1 `Message Text]`

  When they toggle to m_2, the scenario text and m_1 are presented as,

  `scenario: [...]` ***scenario-text*** `[...]`

  `[`m_1 `Sender] : [`m_1 `Message Text]`

  `(Target Message) [`m_2 `Sender] : [`m_2 `Message Text]`

  However, when annotating m_3, annotators are only shown the ***two*** most recent messages (m_1, m_2) as context and target message (m_3). In this way, *limited context* enforces a sliding window of two prior turns.

- **Extended Context**: Similar to *limited context*, a scenario description and explicit conversation start marker were provided. However, in this case, annotators were presented with the **five preceding messages** as context. For example, in a conversation with seven sequential messages (m_1, m_2, m_3, m_4, m_5, m_6, m_7); under this

annotation structure, messages till m_3 follow the same structure as *limited context*, however, from when annotating m_4 they are shown previous three messages, and at m_5 it is four, and at m_6 it is presented as,

```
scenario: [...] scenario-text [...]

[m1 Sender] : [m1 Message Text]

[m2 Sender] : [m2 Message Text]

[m3 Sender] : [m3 Message Text]

[m4 Sender] : [m4 Message Text]

[m5 Sender] : [m5 Message Text]

(Target Message) [m6 Sender] : [m6 Message Text]
```

However, when annotating m_7, the previous context will only be from m_2 to m_6. In this way, *extended context* enforces a sliding window of five prior turns.

The main study design resulted in a total of six annotation tasks, reflecting the combination of three sub-tasks (A: Aggression Labelling, B: Harm Type Labelling, and C: Intent Type Labelling) with two context-window conditions (Limited and Extended Context). Annotation activities in this doctoral study were twofold: (i) annotation of the youth co-created text dataset (the focus of this chapter), and (ii) annotation of images (discussed in Chapter 7). To ensure feasibility under funding limitations, annotator participation was capped at 10 hours per person, with 7 hours allocated to text annotation and 3 hours to image annotation. The tasks were distributed across annotators in a way that balanced workload while maintaining sufficient redundancy for reliability testing. Specifically, each of the six tasks was assigned to a group of three independent annotators. This design allowed for cross-checking of inter-annotator agreement within each task while keeping the total annotator pool manageable. This meant there would be a need to recruit 18 annotators.

Recruitment: Consistent with the pilot study, ethical safeguards were upheld through ethics approval, informed consent, resilience training, support mechanisms, flexible participation, and guaranteed compensation. One key insight from the pilot study concerned the age of participants: much of the dataset's language reflected youth culture. For example, a 28-year-old female participant, a native English speaker noted that phrases such as "`ngl`" were unfamiliar and she had to look them online. This highlighted the need

for younger annotators, given the youth-oriented context of the co-created dataset. However, after consultation with the ethics committee, it was agreed that involving minors or very young participants would carry unacceptable risks like potential re-traumatisation or the normalisation of harmful behaviour, especially outside a controlled school setting. To balance ethical considerations with contextual relevance, the participant recruitment pool was therefore set at university students aged 18–22, primarily undergraduates. To ensure diversity, purposeful sampling was used to include students from both STEM and non-STEM disciplines. Recruitment targeted multiple faculties, including: (a) School of Applied Language and Intercultural Studies (SALIS), (b) School of Theology, Philosophy and Music, (c) Institute of Education, (d) School of Computing, and (e) School of Nursing. The recruitment was open to students of varied cultural backgrounds and genders. We facilitated this process by advertising the annotation study during lectures, in collaboration with course coordinators. Applicants were screened using the witnessing experiences used in the survey phase-1 of SDLC outlined in Table 3.2 and 3.4 in §3.3.1 of Chapter §3. This included,

Q *Thinking about your experience with social media platforms in the last year, have you come across mean or hurtful online content?*

Q *Thinking about your experience with social media platforms in the last year, have you seen any of the following online harm types (See Table 3.4) happening to someone else?*

Q *What was the type of the hurtful or nasty content?*

These selected screening questions were designed to identify participants with recent and relevant exposure to online harm on social media platforms. By asking about witnessing encounters with mean or hurtful content, observed online harm types, and the nature of such content, the questions ensured that participants could provide informed and authentic perspectives for the study. This filtering process (responses $\geq$ `sometimes`) allowed us to select a pool of individuals (18 out of 94 interested annotators) with sufficient experience to meaningfully contribute to the research, while maintaining a standby group to address potential dropout.

5.4.4 Results

Annotator Demographics: As shown in Figure 5.7, the study included 18 annotators, of whom 11 (61.1%) were female and 7 (38.9%) were male. In terms of ethnicity, half of the annotators identified as White (9), followed by Black or African (5), with 2 identifying as Asian and 2 as Middle Eastern or North African. Figure 5.8 shows the age distribution: 6 annotators were 19 years old, while 4 were each 18, 20, and 21 years old. All annotators reported having witnessed harm, rumours, threats, and harassment based on skin colour or body shaming. In comparison, 17 had witnessed harassment based on sexual orientation, 16 based on religion, 14 based on disability, and 13 based on exclusion. Drawing these annotators randomly from a larger recruitment pool helped ensure that participants came from a variety of demographic and cultural backgrounds.

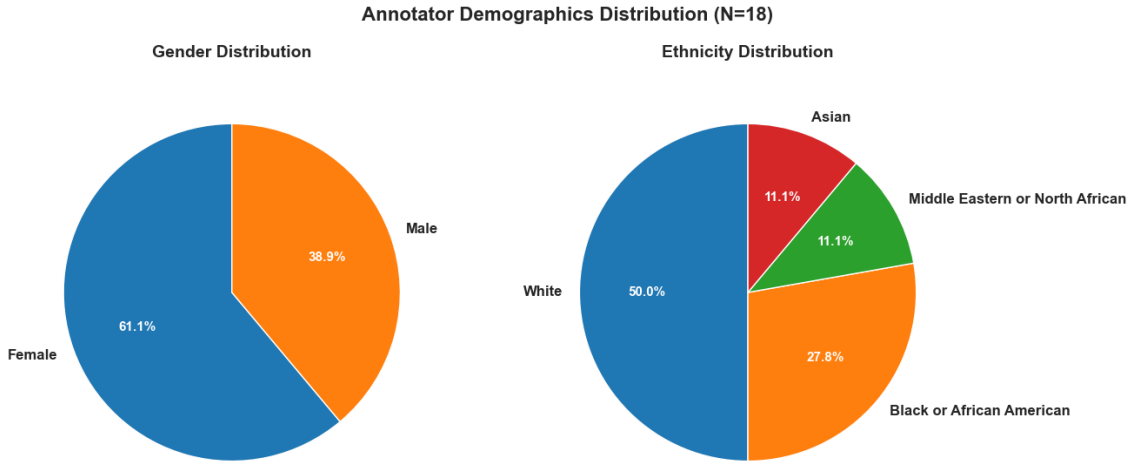


Figure 5.7: Gender and Ethnicity breakdown of the recruited annotators

Task A: Aggression Type Labelling: Table 5.28 presents the inter-rater reliability scores, measured using Krippendorff's alpha, for the annotation sub-tasks. Overall, reliability is higher when annotators are provided with extended conversational context. The improvement in reliability across all tasks is **+0.08** when comparing extended context (five preceding turns) with limited context (two preceding turns).

The inter-annotator reliability in the **limited context** condition was $\alpha = 0.6712$. This translates to 384 out of 553 messages where all three annotators agree, 482 messages where at least two agreed, and 5 messages where all three disagreed. Under majority voting (i.e., at least two annotators agree), the messages were labelled as follows: 230 as

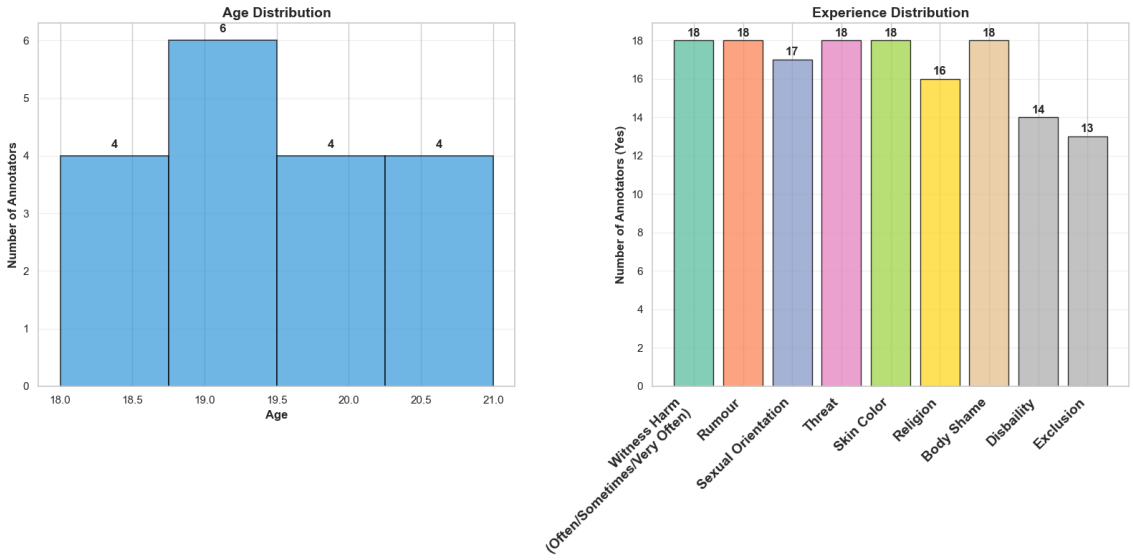


Figure 5.8: Age distribution of recruited annotators and their experience counts with witnessing different types of harm

Task	**Limited Context** α_2	**Extended Context** α_2	$\Delta\alpha$=α_5 - α_2
(A) Aggression Type	0.6712	0.7923	0.1211
(B) Harm Type	0.6693	0.7469	0.0776
(C) Intent Type	0.7259	0.7983	0.0724

Table 5.28: Inter-rater reliability scores presented as Krippendorff's α values for all annotation tasks

`not aggressive`, 230 as `covertly aggressive`, and 88 as `overtly aggressive`. The α = 0.6712 also reflects 96 cases of disagreement specifically between the labels covertly and overtly aggressive (see Table 5.29). In contrast, when annotators were provided with **extended context**, reliability improved to $\alpha = 0.7923$, with disagreements between `covertly aggressive` and `overtly aggressive` reduced to 52. Within this condition, all three annotators agreed on 449 of the 553 messages, at least two annotators agreed on 537 messages, and only 3 messages resulted in complete disagreement. These were:

1. `[ID: 299438]` *Sarah: Lol i knw y , @Tom deserves it !*
2. `[ID: 299484]` *yall think bullying him like this will have ny affect, dont worry TOM i know where @Tadhg and @Roisin stay,*
3. `[ID: 299485]` *Brad: i know where @Jane and @Sarah stay*

A common theme across these disagreed messages is that they overlap between covert

and overt aggression. Each contains elements that can be interpreted either as veiled threats, casual mockery, or explicit intimidation. For instance, the use of humour or sarcasm. ("*Lol [...] deserves it*") may downplay hostility and make the aggression appear covert, while references to knowing where someone lives introduce a more direct and threatening element that feels overtly aggressive. This mixture of ambiguity and threat likely caused annotators to interpret the degree and type of aggression differently, leading to disagreement.

Disagreement Pair	**Limited Context Count**	**Extended Context Count**
('Covertly Aggressive', 'Overtly Aggressive')	96	52
('Covertly Aggressive', 'Not Aggressive')	77	51
('Not Aggressive', 'Overtly Aggressive')	6	7

Table 5.29: Count of disagreement pairs for Task A: Aggression Type Labelling

As depicted in Table 5.29, the pair-wise disagreements between `covertly` and `not` aggressive messages decreased from 77 in the limited context setting to 51 with extended context. However, a small increase was observed in disagreements between `not` and `overtly` aggressive labels, rising from six in the limited context to seven in the extended context condition. In the limited context condition (2 turns), this disagreement pair appeared in 6 unique instances:

- 47] Tadhg: fr3ak 😂
- 60] Jaimie: we hve all seen ur maa's 🍑 jake ... v knw y its big n round
- 61] Jaimie: ur daa 👋 ur maa's 🍑
- 35] Arthur: Anybody who defies the bible in this manner deserves the EVIL coming to them
- 62] Riley: they will 💩 their pants
- 35] Carmen: @Sean who d f said I made this account?

In the extended context condition (5 turns), the disagreement pair appeared in 7 distinct sentences:

- 3] Sarah: Lol i knw y , @Tom deserves it !
- 31] Jake: Tom b ready im comin
- 39] Roisin: see yu in a bit Tom
- 49] Alex: yall think bullying him like this will have ny affect, dont worry TOM i know where @Tadhg and @Roisin stay,
- 50] Brad: i know where @Jane and @Sarah stay
- 44] Paul: leave me alone u cunts !
- 4] Cameron: @Sean who d fk said i made this account

This slight increase in disagreement may be attributed to contextual ambiguity introduced when annotators were given a broader conversational window. In the *limited context*, sentences such as [47] or [62] can be interpreted as humorous, dismissive, or aggressive depending on the annotator's stance, but the shorter context may restrict interpretive flexibility. In contrast, extended context often exposes relational dynamics (e.g., ongoing targeting, threats, or explicit hostility) that may push some annotators toward an "overtly aggressive" interpretation while others still regard the utterance as playful or non-aggressive. For instance, [49] and [50] reference knowing where individuals "stay", which can be perceived either as casual information-sharing (not aggressive) or as a threat (overtly aggressive). Similarly, [44] uses explicit profanity, where annotators diverge between viewing it as a defensive expression versus overt aggression. Thus, the additional instance of disagreement likely reflects greater interpretive variability in the *extended context*: more cues are available, but they can be weighted differently depending on how annotators judge intent (e.g., banter vs. threat).

Also, as depicted in Table 5.30, the majority of labels assigned by annotators differ between the two context window settings. While there is substantial agreement across both limited and extended conditions, especially for `Covertly Aggressive` (188) and `Not Aggressive` (185), many disagreements occur in cases involving subtle aggression. Notably, 43 messages with the majority label as `Not Aggressive` in the limited context condition shift to `Covertly Aggressive`, and two messages are labelled as `Overtly Aggressive`. For example, the message *`Roisin: @Jake I saw Tom at [location] 10`*

minutes ago appears benign in isolation, but with preceding context it becomes clear that Tom's location is being shared to facilitate a planned fight. Similarly, the message *Julie: @Pierre I dont think you are from this school OR this country* 😂 initially seems like a light-hearted remark, but with additional context, it is revealed as part of a bullying exchange targeting Pierre's identity. Conversely, there are also cases where aggression is over-attributed under the limited context. For instance, the message *Carmen: @Sean who d f said I made this account??* is labelled as `Overtly Aggressive` in the limited context condition, as it seems like a hostile attack toward Sean. However, with extended context, it becomes clear that Carmen is defending themselves against an accusation of creating the account, and the label shifts to `Not Aggressive`. These examples highlight how additional conversational history enables annotators to more accurately assess both the intent and social dynamics underlying a message. Extending the context not only helps to uncover covert aggression that may otherwise go unnoticed, but also prevents misclassifying defensive or corrective responses as overtly aggressive.

		Extended Context *Majority Label*		
		Covertly Aggressive	Not Aggressive	Overtly Aggressive
Limited Context *Majority Label*	Covertly Aggressive	188	22	22
	Not Aggressive	43	185	2
	Overtly Aggressive	44	1	46

Table 5.30: confusion matrix showing the results of classification with limited context versus extended context.

Task B: Harm Type Labelling: To evaluate inter-annotator reliability in this multi-label classification task, i.e., annotators could assign one or more labels to a message. We adopted a collapsing strategy to derive a single common label per instance. The procedure prioritised consensus: if all three annotators agreed on a single label, that was selected; in cases with multiple shared labels, the first label most consistently chosen across annotators was retained; and when no consensus existed, the first label provided was used. Applying this method in the **limited context** condition, the inter-annotator reliability

reached $\alpha = 0.6693$ (See Table 5.28). Within this setting, annotators fully agreed on 366 of 553 messages, at least two annotators agreed on 471 messages, and 17 messages had complete disagreement with all three providing distinct labels. As depicted in Table 5.31, within the limited context setting, the largest proportion of messages were labelled as No Harm (258), followed by Personal Attack (123), and Identity-Based Harassment (46). Less frequent but still present categories included General Insult (Fallback Only) (40), Threats of Violence (33), and Body-shaming (15). Rare categories such as Exclusion (10), Sexual Harassment (6), Encouragement to Self Harm (4), and Blackmail (1) were observed only in a small number of cases. Also, within this context condition, there were a total of 221 total disagreement pairs i.e., disagreement between two labels, with the largest number (76 instances) arising between the fallback category `general insult` and `no harm` (See Table 5.32).

Category	**Limited Context Count**	**Extended Context Count**
No Harm	258	275
Personal Attack	123	121
Identity-Based Harassment	46	47
General Insult (Fallback Only)	40	31
Threats of violence	33	34
Body-shaming	15	19
Exclusion	10	7
Sexual Harassment	6	5
Encouragement to Self Harm	4	4
Blackmail	1	1

Table 5.31: Distribution of majority and collapsed annotation labels across limited and extended context conditions for Sub-task B.

When extending the conversational context, reliability improved modestly, with Krippendorff's alpha increasing by 0.0776 to α=0.7469. This increase in agreement corresponds to 413 messages where all three annotators shared at least one common label, and only nine messages exhibited complete disagreement among them. Within this context, at least two annotators agreed on 524 messages. The category distribution for extended context depicted in Table 5.31 shows a slight shift compared to the limited context. No Harm

Disagreement Pair	Limited Context	Extended Context
(General Insult (Fallback Only), No Harm)	76	36
(General Insult (Fallback Only), Personal Attack)	60	40
(No Harm, Personal Attack)	25	16

Table 5.32: Count of disagreement pairs for Task B: Harm Type Labelling. Further details are presented in Table A.15 in Appendix A.13

messages increased to 275 (from 258), suggesting that access to a broader conversational context allowed annotators to recognise more messages as benign. Conversely, `General Insult (Fallback Only)` decreased to 31 (from 40), indicating that extended context helped annotators resolve ambiguity and assign more specific categories. Other categories remained relatively stable, with slight fluctuations: Personal Attack dropped slightly (121); Body-shaming increased (19), and Threats of Violence rose marginally (34). The $\Delta\alpha$ of 0.0776 may seem small, but it corresponds to a reduced 158 disagreement pairs from 221 pairs, with the most frequent (36 instances) again occurring between `general insult` and `no harm`. Another prominent disagreement pair was `general insult` vs `personal attack`, with 60 instances in the limited context and 40 in the extended context. Interestingly, the extended context introduced three disagreement pairs that were not present under the limited condition. For example, one message involved disagreement among `blackmail` and `no-harm`, and `personal attack`. The contested message was: *Alex: d video* 😂😂😂😂, produced divergent interpretations. Within the extended conversational window, the preceding five messages were:

```
Jenny: You shoulda stáyed a 💦 Wendy
Alex: Forever 😂 at you for being so f#king p@thetic
hahahahaha
Jaimie: what did wendy do to you
Alex: Thats what she deserves
Wendy: Delete this video
```

This example illustrates how a short, ambiguous message from Alex, a possible enabler or bully, can be interpreted as blackmail, personal attack, or no harm depending on contextual emphasis. The same message also triggered another new disagreement pair (`blackmail` vs. `personal attack`). Similarly, extended context revealed another previ-

ously unseen disagreement pair between encouragement to self-harm and no harm. The message in question was: *Taylor: u have no right to breathe oxygen in 2025 if u dont like* 🌈, with the preceding conversational context being:

```
Arthur: @Paul youre acc b3nt

Paul: wut?

Ellian: it means you're gay idi0t 😂😂

Arthur: are you gay and slow?? Jeez double h0m0s1de.

Jimmy: cmon we are proud 🌈
```

In this case, one annotator classified the statement as encouragement to self-harm, while the others judged it as no harm. Taken together, these findings indicate that extended context generally reduced disagreement and increased reliability, but also introduced new categories of disagreement not observed in limited windows. The emergence of novel disagreement pairs underscores the complex interplay between context and interpretation: while longer context clarifies some ambiguities, it simultaneously opens up alternative readings that annotators may weigh differently.

As shown in Figure 5.9, across both limited and extended contexts, messages were distributed as follows: 253 labelled as `no harm`, 90 as personal attacks, 41 as identity-based harassment, 32 as threats of violence, 18 as general insults (the fallback label), 16 as body-shaming, 9 as exclusion, 6 as sexual harassment, 4 as encouragement to self-harm, and 1 as blackmail. For blackmail, encouragement to self-harm, and sexual harassment, the labels remained consistent across both context conditions. A notable shift occurred between `personal attack` and `general insult`: 16 messages labeled as `personal attack` in the limited context were relabeled as `general insult` with extended context, while 12 moved in the opposite direction. For example, messages such as "*its not her fault ur stup!d*" or "*ur a bunch of rascist b1tches*", were more clearly identified as personal attacks when additional context clarified that the insults targeted individuals rather than their arguments.

Task C: Intent Type Labelling: Since this was also a multi-label classification task, the same label-collapsing strategy described earlier for Task B was employed here. As depicted in Table 5.28 within the **limited context** condition, Krippendorff's α was 0.7259, with 10 instances of complete three-way disagreement. Within this setting, all



Figure 5.9: Heatmap depicting shift in majority labels across both limited and extended context conditions for Task B.

three annotators agreed on 398 messages, and at least two annotators agreed on 495 messages. As illustrated in Table 5.33, within limited context, the most frequent label was Defend (231 messages), followed by Attack (189) and Abet or Instigate (47). Less common labels included Blaming (39), Sarcasm (26), and Gaslighting (8). Unlike other sub-tasks, the label *Unclear* was used in this sub-task for 14 messages where at least one rater marked it as unclear, and in only three instances did at least two annotators use this label. These three messages, included a first message by the victim as *Penny: take down this account*, another message as *Fatima: @Theo do tell where to? i think i will*, *Kevin: he's gone now*. The limited preceding context for the last two messages made it difficult for annotators to determine the intent, leading to the unclear labelling. Within the limited context setting, the Krippendorff's α was 0.7259 corresponds to a total of 175 disagreement pairs, the most frequent of which occurred between `Abet or Instigate` and `Attack`, with 32 instances (See Table 5.34).

Category	**Limited Context Count**	**Extended Context Count**
Defend	231	231
Attack	189	227
Abet or Instigate	47	35
Blaming	39	17
Sarcasm	26	26
Gaslighting	8	11
Unclear	3	6

Table 5.33: Distribution of majority and collapsed annotation labels across limited and extended context conditions for Sub-task C.

Disagreement Pair	**Limited Context**	**Extended Context**
(Abet or Instigate, Attack)	32	21
(Attack, Sarcasm)	21	15
(Abet or Instigate, Sarcasm)	19	11
(Attack, Blaming)	14	4

Table 5.34: Count of disagreement pairs for Task C: Intent Type Labelling. Further details are presented in Table A.16 in Appendix A.13

In the **extended context** condition, reliability improved modestly, with Krippendorff's α increasing by 0.0724 to α=0.7983. Notably, there were no instances of full three-

way disagreement in this condition. Within this context setting, there were 440 messages where all three annotators fully agreed and all 553 messages where at least two annotators agreed. As illustrated in Table 5.33, the distribution of labels also shifted slightly in the extended context: Defend remained unchanged at 231, Attack increased to 227, and Abet or Instigate decreased to 35. Other categories saw modest changes, with Blaming dropping to 17, Sarcasm remaining stable at 26, and Gaslighting increasing to 11. However, there were 6 instances where all three annotators labelled a message as Unclear. These sentences were:

- ID: 302925] *Jane: looks lyk sm1 left the cinemas*
- ID: 303167] *Paul: thanks folks ! @Eillian*
- ID: 303230] *Group Setting: **** Alex added Rosalyn to the group ****
- ID: 303425] *Kevin: guess what the coach said..*
- ID: 303426] *Roman: wut?*
- ID: 303427] *Kevin: he's gone now*

For the last three messages [ID: 303425–27], annotators noted that even with context, the individual utterances were not clear enough to fall into any intent category. The group setting message was also difficult to categorise. For the first two messages, annotators agreed they did not match any intent type either. Despite the availability of the Unclear label, it is noteworthy that all three annotators independently agreed on assigning it to these six cases in the extended context condition. Within this **extended context** setting, the total number of disagreement pairs also declined sharply to 113, indicating a substantial reduction in annotator divergence when additional conversational information was available. Among these, disagreement between `Abet or Instigate` and `Attack` dropped to 21 instances, representing the largest reduction across all label pairs. An additional point of interest concerns the optional label `Unclear`, which annotators could assign when they were uncertain about the appropriate category. In the limited context window, `Unclear` appeared relatively often as part of disagreement pairs, typically contrasted with more definitive categories such as *Defend, Attack, Gaslighting, Sarcasm, or Blaming* (see Table 5.34). However, in the extended context, the reliance on `Unclear` diminished markedly:

it appeared in only one instance for a starting message by the victim *Penny: take down this account* for a scenario where Penny's photos are circulated via an account. For this instance, two annotators selected `Defend` and one chose `Unclear`.

The disagreement pairs between attack and defend increased notably, rising from 13 instances in the limited context window to 22 in the extended window. Importantly, the messages in each set were entirely different, i.e., none overlapped between the two annotation conditions. Across both the limited and extended contexts, much of the ambiguity stems from the blurred boundary between an *attack* and *defense*. Many defensive responses are phrased aggressively, e.g., "`Kevin: STFU @Cathal ! u dnt knw everythng !`", in the limited context window was labelled as *attack* and *defend* by annotators. Similarly, despite more messages as context in the extended context window, a message such as e.g, "`Tom: stfu evry1` was also labelled as both *attack* and *defend* by annotators. These examples illustrate a recurring difficulty in determining whether the primary function of a message is to protect someone (Defend) or escalate the conflict further (Attack). In the limited context window, disagreements around *Attack* vs. *Defend* were typically linked to short and hostile messages like *Cathal: u so salty @Kevin!* or *stop being a d**khead*. With only minimal surrounding dialogue visible, annotators had to infer intent from tone alone, which often made a defensive message difficult to distinguish from an outright attack. The absence of speaker role information further compounded this issue, since annotators could not reliably determine whether the speaker was siding with a victim or antagonising them. In the extended context window, however, disagreements around attack vs. defend were more often associated with longer, more elaborated hostile statements. For example, messages like *Lexi: u think u can last longer than Wendy? lets try u n i ?* or *Jaimie: What l o s e r s. Spent all their free time just to lie and put Penny down. Dont they have anything better to be doing*, contain both elements of defense and elements of hostility. On one hand, they push back against bullying or call out unfair behaviour (a defensive function), but they also include direct insults, taunts, or challenges that escalate tensions (an attacking function).

As illustrated in Figure 5.10, messages were distributed across intent categories as follows: 210 labelled as `defend`, 143 as `attack`, 15 as `abet or instigate`, 8 as `blaming`, and 7 each as `sarcasm` and `gaslighting`, and 6 as `unclear`. While these labels capture

overall trends across both limited and extended contexts, they also reveal shifts in annotation under the extended context. The most pronounced change occurred in the `abet or instigate` category, where 33 messages labelled as `abet or instigate` in extended were differently labelled in the limited context, most often labelled as `attack`. Similarly, `sarcasm` proved highly unstable, with 9 messages labelled as `attack` in the limited context condition were labelled as sarcasm. `Gaslighting` and `blaming` also showed moderate rates of changes. These shifts highlight that certain intent categories like `sarcasm` and `abet or instigate` remain sensitive to contextual information, whereas others, such as `defend`, remain comparatively robust. This suggests that the source of ambiguity differs by context window size. In the limited window, ambiguity arises primarily from lack of information, i.e., short, blunt messages are stripped of cues about intent or alignment, making them hard to classify. In the extended window, by contrast, ambiguity arises from excess or overlap of functions, i.e., longer defensive statements are interwoven with insults or counterattacks, creating hybrid messages that resist simple categorisation as purely "*Attack*" or "*Defend*".

5.4.5 Conclusion and Insights

The pilot study provided critical insights into the feasibility and limitations of the proposed annotation framework. While it confirmed that the three-task design that captures aggression, harm, and intent type was conceptually robust, message-level pilot results revealed consistently low inter-annotator agreement across all dimensions. This divergence was also shaped by the heterogeneous backgrounds and interpretive lenses of the annotators. The lack of worked examples further amplified interpretive variability, highlighting the need for clearer training materials and more demographically aligned annotators in subsequent phases. Despite its low reliability scores, the pilot fulfilled its role as a diagnostic exercise and helped identify categories of confusion. The lessons learnt from the pilot study directly informed the methodological refinements introduced in the main study, including the use of a context-window strategy, targeted recruitment of youth annotators, and structured support mechanisms.

By introducing a structured *context-window* strategy, the study balanced the interpretive advantages of extended dialogue with the cognitive constraints faced by annotators.

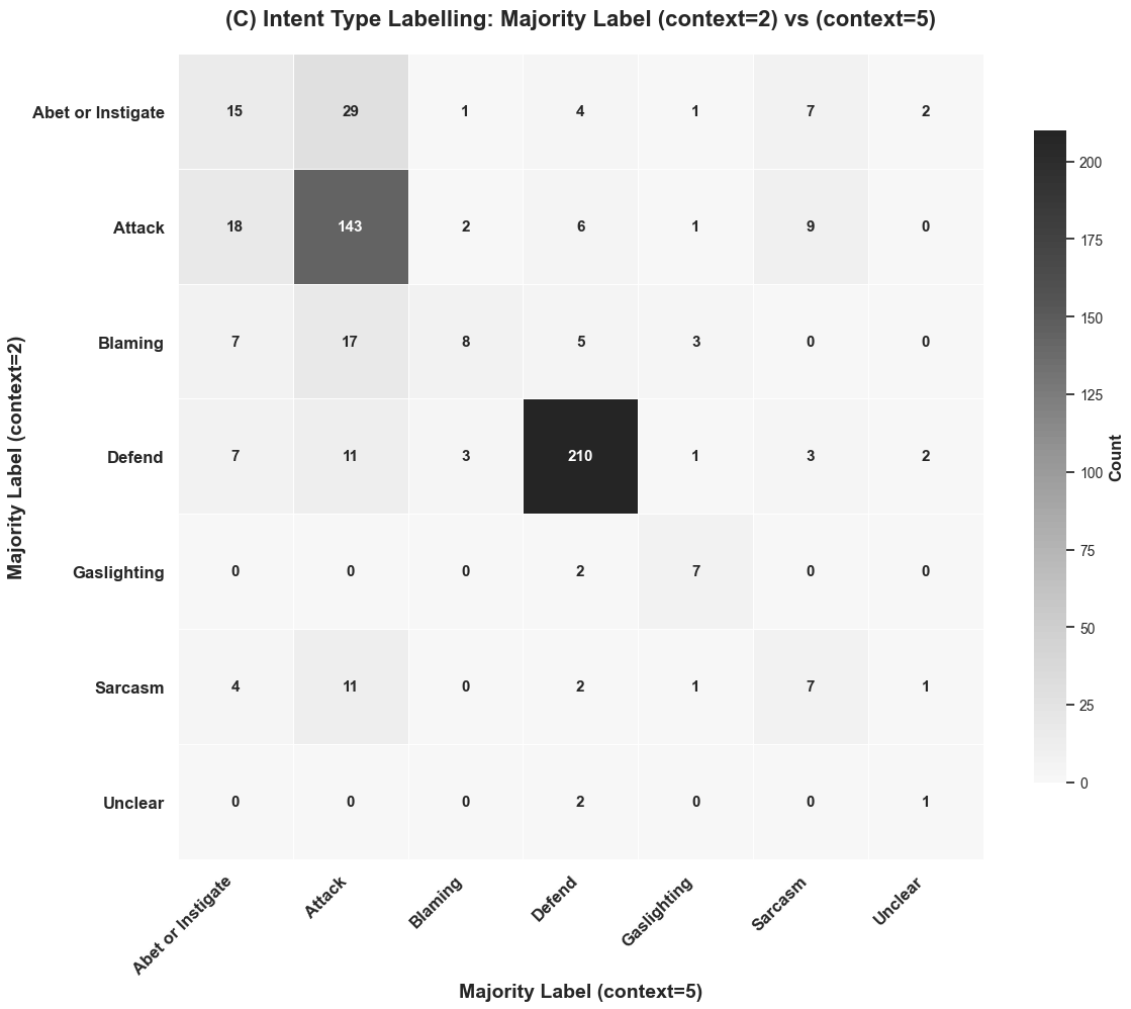


Figure 5.10: Heatmap depicting shift in majority labels across both limited and extended context conditions for Task C.

The results show a consistent and meaningful improvement in inter-annotator reliability when additional conversational history was provided. Across all three annotation tasks Krippendorff's alpha increased when annotators had access to five prior conversational turns instead of two. Thereby underscoring that many forms of online aggression, particularly covert or identity-based harms, cannot be reliably identified in isolation and are best understood within broader interactional dynamics.

At the same time, the results also reveal that extended context does not eliminate ambiguity. In some cases, it introduces new forms of disagreement. Annotators sometimes diverged more when longer histories provided competing cues, for example, when hostility was embedded within defensive responses, or when sarcasm and humour overlapped with genuine aggression. In this sense, context clarifies certain subtleties (e.g., distinguishing playful banter from covert bullying) while complicating others (e.g., teasing apart defense

from attack in hybrid utterances). The fact that disagreement shifted rather than disappeared highlights that annotation of social interactions is not a purely mechanical task, but one that requires interpretive judgment shaped by cultural and social cues.

The recruitment strategy also played a crucial role in ensuring valid annotations. By engaging young adult annotators (aged 18–22) with recent and relevant exposure to online harms, the study achieved both contextual sensitivity and ethical safeguards. The balanced pool of Irish annotators across gender, ethnicity, and disciplinary backgrounds enhanced the diversity of perspectives while ensuring feasibility within the practical constraints of time and resources. Nonetheless, ethical restrictions prevented the recruitment of younger participants who might have been even closer in age and cultural knowledge to the dataset creators.

Overall, the study advances methodological understanding of how to operationalise conversational context in annotation design. The *context-window* framework proved both scalable and effective, offering a middle ground between single-message annotation and overwhelming full-conversation annotation. Importantly, the findings provide empirical evidence that longer context windows enhance reliability and reduce misclassification, particularly for subtle or covert harms. By empirically testing structured context windows, it contributes both methodological tools and empirical insights to the broader challenge of building socially-aware datasets for studying cyberbullying.

5.5 Summary

In this chapter, the first research question RQ1, which seeks to explore alternative methods to create youth-reflective datasets, is addressed through both human-centered and machine-assisted data augmentation strategies. By examining four complementary data augmentation strategies like human translation (HT §5.1.1) yields $\mathcal{D}^{HT}_{IT\rightarrow EN}$, machine translation (MT §5.1.2) yields $\mathcal{D}^{MT}_{IT\rightarrow EN}$, human-aided machine translation (H-MT §5.1.3) yields $\mathcal{D}^{HMT}_{FR\rightarrow EN}$, and large language model (LLM) generation (§5.1.4) yields two datasets - $\mathcal{D}^{\text{Phase-1}}_{\text{LLM}}$ and $\mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$. This chapter builds resources for this low-resource area where ethical and methodological constraints often limit data availability.

For data augmentation, the translation-based strategies demonstrated the strengths

and limitations of human and machine approaches. Human Translation (HT) in collaboration with translators at TransPerfect and validations by both experts and stakeholders yields an English version of the (Sprugnoli et al., 2018) dataset (§5.1.1). Machine Translation (MT), particularly through *back-translation*, proved effective in generating synthetic corpora that preserved domain-specific signals, though challenges remained with non-standard tokens like emojis (§5.1.2). Human-aided MT (H-MT) combined student-led translations with professional post-editing of French cyber-aggression and bullying dialogues by (Ollagnier et al., 2022; Ollagnier, 2024) (§5.1.3), yielding contextually rich English datasets that retained annotation integrity while balancing literal fidelity with colloquial naturalness. Together, these translation strategies established replicable methods for building multilingual and cross-cultural cyberbullying corpora.

In parallel, this chapter investigated LLM-based generation in Section 5.1.4, demonstrating that controlled, prompt-engineered outputs can yield realistic, role-specific bullying conversations. Phase-2 LLM outputs (LLaMA 3.3) produced more natural turn-taking, moderated lexical diversity, and coherent multi-party exchanges compared to Phase-1 outputs, confirming the potential of LLMs as ethical and scalable data augmentation tools when prompted in a structured manner.

Extensive dataset profiling discussed in Section 5.2 revealed complementary strengths. The youth co-created dataset (§4.3), emphasised defender–enabler reciprocity and protective behaviours, youth role-played corpora offered structured interactions and role-based interventions, translated datasets maintained structural fidelity while reducing vocabulary richness, and refined LLM outputs closely approximated youth-authored conversational rhythms and at times placed stronger emphasis on bully–victim antagonism.

A key contribution of this chapter is the intrinsic evaluation protocol (Section 5.3), which assessed datasets across nine conversational dimensions (e.g., conflict, support, humour, power dynamics). Results showed high agreement on concrete behaviours but variability on socially nuanced features. Notably, youth co-created data achieved the strongest consensus across stakeholders, underscoring the value of grounding dataset curation through a cascading co-creation and co-design approach. Subgroup analyses further revealed that including youth as evaluator enhanced reliability on subtle dimensions, while professional moderators provided structured consistency.

Additionally, the three sub-task annotation studies discussed in Section 5.4 highlighted the central role of conversational context. Introducing structured context windows (two vs. five prior turns) significantly improved reliability in labelling aggression, harm, and intent. Although extended conversational history at times introduced new interpretive ambiguities. Recruiting young adults for the annotation study ensured cultural sensitivity and ethical safeguards, while also revealing the interpretive nature of cyberbullying annotation tasks.

Through these multi-layered strategies, we thoroughly explore the first research question posed in this doctoral thesis (§RQ.1) and further cyberbullying detection research by creating two datasets,

$\mathcal{D}_{\textbf{Gold}}$: A benchmark English-language dataset co-created by adolescents, and validated through multi-level annotation capturing aggression, harm type, and intent.

$\mathcal{D}_{\textbf{Silver}}$: A collection of synthetic datasets generated via translation and LLM strategies, designed for scalability while retaining conversational realism.

Chapter 6

Insights into Binary, Multi-task Cyberbullying Detection, and Assessing Classifier Robustness

Abstract

This chapter advances the study of automated cyberbullying (CB) detection by addressing research questions **RQ.2.1** and **RQ.2.2** through four interconnected challenges: task clarity, interpretability, robustness, and conversational context. First, in §6.1 we disentangle CB from other closely related yet distinct phenomena, such as hate speech or offensive language by benchmarking traditional machine learning models (SVM, Random Forest, XGBoost), sequential neural architectures (Bi-LSTM, GRU), and transformer-based models (BERT, HateBERT) on CB-specific datasets. Second, in §6.2 we move beyond performance metrics to examine interpretability and explainability. We assess the interpretability of transformer-based models by analysing their attention and gradient weights to understand whether the use of specific words or phrases guides automatic CB classification. Third, in §6.3 we extend binary classification to a multi-task framework that models conversational dynamics by jointly predicting hurtfulness and participant roles (bully, victim, enabler, defender). We benchmark different learning

objectives and model architectures to jointly detect **who** (user-role) **said what?** (hurtful?) in CB conversations. Finally, in §6.4 we strengthen binary cyberbullying detection by evaluating robustness against adversarial obfuscation. Building on the adversarial examples generated by youth participants described in §3.3.3 of Chapter 3, we construct a novel dataset capturing realistic obfuscation and linguistic perturbations in cyberbullying content. This dataset enables a robust evaluation of character-aware architectures, fine-tuned encoder-based models, and closed-weight systems. Through this analysis, we examine how well current approaches withstand deliberate manipulation and assess progress toward more resilient cyberbullying detection models. Taken together, the chapter provides a comprehensive evaluation of cyberbullying detection systems and helps address research questions **RQ.2.1** and **RQ.2.2**.

6.1 Binary Benchmarking with Appropriate Cyberbullying Datasets

Building on insights from prior studies highlighted in Tables 2.2 and 2.3 (Section 2.4.2, Chapter 2), it is evident that although many approaches employ supervised learning and pre-trained language models for cyberbullying detection, their effectiveness is often limited by dataset quality, annotation inconsistency, and platform specificity (Rosa et al., 2019; Emmery et al., 2021). A further concern is that many of these studies conflate `hate-speech` and `offensive/abusive` content with *cyberbullying* by training binary classifiers on datasets originally developed for these adjacent phenomena rather than for bullying itself (see Table 2.2, Chapter 2). Because these datasets do not capture the relational, contextual, and behavioural components distinctive to cyberbullying, the resulting models generalise poorly across platforms, where linguistic norms, user practices, and platform affordances vary substantially.

In most computational studies, binary cyberbullying detection is operationalised as a simple distinction between "`bullying`" and "`not-bullying`" typically approximated

by classifying posts as *nasty/hurtful* or *benign*. However, as established through both conceptual analyses in Chapter 2 and youth-participatory findings in Chapters 3 and 4, cyberbullying is not a binary phenomenon but a *situated*, *relational*, *multi-stage* process (Slonje et al., 2013; Kowalski et al., 2014; Smith and Steffgen, 2013; Milosevic et al., 2022). Importantly, research shows that `hurtful or nasty` behaviour is often a precursor or early signal of cyberbullying, but does not constitute bullying by itself because it lacks the relational, intentional, and repetitive components that youth emphasise as central to their lived experience (Kowalski et al., 2014; Navarro and Jasinski, 2015). Although binary classifiers cannot capture the full complexity of cyberbullying, developing robust binary models remains valuable. They provide a mechanism for detecting early indicators of `hurtfulness`, which can function as a computational precursor or triage signal for more nuanced, context-aware models of cyberbullying behaviour. In this sense, binary detection is better understood as hurtfulness detection, which can support CB detection but should not be conflated with the identification of genuine cyberbullying episodes.

6.1.1 Problem Formulation

Building on this notion of detecting `hurtfulness` in cyberbullying contexts, we address gaps in existing research by formulating the problem as:

1. **Cross-platform benchmarking of language models**: This study evaluates the effectiveness of different language models ($m \in M$) for downstream text classification tasks, specifically predicting whether a given text (x) corresponds to a binary label $y \in \{0, 1\}$, where $y = 1$ indicates `hurtful` content and $y = 0$ indicates `not-hurtful` content. For cross-platform benchmarking, each language model ($m \in M$) is trained on distinct data sources, `User-Comment` ($x_{uc} \in \mathcal{D}_{UC}$), `Question-Answering` ($x_{qa} \in \mathcal{D}_{QA}$), and `Micro-blogging` ($x_{mb} \in \mathcal{D}_{MB}$) datasets, and evaluated on data from the other sources. For example, a model trained on $\mathcal{D}_{UC}$ is denoted as M_{uc} its prediction for a text is given by $\hat{y} = m(x) = \text{argmax}_{k \in \{0,1\}} P(Y = k \mid X = x; m)$

2. **Evaluation of translation strategies for youth data:** Building on the outcomes of Chapter 5, Sections 5.1.1 and 5.1.2, we first conducted a series of *intrinsic*

evaluations of translation quality using multiple human assessment approaches. Following these intrinsic analyses, we present the corresponding *extrinsic evaluation* in this chapter. Here we assess the suitability of human-translated ($x_{ht} \in \mathcal{D}^{HT}_{IT-EN}$) and machine-translated ($x_{mt} \in \mathcal{D}^{MT}_{IT-EN}$) youth translated datasets for detecting binary `hurtfulness`. For each dataset ($\mathcal{D}^{HT}_{IT\rightarrow EN}$ and $\mathcal{D}^{MT}_{IT-EN}$), a model $m \in M$ is trained and produces predictions $\hat{y}$, where $y = 1$ indicates `hurtfulness` and $y = 0$ indicates `not-hurtful`. This downstream modelling setup provides an ***extrinsic evaluation*** of translation quality, replicating the experimental design of (Corazza et al., 2019) on the original Italian corpus by (Sprugnoli et al., 2018) to measure how translation choices affect model performance.

6.1.2 Experiment Setup - Cross-platform benchmarking

Note: This sub-section is published in (Verma et al., 2022)

Datasets: The datasets used in this study originate from three major categories of social media platforms, illustrated in Table 2.3 (Section 2.4.2 Chapter 2), these include:

- *`User-Comment`* platforms are platforms that allow users to create and share multimedia content that others can like, comment on, and redistribute content. Ashktorab and Vitak (2016) curated binary post-level labels from user-comments collected from Instagram, and Rafiq et al. (2015) also curated a similar user-comment dataset from now-defunct video-sharing platform Vine. Together these datasets are hereafter referred to as D_{UC}.

- *`Question-Answering`* platforms, unlike content-sharing sites enable users to remain anonymous for exchanging questions and answers. Van Hee et al. (2018) curated a cyberbullying binary and multi-labelled corpus from the now-defunct ASK.fm platform, and Reynolds et al. (2011) collected a binary corpus from now-defunct Formspring.me platform. Together these datasets are hereafter referred to as D_{QA}.

- *`Micro-blogging`* platforms refer to Twitter (now X), which is a micro-blogging service where users post brief, blog-like messages called tweets, each limited to 280 characters. Xu et al. (2012) and Salawu et al. (2021) compiled binary and multi-

class labelled datasets sourced from Twitter. Together these datasets are hereafter referred to as D_{MB}.

The token lengths (words) in each phrase or comment across the six datasets reflect the similarity of the platforms, as shown in Table 6.1, indicating that platforms with similar functionalities tend to have comparable token lengths. Label and sentence-level details for each dataset are provided in Table 6.1. Merged datasets by platform category were divided into 70% for training, 20% for validation, and 10% reserved as a hold-out test set. All language models trained in the three merged datasets were individually evaluated on the 10% hold-out test sets of each respective merged dataset.

Dataset	**Source Study**	**# of Sentences**	**Hurtful (%)**	**Avg. # Tokens**
D_{UC}	(Hosseinmardi et al., 2015; Rafiq et al., 2015)	249,123	23.53	IG: 62; Vine: 68
D_{QA}	(Van Hee et al., 2018; Reynolds et al., 2011)	129,501	4.60	Ask.fm: 126; F.me: 153
D_{MB}	(Xu et al., 2012; Salawu et al., 2021)	12,310	6.51	(Xu et al., 2012): 128; (Salawu et al., 2021): 138

Table 6.1: Overview of datasets used for cyberbullying detection, including the number of sentences, proportion of bullying labels, and average sentence length (in tokens) per platform.

Data Imbalance: All datasets exhibit a pronounced imbalance, heavily skewed toward the `(0) on-hurtful` class, as shown in Figure 6.1. To address class imbalance, some techniques such as random under-sampling (Prusa et al., 2015), involve discarding randomly selected `non-hurtful` instances. Whereas, techniques such as synthetic minority oversampling (SMOTE) (Chawla et al., 2002) generate synthetic minority (e.g., `hurtful`) sample embedding representations (x_{new}) by linearly interpolating[85] embedding vector representations between an existing minority sample (x_i) and one of its nearest Neighbors (x_j). Despite this, SMOTE generates non-sequential token representations (words) in a sequence (sentence) through interpolated embedding vector, leading to altering the

[85] General form of linear interpolation: $x_{new} = x_i + \lambda \cdot (x_j - x_i)$, where $\lambda \in [0, 1]$ (Chawla et al., 2002)

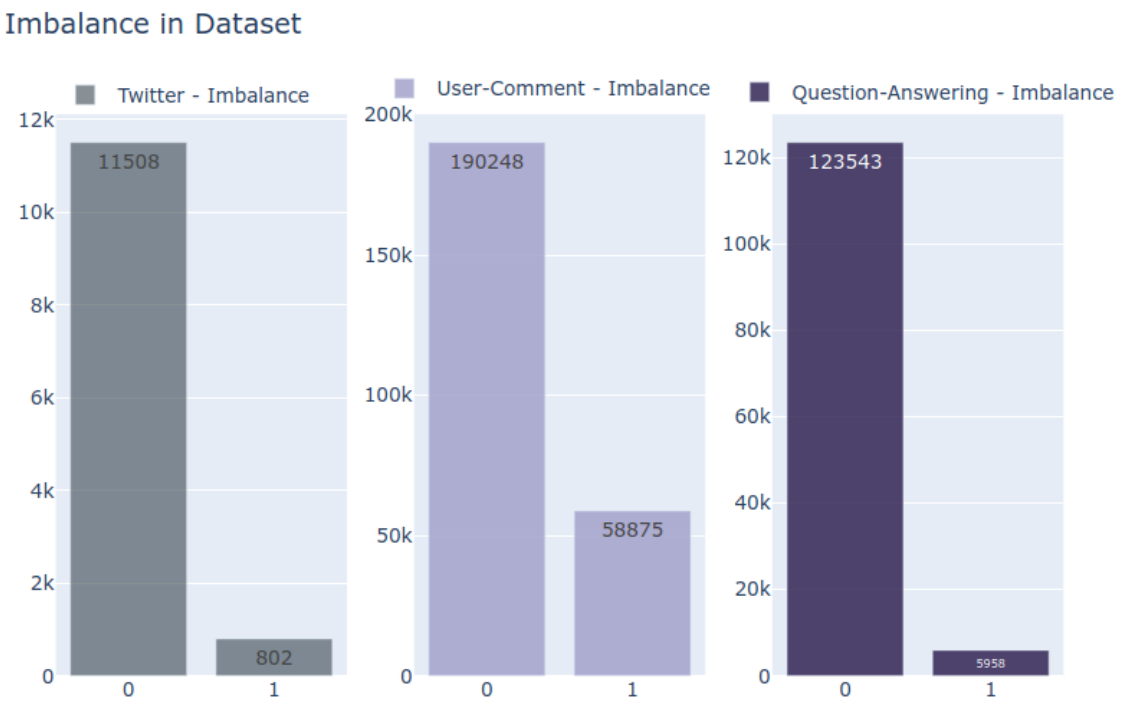


Figure 6.1: Imbalance in Dataset (0: `not-hurtful`; 1: `hurtful`)

structure of the sentence. Both random-undersampling and SMOTE are excluded as they risk either randomly removing sentences or altering the structure of the sentences. To avoid this, we employed random over-sampling (Fernández et al., 2018), which balances the dataset by randomly duplicating examples from the minority class, thereby mitigating class imbalance without altering sentence sequences.

Data Pre-processing: In compliance with the General Data Protection Regulation (GDPR) (Council of European Union, 2016), all datasets were fully anonymised and normalised to remove personally identifiable information (PII)[86] using GATE Cloud (Tablan et al., 2013). Named entities were extracted from the text using the TwitIE API (Bontcheva et al., 2013). Additionally, the following preprocessing steps were applied across datasets: (a) removal of URLs, user mentions, and non-ASCII characters; (b) elimination of retweet (RT) markers for Twitter datasets; (c) conversion of all text to lowercase; and (d) expansion of textual contractions to their formal forms (e.g., *can't* → *cannot*).

[86] Refers to Personally Identifiable Information

Classification Systems: For benchmarking binary `hurtful` vs `not-hurtful` classification as a pre-cursor to detect CB, we compare: (i) traditional machine learning techniques as *traditional baselines*; (ii) deep learning techniques as *neural baselines*; and (iii) transfer learning via pre-trained language models (LMs).

(i) Traditional Baseline: Building on prior work discussed in Section 2.4.2 Chapter 2, Support Vector Machines (SVM), proposed by (Hearst et al., 1998), was selected as a traditional baseline. When textual data is converted into numerical feature representations (such as TF-IDF vectors (Yun-tao et al., 2005)), SVMs aim to identify an optimal hyperplane within this feature space that best separates documents belonging to different classes. SVM does so by maximising the margin between the nearest data points (support vectors) of each class.

(ii) Neural Baseline: Building on prior work discussed in Section Section 2.4.2 Chapter 2, we leverage Bi-directional Long Short Term Memory (Bi-LSTM) (Huang et al., 2015) as the neural baseline. Bi-LSTM processes a sentence in both forward and backward directions, allowing the model to incorporate context from both preceding and succeeding words before making a final classification.

(iii) Pre-trained LMs: As discussed in Section 2.4.2 Chapter 2.4.2, transfer learning techniques (Ruder et al., 2019) with encoder-based transformers has resulted in its applicability to different down-stream tasks. For this task, we leveraging the following two pre-trained LMs:

1. **BERT**: As discussed in Section 2.4.2 Chapter 2, the $BERT_{base-uncased}$ model (Devlin et al., 2019), consists of 12 transformer encoder layers and was pre-trained on two unsupervised tasks: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP), using a very large corpus of unlabelled English text (e.g., BookCorpus (Zhu et al., 2015)).
2. **HateBERT**: Proposed by (Caselli et al., 2020) as a standard $BERT_{base-uncased}$ model, which is further trained specifically for MLM task on comments from Reddit's banned communities (RAL-E). This further pre-training is an efficient strategy for adapting Pre-trained Language Models (PLMs) to domain-specific

tasks such as hate-speech or abusive language detection by specifically exposing it to the linguistic patterns, vocabulary, and slang prevalent in toxic online discourse.

Both these LMs are fine-tuned by adding a fully connected classification layer directly on top, i.e., classification head (See Section 2.4.2 Chapter 2). This layer receives the contextualised token embeddings from the transformer encoder as input, subsequently generating the logits (raw non-normalised class scores) for each class. These logits are then utilised to compute the loss (e.g., binary cross-entropy See Equation 2.21 Chapter 2) required for model training.

Hyper-parameters and Embeddings: All classification systems are optimised with different hyper-parameters.

- **Hate-BERT** & **BERT**: Experiments[87] employed the HuggingFace Transformers library (Wolf et al., 2019) and the HateBERT implementation (Caselli et al., 2021). To adapt the pre-trained models for sequence classification, a fully connected layer was added on top of the transformer encoder to map the contextualised token embeddings to class logits. These logits were then used with a loss function, i.e., binary cross-entropy (Equation 2.21), to train the models. Both BERT and Hate-BERT were fine-tuned separately on each of the three merged datasets for 2–4 epochs, with maximum sequence lengths set between 128 and 256 tokens depending on the dataset. The classification layer used ReLU activation (Equation 2.16), and optimisation was performed with the adaptive moment estimation (Adam) optimiser (Kingma and Ba, 2015) with learning rates[88] ranging from 0.1, 0.001, $1e^{-5}$ to $5e^{-5}$.

- **Bi-LSTM**: Following prior work, the Bi-LSTM was trained with 50-dimensional Glove (Pennington et al., 2014) embeddings and this model was incorporated with a recurrent dropout rate of 0.2. Its fully connected layer consisted of 256 neurons with ReLU activation, while the output layer used softmax (Equation 2.10) for binary classification. The model was also trained using binary cross-entropy (BCE) loss. .

[87] All experiments were conducted on a local system with a 16-core CPU, 16GB RAM, and NVIDIA RTX 2070 GPU (8GB GPU memory)

[88] See Appendix A.14

- **Linear SVM**: Following the methodology of prior studies (Van Hee et al., 2018), a linear Support Vector Machine (SVM) was used for classification. To address the imbalanced nature of the dataset (see Figure 6.1), a cost-sensitive configuration was applied by setting `class_weight='balanced'`, which re-weights the penalty parameter C according to the inverse class frequencies to penalise mis-classifications of the minority class (e.g., `hurtful`) more strongly. The model was trained on document-level TF-IDF features (Yun-tao et al., 2005), with the regularisation parameter C fine-tuned via grid search with five-fold cross-validation over the range $[0.01, 0.1, 1, 10, 100]$. The best-performing configuration was then used to retrain the final model on the full training set.

Evaluation Strategy: For benchmarking, transformer-based models were evaluated alongside Bi-LSTM with 50-dimensional GloVe embeddings and a traditional SVM with TF-IDF features. Model performance was measured using macro average F1 scores (Chinchor and Sundheim, 1993) (Equation 2.14) for both positive (`hurtful`) and negative (`not-hurtful`) classes, as well as overall F1 scores. The F1 score accounts for both precision and recall and represents their average, providing a balanced metric for evaluating classification performance.

6.1.3 Experiment Setup - Evaluating Translation Strategies

The datasets developed in Chapter 5: (i) Human Translated (HT) version of (Sprugnoli et al., 2018) ($\mathcal{D}^{HT}_{IT\rightarrow EN}$) (§5.1.1), (ii) Machine Translated (MT) version of (Sprugnoli et al., 2018) ($\mathcal{D}^{MT}_{IT\rightarrow EN}$) (§5.1.2) are translated version of the Italian youth role-played cyberbullying dataset. As discussed earlier Section 5.1.1 and Section 5.1.2 for **extrinsic** evaluation we investigate potential contextual loss, i.e., whether both human and machine translation of Italian texts (source) and their English translations (target) behave similar. Corazza et al. (2019) evaluated cross-platform hate-speech detection using the Italian youth role-played WhatsApp corpus by (Sprugnoli et al., 2018) alongside other publicly available datasets from social media platforms such as Twitter and Instagram in Italian language. Although Corazza et al. (2019) did not specify how the WhatsApp corpus was adapted for binary classification, correspondence with the authors clarified this process Corazza

(2021). We were informed that a message is labelled as *hate* only if it **does not** include any of the following: (i) *Defence* in entity-type, (ii) *Encouragement-to-the-Harasser* in entity-type, (iii) flagged as *non-offensive*, or (iv) *Victim* in role-type. Applying this label consolidation strategy, which merges multiple original labels into a binary classification of `hate` vs `non-hate`, left only **741** sentences labelled as `hate`. However, we note that collapsing the original labels into a binary scheme **does not strictly** capture the presence or absence of "*hateful*" behaviour. Therefore, we frame the classification task more broadly as `hurtful` versus `not-hurtful`.

Corazza et al. (2019) leveraged a Gated Recurrent Unit (GRU) based classifier for the binary classification. In addition to replicating this classifier, we also conducted additional classification experiments using other classification techniques leveraged in prior cyberbullying detection research. To establish a baseline for model behaviour across the different datasets, we applied traditional supervised machine learning, ensemble learning, and deep learning methods. For feature representation, we leveraged pre-trained Twitter-based fastText embeddings[89] as well as embeddings generated using the Term Frequency–Inverse Document Frequency (TF-IDF) technique (Yun-tao et al., 2005). By comparing the performance of classifiers on the Italian source versus the $\mathcal{D}^{HT}_{IT\rightarrow EN}$ (§5.1.1) and $\mathcal{D}^{MT}_{IT\rightarrow EN}$ (§5.1.2) English texts, we aim to quantify the impact of translation on **binary hurtfulness** detection and to assess whether any semantic or contextual nuances are lost during the translation process. This enables a more nuanced understanding of how cross-lingual differences might affect automated content moderation systems.

Dataset Split: Table 6.2 summarises the dataset splits used for classifier development and evaluation. The data is divided into training, validation, hold-out test, and Scenario-C test sets, with each set containing both `hurtful` (1) and `not-hurtful` (0) instances. The Scenario-C test sets consist of the same conversations previously used for human assessment and annotation by two experts in adolescent bullying behaviour and one adolescent, as described in Section 5.1.1, Chapter 5.

The split, which results in a total of 2,192 samples, follows the label distribution of 30-40% for positive classes per (Corazza, 2021), ensuring consistency with prior work.

[89] `https://fasttext.cc/docs/en/crawl-vectors.html`

Set	**Size**	**Hurtful (1)**	**Not Hurtful (0)**
Training	1320	426	894
Validation	330	141	189
Hold-out Test	412	128	284
Scenario-C Test	130	46	84
Total	**2192**	741	1451

Table 6.2: Dataset Split

Data Pre-processing: Pre-processing steps were as follows: (i) all text was converted to lower-case; (ii) emojis were transcribed into their official English plain-text descriptions; (iii) an additional corpus was created where all emojis were scrubbed.

Classification Systems: In addition to replicating (Corazza et al., 2019)'s GRU classifier to assess translated data usefulness in development of CB moderation systems, we build on prior work and include the classification system discussed in previous Section 6.1.2, i.e., SVM, Bi-LSTM, and $\text{BERT}_{\text{base-uncased}}$. We also include other baseline ensemble learning techniques for fair comparison.

Replicating (Corazza et al., 2019)'s system: For the EVALITA 2018 shared task [90] on Hate-Speech Detection (Bosco et al., 2018), Corazza et al. (2019)'s best performing model was a GRU-based classifier trained with fastText Twitter embeddings[91] for Italian language. Although the authors' trained weights were not publicly available, we successfully reconstructed their classification system based on the detailed fine-tuning strategies described in their research article and subsequent email correspondence (Corazza, 2021). To specifically evaluate cross-platform generalisation using (Sprugnoli et al., 2018)'s WhatsApp corpus, their GRU-based model was trained under the following configurations: (i) Recurrent Layer size: 200; (ii) Batch size: 32; (iii) Embedding size: 300; (iv) Epochs: 5; (v) Loss Function: Binary Cross Entropy; (vi) Activation Function: Sigmoid (Equation 2.15); (vii) Lambda regularisation: 0.1 (Equation 2.20).

Traditional Ensemble Baselines: Expanding our scope of traditional machine learning baselines discussed in §6.1.2, we include ensemble learning techniques such as

[90] `https://www.evalita.it/campaigns/evalita-2018/`
[91] `https://fasttext.cc/docs/en/crawl-vectors.html`

bagging (boot-strapping) (e.g., Decision Tree and Random Forest), and boosting-based methods (e.g., XGBoost). Ensemble learning is a technique within machine learning, where multiple models (weak learners) are trained and combined to produce a single output. Both training and a combination of weak learners are typically done using two methods bagging (bootstrap aggregation) or boosting.

- **Bagging**: Also known as *bootstrap aggregation*, is an ensemble learning technique that trains multiple models on randomly sampled subsets of the training data and aggregates their predictions to reduce variance.
  - **Decision Tree (DT)**: Proposed by (Quinlan, 1986), DT is a non-parametric supervised learning algorithm that recursively partitions the feature space using decision rules based on feature thresholds to maximise class purity.
  - **Random Forest (RF)**: Proposed by (Breiman, 2001), RF extends DT by constructing multiple trees using random subsets of data and features, and aggregates their outputs using majority voting (classification) or averaging (regression).
- **Boosting**: A sequential ensemble learning approach in which models are trained iteratively, with each subsequent learner focusing on minimising the errors of previous models.
  - **Extreme Gradient Boosting (XGBoost):** Proposed by (Chen and Guestrin, 2016), XGBoost is a gradient boosting framework that constructs an ensemble of decision trees sequentially by optimising a differentiable loss function with regularisation to improve generalisation.

Hyper-parameters and Embeddings: Hyper-parameters were optimised using grid search[92] to identify configurations that maximise classification performance, while baseline models were first trained using default library settings. For **DT**, we tuned `max_depth` (5, 10, 20) and `min_samples_leaf` (1, 2, 4). For **RF**, we tuned `n_estimators` (100, 200, 500), `max_features` ('sqrt', 'log2'), and `max_depth`. For **XGBoost**, we tuned

[92] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

`eta` (0.01, 0.1, 0.2), `max_depth` (3, 6, 10), `subsample` (0.7, 1.0), and L2 regularisation (`lambda` = 1, 5, 10). For **SVM**, we tuned the regularisation parameter (`C` = 0.01, 0.1, 1, 10, 100) with L2 penalty and `class_weight='balanced'`. Text representations were generated using **TF-IDF** and pre-trained **fastText** embeddings[93]. TF-IDF provided sparse numerical features, while fastText produced dense semantic vectors. Both representations were used as inputs for traditional machine learning models (with document-level vector averaging for fastText), whereas neural models utilised fastText embeddings as sequential inputs. Hyper-parameters for **Bi-LSTM** and **BERT** are described in §6.1.2.

Evaluation Strategy: Evaluation was conducted on both Scenario-C with additional labels (from expert and youth annotation discussed in §5.1.1) and the hold-out test set (See Table 6.2 to ensure robust assessment across different label configurations and unseen data. We report model performance using the F1-score in line with the previous experiment setup discussed in §6.1.2.

6.1.4 Results and Discussion

Cross-platform Benchmarking Results

The results of cross-platform benchmarking discussed in Section 6.1.2 are illustrated in Table 6.3. The results demonstrate distinct performance patterns across traditional, neural, and transformer-based architectures. Overall, the **HateBERT** (Caselli et al., 2021) exhibits the highest F1-scores across both in-domain and cross-domain evaluations, underscoring the benefits of domain-adaptive pre-training for detecting ***hurtfulness*** as a precursor to cyberbullying detection.

(i) Traditional Baseline: SVM while achieving competitive within-platform accuracy (e.g., 0.73 Avg. F1 on D_{UC}), show a significant degradation under cross-platform settings (dropping to as low as 0.42–0.52 Avg F1). This degradation reflects SVM's limited ability to generalise across domains due to its reliance on static lexical features, such as TF-IDF fails to capture contextual or semantic variability across social media platforms.

[93] `https://fasttext.cc/docs/en/crawl-vectors.html`

Model	**Train-set**	**Test-set**	**Hurtful F-1**	**Not-Hurtful F1**	**Avg F1**
SVM	D_{UC}	D_{UC}	**0.75**	0.71	**0.73**
		D_{QA}	**0.56**	0.58	**0.57**
		D_{MB}	0.51	0.51	0.51
	D_{QA}	D_{UC}	0.34	0.50	0.42
		D_{QA}	0.52	0.54	0.53
		D_{MB}	0.52	0.52	0.52
	D_{MB}	D_{UC}	0.28	0.50	0.39
		D_{QA}	0.48	0.50	0.49
		D_{MB}	0.54	0.54	0.54
Bi-LSTM	D_{UC}	D_{UC}	0.68	0.70	0.69
		D_{QA}	0.30	0.50	0.40
		D_{MB}	0.51	0.51	0.51
	D_{QA}	D_{UC}	0.58	0.60	0.59
		D_{QA}	0.69	0.67	0.68
		D_{MB}	0.52	0.54	0.53
	D_{MB}	D_{UC}	**0.63**	0.61	**0.62**
		D_{QA}	0.61	0.57	0.59
		D_{MB}	**0.69**	0.67	**0.68**
BERT	D_{UC}	D_{UC}	0.65	0.77	0.71
		D_{QA}	0.54	0.58	0.56
		D_{MB}	0.58	0.60	0.59
	D_{QA}	D_{UC}	0.48	0.50	0.49
		D_{QA}	0.62	0.68	0.65
		D_{MB}	0.57	0.61	0.59
	D_{MB}	D_{UC}	**0.54**	0.52	**0.53**
		D_{QA}	0.63	0.61	0.62
		D_{MB}	**0.75**	0.79	**0.77**
Hate-BERT	D_{UC}	D_{UC}	**0.68**	**0.84**	**0.76**
		D_{QA}	0.67	0.59	0.63
		D_{MB}	**0.65**	0.71	**0.68**
	D_{QA}	D_{UC}	**0.73**	0.65	**0.69**
		D_{QA}	**0.73**	0.73	**0.73**
		D_{MB}	0.61	0.65	0.63
	D_{MB}	D_{UC}	**0.71**	0.65	**0.68**
		D_{QA}	0.69	0.65	0.67
		D_{MB}	**0.78**	0.84	**0.81**

Table 6.3: Results for Cross-platform Benchmarking (§6.1.2)

(ii) Neural Baseline: Bi-LSTM benefiting from its bi-directional sequential modelling of text, outperforms SVMs in several cases (e.g., 0.68 Avg. F1 on D_{MB}), but

still suffers from domain sensitivity. Their cross-platform scores (0.59-0.62) suggest that while recurrent models can model sequence dependencies better than static classifiers, they remain vulnerable to domain shifts in slang, user behaviour, and discourse patterns.

(iii) Pre-trained LMs:

- **Fine-tuned BERT** achieves a notable improvement in both within-domain and cross-domain tasks compared to Bi-LSTM and SVM. When trained on D_{MB}, BERT attains an average F1 of 0.77 within-domain and maintains a reasonable 0.59 on D_{QA}, indicating that contextual embeddings partially bridge lexical gaps between platforms. However, the cross-domain penalty persists, showing BERT's limitations when applied to heterogeneous online environments.
- **HateBERT** designed with additional pre-training on abusive and hateful language corpora, demonstrates superior robustness. When fine-tuned on D_{MB} it not only achieves the highest within-platform performance (0.81 Avg. F1) but also sustains strong results on unseen platforms such as D_{UC} (0.68 Avg. F1). Similarly, when trained on D_{QA}, HateBERT maintains balanced F1-scores ($\approx$0.73) across both source and target datasets. This indicates that the inductive bias introduced through hate-speech-focused pre-training provides transferable linguistic features, enhancing cross-platform generalisation.

These findings collectively suggest that domain-adaptive language modelling, such as fine-tuning Hate-BERT, significantly enhances the resilience of binary `hurtful` detection classifiers, making them efficient for triaging cyberbullying detection. In contrast, conventional architectures such as SVM and Bi-LSTM exhibit steep performance declines under domain shift. Nonetheless, even Hate-BERT experiences a measurable cross-platform drop ($\approx$0.08–0.10 F1), pointing to the persistent challenge of domain adaptation in social media abuse detection. These comprehensive experiments suggest that fine-tuning language models across multiple social media platforms is a critical first step toward developing platform-agnostic cyberbullying detection mechanisms. Moreover, the results indicate that further research is necessary to enhance cross-platform generalisation and improve

the robustness of such detection systems.

Classification Results with Translated Data

The results of the experiments described in Section 6.1.3 are presented in two parts. **First**, we report the outcomes of replicating (Corazza et al., 2019)'s GRU-based classifier, which allows us to assess the utility of the Italian-to-English translated datasets discussed in Sections 5.1.1 and 5.1.2 Chapter 5. These findings constitute the **extrinsic evaluation** of translation quality. **Second**, we discuss the binary classification results for all models discussed in Section 6.1.3 across both the held-out test-set and Scenario-C test-set.

Results of Replicating (Corazza et al., 2019)'s systems: Table 6.4 compares the classification performance between the original experiment reported by Corazza et al. (2019) and our replication using the same GRU-based architecture with Twitter-FastText embeddings on the raw Italian corpus. The replication results show only marginal differences across all evaluation metrics. Specifically, the macro-averaged F1-score decreased slightly from 0.754 in the original study to 0.735 in our replication, representing a minor drop of approximately 1.9%. Similarly, class-wise F1-scores for non-hate and hate categories show a negligible variation (`no hate`: 0.814 → 0.79 and `hate`: 0.694 → 0.68, respectively). Such minimal discrepancies can be attributed to implementation-specific factors such as the absence of pre-trained model weights, random initialisation variance, or potential minor preprocessing differences in the text normalisation pipeline. Overall, these results confirm that the GRU model architecture is reproducible with consistent classification behaviour across independent implementations.

	Embeddings	**F1 not-hurtful**	**F1 hurtful**	**F1 Macro Average (Avg)**
Original Scores	Twitter-FastText	0.814	0.694	**0.754**
Replication Scores	Twitter-FastText	0.79	0.68	**0.735**

Table 6.4: Original and Replicated Classification Results of training and testing GRU on the Italian Corpus

Binary Classification Results with both $\mathcal{D}^{HT}_{IT\rightarrow EN}$ and $\mathcal{D}^{MT}_{IT\rightarrow EN}$ corpus: We first discuss the results on the held-out test-set, followed by the results on the Scenario-C test set.

Results on Hold-out Test-set: Table 6.5 reports the binary classification performance on Human Translated ($\mathcal{D}^{HT}_{IT\rightarrow EN}$) and Machine Translated ($\mathcal{D}^{MT}_{IT\rightarrow EN}$) English corpora. Across both datasets, $\text{BERT}_{\text{base}-\text{uncased}}$ consistently achieves the highest macro-averaged F1-scores, demonstrating the advantage of pre-trained Transformer-based contextual embeddings over both classical ML (SVM, Decision Tree, Random Forest, XGBoost) and sequential DL models (Bi-LSTM, GRU).

On the **HT corpus** ($\mathcal{D}^{HT}_{IT\rightarrow EN}$), BERT reaches 0.83 Avg F1, substantially outperforming other models (next best Bi-LSTM/GRU at 0.73). This indicates that the contextualised representations captured by BERT can effectively model subtle semantic cues indicative of hurtful language, especially when augmented with transcribed emojis, which provide important paralinguistic and sentiment signals. Among DL baselines, GRU performs relatively well (0.73), but still underperforms BERT due to limited capacity in capturing long-range dependencies compared to Transformer architectures.

On the **MT corpus** ($\mathcal{D}^{MT}_{IT\rightarrow EN}$), all models experience a slight performance drop, reflecting challenges introduced by machine-translated texts such as lexical inconsistencies, mistranslations, or semantic shifts. Nevertheless, BERT still achieves 0.76 Avg F1, outperforming GRU (0.67) and Bi-LSTM (0.65). The smaller gap between BERT and sequential DL models on MT data suggests that translation noise disproportionately affects contextual embeddings, yet BERT remains robust enough to maintain strong cross-lingual performance.

The **traditional machine learning** models (SVM, Decision Tree, Random Forest, XGBoost) perform moderately well on the `not-hurtful` class but struggle on the `hurtful` class, with F1 scores dropping below 0.60 on both corpora. This underscores the limitations of feature-based approaches (TF-IDF or fastText embeddings) in capturing nuanced, context-dependent hate speech expressions, especially in cross-lingual settings. Overall, these findings indicate that for cross-lingual hurtful detection, fine-tuned Transformer models augmented with emoji representations provide a robust baseline, surpassing both traditional ML and sequential DL approaches, while maintaining resilience to variations

introduced by different translation methods.

Dataset	**Model**	**Emojis**	**Hold-out Test-set**		
			F1 `harm`	F1 `no harm`	F1 Macro Avg.
$\mathcal{D}^{HT}_{IT\rightarrow EN}$	$SVM_{\text{TF-IDF}}$	Removed	0.84	0.42	0.63
	Decision $Tree_{\text{fastText}}$	Transcribed	0.81	0.56	0.69
	Random $Forest_{\text{fastText}}$	Transcribed	0.82	0.57	0.7
	XG-$Boost_{\text{fastText}}$	Transcribed	0.81	0.56	0.69
	Bi-$LSTM_{\text{fastText}}$	Transcribed	0.84	0.62	0.73
	GRU_{fastText}	Transcribed	0.77	0.69	0.73
	$BERT_{\text{base-uncased}}$	Transcribed	**0.84**	**0.81**	**0.83**
$\mathcal{D}^{MT}_{IT\rightarrow EN}$	$SVM_{\text{TF-IDF}}$	Removed	0.80	0.36	0.58
	Decision $Tree_{\text{fastText}}$	Transcribed	0.81	0.34	0.57
	Random $Forest_{\text{fastText}}$	Transcribed	0.81	0.45	0.63
	XG-$Boost_{\text{fastText}}$	Transcribed	0.79	0.46	0.63
	Bi-$LSTM_{\text{fastText}}$	Transcribed	0.81	0.47	0.65
	GRU_{fastText}	Transcribed	0.75	0.59	0.67
	$BERT_{\text{base-uncased}}$	Transcribed	**0.78**	**0.74**	**0.76**

Table 6.5: F1 scores for classification systems on both Human Translated (HT §5.1.1) and Machine Translated (MT §5.1.2) datasets discussed in §6.1.3 on **Hold-out Test set**

Results on Youth and Expert Labelled Scenario-C: The rationale for the domain expert and youth annotations discussed in §5.1.1 in Chapter 5 aids in engineering additional labels for `Scenario-C` and as discussed in §6.1.3, we test all the ML and DL models on these additional binary labels. The Table 6.6 shows that across both Human Translated ($\mathcal{D}^{HT}_{IT\rightarrow EN}$) and Machine Translated ($\mathcal{D}^{MT}_{IT\rightarrow EN}$) English corpora, $BERT_{\text{base-uncased}}$ model outperforms all the other ML and DL models with a F1 macro-average score of **0.81** and **0.75** on the $\mathcal{D}^{HT}_{IT\rightarrow EN}$ and $\mathcal{D}^{MT}_{IT\rightarrow EN}$ English corpus respectively.

On the **HT corpus** ($\mathcal{D}^{HT}_{IT\rightarrow EN}$), the gap between BERT and sequential DL models is notable: Bi-LSTM and GRU achieve 0.79–0.81 Avg. F1, while traditional ML classifiers such as Random Forest, SVM, and Decision Tree score lower (0.65–0.74). This demon-

strates that contextual embeddings capture subtle interdependencies between words and transcribed emojis, which are critical for identifying nuanced harmful content.

The **MT corpus** ($\mathcal{D}^{MT}_{IT\rightarrow EN}$), shows a small but measurable drop in performance across all models, likely due to translation-induced noise, such as lexical shifts or semantic distortions. Despite this, BERT maintains a macro-average F1 of 0.75, outperforming both GRU (0.70) and Bi-LSTM (0.74), highlighting its robustness to variations introduced by machine translation. Interestingly, traditional ML models perform relatively better on the HT corpus than on MT, indicating that feature-based approaches are sensitive to translated texts. Sequential DL models like GRU and Bi-LSTM are more resilient but still fall short of BERT's performance, underscoring the benefits of pre-trained Transformers for expert-labelled fine-grained classification tasks. Overall, these results suggest that fine-tuned Transformers with emoji integration (see pre-processing steps Section 6.1.3) provide a strong and reliable baseline for expert-informed harm detection scenarios, outperforming classical ML and sequential DL models while showing resilience to translation variability.

6.2 Moving Towards Explainable Cyberbullying Detection

6.2.1 Related Work

The self-attention mechanism introduced by (Vaswani et al., 2017), and used in BERT and other models discussed in Section 2.4 of Chapter 2, is often described as allowing models to "attend to" or "look at" specific tokens during processing (Mullenbach et al., 2018; Xie et al., 2017a). This interpretation has been fundamentally questioned in recent debates (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bastings and Filippova, 2020). The self-attention weights do not always correspond to true feature importance or decision causality (Serrano and Smith, 2019), and the identifiability of attention weights is problematic for longer sequences or deep layers (Brunner et al., 2019). While attention may capture linguistic semantics (such as parts of speech, pronouns, noun-phrase links) (Clark et al., 2019; Howard and Ruder, 2018) and support transfer learning by leveraging contextual embedding generalisations (e.g., fine-tuning pre-trained transformer models

Dataset	**Model**	**Emojis**	**Scenario-C Expert Labels**		
			F1 `no harm`	F1 `harm`	F1 Macro Avg.
$\mathcal{D}^{HT}_{IT \rightarrow EN}$	$SVM_{fastText}$	Transcribed	0.74	0.64	0.69
	$Decision\ Tree_{TF\text{-}IDF}$	Transcribed	0.72	0.59	0.65
	$Random\ Forest_{fastText}$	Transcribed	0.8	0.65	0.74
	$XG\text{-}Boost_{fastText}$	Transcribed	0.74	0.76	0.71
	$Bi\text{-}LSTM_{fastText}$	Transcribed	0.85	0.74	0.79
	$GRU_{fastText}$	Transcribed	0.83	0.79	0.81
	$BERT_{base\text{-}uncased}$	Transcribed	**0.86**	**0.79**	**0.82**
$\mathcal{D}^{MT}_{IT \rightarrow EN}$	$SVM_{TF\text{-}IDF}$	Removed	0.68	0.51	0.6
	$Decision\ Tree_{TF\text{-}IDF}$	Transcribed	0.69	0.54	0.62
	$Random\ Forest_{TF\text{-}IDF}$	Transcribed	0.71	0.49	0.6
	$XG\text{-}Boost_{fastText}$	Transcribed	0.67	0.47	0.57
	$Bi\text{-}LSTM_{fastText}$	Transcribed	0.79	0.7	0.74
	$GRU_{fastText}$	Transcribed	0.77	0.64	0.7
	$BERT_{base\text{-}uncased}$	Transcribed	**0.76**	**0.74**	**0.75**

Table 6.6: F1 scores for classification systems on both Human Translated (HT §5.1.1) and Machine Translated (MT §5.1.2) datasets discussed in §6.1.3 on **Scenario-C** with Expert labels

across domains (Devlin et al., 2019; Howard and Ruder, 2018; Ruder et al., 2019)), relying on raw attention as an interpretation mechanism remains contested (Serrano and Smith, 2019; Brunner et al., 2019).

To address this gap, gradient-based attribution methods such as Integrated Gradients (IG) by (Sundararajan et al., 2017) have been proposed. IG estimates feature importance by comparing the model's predictions for a given input x with those for a baseline input x', which represents a point of minimal information or neutral reference (e.g., zero vectors for neutral reference sentence). It computes the average gradients of the model's output $F(x)$ with respect to the input features along a linear interpolation path, i.e., a straight-line transition from the baseline input (x') to the actual input (x), represented as $x(\alpha) = x' + \alpha(x - x')$, where $\alpha \in [0, 1]$. Sundararajan et al. (2017) define the contribution of each input feature i as:

$$\text{IG}_i(x) = (x_i - x'_i) \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

This formulation satisfies key interpretability axioms[94]: **Sensitivity**, ensuring features that alter predictions receive non-zero attribution, and **Implementation Invariance**, guaranteeing consistent attributions across functionally equivalent models. Further details and examples are in Appendix A.15.

From Harm Detection to Cyberbullying - The Interpretability Gap: As discussed in Sections 6.1.2 and 6.1.3, we leveraged attention-based pre-trained LMs such as $\text{BERT}_{\text{base-uncased}}$ and HateBERT, which outperform traditional baselines (Section 6.1.4). These deep multi-layer models assign attention weights to tokens, which have been used for interpretability (Zhang et al., 2019; Adadi and Berrada, 2018; Sundararajan et al., 2017), supported by visualisation tools (Vig, 2019a; Vig and Belinkov, 2019).

However, the notion that "*attention* explains model decisions" remains debated (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bastings and Filippova, 2020). Elsafoury et al. (2021) explored BERT's interpretability for cyberbullying detection via POS tags on datasets originally labelled for **hate speech** (Waseem and Hovy, 2016), **personal attacks** (Wulczyn et al., 2017), and insults[95]. While valuable for reproducibility, their

[94] Axioms are fundamental principles that define how an attribution method should behave to ensure it is theoretically sound and reliable.

[95] `https://www.kaggle.com/competitions/detecting-insults-in-social-commentary`

study exhibits key limitations motivating gradient-based methods such as *Integrated Gradients*, these include: (i) **Cyberbullying Conflation**: The datasets conflate cyberbullying with hate speech and personal attacks, missing context-specific linguistic patterns and relational nuances (Section 2.4.2, 2.4.3); (ii) **Shallow Decision Analysis**: Token attributions across positive and negative instances are not compared, limiting understanding of features driving predictions; (iii) **Limited Model Scope**: Only BERT is analysed, leaving interpretations of domain-adapted models like HateBERT or other harm-detection architectures unexamined.

These gaps highlight a critical question: *Can attention weights from transformer LMs fine-tuned on authentic cyberbullying datasets reliably support interpretable detection?* Beyond POS analysis, we also examine sentiment scores from VADER (Hutto and Gilbert, 2014), hypothesising:

> "*Gradient-based feature importance scores will reveal whether the model attends more to adjectives, verbs, or words with strong positive or negative sentiment identifying hurtful content in cyberbullying contexts*"

6.2.2 Experiment Setup

Datasets: Extending the cross-domain experiments described in §6.1.2 and summarised in Table 6.1, we extend our evaluation by incorporating the dataset introduced by (Bretschneider and Peters, 2016). This dataset consists of posts from multiplayer online gaming forums, including communities such as *World of Warcraft* and *League of Legends* (see Table 2.3). By including this dataset, we broaden the evaluation to gaming communities, which are characterised by highly domain-specific slang, abbreviations, and in-game references. This enables a more robust assessment of model performance on informal and context-dependent language, which is often distinct from language seen in other social media domains. As shown in Table 6.7, the proportion of hurtful, i.e., binary bullying instances, is consistently low across all seven datasets, reflecting the class imbalance typical of real-world social media data. Each of the four grouped datasets was partitioned into training, validation, and test sets, with 80% allocated to training and 10% each for validation and testing. The test set was further subdivided for interpretability analyses: 90% was used for standard model performance evaluation, while the remaining 10% was

reserved to analyse attention weights and gradient-based feature importance scores across model layers and attention heads. This separation ensures that interpretability analyses do not interfere with conventional evaluation metrics.

Platform-Type	**Study**	**% Hurtful**	**# of Sentences**	**# Sentences (Analysis)**
Question-Answering (D_{QA})	(Van Hee et al., 2018; Reynolds et al., 2011)	4.6	129,501	1,395
User-Comment (D_{UC})	(Hosseinmardi et al., 2015; Rafiq et al., 2015)	23.53	249,123	1,103
Micr-blogging (D_{MB})	(Xu et al., 2012; Salawu et al., 2021)	12,310	6.51	140
Gaming (D_G)	Bretschneider and Peters (2016)	2.3	34,229	342

Table 6.7: Dataset Description

Pre-trained Language Models: This study focuses on the interpretability of both general-purpose and domain-adapted language models. Specifically, we examine BERT (Devlin et al., 2019) as a general-purpose model and HateBERT (Caselli et al., 2021) as a model pre-trained on abusive language in online discussions. This dual focus allows us to compare whether domain adaptation improves performance and interpretability on cyberbullying detection tasks (see §6.1.2).

Data Pre-processing, Imbalance Handling, and Hyper-parameters: We apply the same preprocessing pipeline, class imbalance handling techniques, and hyper-parameter settings as outlined in §6.1.2.

Collecting Part-of-Speech (POS) Tags and Sentiment Scores: To formally evaluate our hypothesis about linguistic and affective patterns in cyberbullying, we enriched our datasets with POS tags and sentiment scores. The datasets were originally collected from multiple platforms, curated to align with standard definitions of cyberbullying, and contained informal, highly context-dependent language, but lacked linguistic

and sentiment annotations. For analysis, we applied off-the-shelf NLP tools: spaCy[96] (Honnibal and Montani, 2017) to extract POS tags and VADER (Hutto and Gilbert, 2014) to compute sentiment scores. These annotations were generated for the 10% subset of the test set reserved for interpretability analyses (See Table 6.7), allowing us to investigate how model decisions correlate with syntactic structures and sentiment cues.

Attention Weights and Gradient-Based Feature Importance: To understand if the attention weights (α Equation 2.9) of an encoder-only pre-trained transformer (e.g., HateBERT), when fine-tuned on domain-specific datasets, can help interpret or explain *hurtfulness* in cyberbullying contexts, we extend prior work on transformer interpretability by (Jain and Wallace, 2019; Serrano and Smith, 2019; Vashishth et al., 2019; Sun and Lu, 2020). We first extracted the attention weights of the fine-tuned language models on 10% of the test set reserved specifically for interpretability analyses.

As discussed in the related work (§6.2.1), attention weights are often complemented with gradient-based feature importance scores to provide a more rigorous estimate of the relevance of individual tokens, especially those with known linguistic or semantic roles (Clark et al., 2019; Serrano and Smith, 2019; Sun and Lu, 2020). To operationalise this as discussed earlier, we leveraged the Integrated Gradients (IG) method using Captum's `Layer Integrated Gradients` (Kokhlikyan et al., 2020). This was applied to the same reserved 10% of the test set, providing token-level importance scores complementary to attention weights. Given that the pre-trained language models used in this study consist of 12 transformer layers, each with 12 attention heads, we computed mean attention weights for each head across all layers. Extending the methodology proposed by (Jain and Wallace, 2019), we calculated Pearson's correlation coefficient (PCC) to quantify the linear relationship between mean attention weights and mean gradient-based feature importance scores.

Additionally, to provide a more fine-grained interpretability analysis, we:

1. Computed mean attention weights and gradient-based feature importance scores for every token grouped by its part-of-speech (POS) tag within the reserved test set.

2. Examined tokens with strongly negative sentiment, observing both their attention

[96] `https://spacy.io/usage/linguistic-features#pos-tagging`

weights and gradient-based importance scores to determine whether the model prioritises these tokens when identifying cyberbullying content.

This dual analysis allows us to (i) compare attention and gradient-based measures, and (ii) assess whether model decisions align with linguistic intuition, particularly the importance of adjectives, verbs, and sentiment-relevant tokens.

6.2.3 Results and Discussion

We begin by presenting the classification results for the two pre-trained *encoder-only* language models fine-tuned for cross-platform *hurtful* detection in the context of cyberbullying. Followed with three interpretability analyses: (i) attention-weight distributions, (ii) layer-wise attention patterns for POS tags, and (iii) layer-wise attention patterns for negative-sentiment words.

Classification Results: Optimal configurations for each fine-tuned language model (LM) across all four platform-specific validation sets are summarised in Table A.17 in Appendix A.16. The validation accuracy revealed that the best-performing configurations vary slightly depending on both the model architecture and the dataset domain. In general, the optimal parameters include, `128`-token sequence length; `8-32` batch sizes; learning-rates: `1`e^{-5} to `5`e^{-5} and 2 epochs. These variations observed between datasets suggest that the linguistic and structural characteristics of each platform (e.g., length, tone, and syntactic complexity of messages) influence how the model generalises. Given the severe class imbalance across all datasets (see Table 6.7), F1-score was chosen as the main evaluation metric since it provides a balanced measure that accounts for both precision and recall. Table 6.8 summarises the classification performance of both BERT and HateBERT models across the **four** platform-specific datasets. The table reports F1-scores for the `hurtful` and `not-hurtful` classes, along with their macro-average. Several key observations can be drawn from these results:

- **HateBERT** consistently outperforms BERT across all datasets.

  The domain-adaptive pre-training of HateBERT on abusive and hate-related corpora leads to improved detection of bullying content. For example, in the Question-Answering dataset (D_{QA}), HateBERT achieves an average F1 of 0.73, outperforming

BERT's 0.65 by a notable margin. Similar improvements are observed in the User-Comment and Gaming datasets, where HateBERT yields higher F1-scores for both classes, demonstrating its robustness across diverse linguistic contexts.

- **Performance gains** are most significant in linguistically informal domains.

  The improvement margin is particularly evident in the Micro-Blogging dataset (D_{MB}), where the short, slang-heavy nature of posts makes classification challenging. HateBERT achieves an average F1 of 0.81, compared to 0.77 for BERT, indicating that domain-adaptive language representations better capture subtle cues of aggression and sentiment polarity typical of social media discourse. A similar trend is visible in the Gaming dataset (D_G), where HateBERT again surpasses BERT (0.76 vs. 0.72 average F1), confirming its ability to handle the noisy, context-dependent slang prevalent in gaming forums.

- **BERT performs competitively** on less toxic, conversational datasets.

  On the User-Comment corpus (D_{UC}), which contains more balanced and less explicitly aggressive discussions, BERT achieves a relatively strong average F1 of 0.71, only 0.05 points lower than HateBERT's 0.76. This suggests that while BERT's general-purpose language representations suffice for structured or civil discourse, HateBERT's domain specialisation offers clear advantages as linguistic informality and toxicity increase.

Overall, the results in Table 6.8 confirm that domain-adaptive pre-training significantly enhances cyberbullying detection performance, particularly in datasets characterised by informal, emotionally charged, and context-specific language use—attributes that closely mirror real-world online environments.

Attention-weights and Gradient-based feature analysis: We explore and test our hypothesis, i.e., *gradient based feature importance scores will reveal whether model attends more to adjectives, verbs, or words with strong positive or negative sentiment when identifying hurtful content in cyberbullying contexts*". For this we used Integrated Gradients (IG) from Captum's Layer Integrated Gradients implementation, and we computed token-level importance scores for the same subset of test data reserved for interpretability

Model	Dataset	Hurtful F1	Not-Hurtful F1	Average F1
BERT+	D_{QA}	0.62	0.68	0.65
	D_{UC}	0.65	0.77	0.71
	D_{MB}	0.75	0.79	0.77
	D_G	0.68	0.79	0.72
Hate-BERT+	D_{QA}	0.73	0.73	0.73
	D_{UC}	0.68	0.84	0.76
	D_{MB}	0.78	0.84	0.81
	D_G	0.74	0.78	0.76

Table 6.8: Classification Results

analysis. Unlike previous studies (e.g., (Elsafoury et al., 2021)), we compute Pearson's Correlation Coefficient (PCC) between mean attention weights and mean gradient-based feature importance scores across each transformer layer and its attention heads. As shown in Figure 6.2, the degree of alignment between attention and gradient-based measures is generally weak. For the best-performing HateBERT trained on D_{MB}, the PCC ranges from r =0.0129 (in D_{UC}) to r =0.1202 (in D_{UG}). For the best performing BERT also trained on D_{MB}, the correlation ranges from r =0.0042 (in question-answering) to r =0.19 (in micro-blogging datasets). These correlations, which hover close to zero in most cases, suggest that attention weights alone do not consistently reflect the true importance of tokens in the model's decision-making process. Even for datasets where the model achieves strong F1-scores, such as HateBERT fine-tuned on micro-blogging Twitter data, the near-zero correlation indicates that attention mechanisms, though effective for performance, are not inherently interpretable. This finding reinforces claims made in prior studies (Jain and Wallace, 2019; Serrano and Smith, 2019; Sun and Lu, 2020; Vashishth et al., 2019), which argue that attention weights improve classification accuracy but cannot be reliably used to explain model behaviour.

Layer-wise Attention and POS Analysis: We further explored layer-wise attention patterns for both BERT and HateBERT trained on D_{MB} data, the two models that achieved high classification performance and relatively stronger (though still weak) attention–gradient correlations. Figures 6.3 and 6.4 illustrate how attention weights distribute across POS categories at each transformer layer. Contrary to our hypothesis,

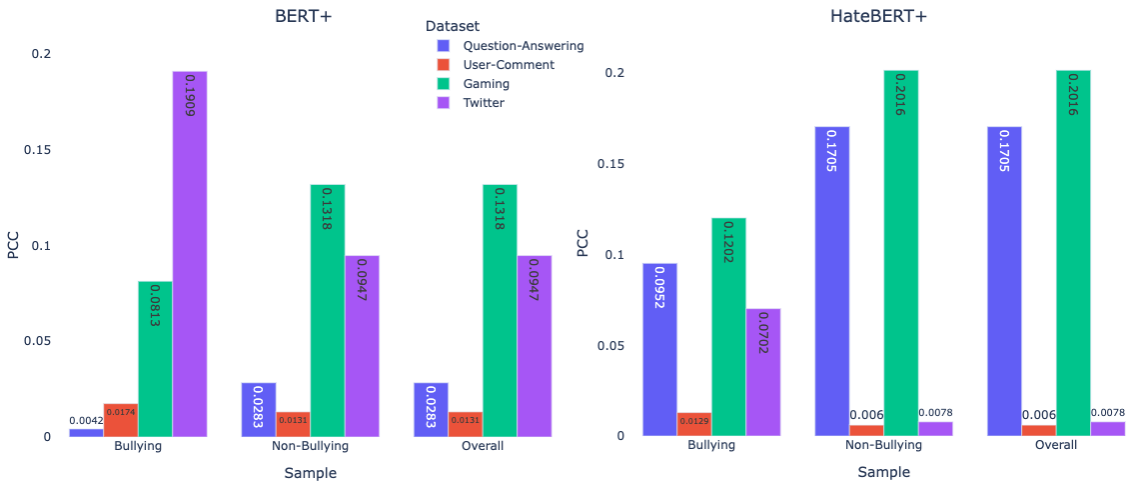


Figure 6.2: Pearson's Correlation between Mean Attention Weights and Mean Gradient-Importance Score for all LMs and Datasets

adjectives (`ADJ`), verbs (`VERB`), and nouns (`NOUN`), did not consistently receive higher attention weights. Instead, similar to (Elsafoury et al., 2021), we found that auxiliary verbs (`AUX` e.g., *is, have, do*), determiners (`DET` e.g., *this, the, my*), adpositions (`ADP` e.g., *in, on, over*), conjunctions (`CONJ` e.g., *and, but, because*), and even interjections (`INTJ` e.g., *hey!, wow!*) often attracted comparable or higher attention scores across layers. For instance, in the fine-tuned BERT model, verbs in `hurtful` samples peaked around layer 6 with a mean attention weight of ≈0.1, then decreased toward the final layer. In contrast, HateBERT exhibited more uniform attention across POS categories, with nouns initially receiving higher attention (≈0.14) that tapered off in deeper layers. This pattern suggests that while transformers distribute attention broadly across syntactic roles, their learned focus does not necessarily align with human linguistic intuition about which words "matter most" for detecting abusive intent.

Layer-wise Attention and Negative Sentiment Analysis:

We also compared attention distributions for tokens with strongly positive and strongly negative sentiment, using VADER-generated sentiment scores for all tokens in the reserved test subset. Representative negative words (e.g., ***f**k, as*, b**ch, stupid***) and positive words (e.g., ***like, cute, pretty, truth***) were tracked layer by layer (Figures 6.5 and 6.6). Interestingly, attention weights for positive words such as pretty were comparable to, or even higher than, those for negative words like ***b**ch*** and ***f**k***. This

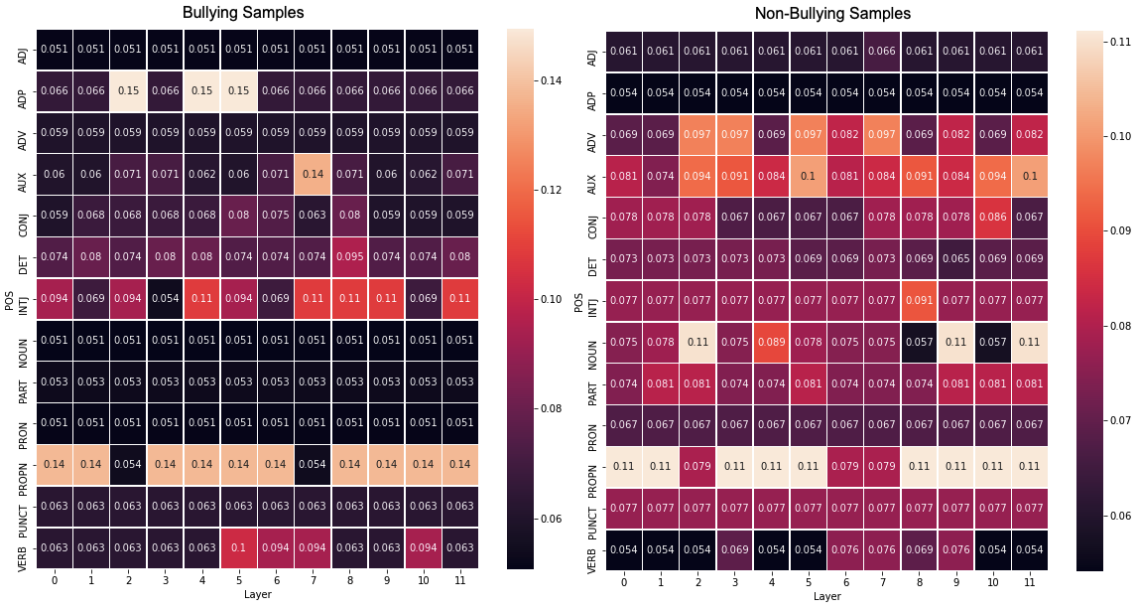


Figure 6.3: BERT+ Mean-attention weights per Layer POS-tag wise. Please note, *Bullying* samples mean *Hurtful* samples and *Non-Bullying* samples mean *Not-Hurtful* samples.

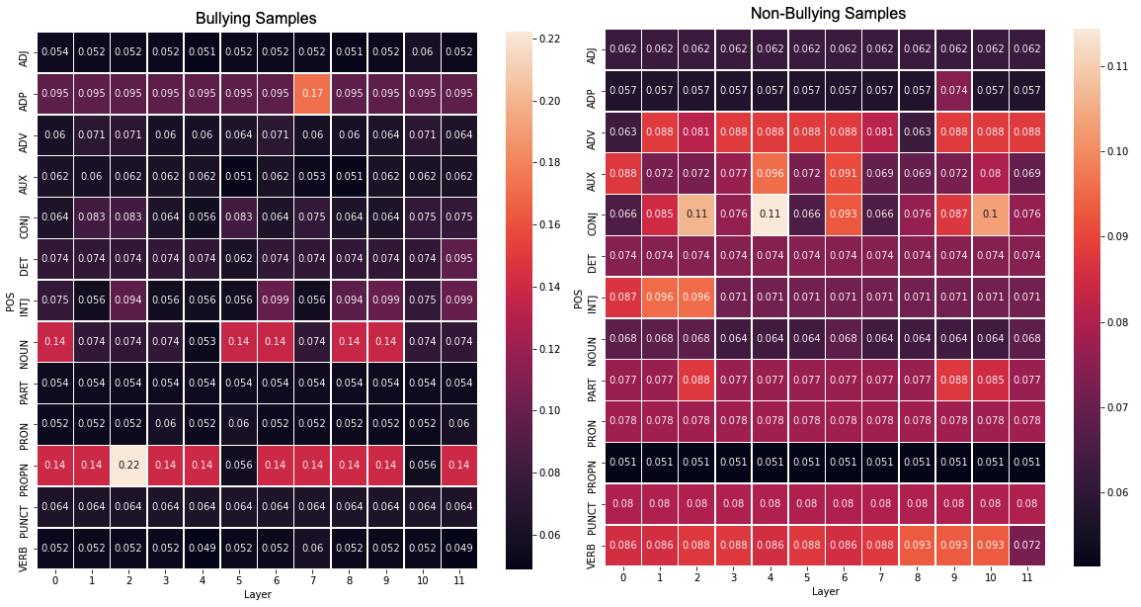


Figure 6.4: HateBERT+ Mean-attention weights per Layer POS-tag wise. Please note, *Bullying* samples mean *Hurtful* samples and *Non-Bullying* samples mean *Not-Hurtful* samples.

directly contradicts our hypothesis that negative sentiment tokens would receive stronger attention as cues of *hurtfulness* in cyberbullying contexts. To visualise these trends, we examined two short sentences from existing datasets "*f*ck the st*pid b*tch*" (negative sentiment), and "*i love this person*" (positive sentiment), and visualised the sequence-to-sequence attention by leveraging BertViz (Vig, 2019b). As shown in Figures 6.7 and 6.8,

attention distribution does not concentrate solely on sentiment-bearing words. Instead, tokens like "*the*" and "*person*" often share or exceed the attention intensity of sentiment terms such as "*love*" or "*b**ch*". These results demonstrate that attention does not necessarily highlight sentiment or syntactically salient words, implying that interpretability in transformer-based models requires more nuanced, gradient-informed analysis rather than reliance on raw attention maps.

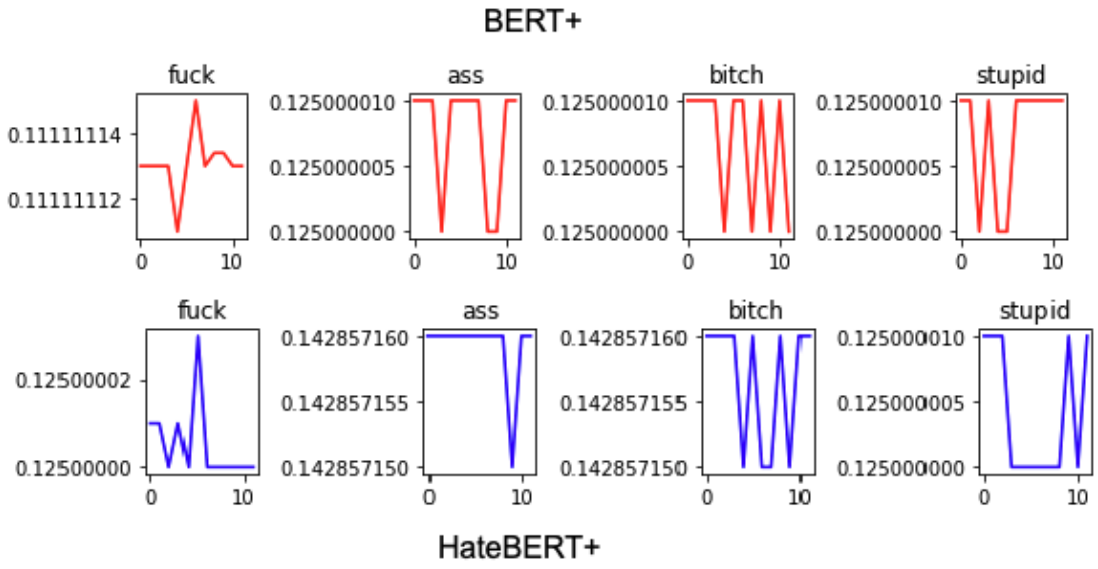


Figure 6.5: Attention per layer for Negative Sentiment words. Please note X-axis denotes layers 0-11 and y-axis denotes range of attention scores.

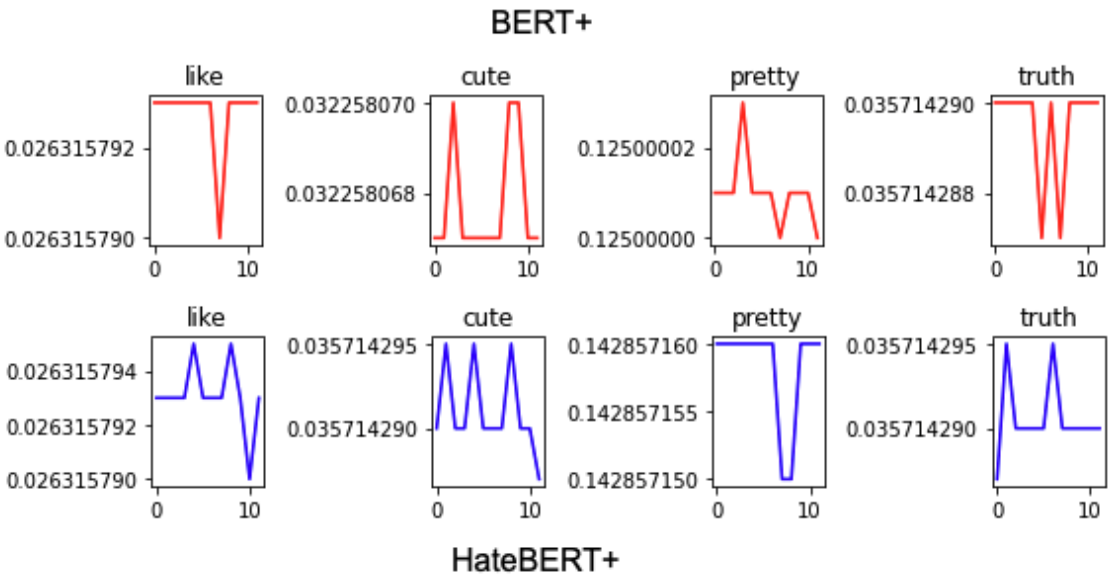


Figure 6.6: Attention per layer for Positive Sentiment words Please note X-axis denotes layers 0-11 and y-axis denotes range of attention scores.

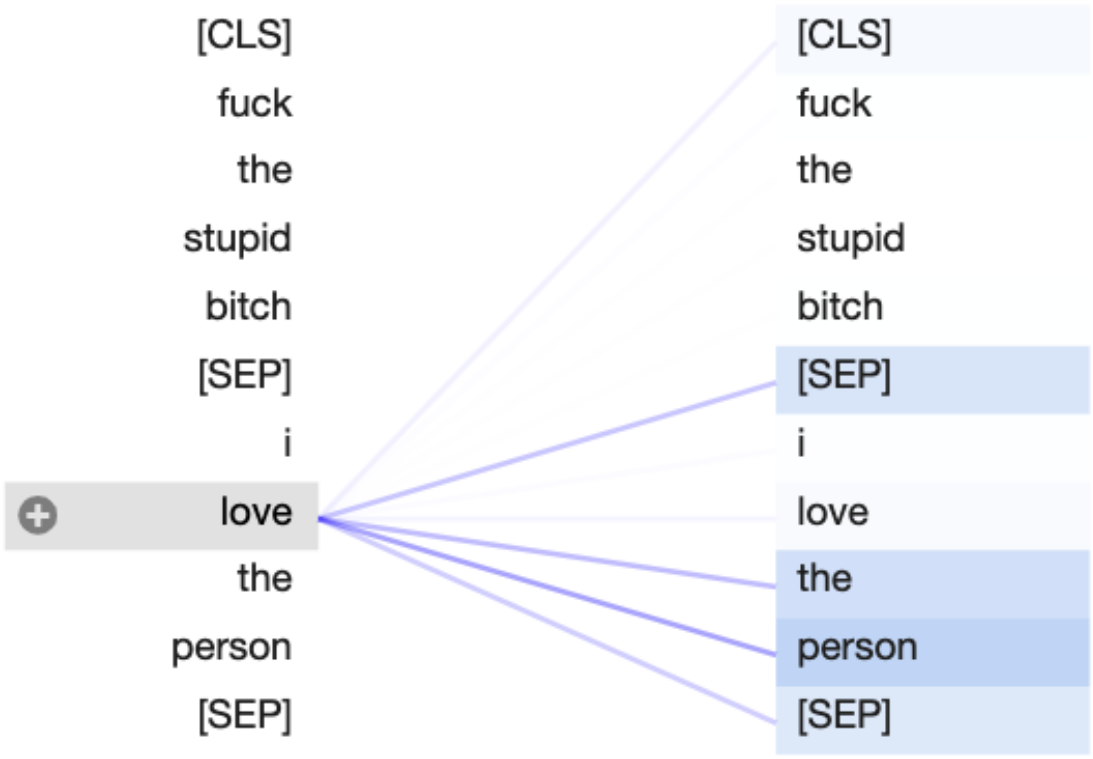


Figure 6.7: Attention analysis for Positive Sentiment Word

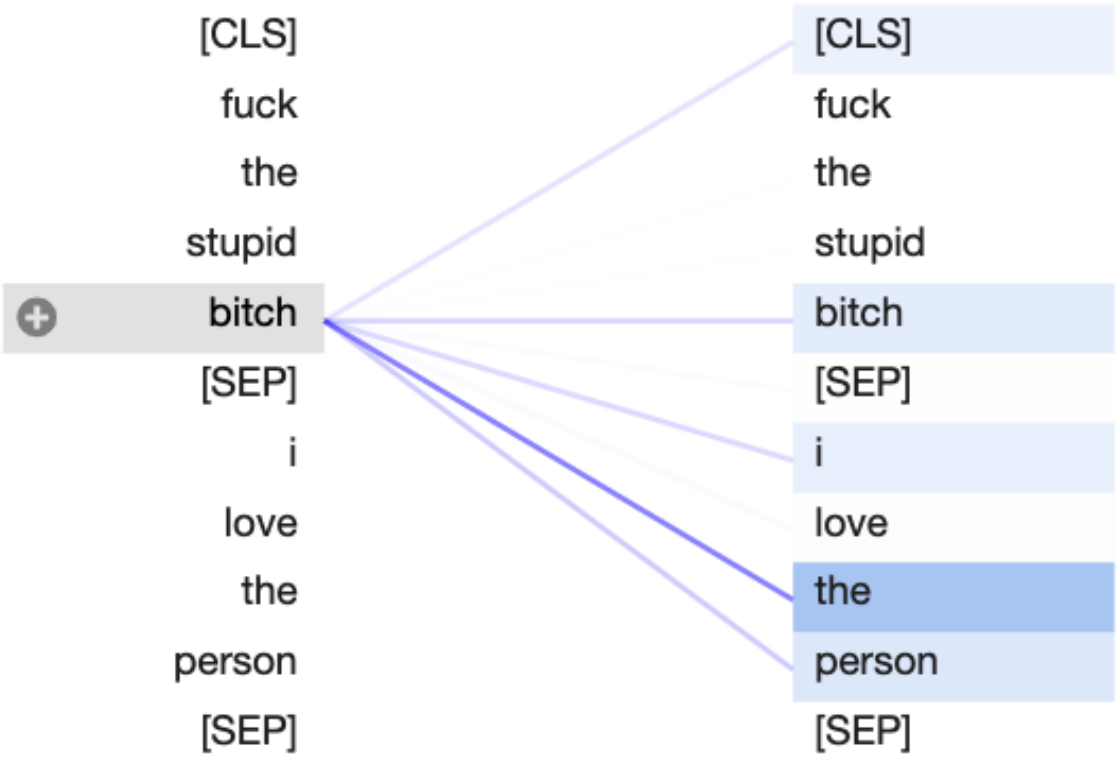


Figure 6.8: Attention analysis for Negative Sentiment Word

6.3 Towards Multi-task Classification - Is it hurtful? By whom?

Up to this sub-section, in addressing **RQ2.1**, we have examined multiple classification paradigms to assess their ability to perform a binary task of detecting `hurtful` vs. `not-hurtful` content across platforms (Section 6.1.2) and within translated youth-role-played datasets (Section 6.1.3). While these experiments showed that fine-tuning pre-trained language models can outperform traditional machine learning and neural baselines for binary hurtfulness detection, our analysis in Section 6.2 demonstrated that model "*attention*" alone is insufficient for interpreting or explaining model behaviour. We further explore **RQ2.1** from a multi-task perspective, i.e., whether we can identify participant roles in cyberbullying conversations and determine whether their messages are hurtful? To that effect, we extend our investigation to two additional tasks: (i) multi-class role identification and (ii) multi-task learning combining `hurtful` detection with multi-class role identification. Through this expanded analysis, we aim to evaluate whether modelling conversational roles alongside hurtfulness leads to more context-aware and behaviourally grounded cyberbullying detection.

6.3.1 Multi-class Role Identification

The first experimental setting aims to predict the **social role** of each participant in an online bullying conversation.

Problem Formulation: Formally, given an input text sequence $x_i \in \mathcal{X}$, the goal is to predict its corresponding role label $y_i \in \mathcal{Y}$, where $\mathcal{Y} = \{bully, victim, enabler, defender\}$. We model role identification as a probabilistic classification task, where the function, $f_r : \mathcal{X} \rightarrow \Delta^{|y_r|}$, maps each input message x_i to a vector of probabilities over all possible roles. Here, $\Delta^{|y_r|}$ denotes the set of all valid probability distributions over the role classes $\mathcal{Y}_r$.

Multi-Source Dataset: Building on the outcomes presented in Chapter 5, we leverage multiple datasets for this experiment. Specifically, we use: (i) $\mathcal{D}^{HT}_{IT\rightarrow EN}$ the human-translated Italian-to-English corpus discussed in Section 5.1.1; (ii) $\mathcal{D}^{HMT}_{FR\rightarrow EN}$ the French-

to-English human-aided machine-translated dataset described in Section 5.1.3; and (iii) $\mathcal{D}_{\textbf{LLM}}^{\textbf{Phase-2}}$, the large language model (LLM)-generated conversations presented in Section 5.1.4. For the experiment, we merge these sources into a single dataset $\mathcal{D} = \mathcal{D}_{IT\rightarrow EN}^{HT} \cup \mathcal{D}_{FR\rightarrow EN}^{HMT} \cup \mathcal{D}_{\text{LLM}}^{\text{P2}}$ and evaluate the models on a 10% hold-out subset of the merged data, as well as on the youth co-created gold-standard dataset introduced in Chapter 5 ($\mathcal{D}_{\textbf{Gold}}$). The label distributions across the four datasets, as shown in Table 6.9, indicate that certain roles, such as `conciliator` and `silent bystander` are absent from some sources. To ensure a fair comparison, we exclude all sentences corresponding to these roles from the merged dataset $\mathcal{D}$. Table 6.10 illustrates the size of the training, validation, and both test-sets, highlighting a moderate class imbalance across different roles.

Role	$\mathcal{D}_{IT\rightarrow EN}^{HT}$ §5.1.1)	$\mathcal{D}_{FR\rightarrow EN}^{HMT}$ §5.1.3)	$\mathcal{D}_{\textbf{LLM}}^{\textbf{Phase-2}}$ (§5.1.4)	$\mathcal{D}_{\textbf{Gold}}$ (§4.3.2)
Enabler	27.08%	28.75%	17.36%	33.99%
Defender	28.76%	19.56%	22.87%	43.94%
Bully	29.65%	20.39%	36.85%	17.54%
Victim	15.78%	22.06%	20.29%	4.52%
Conciliator	-	10.97%	-	-
Silent Bystander -	-		0.83%	-

Table 6.9: Role Label Distribution across different synthetic datasets and youth co-created $\mathcal{D}_{\text{Gold}}$

Role	**Train**	**Validation**	**Test**	$\mathcal{D}_{\textbf{Gold}}$
Bully	2449 (32%)	617 (32.2%)	234 (33%)	97 (17.6%)
Enabler	1781 (23.2%)	447 (23.3%)	172 (24.3%)	187 (33.9%)
Defender	1814 (23.7%)	458 (23.9%)	134 (18.9%)	243 (44.1%)
Victim	1618 (21.1%)	393 (20.5%)	169 (23.8%)	24 (4.4%)
Total	7662	1915	709	551

Table 6.10: Distribution of role labels across Train, Validation, and Test sets. Percentages indicate the relative proportion per split.

Classification Paradigms: Building upon previous work by (Rathnayake et al., 2020; Jacobs et al., 2022; Ollagnier et al., 2023), we compare following classification approaches:

- **Direct Multiclass (Baseline):** A single model $f : \mathcal{X} \rightarrow \mathcal{Y}$ is trained to predict one of the four mutually exclusive role labels - $\mathcal{Y}$.

- **Hierarchical Pairwise:** Extending the hierarchical classification setups discussed in (Rathnayake et al., 2020; Ollagnier et al., 2023), we implement a two-level hierarchical classifier as depicted in Figure 6.9. This involves, **Level-1**: a binary classifier ($f_1 : \mathcal{X} \rightarrow \{0, 1\}$) distinguishes between primary roles: bully vs victim. **Level-2**: Conditional classifiers refine each side: $f_{2\text{a}} : \mathcal{X} \rightarrow \{bully, enabler\}$ (bully-side), and $f_{2\text{b}} : \mathcal{X} \rightarrow \{victim, defender\}$ (victim-side). We compute full joint role probabilities as: $P(role \mid x) = P(side \mid x) \cdot P(role \mid side, x)$ with $P(side \mid x)$ from f_1, and $P(role \mid side, x)$ from the appropriate Level-2 classifier. Both Level-2 models are applied to all instances, yielding a normalized distribution over: *bully*, *enabler*, *defender*, and *victim*. The predicted role is: $\hat{y} = \arg\max_{role} P(role \mid x)$. We abstain from prediction if $\max_{\text{role}} P(\text{role} \mid x) < \tau$, where τ is a tunable threshold.

- **Prompt-based LLM classification:** Given LLMs' comparable performance to transfer learning in low-resource offensive language detection Riabi (2025); Riahi Samani et al. (2025); Plaza-del arco et al. (2023), we benchmark prompt-based classification using zero-, one-, and few-shot settings.

Language models: `RoBERTa-base` (Liu et al., 2019) and `GPT2-medium` (Radford et al., 2019) have been widely studied for abusive language tasks (Wei et al., 2021; Philipo et al., 2024; Kancharla et al., 2025) and achieved near state-of-the-art performance. We also evaluate zero- and few-shot classification using `LLaMA-3.3-70B` (Touvron et al., 2023).

Learning Objective: As seen in Table 6.10, the class distributions are skewed, so we benchmark both loss functions discussed in Section 2.4.2 Chapter 2, Cross Entropy ($\mathcal{L}_{CE}$) (Equation 2.21) and Focal Loss ($\mathcal{L}_{FL}$) (Equation 2.22) to see which learning objective is best suited for this task.

Training and Evaluation Configurations: Across $\mathcal{D}^{HT}_{IT \rightarrow EN}$ and $\mathcal{D}^{HMT}_{FR \rightarrow EN}$ consecutive messages from the same user were merged into single utterances to reflect conversational flow. Also short greeting messages[97] were filtered to reduce conversational noise. We evaluate cross-domain transfer learning using two training setups with 5-fold

[97]text starting with "hi" or "hey" and containing tokens < 3

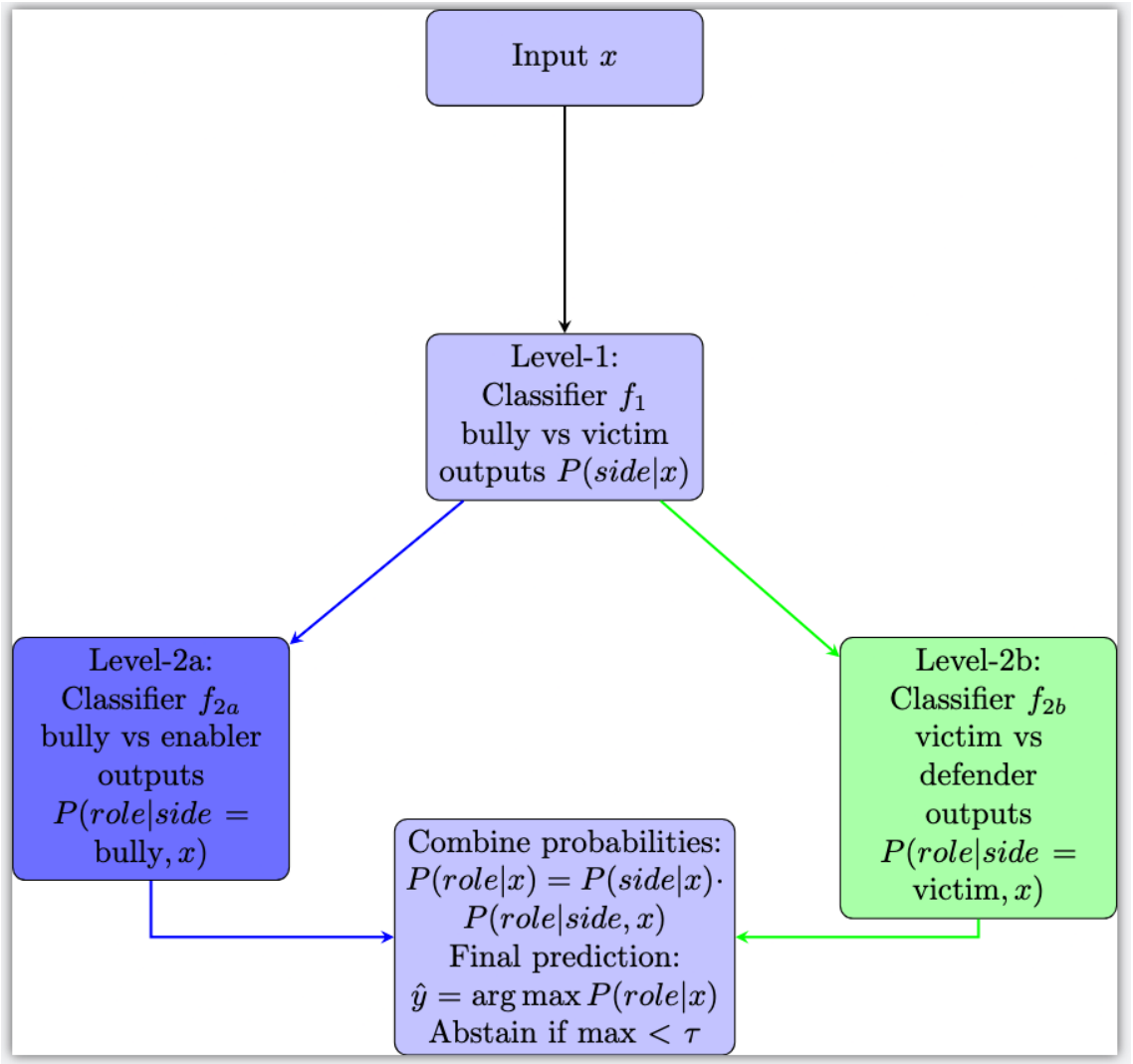


Figure 6.9: Hierarchical Pair-wise Classification Method

cross validation with no-sampling and random up-sampled class distributions[98], on (i) translated-only data ($\mathcal{D}^{HT}_{IT\rightarrow EN} \bigcup \mathcal{D}^{HMT}_{FR\rightarrow EN}$) and (ii) all-combined multi-source dataset ($\mathcal{D}^{HT}_{IT\rightarrow EN} \bigcup \mathcal{D}^{HMT}_{FR\rightarrow EN} \bigcup \mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$. All models are trained for early stopping based on both validation macro-F1 and validation loss with a patience of 3 epochs and threshold of 0.001. We tune hyper-parameters to maximise macro-average F1 on validation data. For comparing this baseline with `Llama-3.3-70B-Instruct` we followed hyper-parameters and classification prompt-structures as previously studied in Kazemi et al. (2025)[99]. We benchmark the models on $\mathcal{D}_{\text{Gold}}$ as our test-set and report macro-average F1 score across all roles.

Hyper-parameter Optimisation: To tune hyper-parameters, we maximise on macro-

[98] we randomly copy sentences at seed=42

[99] `https://github.com/kanishk-r-verma/bully-bench/blob/main/supplementary_material/prompt_classify.txt`

F1 for the validation data. Search spaces include: (i) Learning rate: log-uniform [2e-5, 5e-3]; (ii) Batch size: {8, 16, 32}; (iii) Epochs: {2, 3, 5}; (iv) Focal loss parameters: $\alpha \in [0.25, 1.0]$, $\gamma \in [1.0, 3.0]$; (v) Regularization: L2 $\lambda \in$ [1e-5, 1e-3], dropout $\in [0.1, 0.5]$; (vi) Weight decay: [1e-4, 1e-2], warmup steps: {100, 500, 1000}; (vii) τ: [0.2, 0.3]

6.3.2 Multi-task Hurtful + Role Identification

Problem Formulation: This setting formulates cyberbullying detection as a multi-task learning problem, jointly modelling **(a) hurtfulness** detection, and **(b) role identification**. Given an input text $x \in \mathcal{X}$, we aim to predict,

(a) Binary Hurt Label: $y_h \in \mathcal{Y}_h = \{$ `hurtful`, `not-hurtful`$\}$

(b) Multi-class Role Labels: $y_r \in \mathcal{Y}_r =$ `bully`, `victim`, `enabler`, `defender`

Multi-Source Dataset: We extend the multi-source dataset introduced in §6.3.1 to support multi-task learning for ***hurtfulness*** and ***role*** identification. As $\mathcal{D}_{\text{LLM}}^{Phase-2}$ (LLM-generated conversations; §5.1.4) contains only *role* labels, it is used exclusively for the *role* task. To broaden the contextual and linguistic coverage of cyberbullying instances, we integrate the **Amica** dataset by (Van Hee et al., 2018) ($\mathcal{D}_{\text{AMiCA}}$), which includes 113K English messages across 6,610 Ask.fm posts and introduces `negative` (non-role) labels. Because labelled role instances in $\mathcal{D}_{\text{AMiCA}}$ are scattered with unlabelled messages, we preserve conversational context using a sliding window of two preceding and two following messages, merging overlapping windows. This produces 16,559 messages across 2,246 conversations, annotated for both *hurtfulness* and *role*.

Hurtfulness Task Data: Table 6.11 summarises the datasets used for binary `hurtfulness` classification. Since $\mathcal{D}_{\text{LLM}}^{Phase-2}$ lacks hurtfulness labels, only $\mathcal{D}_{IT\rightarrow EN}^{HT}$ (§5.1.1), $\mathcal{D}_{FR\rightarrow EN}^{HMT}$ (§5.1.3) and $\mathcal{D}_{\text{AMiCA}}$ are included. Together, this yields 22,267 annotated messages. The class distribution is moderately imbalanced: ≈68% `not-hurtful` and ≈32 `hurtful`.

User-Role Identification Task Data: Table 6.12 reports the distribution of four roles—`Victim`, `Defender`, `Enabler`, and `Bully`. Only explicitly annotated role instances are retained, resulting in 16,023 messages. Despite contextual expansion, $\mathcal{D}_{\text{AMiCA}}$ has many neutral (non-role) messages and relatively few labelled roles, a pattern also observed in $\mathcal{D}_{IT\rightarrow EN}^{HT}$.

Across all sources, `Bully` is the most frequent role (6,891; 43%), followed by `Victim` (3,639; 23%), `Defender` (2,978; 19%), and `Enabler` (2,515; 16%).

Source	**Not-hurtful**	**Hurtful**
$\mathcal{D}^{HT}_{IT\rightarrow EN}$	1443 (65%)	749 (35%)
$\mathcal{D}^{HMT}_{FR\rightarrow EN}$	1674 (48%)	1842 (52%)
$\mathcal{D}_{\text{AMiCA}}$	12090 (73%)	4469 (27%)
Total	15207 (68%)	7060 (32%)

Table 6.11: Distribution of binary hurtfulness labels across the multi-source dataset. $\mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$ is excluded since it does not contain hurtfulness annotations. This distribution highlights class imbalance across the combined dataset.

Source	**Bully**	**Enabler**	**Defender**	**Victim**
$\mathcal{D}^{HT}_{IT\rightarrow EN}$	330	332	310	147
$\mathcal{D}^{HMT}_{FR\rightarrow EN}$	731	979	692	668
$\mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$	2505	1180	1555	1499
$\mathcal{D}_{\text{AMiCA}}$	3325	24	421	1325
Total	6891	2515	2978	3639

Table 6.12: Distribution of role labels across the multi-source dataset for multi-task learning. $\mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$ only contributes to role labels, while $\mathcal{D}_{\text{AMiCA}}$ includes both role and hurtfulness labels.

Classification Paradigms: We evaluate two paradigms:

- **Hard Parameter Sharing**: A single backbone (θ) (e.g., BERT) is shared across both tasks, with separate classifier heads $h_h(x)$ for *hurtfulness* detection, and $h_r(x)$ for *role* identification. Each head uses its own weight matrix and bias (W_h, b_h and W_r, b_r), with outputs computed as: $h_h(x) = softmax(W_h.h_\theta(x) + b_h)$ and $h_r(x) = softmax(W_r.h_\theta(x) + b_r)$. Gradients from both tasks jointly update the shared parameters θ, optimizing the combined loss $\mathcal{L} = \alpha\mathcal{L}_h + (1-\alpha)\mathcal{L}_r$,where $\mathcal{L}_h$ and $\mathcal{L}_r$ are the task-specific losses, and α controls the relative importance of each task. This setup enables the backbone to learn representations that simultaneously benefit both tasks.

- **AdapterFusion (AF):** Task-specific adapters (A_h, A_r) are small networks inserted into the backbone. The base model parameters are frozen, and only the adapters

are trained. Here, $A_h(x)$ and $A_r(x)$ produce task-specific features for hurtfulness and role identification, respectively. Following (Pfeiffer et al., 2021), a fusion layer combines the adapter outputs:

$$h_F(x) = F(A_h(x), A_r(x)) = \sum_{t \in \{h,r\}} \beta_t A_t(x)$$

where β_t are learnable fusion coefficients. Only the fusion parameters are fine-tuned during joint training, while the adapters are pre-trained individually. This design mitigates catastrophic forgetting and allows the model to leverage specialised knowledge from each task efficiently.

Language Models: For the multi-task setup, we evaluate both encoder-only and decoder-only architectures to assess how well multi-task training adapts to different language model types. Specifically, we experiment with:

- **BERT** (Devlin et al., 2019), an encoder-only transformer pre-trained with masked language modelling, and
- **GPT-2 Medium** (Radford et al., 2019), a decoder-only transformer pre-trained auto-regressively.

This comparison allows us to examine whether joint learning of hurtfulness and role identification benefits more from bidirectional context (encoder) or causal decoding (decoder) models.

Learning Objective: As in the previous experiments, we benchmark both **Cross-Entropy (CE)** and **Focal Loss (FL)**. CE treats all samples equally, while FL (Lin et al., 2020) emphasises hard-to-classify, minority-class examples, making it particularly suited for the imbalanced role and hurtfulness distributions.

Training and Evaluation Configuration: Hyper-parameter optimisation for the single-task experiments, extended to account for multi-task weighting. The key training configurations are: (i) Learning rate: 1×10^{-5} – 3×10^{-5}; (ii) Batch size: 16; (iii) Task weight α: 0.3-0.7 (controls relative importance of hurtfulness vs. role tasks); (iv) Warmup steps: 100; (v) Optimizer: AdamW; (vi) Epochs: 3-5; (vii) Dropout: 0.1; (viii) Adapter

configuration: Pfeiffer (default); (ix) Fusion configuration: Static. Training is conducted on a single RTXa6000 GPU (40GB), with model selection based on the best validation macro-F1 score, computed as the mean across both tasks.

6.3.3 Results and Discussion

In this section, we discuss the results from both the multi-class role identification experiments (Section 6.3.1) and the multi-task hurtfulness–role identification experiments (Section 6.3.2).

Multi-class Role Identification: Across all three classification paradigms (*Direct Multiclass*, *Hierarchical Pairwise*, *Prompt-based LLM classification*) on $\mathcal{D}_{\textbf{Gold}}$, few-shot Llama-3.3-70B-Instruct achieves a macro-average F1 of 0.48, competitive with supervised baselines (See Table 6.13). **Role-wise** analysis in Table 6.14 shows that for the 187 (33.9%) `enabler` messages in $\mathcal{D}_{\textbf{Gold}}$, GPT2-Medium (multi-class) achieves the highest F1 of 0.5234, followed by Llama-3.3-70B-Instruct (few-shot) (F1=0.5149), and RoBERTa (pairwise, F1=0.4422). It also reveals similar weaknesses in identifying `victim` messages (minority class N =24), GPT2-Medium performs poorly (F1=0.1905), followed by Llama-3.3-70B-Instruct (few-shot) (F1=0.2222) and RoBERTa (pairwise) (F1=0.2243). For the majority class in $\mathcal{D}_{\textbf{Gold}}$, i.e., `defender` messages ($N$ =243 44.1%), Llama-3.3-70B-Instruct (few-shot) achieves fairs better with F1 of 0.7381, followed by RoBERTa (pair-wise) (F1=0.6052).

Model	**Strategy**	**Configuration**	**$\mathbf{D_{Gold}}$ Macro-F1**
Llama-3.3-70B-Instruct	**Few-shot**		**0.4870**
RoBERTA-base	Pairwise	Fold-2 + ($\mathcal{D}^{HT}_{IT\rightarrow EN}$ + $\mathcal{D}^{HMT}_{FR\rightarrow EN}$ + $\mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$) + Focal Loss + Upsampling	0.4034
GPT2-medium	Multi-class	Fold-4 + ($\mathcal{D}^{HT}_{IT\rightarrow EN}$ + $\mathcal{D}^{HMT}_{FR\rightarrow EN}$ + $\mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$) + Focal Loss + Upsampling	0.3807

Table 6.13: Top-3 baseline benchmarking on $\mathcal{D}_{\textbf{Gold}}$

Model	Strategy	Enabler Macro-F1	Defender Macro-F1	Bully Macro-F1	Victim Macro-F1
Llama3.3-70B-Instruct	Few-shot	0.5149	**0.7381**	**0.4742**	0.2222
RoBERTA-Base	Pair-wise	0.4422	0.6052	0.3420	**0.2243**
GPT2-Medium	Multi-class	**0.5234**	0.5775	0.2317	0.1905

Table 6.14: Detailed breakdown of baseline benchmarking on $\mathcal{D}_{\textbf{Gold}}$

Error analysis reveals that all 3 baselines fail on 94 sentences in $\mathcal{D}_{\textbf{Gold}}$ indicating that messages like "*b1tches ik ur stup!d but like thats so low*" written by `defenders`, due to its obfuscation and use of profanities are predicted either as *bully* (Llama-3.3 and RoBERTa) or *enabler* (GPT2-medium) hinting that additional context is needed and sequence-classification for this task might not be ideal for such messages. The most frequent error across all baselines (LLaMA-3.3, RoBERTa, GPT-2 Medium) is misclassifying *Bully* as *Enabler* and vice versa (see Table 6.15). Across all models, Llama-3.3-70B-Instruct, RoBERTa, and GPT2-medium, victim identification on $\mathcal{D}_{\textbf{Gold}}$ proves challenging given that it is a minority class. However, brief victim statements such as "*stop it now*" and "*leave me alone*" are frequently misclassified as `bully` or `defender` roles. Similarly, messages employing veiled sarcasm, like "*sure, great joke*", evade correct classification across all models in our sequence-level paradigm. These limitations indicate that sequence-based classification may be insufficient for capturing the nuanced dynamics of cyberbullying (CB) role identification.

Model	Confusion Labels	Error rate $\mathcal{D}_{\textbf{Gold}}$
Llama-3.3		0.6080
RoBERTa	Bully ↔ Enabler	0.6279
GPT2-Medium		0.7026

Table 6.15: Most common error across best performing models on $\mathcal{D}_{\text{Gold}}$

Table 6.16 summarises performance across different models and strategies across both the test-set ($N = 709$ See Table 6.10 for details) and $\mathcal{D}_{\textbf{Gold}}$. On the held-out test set, fine-tuned models substantially outperform the prompt-based approach. RoBERTa (pair-wise)

achieves the highest macro-average F1 score (0.52), followed by GPT2-medium (multi-class) with a F1 score of 0.49, while Llama-3.3-70B-Instruct with few-shot learning lags behind at 0.41. Examining individual role classification RoBERTa (pair-wise) demonstrates the strongest performance on bully detection (F1 = 0.61) and victim identification (F1 = 0.48), while maintaining competitive performance on enabler (F1 = 0.46) and defender (F1 = 0.55) roles. GPT2-medium demonstrates comparable bully detection capability (F1 = 0.57) and performs reasonably well on enabler classification (F1 = 0.46), though it is relatively weaker at defender detection (F1 = 0.47) compared to the other models. Llama-3.3-70B-Instruct (few-shot) achieves strong bully classification (F1 = 0.52) but struggles with both enabler (F1 = 0.29) and victim (F1 = 0.30) detection.

Key Take-away: The performance differences between $\mathcal{D}_{\textbf{Gold}}$ and the test-set reveal important distinctions in model generalisation capabilities. Although Llama-3.3-70B-Instruct (few-shot) achieves competitive performance on $\mathcal{D}_{\textbf{Gold}}$ (Macro-F1 = 0.48), its performance drops notably on the test-set (Macro-F1 = 0.41), with particularly pronounced declines in defender detection (from F1 = 0.74 to 0.53) and victim identification (from F1 = 0.52 to 0.30). In contrast, fine-tuned models demonstrate more robust generalisation: RoBERTa (pair-wise) achieves the highest test-set performance (Macro-F1 = 0.52), substantially outperforming the prompt-based approach by 0.11 points. This divergence in performance can be attributed to the compositional differences between datasets, $\mathcal{D}_{\textbf{Gold}}$ as discussed in Chapter 4 and Chapter 5 comprises of youth-curated conversations containing authentic linguistic patterns, obfuscation, and profanities, while the test-set consists of solely of human-translated ($\mathcal{D}^{HT}_{IT\rightarrow EN}$ §5.1.1), human-aided machine translation ($\mathcal{D}^{HMT}_{FR\rightarrow EN}$ §5.1.3), and LLM-generated conversations ($\mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$ §5.1.4). The superior generalisation of encoder models on the test-set suggests they learn more robust features that transfer effectively across translated and synthetic text, whereas Llama-3.3-70B-Instruct appears more sensitive to the stylistic and linguistic differences between authentic youth language and translated or generated content.

Multi-task Hurt + Role Identification: Figure 6.10 shows macro-average F1 scores for BERT (encoder-only) and GPT-2 (decoder-only) across binary hurt detection and multi-class role classification. On the test set, BERT with Adapter Fusion (AF) performs slightly below Hard Parameter Sharing (HP) (Hurt: AF-CE=0.695 vs. HP-

Results on $\mathcal{D}_{\text{Gold}}$						
Model	**Strategy**	**Macro-F1**	**Enabler-F1**	**Defender-F1**	**Bully-F1**	**Victim-F1**
Llama-3.3-70B	Few-shot	0.48	0.22	0.74	0.47	0.52
Llama-3.3-70B	One-shot	0.31	0.00	0.51	0.39	0.32
Llama-3.3-70B	Zero-shot	0.38	0.14	0.60	0.42	0.35
RoBERTa	Pair-wise	0.40	0.44	0.61	0.34	0.22
RoBERTa	Multi-class	0.39	0.52	0.57	0.31	0.17
GPT2-medium	Multi-class	0.38	0.52	0.58	0.23	0.19
GPT2-medium	Pairwise	0.38	0.50	0.55	0.29	0.16
Results on $\mathcal{D}_{\text{Test}}$						
Model	**Strategy**	**Macro-F1**	**Enabler-F1**	**Defender-F1**	**Bully-F1**	**Victim-F1**
Llama-3.3-70B	Few-shot	0.41	0.29	0.53	0.52	0.30
RoBERTa	Pair-wise	0.52	0.46	0.55	0.61	0.48
GPT2-medium	Multi-class	0.49	0.46	0.47	0.57	0.45

Table 6.16: Model performance comparison on Gold and Test datasets. All Llama-3.3-70B models use Temperature=0.1, do_sample=True, max_tokens=50, top_p=50, top_k=50. RoBERTa and GPT2-medium models trained on $\mathcal{D}^{HT}_{IT\rightarrow EN} \bigcup \mathcal{D}^{HMT}_{FR\rightarrow EN} \bigcup \mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$ with upsampling and focal loss (Gold) or combined data (Test).

CE=0.718; Role: AF-CE=0.548 vs. HP-CE=0.555). On the youth co-created $\mathcal{D}_{\text{Gold}}$, AF yields marginal gains for binary hurt detection (AF-FL=0.626; HP-FL=0.593), but not for multi-class role classification (AF-FL=0.356; HP-FL=0.386). GPT-2 performs comparably to BERT, with Cross Entropy (CE) yielding more stable and generalizable results than Focal Loss (FL) across both datasets (binary HP-CE=0.721–0.641 vs. HP-FL=0.732–0.595). Similar stability is observed for multi-class tasks, where HP with CE consistently performs best.

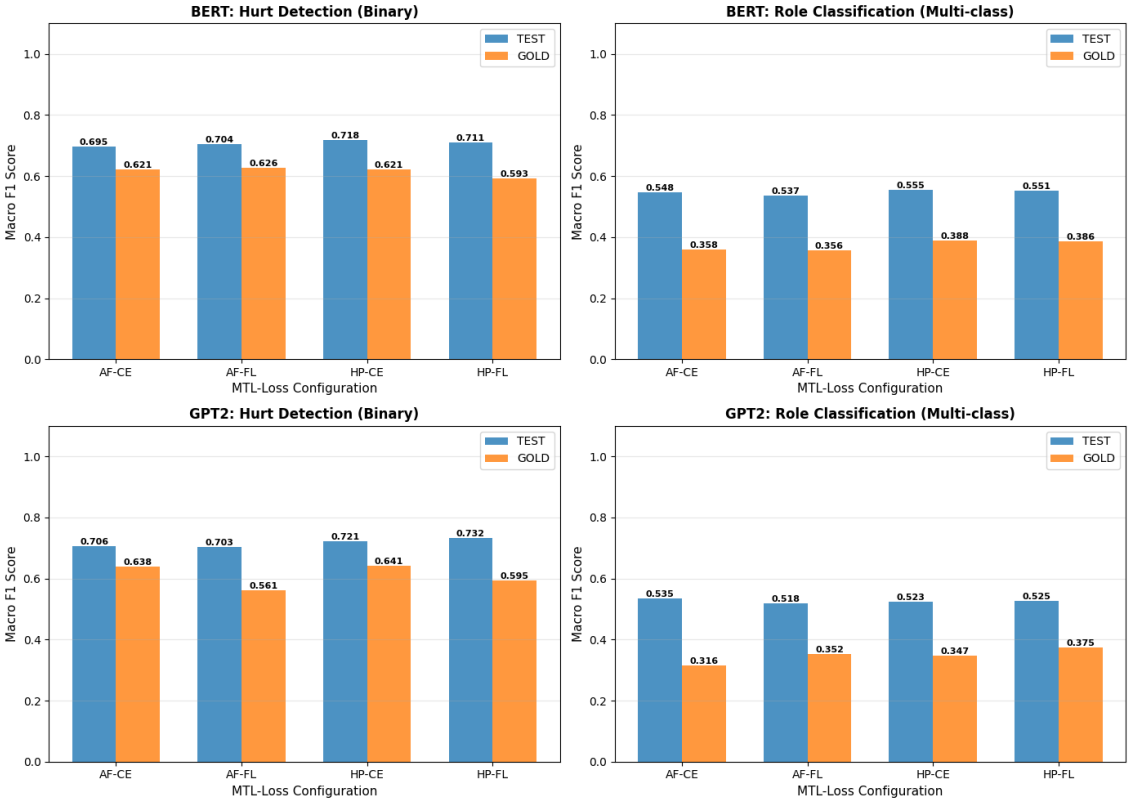


Figure 6.10: BERT and GPT2 multi-task results with different configurations

As shown in Figure 6.11, Δ values indicate greater generalisation difficulty with FL, regardless of the training strategy. For binary detection (left), AF models show smaller performance drops on $\mathcal{D}_{\text{Gold}}$ (BERT: HP-CE: –0.098; AF-CE: –0.075; GPT-2: HP-CE: –0.08; AF-CE: –0.068), suggesting AF enhances cross-dataset generalisation. In contrast, for multi-class classification (right), HP tends to yield more stable results, even though FL occasionally benefits minority role detection. Figure 6.12 further details label-wise F1 results. The best configuration, BERT-HP-CE, achieves strong and balanced performance on the test set (`not-hurt`: 0.846, `hurt`: 0.591, `bully`: 0.762, `defender`: 0.542, `enabler`: 0.457), with minimal drops on $\mathcal{D}_{\text{Gold}}$ (–0.004 for `victim`, –0.018 for `enabler`). For GPT-

2, no single configuration consistently achieves strong and balanced performance across both test sets. GPT2 with Hard Parameter Sharing and Cross Entropy (GPT2-HP-CE) ranks second in macro-average F1 for the binary task on the test set (F1= 0.721) but performs comparatively better on the gold set (F1 = 0.641). In contrast, GPT2 with Hard Parameter Sharing and Focal Loss (GPT2-HP-FL) attains the best performance for multi-class role detection on both the test and gold sets (Test F1 = 0.525; Gold F1 = 0.375) and shows particular advantages for minority roles, such as `victim`. Overall, as an encoder baseline, BERT with Hard Parameter Sharing and Cross Entropy (BERT-HP-CE) consistently performs best across tasks. For the decoder baseline, GPT-2 with HP-CE is generally the most stable choice, while using Focal Loss (FL) can enhance performance for specific minority roles.

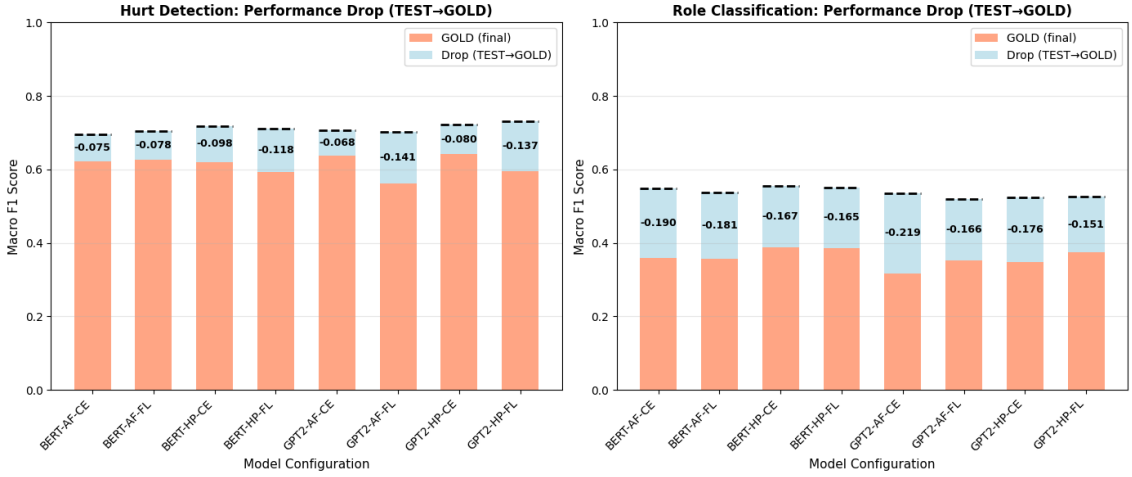


Figure 6.11: F1 score drop when measuring performance of different *encoder* and *decoder* transformers on $\mathcal{D}_{\text{Test}}$ to $\mathcal{D}_{\text{Gold}}$

Error Analysis: We discuss the error rates for the best-performing encoder-only and decoder-only models in the multi-task setup. The best decoder-only model, **GPT-2-HP-FL**, trained with Hard Parameter Sharing (HP) and Focal Loss (FL) as the learning objective, achieves an average macro-F1 of 0.6285 across both tasks on the test set (hurt: 0.732; role: 0.525) and 0.48 on the $\mathcal{D}_{\text{Gold}}$ set. The best encoder-only model, **BERT-HP-CE**, trained with Hard Parameter Sharing (HP) and Cross-Entropy (CE) as the learning objective, yields a slightly higher average macro-F1 of 0.6365 across both tasks on the test set (hurt: 0.718; role: 0.555) and 0.48 on the $\mathcal{D}_{\text{Gold}}$ set. The detailed error rates for **GPT-2-HP-FL** and **BERT-HP-CE** are reported in Table 6.17 and Table 6.18, respectively. Both models exhibit similar confusion patterns, particularly in misclassifying `Bully`

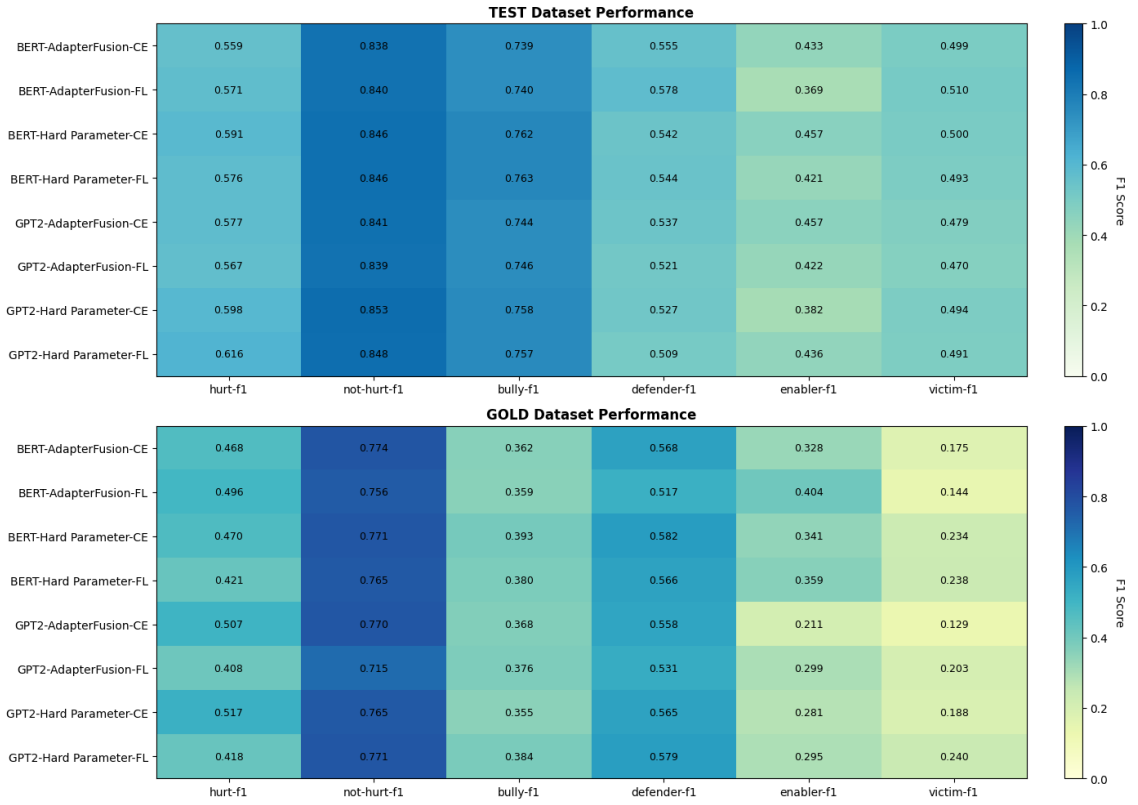


Figure 6.12: F1 scores across different *encoder* and *decoder* transformers on $\mathcal{D}_{\text{Test}}$ and $\mathcal{D}_{\text{Gold}}$

as `Enabler` (BERT: 30.33% in Test, 57.75% in Gold; GPT-2: 27.87% in Test, 59.36% in Gold) and `Bully` as `Victim` (BERT: 39.46% in Test, 20% in Gold; GPT-2: 38.38% in Test, 24% in Gold). Across both models, `Enabler` remains the most challenging role to classify accurately. **BERT-HP-CE** demonstrates modest improvements over **GPT-2-HP-FL**, particularly on the Test-Set, with lower cross-class errors (e.g., `Defender` → `Enabler`: 12.86% vs. 17.14%). Misclassification patterns remain consistent between models, with `Victim` often confused with `Bully` and `Defender`. However, **BERT-HP-CE** slightly reduces errors for `Victim` → `Enabler` in the Test-Set. The qualitative error trends align with those observed in the baseline multi-class classification experiments discussed in §6.3.1. Persistent mis-classifications often stem from nuanced or sarcastic language use. For instance, earlier discussed messages such as "*b1tches ik ur stup!d but like thats so low*" (written by a defender) or bullying comments like "*look at that belgua whale*" continue to be incorrectly labelled. These examples highlight the limitations in detecting subtle linguistic cues.

Key Takeaway: Across both binary hurt detection and multi-class role classification,

BERT (encoder-only) consistently outperforms GPT-2 (decoder-only). For binary hurt detection on the test set, BERT trained for Hard Parameter (HP) sharing with Cross Entropy (CE) achieves an F1 score of 0.718 compared to GPT-2-HP-CE at 0.721, with a slightly lower macro-average on the gold set (0.641). In multi-class role classification, BERT-HP-CE again outperforms GPT2-HP-FL, with test F1 scores of 0.555 versus 0.525, indicating that Hard Parameter (HP) sharing combined with Cross-Entropy (CE) is the most stable configuration for both tasks. Training strategies also impact performance: Adapter Fusion (AF) reduces cross-dataset drops for binary hurt detection on the gold set, with BERT showing an AF-CE drop of –0.075 versus –0.098 for HP-CE, and GPT-2 showing –0.068 versus –0.08. Focal Loss (FL), while beneficial for detecting minority roles, can reduce generalisation stability; for instance, GPT2-HP-FL achieves a gold F1 of 0.375 for multi-class role classification compared to 0.356 for HP-CE. Error analysis reveals persistent challenges in classifying certain roles. The `Enabler` role remains the most difficult to predict accurately. Misclassification patterns show that `Bully` is frequently confused with `Enabler` (BERT Test = 30.33%, Gold = 57.75%; GPT-2 Test = 27.87%, Gold = 59.36%) or `Victim` (BERT Test = 39.46%, Gold = 20%; GPT-2 Test = 38.38%, Gold = 24%), while `Defender` is occasionally misclassified as `Enabler` (BERT Test = 12.86%, GPT-2 = 17.14%). These patterns highlight that certain roles, particularly minority roles, remain challenging despite overall high model performance.

Test-Set				
	True Label			
Predicted ↓	**Bully**	**Enabler**	**Defender**	**Victim**
Bully	-	7.10%	5.76%	8.43%
Enabler	27.87%	-	22.95%	23.77%
Defender	25.00%	17.14%	-	19.29%
Victim	38.38%	8.65%	12.97%	-
$\mathcal{D}_{\textbf{Gold}}$				
Bully	-	6.19%	12.37%	11.34%
Enabler	59.36%	-	11.76%	9.09%
Defender	29.75%	8.26%	-	14.46%
Victim	24.00%	4.00%	24.00%	-

Table 6.17: Error rates (%) for multi-class classification using GPT-2 with Hard Parameter Sharing and Focal Loss (HP-FL). Rows represent predicted labels and columns represent true labels, shown for both the Test-Set and $\mathcal{D}_{\text{Gold}}$ youth co-created dataset.

	Test-Set			
	True Label			
Predicted ↓	**Bully**	**Enabler**	**Defender**	**Victim**
Bully	-	5.54%	5.76%	9.98%
Enabler	30.33%	-	16.39%	18.85%
Defender	25.71%	12.86%	-	17.86%
Victim	39.46%	5.41%	12.43%	-
	$\mathcal{D}_{\textbf{Gold}}$			
Bully	-	11.34%	5.15%	11.34%
Enabler	57.75%	-	11.23%	6.95%
Defender	31.40%	7.85%	-	14.05%
Victim	20.00%	8.00%	28.00%	-

Table 6.18: Error rates (%) for multi-class classification using BERT with Hard Parameter Sharing and Cross-Entropy loss (HP-CE). Rows represent predicted labels and columns represent true labels, shown for both the Test-Set and $\mathcal{D}_{\text{Gold}}$ youth co-created dataset.

6.4 Assessing Model Robustness

6.4.1 Related Work

Addressing the challenge of obfuscation discussed in **C.4**, we investigate the research question **RQ2.2** i.e., *How can classification techniques be adapted to detect* `hurtfulness` *in cyberbullying contexts within obfuscated or perturbed textual social media content?*. By exploring methods to improve text classification robustness against intentionally deceptive language patterns, including obfuscation and character-level perturbations (e.g., `stupid` → `stp!d`). Traditional machine-learning or deep-learning approaches typically apply normalisation or de-obfuscation as preprocessing steps, attempting to map perturbed words back to canonical forms using dictionaries, edit-distance heuristics, or phonetic matching algorithms (Han and Baldwin, 2011; Eisenstein, 2013; Aggarwal and Zesch, 2022). However, such normalisation pipelines introduce *cascading error risks*, i.e., when normalisation fails to identify and restore obfuscated text correctly, downstream classification performance degrades substantially (Pruthi et al., 2019).

Character- and byte-level representations offer a more principled alternative. For instance, character-level CNNs are effective at identifying misspellings and informal text

patterns (Zhang et al., 2015), and hybrid architectures that fuse CNN character features with BERT embeddings have shown improvements on sentiment and fake news detection (Bao et al., 2021; Huang et al., 2021). CharBERT (Ma et al., 2020) goes further by replacing BERT's WordPiece embeddings with character-derived representations, pre-training on both masked language modelling and character reconstruction, yielding improved robustness on informal, user-generated text. At the byte level, models such as ByT5 (Xue et al., 2022) and CANINE (Clark et al., 2022b) operate directly on UTF-8 bytes or character sequences, eliminating fixed vocabularies and handling rare symbols, emojis, and code-switched text that BPE-based models struggle with (Kudo and Richardson, 2018). CharFormer (Tay et al., 2021) extends this paradigm through gradient-based subword tokenisation (GBST), dynamically learning latent subword blocks end-to-end, making it particularly suitable for noisy or offensive language. Google Jigsaw's Perspective API has publicly documented use of CharFormer-based architectures (Lees et al., 2022), underscoring its practical relevance. Despite these defences, adversarial attacks continue to expose model vulnerabilities through gradient-based token swaps (Ebrahimi et al., 2018), contextual synonym replacement (Jin et al., 2020), and emoji-based perturbations (Kirk et al., 2022), among others.

Moreover, current research has largely overlooked the role of **youth** (adolescents and young adults) as potential curators of adversarial samples. This oversight is noteworthy because young people are often early adopters and innovators of digital language, generating novel slang, memes, and intentionally obfuscated text that evolves rapidly across online platforms (Danescu-Niculescu-Mizil et al., 2013). These characteristics mirror the hallmarks of effective adversarial text: creativity, contextual fluency, and the ability to circumvent detection systems. Investigating the natural capacity of youth to produce "*evasion tactics*" can provide critical insights to (a) assess robustness of language models and (b) design appropriate defense mechanisms.

6.4.2 Adversarial Sample Curation

Procedure and Ethics: To generate adversarial samples for content related to cyberbullying, we selected 100 random sentences that were consistently misclassified as `non-bullying` by all models discussed in Section 6.2. These sentences were sourced from

user-comment platforms, including Instagram and Vine (Hosseinmardi et al., 2015; Rafiq et al., 2015) ($\mathcal{D}_{UC}$), and ASK.fm (Van Hee et al., 2018) ($\mathcal{D}_{QA}$). During the Young Research Assistantship Programme (Section 3.4, Chapter 3), students were invited to participate in an optional task to create variations of the 100 sentences. This activity was framed as a "*masking of words*" task, in which participants were asked to modify elements of each sentence such as profanities, verbs, adjectives, or phrases, using punctuation, intentional misspellings, or youth-specific slang to mimic realistic social media obfuscation. All potentially offensive words were pre-masked (e.g., `sh*t`) before participants modified them, and students could further obfuscate them using substitutions (e.g., `$h*t` or 💩).

This study was conducted after university ethics approval and follows similar ethical protocols discussed in Section 3.4 Chapter 3. Participation was voluntary, and students could choose how many sentences to work on, up to the full set of 100. The task was conducted in-person under supervision, ensuring a controlled environment and minimising exposure to unfiltered offensive content. To further reduce potential emotional distress, each participant had one-to-one discussion sessions during and after the activity, allowing them to reflect on the content and ask questions about any negative reactions. Following the activity, a debriefing session was held to review obfuscation strategies employed by participants and to provide guidance on managing potentially negative attitudes related to the task. Through this approach, we were able to collect a diverse set of youth-authored adversarial samples for 61 sentences. In total, only seven students (4M; 3F) opted for this task and were able to produce a total of 423 unique variations corresponding to the 61 original sentences. These curated samples included a range of linguistic transformations, such as slang, emoji substitutions, intentional character obfuscation, and informal syntax that reflect realistic adversarial manipulations in user-generated content.

Obfuscation Pattern Analysis: The 61 original sentences contained 499 unique words, of which seven student participants transformed 307 words (61.5%) into 1005 character- and word-level variations, yielding an average of 3.25 obfuscation variants per targeted word. This substantial variation rate demonstrates the creative adaptability youth might employ to circumvent content moderation systems. Building on (Gröndahl et al., 2018; Aggarwal and Zesch, 2022), we implemented a regular expression (regex) based pattern analyser in Python, supplemented with phonetic matching via the Meta-

phone algorithm (Philips, 1990) through the `phonetics` python implementation[100], to classify phonetic replacements. Our analysis reveals a hierarchical distribution of evasion strategies, as detailed in Table 6.19.

- **Leet speak** substitutions (n=219, 21.79%), where alphanumeric characters replace letters (e.g., "1" → "i", "3" → "e"), represented as the most common technique employed by youth to obfuscate texts.
- **Phrase transformations** (n=196, 19.5%), which involved changing the sub-phrases in the sentence, was another dominant category. This included replacements such as *kill* → *unalive*, *shutup b***h* → *shurup karen*.
- **Emoji replacement** (n=173, 17.2%) appeared as another category for obfuscation, this included noun or verb level replacement such as "sh*t" → 💩 or "cut" → 🗡️.
- **Special Characters** (n=110, 10.95%) involved inserting punctuation or symbols within words to disrupt pattern matching (e.g., "*f##k*", "*k!ll*", "*sh*t*").
- **Character Spaces** (n=107, 10.65%) employed spatial obfuscation by inserting spaces or hyphens between letters (e.g., *kill* → *k i l l*).
- **Phonetic Replacement** (n=102, 10.15%) captured sound-preserving substitutions identified through Metaphone phonetic matching (e.g., "you" → "u", "*f**k*" → "*phck*").
- **Character Removal** (n=64, 6.36%) involved deleting specific letters while preserving word structure (e.g., "*kill*" →"*kll*").
- **Word Deletion** (n=30, 2.99%) represented instances where targeted words were entirely removed from sentences rather than transformed, indicating complete avoidance as an evasion strategy.
- **Character Repetition** (n=4, 0.40%) comprised elongated words with repeated character (e.g., "ssssssshhhh**t").

[100] `https://github.com/ZackDibe/phonetics`

Re-annotation of Original Messages: During the review of the 61 original sentences obfuscated by the seven student participants, we observed that not all messages initially labelled as cyberbullying-related were inherently offensive. For instance, the sentence "`b***ch please i'm a pop star`" was originally labelled by (Van Hee et al., 2018) as cyberbullying-related under the binary scheme, yet on closer inspection, it appears to be a non-aggressive, casual expression containing profanity used in a playful or self-referential context rather than as an insult. To ensure a more accurate categorisation, we conducted a brief re-annotation exercise involving three annotators (2 Male; 1 Female) during their participation in the main annotation study (see Section 5.4.3 Chapter 5.4.3). Each annotator was asked to independently label the 61 sentences as either (a) **`hurtful`** or (b) **`not hurtful`**. The category **`hurtful`** encompassed both `covertly aggressive` and `overtly aggressive` expressions, consistent with the definitions used in Section 5.4.3 Chapter 5.4.3. Inter-rater reliability for this re-annotation task was assessed using Krippendorff's alpha (Krippendorff, 2011), which yielded a score of 0.71, indicating substantial agreement among the annotators. Out of the 61 sentences, full agreement (all three annotators assigning the same label) was reached for 52 sentences: 43 were unanimously classified as **`hurtful`**, and 9 as **`not hurtful`**. For the remaining 9 sentences, partial agreement was observed, with at least two annotators agreeing on one label. Among this, 5 were labelled as **`not hurtful`** and 4 were labelled to be **`hurtful`**. We consider these re-annotated labels as ground truth for further evaluations.

Pattern	**Count**
Leet Speak	219
Phrase Transformation	196
Emoji Replacement	173
Special Characters	110
Character Spaces	107
Phonetic Replacement	102
Character Removal	64
Word Deletion	30
Character Repetition	4

Table 6.19: Distribution of obfuscation patterns employed by seven participants (N=1,005 total variations). Patterns are mutually exclusive with each variation assigned to exactly one category.

6.4.3 Experiment Setup

As discussed previously, industry-led moderation APIs such as the OpenAI Moderation API (OpenAI, 2025) operate as closed systems, with their underlying architectures undisclosed. In contrast, other commercial moderation tools like Google Jigsaw's Perspective API (Jigsaw and Google's Counter Abuse Technology Team, 2025) have publicly shared details about their models, noting the use of a CharFormer (Tay et al., 2021) variant for toxicity detection (Lees et al., 2022). To examine model robustness against youth-designed perturbations and obfuscations in cyberbullying detection, we benchmark different character-level, sequential and hybrid architectures for detecting *hurtfulness* (as formulated in Section 6.3.2) in youth-designed obfuscated samples.

Model Architectures: We evaluate **three model** categories to understand how character-level, sequential, and hybrid architectures handle obfuscated text for binary classification. These include:

1. **Traditional character-level models**, that serve as traditional baselines to assess the limitations of vocabulary-dependent architectures when encountering obfuscated text.

    - **Char-CNN**: As described in Figure 6.13 we build on the character-level convolutional architecture by (Kim et al., 2016). After character-level tokenisation, a character embedding layer converts each raw character into a continuous vector representation, allowing the model to learn semantic and syntactic properties of characters (vocabulary size ≈70-100 characters; 50-dimensional learned embeddings). We then apply four parallel one-dimensional (1D) convolutional layers with kernel sizes of 3, 4, 5, and 7 characters, each comprising 256 filters. Each filter spans the entire embedding dimension (e.g., 3×50) and slides along the sequence axis, enabling the model to capture local character *n-gram* features such as prefixes, suffixes, and root fragments. The resulting feature maps are passed through a ReLU activation for non-linearity, followed by max-pooling over the sequence length to retain the most salient (maximum value) features and obtain a fixed-length representation independent of input size. The pooled outputs from all four convolutional layers are concatenated

into a single comprehensive feature vector. Finally, dropout regularization is applied to reduce over-fitting, and a dense output layer with a sigmoid activation produces the final prediction scores for the target classes.

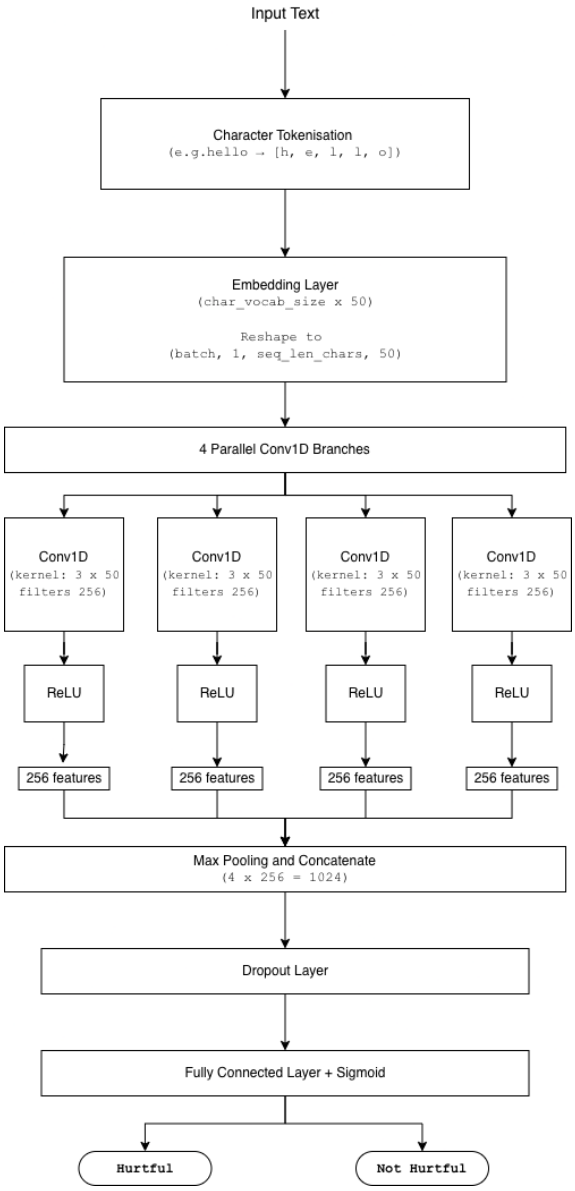


Figure 6.13: Architectural Flow of Char-CNN

- **Char-BiLSTM**: As described in Figure 6.14, we build on the character-level recurrent architecture proposed by Zhou et al. (2016), using the same character-level tokenizer and embedding setup as in Char-CNN. To process the character embeddings sequentially in both forward and backward directions, we employ two stacked layers of bidirectional LSTMs. The first layer processes the raw character embeddings, while the second layer captures higher-level

temporal dependencies from the outputs of the first. To improve generalisation and mitigate over-fitting, dropout is applied only to the outputs of the first BiLSTM layer before feeding them into the second layer. The final forward and backward hidden states from the top layer (each 256-dimensional) are concatenated to form a single 512-dimensional sequence representation summarising the entire input. This setup allows the model to learn increasingly abstract and hierarchical representations of character sequences. Finally, a fully connected layer with a sigmoid activation produces the binary classification scores.

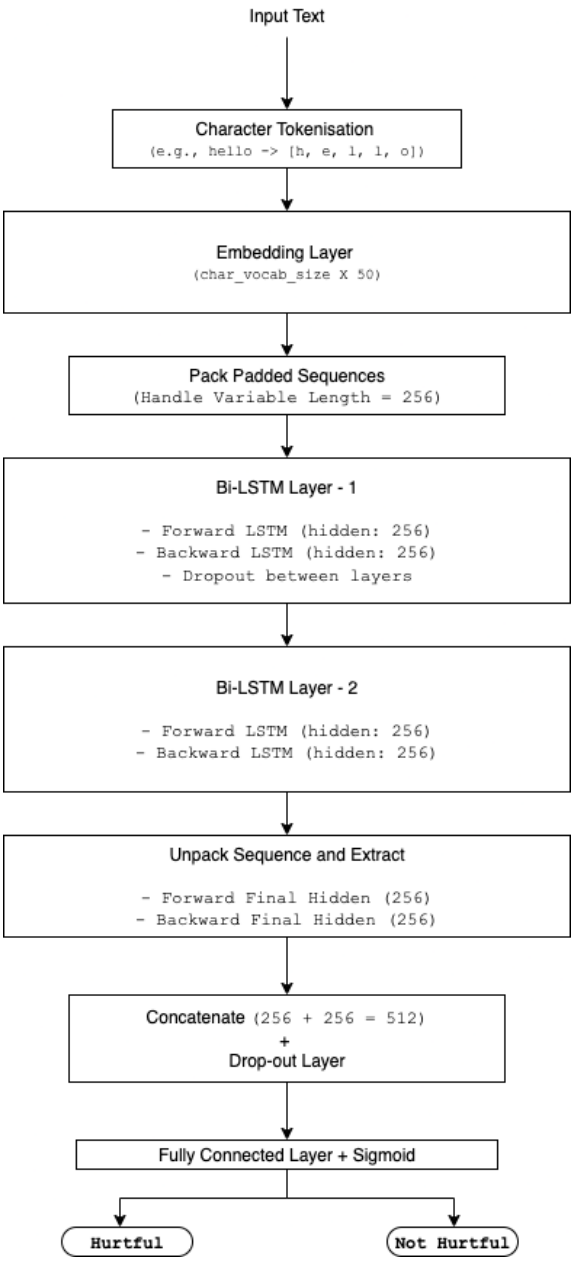


Figure 6.14: Architectural Flow of Char-LSTM

2. **Pre-trained BERT** as a transformer baseline for comparison with standard contextualised representations, we use the BERT model by (Devlin et al., 2019) for fine-tuning on the cyberbullying detection task, serving as the contextual baseline without explicit character-level modelling.

3. **Hybrid Character-Word Fusion**: Following the design of hybrid architectures in prior work (Bao et al., 2021; Alghamdi et al., 2022), we construct a **character-CNN-BERT** model that integrates local character-level features with contextual word-level embeddings, as illustrated in Figure 6.15. The architecture consists of two parallel encoding pathways:

   (A) **Character Pathway**: Each character is first mapped to a continuous embedding vector, using the same character embedding setup as in Char-CNN. Multi-kernel one-dimensional (1D) CNNs with kernel sizes of 3, 4, and 5 characters (256 filters each) are applied to extract local character patterns, effectively capturing character n-grams such as prefixes, suffixes, and morphemes. To aggregate features from variable-length character sequences into a fixed-size representation, adaptive max pooling is applied over the sequence length. Unlike average pooling, which computes the mean across the sequence, adaptive max pooling selects the maximum activation over the sequence for each filter. Thereby, producing a fixed-length vector regardless of the input sequence length that summarises the most salient character-level patterns across the entire sentence. These pooled vectors are then processed through a multi-head self-attention mechanism (eight heads), enabling the model to weigh the relative importance of different character-level features across the sequence and capture long-range dependencies that CNNs alone may miss. Positional information is implicitly preserved through the CNN convolution operations, providing the attention mechanism with a sense of character order. The output of this pathway is a 256-dimensional character-level feature vector for each input sequence.

   (B) **Word Pathway**: Simultaneously, the same text sequence is tokenised into words and processed through the pre-trained BERT-base-uncased model, pro-

ducing 768-dimensional contextualised embeddings for each token. These embeddings capture semantic and syntactic dependencies over the full sequence, complementing the local character-level patterns extracted in the character pathway. A pooling operation (e.g., [CLS] token representation or mean pooling) is used to convert BERT outputs into a fixed-length 768-dimensional sequence vector.

The outputs from the two pathways are concatenated, combining the local character-level features with the contextual word-level representations into a unified 1024-dimensional vector. This concatenated vector is projected to 256 dimensions through a linear feature fusion layer, followed by layer normalisation, ReLU activation, and dropout regularisation to improve generalisation. While this reduces dimensionality and may discard some fine-grained information, it allows the model to focus on the most salient combined features and improves generalisation. Finally, a two-layer fully connected classification head maps the fused features to the target classes. The first layer reduces the feature dimension with ReLU activation and dropout, while the second layer outputs raw logits corresponding to the number of target classes. No activation is applied within the model itself; instead, the logits are passed to the loss function (Binary Cross-Entropy) for binary classification.

Dataset and Evaluation Method: We extend the binary hurtfulness detection task from the multi-task classification framework described in Section 6.3.2, and utilise the same multi-source datasets: ($\mathcal{D}^{HT}_{IT\rightarrow EN}$) human translation (Section 5.1.1); ($\mathcal{D}^{HMT}_{FR\rightarrow EN}$) human-aided machine translation of French cyberbullying content to English (Ollagnier et al., 2022; Ollagnier, 2024) (Section 5.1.3); and ($\mathcal{D}_{\text{AMiCA}}$) English dataset from Van Hee et al. (2018). We maintain consistent training, validation, and test splits as detailed in Table 6.11 of Section 6.3.2. Model performance is evaluated using macro-average F1 score on both the held-out test set and the youth co-curated $\mathcal{D}_{\text{Gold}}$ dataset. To benchmark these architectures against industry moderation tools, including transparent systems like Google Jigsaw's Perspective API and proprietary solutions such as OpenAI's Moderation API, we adopt a robustness-focused evaluation framework. Goyal et al. (2023) emphasise the critical need for **sensitivity metrics** in machine learning models, particularly for

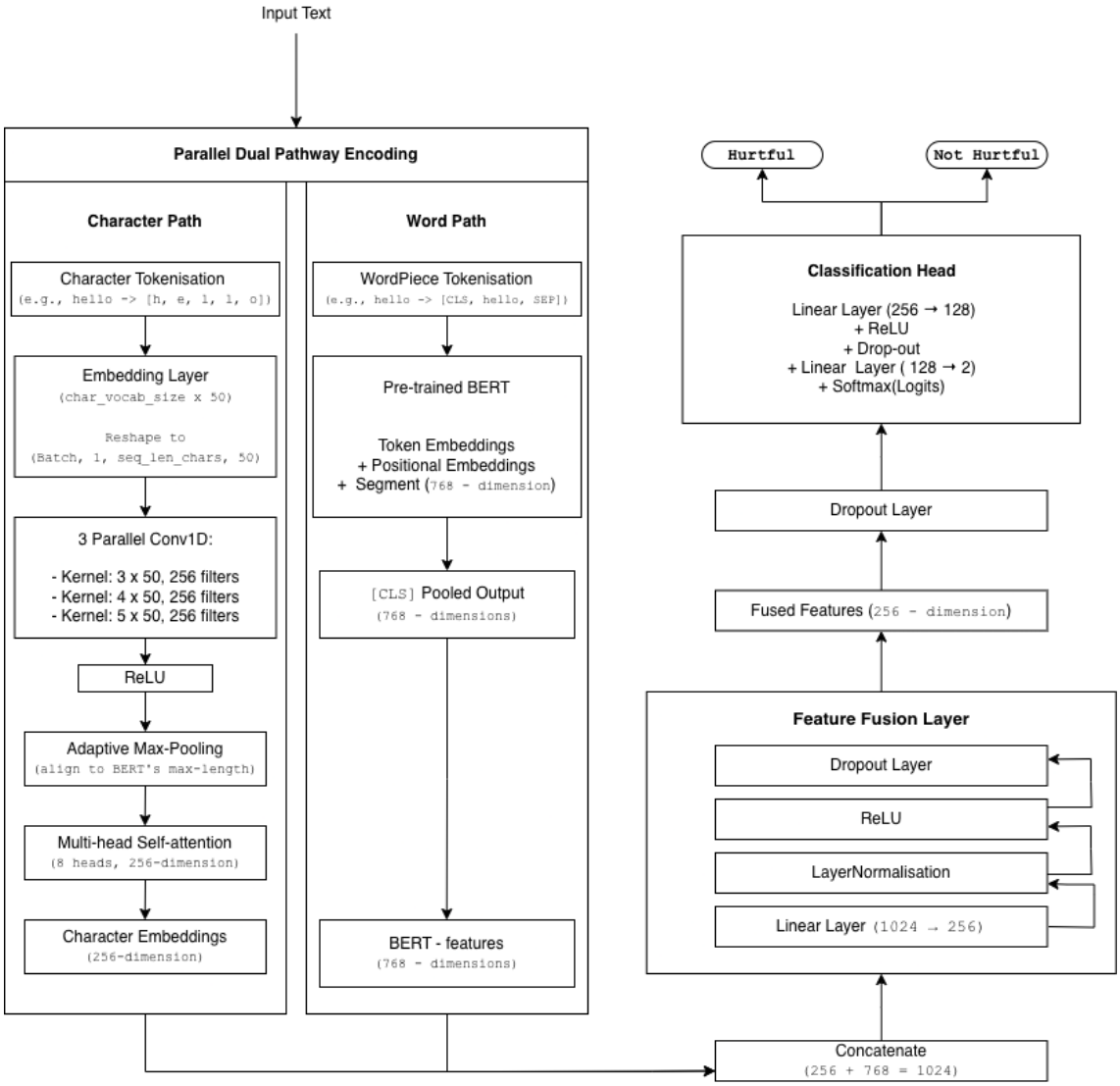


Figure 6.15: Architectural Flow of Character-CNN-BERT

assessing their resilience to adversarial perturbations. They further note the scarcity of comprehensive evaluation methodologies that account for both perceptibility and the naturalness of adversarial examples. Following established practices in adversarial robustness research (Carlini and Wagner, 2017; Zhang et al., 2020b), we employ the 61 adversarially perturbed sentences as our primary evaluation set and analyse model behaviour through the following metrics:

- **Flip Rate** (Alzantot et al., 2018; Jin et al., 2020): The proportion of samples whose predicted labels change following text obfuscation, computed as:

$$\text{Flip Rate} = \frac{\#\text{ changed predictions}}{\#\text{ total samples}} \tag{6.1}$$

This metric quantifies prediction instability and can be disaggregated by class to identify differential vulnerabilities across label categories.

- **Robustness Score** (Jin et al., 2020): : Computed as $1 - \text{Flip Rate}$, providing an interpretable measure of model stability where higher scores indicate greater resilience to perturbations. This metric is reported both overall and per-class.
- **Confidence Drop**: The average decrease in model confidence after obfuscation, measured as the mean absolute difference between original and perturbed prediction probabilities:

$$\text{Confidence Drop} = \text{mean}\left(|\text{confidence}_{\text{original}} - \text{confidence}_{\text{obfuscated}}|\right) \tag{6.2}$$

  Here, confidence is the predicted probability of the top class. This captures uncertainty shifts even when the predictions remain unchanged.

- **Attack Success Rate (ASR)** (Carlini and Wagner, 2017; Zhang et al., 2020b): Following adversarial machine learning literature, we define ASR as the proportion of originally correct predictions that become misclassified after obfuscation:

$$\text{ASR} = \frac{\#\text{ correct} \rightarrow \text{incorrect}}{\#\text{ originally correct}} \tag{6.3}$$

  This metric directly quantifies the effectiveness of text obfuscation as an adversarial attack strategy.

Training Configuration:

- **Character-Level Models (Char-CNN & Char-BiLSTM):** Search space includes learning rate $[10^{-5}, 10^{-3}]$, batch size $\{16, 32, 64\}$, filters $\{50, 100, 150\}$ (Char-CNN), dropout $[0.1, 0.5]$, and weight decay $[10^{-6}, 10^{-3}]$, with fixed embedding (50) and hidden dimension (256).
- **Pre-trained BERT Baseline:** Fine-tuning explores learning rates $\{1e^{-5}, 2e^{-5}, 3e^{-5}, 5e^{-5}\}$, weight decay $\{0.001, 0.01\}$, warmup $\{0.1, 0.2\}$, epochs $\{3, 5\}$, and batch sizes $\{16, 32\}$.
- **Hybrid Character-CNN-BERT:** Uses similar optimisation as BERT, additionally tuning character embeddings $\{64, 128\}$, fusion hidden size $\{128, 256\}$, and dropout

$\{0.1, 0.2, 0.3\}$.

- All models use binary cross-entropy loss (`BCEWithLogitsLoss`)[101], AdamW optimisation with gradient clipping (1.0), and early stopping (patience = 10) based on validation F1-macro. BERT-based models use linear warmup/decay, while character models apply `ReduceLROnPlateau`[102].

6.4.4 Results and Discussion

Performance on Test-set and $\mathcal{D}_{\text{Gold}}$: The performance of the four models discussed in Section 6.4.3 on both the test-set and $\mathcal{D}_{\text{Gold}}$ are reported in Table 6.20. On the test set, the char-CNN-BERT hybrid architecture achieves the highest macro-average F1 score of 0.7920, demonstrating superior performance compared to individual architectures. BERT-based models (BERT: 0.7766, char-CNN-BERT: 0.7920) substantially outperform character-level models (char-CNN: 0.7507, char-BiLSTM: 0.7509), with improvements of approximately 3-4 percentage points in macro-average F1. This performance gap is particularly pronounced in detecting hurtful content, where char-CNN-BERT achieves an F1 score of 0.7172 compared to 0.6493 for char-CNN. However, a substantial performance degradation is observed across all models when evaluated on $\mathcal{D}_{\text{Gold}}$, with macro-average F1 scores dropping by 21-24 percentage points. The char-CNN-BERT model, while maintaining relative superiority (F1: 0.5762), exhibits a notable increase in false positives (32.69%) and a high false negative rate of 47.9%. This suggests that youth-generated content contains linguistic patterns and contextual nuances that differ substantially from the training distribution. The elevated false negative rates across all models (ranging from 47.9% to 66.95%) indicate misclassification of harmful content in youth discourse, highlighting a critical gap in model generalisation to adolescent communication styles.

Robustness to Adversarial Obfuscation: Table 6.22 shows the model performance on 61 original sentences and their 423 adversarially obfuscated variations, including comparisons with industry-standard APIs. On the original sentences, OpenAI's

[101] `https://docs.pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html`

[102] `https://docs.pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html`

Moderation API achieves the highest macro-average F1 score (0.7964), followed closely by BERT (0.7621) and char-BiLSTM (0.7553) and the hybrid architecture (char-CNN-BERT: 0.7458). Notably, all custom-trained models outperform Google's Perspective API (0.6020), which exhibits a substantially elevated false negative rate of 44.68%. When subjected to adversarial obfuscation, performance degradation is observed across all systems, though with varying severity. OpenAI's Moderation API demonstrates the most robust performance, maintaining a macro-average F1 of 0.6758 (a decline of 15.1%), while BERT shows relative resilience with F1 of 0.6507 (14.6% decline). In contrast, Perspective API suffers catastrophic performance collapse, dropping to an F1 of 0.4440 (26.2% decline) with false negatives increasing to 69.32%. Among character-level models, char-BiLSTM exhibits marginally better robustness (F1: 0.6191) compared to char-CNN (0.6029) and char-CNN-BERT (0.6098). The false positive and false negative trade-off shifts notably under adversarial conditions. While Perspective API maintains low false positive rates (9.27%), its inability to detect obfuscated harmful content results in near-random classification performance. Conversely, OpenAI's API maintains higher false positives (28.86%) but achieves substantially better recall on adversarial samples, suggesting a more conservative classification strategy that prioritises harm detection over precision.

Models	**Hurtful F1**	**Not-Hurtful F1**	**Macro-Average F1**	**False Positive %**	**False Negative %**
			Test		
char-CNN	0.6493	0.8522	0.7507	15.6%	38.41%
char-BiLSTM	0.6348	0.8581	0.7509	11.4%	40.23 %
BERT	0.6960	0.8571	0.7766	15.8%	27.51%
char-CNN-BERT	0.7172	0.8667	0.7920	15%	25.14%
			$\mathcal{D}_{\text{Gold}}$		
char-CNN	0.5057	0.5470	0.5264	24.52%	61.11%
char-BiLSTM	0.5578	0.4575	0.5076	18.75%	66.95%
BERT	0.5607	0.5444	0.5526	29.32%	54.09%
char-CNN-BERT	0.6054	0.5469	0.5762	32.69%	47.9%

Table 6.20: Performance comparison of models on held-out test set and youth co-created $\mathcal{D}_{\text{Gold}}$ dataset. FP: False Positive rate, FN: False Negative rate.

Prediction Stability Analysis: Table 6.22 presents flip rates and robustness scores, quantifying prediction consistency across original and obfuscated text pairs. Robustness scores reveal vulnerabilities across all models, with hurtful content consistently more susceptible to adversarial perturbations than benign content. For hurtful samples, robustness scores range from 0.7116 (char-CNN-BERT) to 0.7975 (char-CNN), indicating that 20-29% of hurtful predictions flip under obfuscation. In contrast, not-hurtful samples demonstrate substantially higher stability, with robustness scores exceeding 0.82 across all models. Perspective API exhibits the most pronounced asymmetry: while achieving exceptional stability on not-hurtful content (robustness: 0.9587, flip rate: 4.12%), it demonstrates moderate instability on hurtful content (flip rate: 26.38%). This pattern suggests the API's aggressive filtering may be circumvented through relatively simple obfuscation techniques. OpenAI's Moderation API shows more balanced robustness across classes (hurtful: 0.7914, not-hurtful: 0.8659), though with higher flip rates for benign content (13.4%) compared to other models. Char-CNN demonstrates the highest robustness

Models	**Hurtful F1**	**Not-Hurtful F1**	**Macro-Average F1**	**False Positive %**	**False Negative %**
Original 61 sentences					
char-CNN	0.8235	0.5946	0.7091	21.43%	25.53%
char-BiLSTM	0.8636	0.6471	0.7553	21.43%	19.14%
BERT	0.8791	0.6452	0.7621	28.57%	14.89 %
char-CNN-BERT	0.8667	0.6250	0.7458	28.57%	17.02%
Perspective API	0.6933	0.5106	0.6020	14.28%	44.68 %
OpenAI Mod. API	0.9032	0.6897	0.7964	28.57%	10.63 %
Adversarial Samples (423 sentences)					
char-CNN	0.7090	0.4968	0.6029	20.61%	41.71%
char-BiLSTM	0.7236	0.5147	0.6191	18.55%	40.18%
BERT	0.7572	0.5442	0.6507	17.52%	35.88 %
char-CNN-BERT	0.7011	0.5185	0.6098	13.40%	43.86%
Perspective API	0.4598	0.4282	0.4440	9.27%	69.32 %
OpenAI Mod. API	0.8061	0.5455	0.6758	28.86%	26.68 %

Table 6.21: Performance comparison on original sentences ($n = 61$) and adversarially obfuscated variations ($n = 423$). Industry APIs (Perspective, OpenAI) are benchmarked against custom-trained models.

for hurtful content (0.7975), suggesting that character-level representations may capture orthographic variations more effectively than subword tokenization. However, the hybrid char-CNN-BERT architecture paradoxically exhibits the lowest robustness score (0.7116), with flip rates of 28.83% for hurtful content, indicating that ensemble approaches do not necessarily enhance adversarial resilience.

Models	**Label**	**Flip Rate**	**Robustness Score**
char-CNN	`Hurtful (1)`	0.2024	0.7975
	`Not-Hurt (0)`	0.0515	0.9484
char-BiLSTM	`Hurtful (1)`	0.2484	0.7515
	`Not-Hurt (0)`	0.0515	0.9484
BERT	`Hurtful (1)`	0.2484	0.7515
	`Not-Hurt (0)`	0.1752	0.8247
char-CNN-BERT	`Hurtful (1)`	0.2883	0.7116
	`Not-Hurt (0)`	0.1546	0.8453
Perspective API	`Hurtful (1)`	0.2638	0.7361
	`Not-Hurt (0)`	0.0412	0.9587
OpenAI Mod. API	`Hurtful (1)`	0.2085	0.7914
	`Not-Hurt (0)`	0.1340	0.8659

Table 6.22: Per-class flip rates and robustness scores across models. Robustness score is computed as $1 - \text{Flip Rate}$

Attack Success Rate: Table 6.23 quantifies the vulnerability of correctly classified samples to be mis-classified through the Attack Success Rate (ASR) metric. Overall ASR values range from 0.1856 (OpenAI) to 0.3144 (Perspective API), indicating that 18-31% of originally correct predictions become misclassified after obfuscation. Perspective API demonstrates the highest vulnerability (ASR: 0.3144), with 46.11% of correctly identified hurtful samples flipping to incorrect predictions under adversarial conditions. This substantially exceeds other models and confirms the API's brittleness to text perturbations. For hurtful content, ASR values range from 0.2061 (OpenAI) to 0.4611 (Perspective), while not-hurtful content shows near-zero ASR for several models (char-CNN-BERT: 0%, Perspective: 0%). This suggests that the obfuscation techniques employed by the young research assistants result in higher attack success rates for the hurtful class. OpenAI's Moderation API achieves the lowest overall ASR (0.1856) and demonstrates balanced resilience across classes, though with 10% ASR for not-hurtful samples indicating some false

alarm vulnerability. Char-CNN-BERT exhibits the highest vulnerability (overall ASR: 0.2676, hurtful ASR: 0.3357), while char-CNN shows marginally better robustness (overall ASR: 0.1938). Char-CNN-BERT and Perspective API reamin robust to attacks on not-hurtful samples (ASR: 0%) suggesting that these models may employ conservative classification boundaries that resist false positive flips, at the cost of higher vulnerability to false negatives in the hurtful class.

Models	**Hurtful ASR**	**Not-Hurt ASR**	**ASR**
char-CNN	0.2459	0.0263	0.1938
char-BiLSTM	0.2840	0.0131	0.2235
BERT	0.2661	0.0434	0.2219
char-CNN-BERT	0.3357	0	0.2676
Perspective API	0.4611	0	0.3144
OpenAI Mod. API	0.2061	0.1	0.1856

Table 6.23: Attack Success Rate (ASR) per class and overall, computed as the proportion of originally correct predictions that become incorrect after obfuscation.

Confidence Degradation Under Adversarial Perturbation: Table 6.24 presents the average confidence (probability) drop across model predictions, capturing uncertainty shifts even when classifications remain unchanged. Character-level models exhibit substantially larger confidence perturbations, with char-BiLSTM showing the highest drops for hurtful content (0.1650) and char-CNN demonstrating significant instability for not-hurtful samples (0.0422). In contrast, BERT-based architectures display remarkable confidence stability, with average drops below 0.08 across both classes, suggesting that contextual embeddings provide more robust probability calibration under adversarial conditions. The confidence drop patterns reveal distinct architectural behaviours under perturbation. Transformer-based models (BERT: 0.0627/0.0542, char-CNN-BERT: 0.0757/0.0642) maintain relatively consistent confidence across label types, indicating balanced uncertainty responses. Conversely, character-level models show pronounced asymmetry, with char-BiLSTM exhibiting nearly four times higher confidence drops for hurtful content compared to not-hurtful content (0.1650 vs. 0.0422). This suggests that character-level architectures are more sensitive to surface-form variations in harmful language, where

obfuscation techniques directly manipulate orthographic features. Industry APIs demonstrate intermediate confidence stability. OpenAI's Moderation API shows the highest confidence drop for hurtful content (0.1950), potentially reflecting its conservative classification strategy that maintains high uncertainty under ambiguous inputs. Perspective API exhibits moderate drops (hurtful: 0.1443, not-hurtful: 0.0938), though these values must be interpreted alongside its catastrophic flip rate and ASR metrics. Notably, lower confidence drops do not necessarily indicate superior robustness, BERT maintains stable confidence while still experiencing substantial performance degradation (F1 decline: 14.6%), suggesting that models can be confidently wrong under adversarial conditions.

Models	**Label**	**Avg. Confidence Drop**
char-CNN	`Hurtful (1)`	0.2024
	`Not-Hurt (0)`	0.0515
char-BiLSTM	`Hurtful (1)`	0.1650
	`Not-Hurt (0)`	0.04222
BERT	`Hurtful (1)`	0.0627
	`Not-Hurt (0)`	0.0542
char-CNN-BERT	`Hurtful (1)`	0.0757
	`Not-Hurt (0)`	0.0642
Perspective API	`Hurtful (1)`	0.1443
	`Not-Hurt (0)`	0.09378
OpenAI Mod. API	`Hurtful (1)`	0.1950
	`Not-Hurt (0)`	0.0733

Table 6.24: Average confidence (probability) drop per class across models, computed as the mean absolute difference between original and obfuscated prediction probabilities.

Error Analysis: The original sentence, "⚫ *your face after I put my hands on you*" which received a unanimous `hurtful` label from all three annotators, is consistently misclassified as `not-hurtful` by all systems, including both the OpenAI Moderation API and Google Jigsaw Perspective API. Similarly, the sentence "*you have downs*" is also consistently misclassified as `not-hurtful` across all models. Minor variations, such as "*you have* ⬇️" or "*you have dwnS*" are also consistently misclassified by all models. However, its other variants, such as "*your a downie*" and "*you have dOvvns*", though identified as `hurtful` by the OpenAI Moderation API, yield relatively low confidence scores of 0.523

and 0.398 by the moderation API, respectively. Notably, the phrase "*your a downie*" is correctly classified by both the char-CNN-BERT and base-BERT models, whereas all other models fail to do so. These results highlight that all models struggle to detect subtle, context-dependent, or implicitly hurtful expressions, particularly those involving coded language. The consistent misclassification of phrases referencing disabilities suggests that these models may lack sufficient sensitivity to marginalised identity, possibly due to imbalanced or insufficiently diverse training data. Moreover, the low confidence scores for obfuscated or misspelt slurs indicate that, though effective at identifying overt hurtful speech, the models remain vulnerable to linguistic variation, creative spelling, and contextual nuance, thereby underestimating the real-world harm of subtle or implicit hurtfulness.

Overall, we find **no model is truly robust**. All models show vulnerability to youth-created obfuscations, with performance drops of 14-26% in F1 scores. Even the best-performing system (OpenAI Moderation API) has an Attack Success Rate (ASR) of 18.56%, meaning nearly 1 in 5 correct predictions can be flipped through simple text perturbations. Character-level models are robust to hurtful content obfuscation (char-CNN: 0.7975 robustness score), but with high confidence drops (0.1650), suggesting they detect surface perturbations but struggle to maintain stable predictions. BERT, when fine-tuned for hurtful detection, has best F1 performance and confidence stability, but for hurtful content has moderate flip rates (24-29%). The hybrid char-CNN-BERT architecture shows the worst robustness (flip-rate: 28.83%), suggesting that simply combining architectures does not enhance adversarial resilience. We also find that Industry tools are not superior, Perspective API fails under obfuscation (F1: 0.4440, ASR: 31.44%, FN: 69.32%), making it essentially unreliable for detecting obfuscated harmful content. OpenAI API performs best but is still vulnerable, particularly maintaining high false positives (28.86%) on youth-created obfuscated samples.

6.5 Summary

This chapter investigates state-of-the-art techniques for detecting cyberbullying and harmful online content among youth, directly addressing following research gaps:

C.3, i.e., *Limited benchmarking of classification approaches (e.g., traditional machine learning, pre-trained language models, role-aware models) for (i) detecting whether the content is* `hurtful` *and (ii) identifying the* `user roles` *in cyberbullying contexts*; and **C.4**, i.e., *Content altered through obfuscation or perturbation frequently bypasses moderation and limits the robustness of detection systems.* This chapter fills critical gaps by benchmarking classification paradigms for CB-specific detection **RQ2.1** and adapting techniques to handle obfuscated content **RQ2.2**.

The chapter explores **four** interconnected themes, examining both **what is said** (`hurtful` vs. `not-hurtful`) and **who says it** (`user-roles`), alongside model robustness:

Theme 1: Binary Benchmarking To explore **RQ2.1**, we extend Ireland's (National Advisory Council for Online Safety, 2021) suggestion to phrase cyberbullying as repeated `nasty or hurtful` behaviour and apply it to formulate it as a binary task as precursor to detect CB. We evaluate traditional ML (SVM, Random Forest, XGBoost), sequential deep learning (Bi-LSTM, GRU), and transformer models (BERT, HateBERT) across different social media categories, such as user-comment (e.g., Instagram/Vine), question-answering (e.g., Ask.fm), micro-blogging (e.g., X). Transformer-based models, particularly HateBERT with domain-adaptive pre-training when fine-tuned for binary classification task consistently outperform classical and sequential approaches in both within-platform and cross-platform settings. When assessing **translation strategies** using Italian youth WhatsApp data by (Sprugnoli et al., 2018) and its human and machine translations, Section 5.1.1 and Section 5.1.2, respectively, BERT achieves the highest performance on both human-translated ($\mathcal{D}^{HT}_{IT\rightarrow EN}$) (macro F1: 0.81) and machine-translated ($\mathcal{D}^{MT}_{IT\rightarrow EN}$) (F1: 0.75) corpora, demonstrating superior capability to capture nuanced linguistic cues despite translation-induced noise. This result is stable even when tested with translation text as seen in Section 6.1.4, suggesting widespread utility for platforms which typically have global user-bases.

Theme 2: Moving Towards Explainable CB Detection This extends evaluation in Theme-1 to gaming forums with domain-specific slang, comparing general-purpose BERT against domain-adapted HateBERT. While HateBERT consistently outperforms

BERT across all datasets (achieving F1 of 0.73 versus 0.65 on question-answering data, and 0.81 versus 0.77 on micro-blogging platforms), interpretability analyses reveal critical limitations in using attention mechanisms to explain model decisions. Following established transformer interpretability methodologies, attention weights were extracted and complemented with gradient-based feature importance scores using Integrated Gradients on a reserved 10% test subset. Pearson correlation coefficients between mean attention weights and gradient-based importance scores prove consistently weak across all models (ranging from 0.0129 to 0.19), indicating that attention weights alone do not reliably reflect token importance even when models achieve strong F1-scores. Layer-wise POS analysis reveals that attention does not prioritise linguistically salient features—adjectives, verbs, and nouns receive comparable or lower attention than auxiliary verbs, determiners, and conjunctions. Sentiment-based analysis further contradicts expectations: positive words (like, cute, pretty) receive attention weights comparable to or exceeding negative sentiment tokens (f**k, b**ch, stupid). These findings confirm prior research demonstrating that while domain-adapted models like HateBERT achieve superior detection performance, their decision-making processes remain opaque, highlighting a critical gap between model effectiveness and explainability in safety-critical cyberbullying detection applications.

Theme 3: Multi-class and Multi-task Classification We address role identification (`bully`, `victim`, `enabler`, `defender`) using merged multi-source datasets. Direct multiclass, hierarchical pairwise, and prompt-based LLM approaches are compared. Few-shot LLaMA-3.3-70B achieves competitive performance (macro F1: 0.48) with supervised baselines, though all models struggle with minority classes (victim F1: 0.19-0.22) and obfuscated defender messages. Multi-task learning jointly models hurtfulness detection and role identification using hard parameter sharing and AdapterFusion strategies. BERT with hard parameter sharing and cross-entropy (BERT-HP-CE) consistently outperforms GPT-2 across tasks (test F1: 0.718 for hurtfulness, 0.555 for roles). Error analysis reveals persistent confusion between bully and enabler roles (30-60% misclassification rates), highlighting challenges in capturing subtle linguistic cues and sarcasm.

Theme 4: Adversarial Robustness To evaluate model resilience against youth-created obfuscation strategies, seven students generated 423 variations of 61 sentences

using leet speak (21.79%), phrase transformations (19.5%), emoji replacements (17.2%), and special characters (10.95%). Results demonstrate that no model is truly robust: all systems show 14-26% F1 degradation under obfuscation. The hybrid char-CNN-BERT architecture, while achieving the highest test performance (F1: 0.7920), exhibits the worst robustness (flip rate: 28.83%, attack success rate: 26.76%). Character-level models demonstrate better stability for hurtful content (char-CNN robustness: 0.7975) but suffer large confidence drops. Industry APIs show mixed results—Perspective API catastrophically fails (F1 drops from 0.60 to 0.44, false negatives: 69.32%), while OpenAI's Moderation API maintains relative resilience (ASR: 18.56%) but with elevated false positives (28.86%).

Overall, this chapter demonstrates that transformer-based models, particularly domain-adapted variants, outperform traditional and sequential architectures for cyberbullying detection, though attention mechanisms remain unreliable for interpretability. Multi-task learning with hard parameter sharing effectively balances hurtfulness and role detection, with BERT-based encoders proving more stable than GPT-2 decoders. However, critical limitations persist: minority role detection remains challenging, obfuscation techniques substantially degrade all model performance, and even state-of-the-art systems are vulnerable to youth-created perturbations. These findings underscore the need for more robust architectures and richer contextual modelling to address the dynamic, nuanced nature of youth cyberbullying communication.

Chapter 7

Vision-Language Models for Meme-Based Harm Detection

Abstract

Building on the youth-centered research introduced in earlier chapters, this chapter investigates the potential of vision–language models (VLMs) to detect meme-based toxicity. Given that memes can serve as tools for bullying and harassment, evaluating their harmful or hateful nature is crucial for detecting malicious intent. Moreover, as discussed in §2.4.5 of Chapter 2, controlling audience exposure, i.e., deciding **who can see what**, has become a key strategy through which very large online platforms (VLOPs) seek to protect younger users from psychologically distressing content. To this end, the chapter examines whether VLMs can generalize to age-differentiated moderation decisions through a three-phase methodology. First, in §7.1.1 we introduce a newly curated dataset of 300 previously unseen memes, annotated under (Röttger et al., 2022)'s "descriptive paradigm" to capture binary judgments about whether content should be restricted for adults versus adolescents. Second, in §7.1.2 we benchmark both unimodal and multimodal models, including both early and late fusion multimodal models to evaluate their ability to generalise across differing definitions of harm. Third, in §7.1.3 we develop a novel human preference dataset to compare the reasoning pro-

cesses and alignment of closed-source models such as GPT-4o and Gemini-2.5 in making audience-specific visibility decisions. These contributions advance understanding of multimodal AI for safer digital spaces and provide practical resources for researchers and platforms to implement audience-specific cyberbullying (CB) content moderation effectively.

7.1 Methodology

To explore the research question **RQ.3**, this study follows **three** phases (See Figure 7.1) that places benchmarking within a broader pipeline for robust evaluation and human-grounded interpretation. These include,

- **Phase-1 (Unseen data curation)**: Modern VLMs are often trained on web-scale data, where overlap with public benchmarks can inflate evaluation scores (Brown et al., 2020; Dodge et al., 2021; Song et al., 2025). To approximate real-world generalisation and avoid optimistic assessments (Brown et al., 2020; Song et al., 2025), a new unseen test set was collected from Imgur[103] using keywords provided by young researchers through the Young Researcher Assistant Programme (Section 3.4, Chapter 3). Building on the "**descriptive**" paradigm (Röttger et al., 2022) discussed in Section 2.3.3 Chapter 2, we implement a novel data annotation protocol designed to seek if ***content should be `limited` or `not-limited`*** for different user groups (e.g., `adult-users`: Facebook; `young-users`: TikTok). These annotations provide a coarse-grained signal of audience suitability for subsequent evaluation.

- **Phase-2 (Baseline Benchmarking)**: Extending the baseline benchmarking by (Zhao et al., 2023; Pramanick et al., 2021a) for multimodal content `hateful` and `harmful` classification of memes, we evaluate the following multi-tiered model suite on Imgur memes collected in part of Phase-1:

  1. Unimodal Image-only ($\mathcal{U}^I$): Vision Transformer ($\mathcal{U}^I_{ViT}$ (Dosovitskiy et al., 2021); CLIP-image-only encoder (Radford et al., 2021) $\mathcal{U}^I_{LIP}$).

[103] `https://imgur.com`

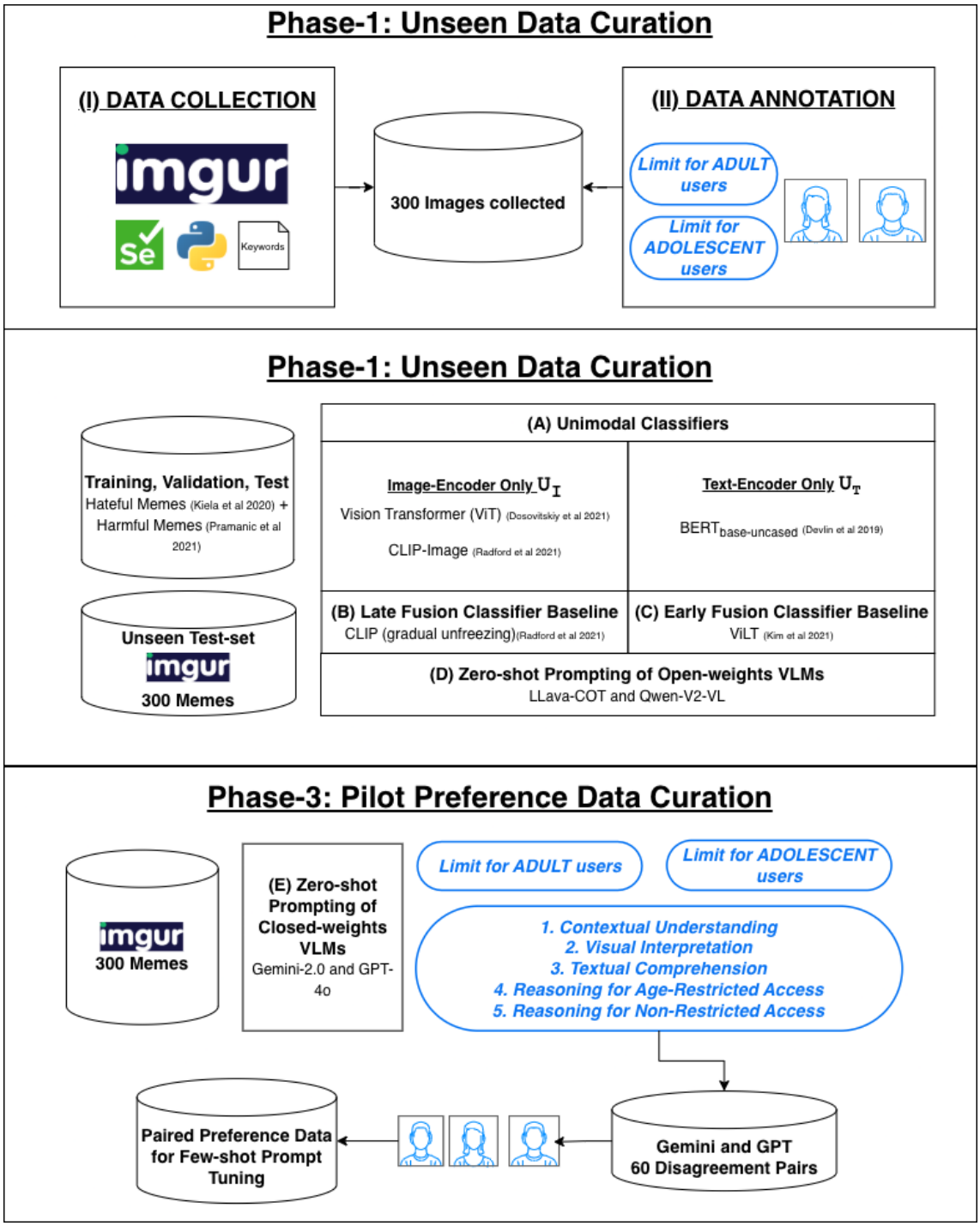


Figure 7.1: Methodological Framework for **RQ2.2** Section mall*Note: '*' indicates values differing from (Kiela et al., 2020), as original test-set labels are unavailable (total 9,000 images). 80% is used for training (90% of original training), 10% of training plus half of the dev set for validation, and the remaining 500 samples for testing.*

2. Unimodal Text-only ($\mathcal{U}^T$): BERT (Devlin et al., 2019).

3. Multimodal Late Fusion (MF^L) : CLIP-based classifier (Radford et al., 2021).

4. Multimodal Early Fusion MF^E: Vision Language Transformer (ViLT) (Kim

et al., 2021c).

5. Large vision-language models: Llava-COT (Xu et al., 2024), Qwen-V2-VL (Bai et al., 2023; Team et al., 2024).

Conducted under a binary classification protocol, paired with qualitative error analysis on **unseen** data collected in **Phase-1** to detect if they are `hateful` or `harmful` or `neither`. Although binary labels provide a straightforward comparative metric, the error analysis (e.g., categorising failures into **explicit** vs **implicit**) with the unseen data helps analyse whether models truly rely on multimodal context or merely exploit surface cues.

- **Phase-3 (Preference data via closed-source VLMs)**: To capture nuanced judgments beyond binary labels, closed-source reasoning VLMs (GPT-4o[104], Gemini-2.5-Flash (Comanici et al., 2025)) were prompted to perform the same annotation task as **Phase-1** through APIs to provide reasoning across **four** aspects: (i) contextual understanding, (ii) visual interpretation, (iii) textual comprehension, and (iv) rationale for **`limiting`** or **`not limiting`** distribution. Through a pilot study, cases where the closed-source models disagreed, were reviewed by three human annotators, who indicated preferred reasoning and outcomes. This process produces a **preference dataset** that can (a) supplement future few-shot or prompt-based tuning of open-source models, (b) align evaluation with human-grounded interpretation and (c) support investigation of both explicit and implicit harmful/hateful semantics (Ouyang et al., 2022; Rafailov et al., 2023).

7.1.1 Phase-1: Unseen Data Curation

Data Collection: Building on (Sharma et al., 2022)'s recommendation to integrate user perspectives in multimodal dataset design, this study leverages the keywords and themes provided by young teens through the Young Research Assistant Programme (Chapter 3, Section 3.4). Under the *Participation with a Purpose* framework, each student provided keywords or themes reflecting *controversial* or *noteworthy* memes they had encountered online. Contributions were collected sequentially, with each input informing

[104] `https://openai.com/index/hello-gpt-4o/`

the next, ultimately yielding eight thematic categories that represent commonly encountered memes: (i) Sexual/Adult Humour; (ii) Sarcasm/Satire/Dark Humour; (iii) Religion/Race/Offensive; (iv) Relationship/Gender/LGBT; (v) Political Satire; (vi) Harmless Humour/Wordplay/Stereotypes; (vii) Informative/Neutral; (viii) Body Shaming/Appearance. These categories were then used as tags to guide data collection, with programmatic queries employing logical `OR` and `AND` operators to retrieve images matching one or multiple themes, ensuring comprehensive coverage of targeted meme types. To compile a dataset corresponding to the categories outlined above, only a few platforms currently allow scraping of user-generated multimodal content, namely *Reddit*[105], *X*[106], and *Imgur*[107]. Insights from the **Phase-1 Requirement Gathering** (Section 4.1 in Chapter 5), indicated that Reddit and X were generally not preferred by youth participants. While Imgur, an image-sharing platform, was not directly surveyed, it hosts content similar to what is commonly found on TikTok and Instagram (highly used by youth participants), including memes and other user-generated images. This overlap made Imgur the most suitable and accessible platform for collecting the target multimodal content.

Data Annotation: The task of annotating memes or images into one or more harmful categories is inherently subjective, particularly when accounting for the intended audience and platform-specific context (Uma et al., 2021). Kalliris (2024) note that a single, universally applicable definition of "*harmful*" is rarely sufficient to capture the diversity of policies, ethical standards, and cultural norms that govern content across digital platforms. To address this inherent subjectivity, (Röttger et al., 2022) note annotation tasks in the online safety research domain can be broadly categorised as: (a) ***Descriptive paradigm***, which embraces annotator subjectivity to generate datasets that capture granular, individual judgments, and (b) ***Prescriptive paradigm***, which minimises subjectivity by instructing annotators to encode a single, predefined belief as specified in the annotation guidelines.

Up to this point in this doctoral dissertation, all data annotation and evaluation tasks (Section 5.4 and 5.3, discussed in Chapter 5) have followed the ***Prescriptive Paradigm*** (Röttger et al., 2022), where annotators and evaluators were required to assess content

[105] `https://reddit.com`
[106] `https://x.com/?lang=en`
[107] `https://imgur.com/`

according to specific, predefined guidelines with examples. As discussed in Section 2.4.5, both the *Hateful Memes* (Kiela et al., 2020) and *Harmful Memes* (Pramanick et al., 2021a) datasets define antisocial online behaviour as psychological, social, or reputational harm directed at individuals or organisations. While *harmful memes* broadly encompass content that causes or normalises such harm, *hateful memes* specifically target groups based on attributes like gender, ethnicity, religion, or sexual orientation. However, in practice, distinguishing between these categories remains challenging due to overlapping intent, context, and multimodal ambiguity. For example, the meme collected from Imgur in the Figure 7.2 depicts one girl reading a Bible and another a Quran, with the implication that either may go to "hell". While some viewers might interpret this as satirical, the boundaries of **hatefulness** or **harmfulness** are not clearly defined, illustrating the subjective nature of meme interpretation and the limitations of strictly rule-based annotation frameworks.

Kaushal et al. (2024) in their study of the EU's Digital Services Act (briefly discussed in Chapter 1), analysed statements of reasons content moderation actions submitted in November 2023 and found that the majority of the moderation actions (93.4%) focused on **limiting content visibility** rather than outright removal. While the study does provide insights into the specific categorisation criteria or moderation practices underlying these visibility restrictions, their findings underscore a broader trend in contemporary content moderation, i.e., the emphasis on **controlling who can see content** rather than censoring it entirely. Motivated by this understanding and guided by the ***Descriptive Paradigm*** (Röttger et al., 2022), we designed the annotation task for the 300 collected images/memes as a survey-style evaluation task. Annotators were asked to provide binary judgments for each meme in response to two platform-contextual questions:

1. "*In your opinion, this image should or should not be limited in distribution on platforms with a* ***mixed audience of adults and adolescents****?*"

2. "*In your opinion, this image should or should not be limited in distribution on platforms with* ***more adolescent users****?*"

For each question, annotators could choose between two options: (a) ***yes, limit the image*** and (b) ***no, do not limit the image***. To explain what we mean by "*limited in distribution*", questions were followed by a statement,



Figure 7.2: Image Collected from Imgur as part of the dataset

> *When we say "**`limit content`**" we mean that the image should be restricted, and when we say "**`do not limit`**", we mean that the image should be freely accessible to users on the platform.*

To further capture the reasoning behind these judgments, annotators were also asked to provide an explanation for their choice followed by a descriptive sentence describing the content of the image. This allows to capture (a) their **implicit** understanding of the image content and (b) the **explicit** content of the image. Although the terms *implicit* and *explicit* do not appear in the annotation schema, the pre-annotation briefing included a sample image (Figure 7.2) to illustrate how annotators might select options and provide reasons for their choices. For instance, annotators could select: "`no (limit adult); yes (limit youth)`", with the reasoning that "*adults may interpret the content as satire*

or sarcasm, while youth are more impressionable". Alternatively, they could select: "`no (limit adult); no (limit youth)`", reasoning that the "*sarcasm targets religion and is not inherently nasty, hurtful, harmful, or inappropriate*". To illustrate how annotators might describe the content in a sentence, an example for the sample image was provided: "*image shows two girls reading a Bible and Quran (religious books) and telling one another that both books say they are going to hell*".

Ethical Considerations and Participant Recruitment: The study was approved by the university ethics board and implemented safeguards for annotators exposed to sensitive content, including briefing sessions, mandatory psychologist-led training on resilience, optional counselling, and regular check-ins. Participants could withdraw at any time and choose remote or on-site work. Compensation was €13/hour (up to 10 hours), provided via vouchers to preserve anonymity. As described in Section 5.4.3, two of twenty recruited participants, a 21-year-old male in Computational Science and a 21-year-old female in Social Sciences, completed the meme annotation task. Both were active social media users familiar with hurtful online interactions and prior research activities, providing informed perspectives. Annotation was conducted via a custom AWS-hosted survey platform for secure, structured categorisation and feedback. While the small sample limits generalisability, it was chosen to maximise consistency, reliability, and qualitative depth.

7.1.2 Phase-2: Baseline Benchmarking

Problem Formulation: As discussed earlier in Section 2.4.5 Chapter 2, despite the availability of multimodal cyberbullying and harmful meme datasets, challenges remain, including coarse-grained labels, unclear annotation guidelines, and limited reproducibility (Hosseinmardi et al., 2015; Maity et al., 2022), which restrict the interpretability of vision–language classification systems. To address this, publicly available datasets such as the *Hateful Memes Challenge* (Kiela et al., 2020) ($\mathcal{D}_{\text{Hate}}$) and *Harmful Memes* (Pramanick et al., 2021a) ($\mathcal{D}_{\text{Harm}}$) are used. Both datasets capture multimodal signals of online toxicity but differ in scope and labelling philosophy. $\mathcal{D}_{\text{Hate}}$ dataset provides binary labels (`hateful` vs. `not hateful`), while $\mathcal{D}_{\text{Harm}}$ captures a broader notion of `harm` with multi-level annotations and `target entities`. For analytical consistency, the task is reformulated as

a binary classification problem: given a meme $x_i \in \mathcal{X}$ (image + text), predict whether it expresses or promotes harmful or hateful content, i.e., $y_i \in \mathcal{Y} = 0, 1$. Merging both datasets as previously done by (Kumar and Nandakumar, 2022; Zhao et al., 2023), among others enables a multi-source learning setup, allowing models to generalise across varying definitions of harm and assess whether visual–textual alignment mechanisms are robust to semantic and contextual shifts between datasets.

Multi-Source Dataset: Due to the unavailability of official unseen test labels for $\mathcal{D}_{\text{Hate}}$ (Kiela et al., 2020), its training and validation sets were merged and re-split as follows: 90% of the combined data was allocated for training, 10% of the training set and 50% of the validation set formed the new validation subset, while the remaining 50% of the validation set was reserved for testing. Following (Pramanick et al., 2021a), $\mathcal{D}_{\text{Harm}}$ fine-grained harmfulness levels were collapsed into binary categories: `harm (1)` = `very harmful` + `partially harmful`, and `not-harm (0)` = `not harmful`. Table 7.1 summarises the distribution of the training, validation, and test splits across $\mathcal{D}_{\text{Harm}}$ and $\mathcal{D}_{\text{Hate}}$. Building on (Kumar and Nandakumar, 2022; Zhao et al., 2023) to evaluate how well models trained on one or both datasets generalise across domains, experiments are conducted under three training regimes (i) training and validation on $\mathcal{D}_{\text{Hate}}$, (ii) training and validation on $\mathcal{D}_{\text{Harm}}$, (iii) training and validation on $\mathcal{D}_{\text{Hate}} \bigcup \mathcal{D}_{\text{Harm}}$. Final evaluations are conducted on the unified test set and unseen test set ($\mathcal{D}_{\text{Imgur}}$).

Source	**Train**		**Validation**		**Test**	
	`0`	`1`	`0`	`1`	`0`	`1`
$\mathcal{D}_{\text{Harm}}$	1949	1064	116	61	230	124
$\mathcal{D}_{\text{Hate}}$	4713	2487	737	563	250	250

Table 7.1: Distribution of Training, Validation, and Test-set. Note: `0` denotes (not-hateful or not-harmful) and `1` denotes (hateful or harmful)

Multimodal Architectures: To investigate the architectural factors contributing to **`harmful`** or **`hateful`** meme detection, building on prior work (Kiela et al., 2020; Pramanick et al., 2021a; Qiu et al., 2022; Maity et al., 2022; Zhao et al., 2023; Kumar and Nandakumar, 2022), as outlined in Figure 7.1, we introduce new baseline configurations that assess the performance of CLIP (Radford et al., 2021) (image-only and image–text with late

fusion) and ViLT (Kim et al., 2021c) through its image-only ViT encoder (Dosovitskiy et al., 2021), text-only BERT encoder (Devlin et al., 2019), and the integrated late fusion ViLT model. This includes,

1. **Unimodal Image-only ($\mathcal{U}^I$)**: As depicted in Figure 7.3, we investigate two vision-only architectures. First, we fine-tune Vision Transformer (ViT) (Dosovitskiy et al., 2021) by adapting the pre-trained `google/vit-base-patch16-224`[108] model with a task-specific classification head for binary classification. The architecture processes raw images through the standard ViT patch embedding layer (16×16 patches), followed by the transformer encoder blocks, and culminates in a linear classifier that projects the `[CLS]` token representation to two output logits. All transformer layers are fine-tuned end-to-end with AdamW optimisation (Loshchilov and Hutter, 2017). Second, we fine-tune $\mathcal{U}^I_{CLIP}$ (image-only) using the `openai/clip-vit-base-patch32`[109] model, where only the visual encoder is leveraged. The pre-trained CLIP vision transformer processes input images independently of text, and a binary classification head is attached to the pooled image embeddings. This configuration allows us to assess whether CLIP's vision-language pre-training provides transferable visual representations for content detection even in the absence of textual input.

2. **Unimodal Text-only ($\mathcal{U}_T$)**: As depicted in Figure 7.3, we fine-tune BERT (Devlin et al., 2019) using the `bert-base-uncased` checkpoint with a sequence classification head for binary prediction. The architecture processes tokenised meme text through BERT's 12-layer bidirectional transformer encoder, extracting contextualized representations. The `[CLS]` token's final hidden state is projected through a linear classification layer to produce binary logits for binary classification. Following baseline development for mulitmodal classification by (Kiela et al., 2020; Pramanick et al., 2021a; Zhao et al., 2023), training is conducted with gradient clipping (max norm 0.5) and linear warm-up scheduling over the first 10% of training steps, enabling us to establish a text-only baseline for comparison with multimodal approaches.

3. **Multimodal Late Fusion (MF^{CLIP}_L)**: As depicted in Figure 7.4 (A), we im-

[108] `https://huggingface.co/google/vit-base-patch16-224`
[109] `https://huggingface.co/openai/clip-vit-base-patch32`

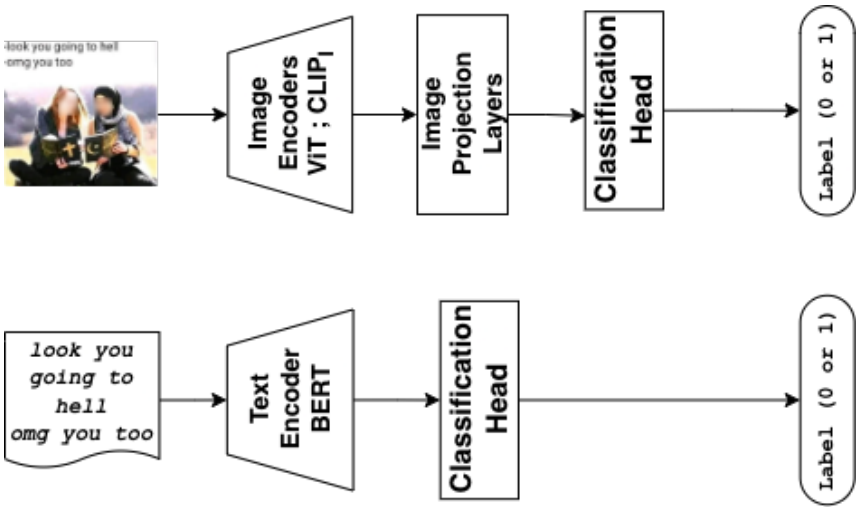


Figure 7.3: Unimodal baseline architectures includes (a) image-encoder based classifiers ($\mathcal{U}_I$): ViT and CLIP$_I$; (b) text-encoder base classifiers ($\mathcal{U}_T$): BERT

plement a late fusion architecture, which leverages CLIP's pre-aligned vision and language encoders (`openai/clip-vit-base-patch32`[110]). Following (Kumar and Nandakumar, 2022), we employ a selective layer-wise fine-tuning strategy wherein encoder layers 0-7 are frozen for both modalities to preserve robust pre-trained representations, while layers 8-11 remain trainable to adapt high-level features to the detection task. Image and text are processed independently through their respective CLIP encoders to obtain projection-normalized embeddings (512-dimensional). These modality-specific representations are concatenated and processed through a multi-layer perceptron fusion network comprising three fully-connected layers with ReLU activations and dropout (p=0.3), ultimately producing binary classification logits. This late fusion strategy, is trained with layer-wise learning rates to enable the model to learn cross-modal interactions while maintaining the semantic alignment established during CLIP pre-training.

4. **Multimodal Early Fusion (MF_E^{ViLT})**: As depicted in Figure 7.4 (B), we fine-tune the Vision-and-Language Transformer (ViLT) (Kim et al., 2021c) using the `dandelin/vilt-b32-mlm`[111] checkpoint. Unlike late fusion approaches, ViLT performs early fusion by jointly processing visual patches and text tokens within a single transformer architecture. This model by (Kim et al., 2021c) employs a unified co-

[110] `https://huggingface.co/openai/clip-vit-base-patch32`
[111] `dandelin/vilt-b32-mlm`

attention mechanism where visual embeddings (obtained via patch projection) and textual embeddings (via word embedding layer) are concatenated and processed through 12 shared transformer layers with cross-modal attention. We implement a classification head which extracts the pooled `[CLS]` representation from ViLT's encoder and projects it through a two-layer classifier (with RELU activation and dropout p=0.1) for binary prediction. The entire architecture, including all transformer parameters, is fine-tuned end-to-end to enable deep multimodal fusion where image-text interactions are learned at all representational levels rather than only at the final classification stage.

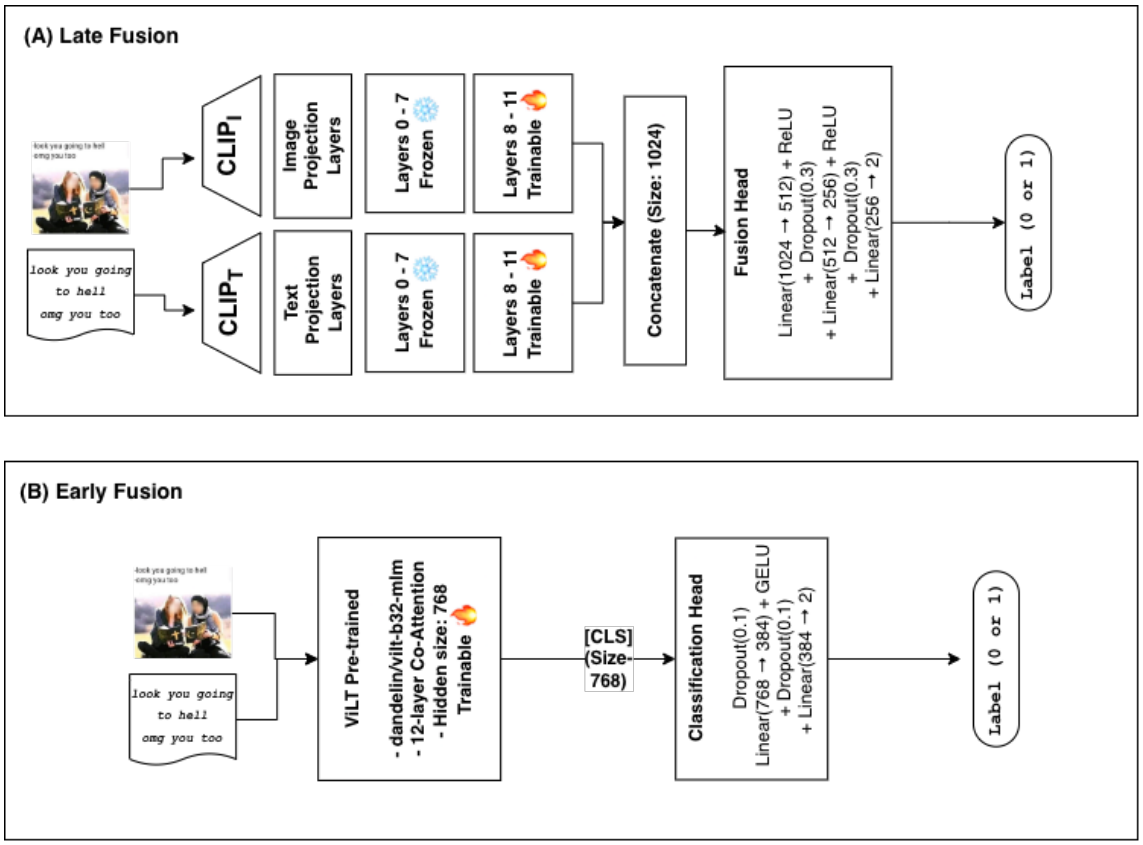


Figure 7.4: Multimodal fusion baseline architectures includes (a) Late Fusion: Fine-tuning CLIP with frozen and unfrozen layers and concatenation; (b) Early Fusion: Fine-tuning ViLT with a classification head

5. **Prompting Large Vision-Language Models (LVLMs)**: As depicted in Figure 7.5, we employ zero-shot prompting strategies with two state-of-the-art large vision-language models: Qwen2-VL (Team et al., 2024) and LLaVA-CoT (Xu et al., 2024). Rather than fine-tuning model parameters, we design structured prompts that guide the models through a chain-of-thought reasoning process for meme classification.

Drawing on the prompt styles summarised in Section 5.1.4 of Chapter 5, we adopt a hybrid prompting strategy for vision–language models that combines (a) persona-driven framing and (b) structured task decomposition. Concretely, persona-driven prompts (e.g., opening with phrases such as "*You are a content moderation AI assistant*" or "*You are a content-moderation assistant specialised in analysing images for potential harm*") provide a stable role-context that focuses the model's task and output style (Ge et al., 2024; Hui et al., 2024; Olea et al., 2024). Complementing this, structured task decomposition, i.e., decomposing the judgement into explicit sub-problems and helps guide the model to separate visual, contextual, and textual reasoning steps (Khot et al., 2022). Together, these two elements improve the reproducibility of the model's reasoning and make downstream annotations easier to interpret. Our prompt template detailed in Appendix A.17elicits four sequential analytical components: (a) *visual interpretation*: requesting detailed description of visual elements, symbols, and depicted entities; (b) *textual interpretation*: analysing the semantic content, tone, and linguistic patterns of overlaid text; (c) *contextual interpretation*: reasoning about the interaction between visual and textual modalities, including detection of irony, sarcasm, or coded hateful messaging; and (d) *final classification*: producing a binary label with justification. This prompting approach allows us to assess whether large-scale vision-language models with broad world knowledge and reasoning capabilities can perform a binary detection task without task-specific training or in-context learning, providing an important comparison point for our fine-tuned architectures.

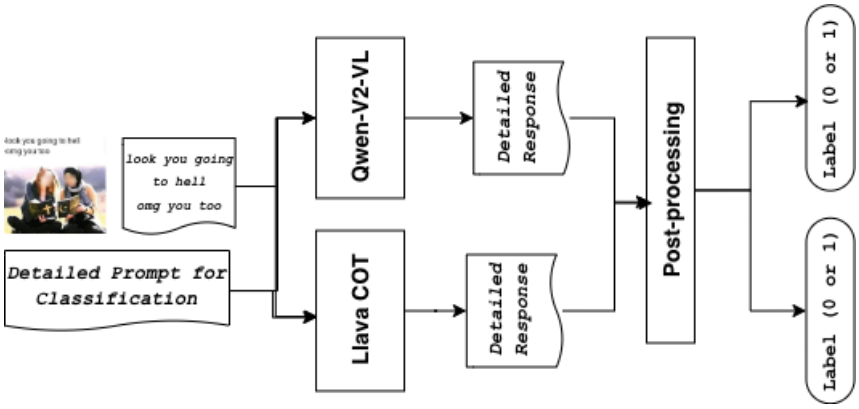


Figure 7.5: Prompting Large Vision-Language Models (LVLMs) such as Qwen-V2-VL and Llava-COT for binary classification

Hyper-parameter Optimisation: To ensure fair comparisons, we perform architecture-specific grid search optimisation via Weights & Biases (Biewald, 2020), with search spaces tailored to each model's properties and pre-training. For details, see Appendix A.18.

7.1.3 Phase-3: Pilot Human Preference Dataset Curation

The motivation for constructing a pilot human preference dataset stems from the central role of human feedback in aligning vision–language systems with human moral and contextual judgments. Techniques such as Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022) and Direct Preference Optimisation (DPO) (Rafailov et al., 2023) have shown that human preference data, particularly pairwise comparisons of model outputs can substantially enhance model alignment and reduce harmful generations (Zhang et al., 2025). Our approach extends this paradigm to multimodal harm detection by generating structured preference data that reflects both explicit visual features (e.g., identifiable objects, personalities, or text overlays) and implicit cues (e.g., sarcasm, cultural context, or innuendo). We select OpenAI's GPT-4o[112] and Google's Gemini-2.5-Flash (Comanici et al., 2025) as representative state-of-the-art **closed-weight** vision–language systems. Both models integrate advanced multimodal reasoning pipelines and industry-grade safety alignment mechanisms, making them ideal for probing the boundaries of responsible multimodal content interpretation (Li et al., 2025; Lou et al., 2025). GPT-4o provides unified text–image–audio processing with strong cross-modal grounding and contextual sensitivity (Hurst et al., 2024), while Gemini-2.5-Flash offers efficient multimodal inference and robust safety enforcement through enhanced content moderation (Comanici et al., 2025). Including these systems enables us to examine how the most capable commercial VLMs operationalise moral and contextual alignment under strict moderation constraints, serving as a practical benchmark against open-weight research models such as LLaVA-COT and Qwen-V2-VL.

When applying the same prompt template introduced in Section 7.1.2 for closed-weight vision–language systems (e.g., GPT-4o, Gemini-2.5-Flash), the models almost invariably refused to respond, producing statements such as "*I cannot assist with this request*" or "*as an AI model I cannot process or evaluate offensive content*".

[112] `https://openai.com/index/hello-gpt-4o/`

These refusals, although reflective of responsible moderation practices, limited our ability to probe fine-grained moral and contextual reasoning in harmful multimodal domain. To mitigate these refusals while preserving ethical rigour, we adopted Mulitmodal LLMs-as-a-judge (Chen et al., 2024a) benchmarking, by implementing the annotation-task (Section 7.1.1) with these two closed-weight VLMs. Instead of directly prompting models to assign `harmfulness` or `hatefulness` labels, we reformulated the task as assessing a meme's "**suitability for distribution**" given specific audience compositions, similar to our annotation task. This re-framing allowed models to engage in structured reasoning about social appropriateness without violating built-in safety constraints, while still eliciting nuanced assessments of potential harm.

In line with **RQ.3**, which examines the capacity of vision–language systems to interpret both explicit and implicit harmful semantics, this phase introduces a pilot human evaluation study designed to leverage human–model disagreement as an informative supervision signal. Specifically, we curate cases in which closed-weight models diverge in their judgments, enabling us to identify the boundary conditions of multimodal harm interpretation and to establish a foundation for future large-scale preference-alignment datasets focused on social and ethical reasoning in memes. To this end, we use the prompt template detailed in Appendix A.19 to query GPT-4o and Gemini-2.5-Flash through their respective APIs on the unseen test set collected as part of Phase-1 (Section 7.1.1). Notably, both APIs provide only thirty-day temporary storage and do not use submitted prompts or data for training their base models (See OpenAI Policy[113] and Gemini API terms[114]). This also ensures that the models are evaluated on truly unseen content and that our assessments reflect their generalisation and alignment capabilities rather than prior exposure.

Pilot Data Setup: To address **RQ2.2**, we select instances where both GPT-4o and Gemini-2.5-Flash disagreed on harm assessments for human preference data curation. Targeting disagreement enriches the dataset with challenging examples, following prior work showing that disagreement-based sampling improves supervision quality and generalization (Askell et al., 2021; Touvron et al., 2023). We sampled 60 binary-disagreement cases

[113] "Data submitted through the API is not used to train OpenAI models [..] unless you explicitly opt in."`https://openai.com/consumer-privacy/`

[114] "Prompts and tuning data [..] are never used to train or enhance models [..]" `https://docs.cloud.google.com/vertex-ai/generative-ai/docs/vertex-ai-zero-data-retention`

between the models, balancing meme diversity with feasible human annotation.

Human Annotator Setup: Three annotators (2 Males: aged 26 and 24; 1 Female aged 25) participated in this pilot. All annotators were doctoral students (2: Computing (1M/1F); 1: Linguistic (1M)), had prior exposure to online content moderation contexts, and were familiar with harmful or offensive internet material. Prior to the annotation phase, each participant received a briefing and written guidelines explaining the evaluation criteria, ethical considerations, and confidentiality procedures. The annotators were only informed that the task involved evaluating the quality and appropriateness of two system responses (System-A and System-B) and were told not to produce their own harm judgments independently. The aim was to determine which model output (from GPT-4o or Gemini-2.5-Flash) better aligned with a human understanding of moderation for meme content.

Evaluation Protocol: Each of the 60 meme–response pairs consisted of (a) the meme image; (b) corresponding outputs from Gemini-2.5-Flash (System-A) and GPT-4o (System-B). The annotators were not made aware of the systems instead it was only referred to as System-A for Gemini-2.5-Flash and System-B for GPT-4o. For every meme, annotators were instructed to make a pairwise preference judgment using one of four options: (i) `System-A` preferred, (ii) `System-B` preferred, (iii) `Neither` preferred, and (iv) `Both` preferred. The preference was determined based on **five** key evaluation dimensions designed to capture multimodal reasoning quality:

1. **Contextual Understanding**: Does the model appropriately interpret the situational or cultural context of the meme?

2. **Visual Interpretation**: Does the model correctly identify and describe key visual elements, symbols, or objects in the meme?

3. **Textual Comprehension**: Does the model accurately understand any embedded text, captions, or linguistic cues within the meme?

4. **Reasoning for Age-Restricted Access (Young Users)**: Is the justification for limiting access to younger audiences logically sound and contextually supported?

5. **Reasoning for Age-Restricted Access (Adult Users)**: Does the model correctly justify when content may be suitable only for adults, considering potential harm or sensitivity?

Annotators were encouraged to record short notes explaining their decision in cases where the preference was non-obvious or based on nuanced contextual interpretation. These qualitative notes later supported thematic analysis (Braun and Clarke, 2006) of reasoning patterns across models. To ensure the reliability of annotations, inter-annotator agreement was computed after the evaluation phase. Agreement was assessed using Fleiss' κ (kappa) (Fleiss, 1971), which quantifies the degree of consensus among multiple raters on categorical judgments.

7.2 Results and Discussion

7.2.1 Data Collection

The data collection described in Section 7.1.1 was conducted between late 2023 and mid-2024, resulting in a total of **300** images collected, excluding user comments. Collecting comments proved challenging because the Imgur API[115] did not reliably render them, and attempts at manual scraping using Selenium were constrained by platform-imposed limits. It is important to note that in 2024 Imgur introduced new Terms of Service policy restrictions[116] on the use of scraped content. In accordance with Imgur's 2024 policy changes, the collected dataset is subject to a Quitclaim License Agreement signed with Imgur's legal team. Under this agreement:

1. **Rights Transfer** – Imgur grants the researcher a non-exclusive, "as-is" transfer of whatever rights they may have in the images, **strictly for research and evaluation purposes**.

2. **No Warranties** – The images are provided without any guarantees regarding ownership, originality, or freedom from third-party claims.

[115] `https://apidocs.imgur.com/`
[116] `https://imgur.com/tos#:~:text=In%20connection%20with%20your%20use,learning%2Fuse%20on%20AI%20platforms.`

3. **Usage Restrictions** – The dataset **cannot be redistributed**, making it unavailable for training or external use.

This legal agreement ensures that the dataset is used ethically and in compliance with platform policies, while clearly defining the scope of permissible use.

7.2.2 Data Annotation

The results of the descriptive-style annotation task described in Section 7.1.1 provide valuable insights into how annotators interpret potential harm in multimodal online content, particularly in relation to audience composition. As outlined in Section 7.1.1, although this activity involved only two annotators, their distinct demographic backgrounds (Annotator-1: 21-year-old female of European descent; Annotator-2: 21-year-old male of Irish descent) provided a measure of diversity in perspectives.

Agreement Levels and Interpretative Variability: As shown in Table 7.2, the two annotators differed slightly in their interpretations. For the question, "*In your opinion, this image should or should not be limited in distribution on platforms with a mixed audience of adults and adolescents?*", the Cohen's κ was 0.6259. For the question, "*In your opinion, this image should or should not be limited in distribution on platforms with more adolescent users?*", the Cohen's κ was 0.5751. Both values fall within the weak-to-moderate agreement range (McHugh, 2012), indicating a generally shared interpretative framework between the annotators but with some divergence in judgments, particularly when audience composition was more youth-oriented. The raw agreement rate followed a similar trend, i.e., 86.67 % for mixed (adult & youth) audiences versus 79.67 % for youth audiences. The observed difference ($\Delta\kappa$, i.e., $\kappa_{\text{mix}} - \kappa_{\text{youth}}$) is 0.05, corresponding to an additional 21 memes in which annotators differed on the youth-oriented task. This suggests that the introduction of a more vulnerable audience segment increased interpretative caution and reduced alignment.

Thematic Divergences in Reasoning: The thematic analysis following (Braun and Clarke, 2006) uncovered three recurring domains of disagreement, each in both mixed-audience and adolescent-audience contexts, which help understand harmfulness judge-

Limited distribution for?	Cohen's κ	Raw Agreement Rate	Raw Disagreement Count
Youth Users	0.5751	79.67%	61
Adult & Youth Users	0.6259	86.67%	40

Table 7.2: IAA statistics for data annotation discussed in §7.1.1

ments. In mixed audience contexts,

- **Offensiveness vs. Humour**: Annotator 1 consistently prioritised potential *impact* i.e., whether a meme might offend or perpetuate harm. Annotator 2 prioritised *intent*, i.e., whether a meme is intended as humour or satire. For example, on a white background text overlay image which reads as "*'You the bomb':'No, you the bomb' A compliment in America. An argument in the Middle East.*", Annotator-1 noted it should be limited and provided a short reason - "*adult and youth is dark humour on word bomb, its not funny and can be offensive*", whereas Annotator-2 noted it should not be limited and provided a short reason - "*dark humour sarcasm for adults*". This aligns with the broader discourse on online content moderation, where the boundary between harmful and humorous expression is fluid and often culturally mediated (Röttger et al., 2022; Uma et al., 2021). The disagreement demonstrates how an annotator's individual tolerance thresholds for irony or mockery significantly affect labelling outcomes.

- **Sensitivity to Political or Religious Satire**: Another divergence in annotation relates to how political or religious satire was interpreted. For instance, one meme featured a bird background with the overlaid text: *women who feel liberated wearing the hijab shouldn't advocate it because it helps the religious leaders, government bodies and men in our families force it on to the rest of us*, Annotator 1 judged that this image should be limited, citing it as "*offensive to adults*" whereas Annotator 2 believed it should not be limited, reasoning that,"*this is more like a political opinion, so in adult social media it is fine*". On another meme depicting Hillary Clinton with the caption "*what do you mean they missed*", a reference to an attack on Donald Trump, Annotator-1 indicated it should not be limited with reason "*meme is a political satire*", whereas Annotator-2 indi-

cated it should be limited citing "*it is dark satire so it should be limited, can be politically bad*". These examples illustrate that annotators differ in their thresholds for what constitutes acceptable satire, particularly when political or religious themes are involved. Such differences highlight how personal beliefs, moral sensitivities, and perceptions of harm intersect with interpretive judgment. In this context, sensitivity to political or religious satire reflects not only disagreement over offensiveness but also differing assumptions about the boundary between free expression and harmful discourse. This variability underscores how annotation of multimodal content is influenced by individual evaluative frameworks, especially when content invokes ideological or moral dimensions.

- **Topic Sensitivity: Gender and Sexual Orientation**: Another recurring point of disagreement emerged in relation to content addressing gender and sexual orientation. For instance, one meme displayed the cartoon character Jerry with the caption "*Gay people rely on straight people to produce more gay people*" on the top half, and "*Jerry will be back soon with more accurate*" over an image of railway tracks on the bottom half. Annotator 1 judged that this image should be limited, reasoning that it was "*offensive to adults*" whereas Annotator 2 felt it should not be limited, describing it as "*it is sarcasm on LGBT*". A similar pattern appeared in an image featuring Andrew Tate (a public figure known for anti-feminist rhetoric), accompanied by the caption "*Crying in front of a woman is like, bleeding next to a shark. Only a few will understand.*", Annotator 1 classified it as should be limited with reasoning "*it is sexist*". At the same time, Annotator 2 viewed it as permissible, explaining that "*it is from Andrew Tate but it is not like very sexist*". This highlights differing levels of tolerance for content that invokes sarcasm or satire on gender or sexual orientation. Annotators appear to apply distinct thresholds for what constitutes harmful stereotyping or legitimate social commentary. In this context, topic sensitivity reflects how personal beliefs, social values, and interpretive framing mediate perceptions of offensiveness. Content related to gender or sexual orientation often triggers stronger evaluative responses, making such themes particularly prone to disagreement. This underscores that judgments about harm in multimodal content are not purely objective but are shaped by varying moral

and cultural standpoints surrounding gender discourse.

In adolescent audience contexts,

- **Humour and Perceived Impressionability in Youth Contexts**: A further point of disagreement arose around content that was humorous but potentially sensitive for adolescent audiences. For example, one image depicted a printed T-shirt with the text "*I don't know where I'm going but I'm gay*", Annotator 1 judged that the image should be limited for youth audiences, reasoning that "*it is impressionable*", whereas Annotator 2 considered it suitable, explaining that "*it is a joke, nothing here that says it should not be suitable*". This divergence highlights how perceptions of adolescent impressionability influence judgments of what constitutes appropriate humour. Although both annotators recognise the image as humorous, their disagreement reflects differing interpretations of potential influence, i.e., whether light-hearted self-expression could be seen as benign or as content that shapes identity perceptions among younger viewers. In this sense, **humour and impressionability** become key factors in content moderation for youth-oriented platforms, illustrating the challenge of balancing freedom of expression with developmental sensitivity.

- **Profanity and Wordplay in Memes**: Another disagreement theme involved the use of subtle profanity embedded with creative wordplay. One example featured an image of a mountain with a moon in the background, overlaid with the text "*Sometimes, I look at the sky at night and think of you..*" where the crescent moon "C" visually formed part of a profanity. Annotator 1 judged that this image should be limited for youth audiences, reasoning that "*it is impressionable for youth*", whereas Annotator-2 indicated it should not be limited, explaining that "*it is a word play by using the moon to spell out a profanity, nothing that suggests it should be limited*". This divergence illustrates differing perceptions of whether implicit profanity should be or should not be limited to vulnerable audiences. While one annotator viewed even indirect exposure to profane language as potentially inappropriate or influential, the other interpreted it as a creative and harmless joke. Profanity and wordplay in memes, therefore, represent a nuanced grey area in adolescent content moderation.

- **Sarcasm vs. Offensive**: In adolescent contexts, another area of disagreement was about the interpretation of sarcasm and its boundary with offensiveness. For example, on an edited image depicting a newborn being baptised in a church with a can of Monster energy drink overlaid with the text "*This is how all white dudes named kyle are born*", Annotator-1 judged that it should be limited for youth audiences, reasoning that it "*can be offensive and impressionable to youth*". However, Annotator 2 indicated that it should not be limited, explaining that it was simply "*sarcasm on monster energy drink and everyone named kyle* ". This divergence underscores how annotators differ in perceiving the intent and potential social impact of sarcastic humour. While one viewed the meme as reinforcing a stereotype with possible negative implications for impressionable audiences, the other interpreted it as harmless cultural parody. The disagreement illustrates how sarcasm when intertwined with group identifiers such as race, name, or cultural contexts can blur the line between humour and harm. In the context of youth-oriented platforms, such differences are common amongst the two annotators, i.e., sarcasm that seems benign to adults may still convey implicit stereotypes or normalise mockery for younger viewers. This represents a nuanced interpretive challenge, revealing how assessments of humour depend on both perceived intent and audience vulnerability.

Implicit and Explicit Themes: Given the **descriptive paradigm** (Röttger et al., 2022) employed in this annotation task, we analysed annotator responses to two core questions: (i) the reason for limited or non-limited distribution, and (ii) the presence of *implicit* and *explicit* meanings in content. It is worth noting that both annotators resorted to short keyword-like descriptions. Based on the analysis of the descriptive phrases by two annotators, we categorise the memes collected from Imgur into three overarching thematic domains: **Explicit Meanings**, **Implicit Meanings**, and **Benign/Normal Meanings** (See Table 7.3),

(Theme-1) Explicit Meanings: These are direct and overt interpretations or meanings that are clearly stated or explicitly referenced in the annotator's descriptions. They often involve unambiguous expressions of bias, sexuality, or discrimination. This has following **five** sub-themes:

(E1) Direct-Religion: Content that explicitly is directed towards religious groups or symbols.

(E2) Direct-LGBT: Content explicitly references or is directed towards LGBTQ+ individuals or communities.

(E3) Direct-Racism/Immigration: Content that explicitly addresses racial or ethnic groups, or makes references to migration and related stereotypes.

(E4) Direct-Sexual Reference: Content containing explicit sexual references, imagery, or language.

(E5) Direct-Sexism/Gender: Content explicitly expresses sexist or misogynistic ideas, or engages directly with gender-related discourse.

(Theme-2) Implicit Meanings: These themes encompass underlying, inferred, or suggestive meanings or interpretations that rely on tone, cultural context, or connotation rather than direct statements. Implicit themes often use humour, sarcasm, or symbolism to mask or soften sensitive or discriminatory ideas. This has following **eleven** sub-themes:

(I1) Double Entendres: Wordplay involving implicit adult or sexual references.

(I2) Political Sarcasm: Use of irony or humour to critique or mock political figures or ideologies (e.g., Trump, Putin).

(I3) Adult Humour: Content framed as humour but implicitly contains sexual or mature themes inappropriate for younger audiences.

(I4) Sarcasm–Prejudice: Use of humour or sarcasm to disguise either discriminatory/prejudiced viewpoints, or sarcastic responses or retorts.

(I5) Edgy Content: "Dark Humour" or borderline offensive material invoking themes of offensive content.

(I6) Word Play: Pun-based or linguistic jokes that implicitly reference sexual or taboo topics.

(I7) Stereotypes: Humorous depictions rooted in cultural, racial, or social stereotypes.

(I8) Body Shaming: Implicit mockery of body types or physical appearance through jokes or visual cues.

(I9) Gender Identity: Sarcasm or commentary around gender identity, pronouns, or fluidity, often implying disapproval or ridicule.

(I10) Association–Terrorism: Indirect associations between particular groups and violence or terrorism.

(I11) Self-harm: Indirect or direct references to self-harm.

(Theme-3) Benign / Normal Meanings: This category includes non-harmful or informational content that carries no apparent offensive, discriminatory, or adult connotations. Such material typically consists of ordinary, educational, or humorous images without implicit or explicit negative intent, for example, memes related to daily life, general knowledge, or cosplay.

Theme	**Sub-theme**	**Count**
Explicit Meanings	E1-Direct-Religion	28
	E2-Direct-LGBT$_+$	23
	E3-Direct-Racism	18
	E4-Direct-Sexual	28
	E5-Direct-Sexism	9
Implicit Meanings	I1-Double-Entendres	21
	I2-Political-Sarcasm	20
	I3-Adult-Humour	54
	I4-Sarcasm-Prejudice	25
	I5-Edgy-Content	9
	I6-Word-Play	4
	I7-Stereotype	12
	I8-Body-shame	3
	I9-Gender-Identity	13
	I10-Association-Terrorism	2
	I11-Self-harm	4
Benign & Normal	–	27

Table 7.3: Themes and sub-themes based on descriptive annotator responses

Following the thematic categorisations, Table 7.4 illustrates how the annotators evaluated each theme in terms of its **suitability for distribution** among different audience groups. The results indicate that memes categorised under *Theme-1: Explicit Meanings*

were overwhelmingly judged as unsuitable for youth audiences, while *Theme-2: Implicit Meanings* were more ambiguous, often leading to disagreement were therefore represented as "`Not-Sure`". This uncertainty reflects the nuanced and context-dependent nature of implicit content, where harmful or suggestive meanings are conveyed indirectly through humour, sarcasm, or coded language. In contrast, *Theme-3: Benign/Normal* content showed strong agreement among annotators as being suitable for unrestricted distribution across both audience groups.

Theme	**Limited Distribution for Youth**			**Limited Distribution for Adults**		
	`Yes`	`No`	`Not-Sure`	`Yes`	`No`	`Not-Sure`
Explicit	84	2	20	44	42	20
Implicit	77	51	39	5	143	19
Benign	-	25	2	-	26	1

Table 7.4: Distribution of annotator judgments across implicit, explicit, and benign themes, showing the majority consensus between two raters on whether the content should be limited for youth or adult audiences. "`Not-Sure`" indicates cases where annotators disagreed or expressed uncertainty.

7.2.3 Vision-language Benchmarking

To address our **RQ.3**, we evaluated multiple model architectures across two distinct datasets: $\mathcal{D}_{Hate}$ for *hateful* meme detection and $\mathcal{D}_{Harm}$ for *harmful* meme detection. Our experimental design discussed earlier in Section 7.1.2 included unimodal models (text-only and image-only), multimodal fusion approaches (late and early fusion), and prompt-based vision-language models (VLMs). Models were trained on three different data configurations: (i) $\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$; (ii) $\mathcal{D}_{Hate}$; and (iii) $\mathcal{D}_{Harm}$ alone.

Optimal Hyper-parameters: Table A.18 in Appendix A.20 presents the optimal hyper-parameter configurations identified for each vision-language model across the validation sets for $\mathcal{D}_{HARM}$ and $\mathcal{D}_{HATE}$. The uni-modal BERT ($\mathcal{U}T$) uses a consistent configuration across datasets (batch size 16, learning rate 1.66×10^{-5}, weight decay 0.039). In contrast, image-only models, i.e., ViT ($\mathcal{U}^I ViT$) and CLIP-image ($\mathcal{U}^I_{CLIP}$) show dataset-specific settings, with smaller batch sizes and higher learning rates for $\mathcal{D}HATE$. Multimodal models ($MF^L CLIP$ and MF^E_{ViLT}) also vary across datasets. ViLT generally

adopts higher learning rates for mixed data while keeping a fixed sequence length of 128 tokens. Early stopping patience also differs across models.

Results on `Harmful` and `Hateful` Test-set: As outlined in Table 7.1 (Section 7.1.2), the test-set for $\mathcal{D}_{Harm}$ is imbalanced (`0`: 230; `1`: 124), whereas $\mathcal{D}_{Hate}$ is balanced with 250 memes each. The results for model performance in terms of class-wise macro-average F1 scores are illustrated in Table 7.5, this helps understand balance in prediction across both `0` and `1` classes. To put this in context, we also discuss the false negative rate (FNR), i.e., instances where content is `hateful (1)` or `harmful (1)` but model predicts as `not-hateful (0)` or `not-harmful (0)` and false positive rate (FPR), i.e., instances where content is `not-hateful (0)` or `not-harmful (0)` but predicted as `hateful (1)` or `harmful (1)`, are outlined in Table 7.6. Higher values in Table 7.5, and lower values in Table 7.6 indicate better performance.

Model Type	**Model**	**Training Data**	$\mathcal{D}_{Harm}$		$\mathcal{D}_{Hate}$	
			`0`-F1	**`1`-F1**	**`0`-F1**	**`1`-F1**
Unimodal	$\mathcal{U}_T$ (BERT)	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.7552	0.6237	0.6613	0.4474
		$\mathcal{D}_{Hate}$	0.6805	0.6192	0.6306	0.5393
		$\mathcal{D}_{Harm}$	**0.8281**	**0.7143**	0.6667	0.3353
	$\mathcal{U}_I$ (CLIP$_I$)	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.7901	0.6491	0.6179	0.3896
		$\mathcal{D}_{Hate}$	0.7315	0.4596	0.5341	0.4788
		$\mathcal{D}_{Harm}$	0.8451	0.7266	0.6505	0.3186
	$\mathcal{U}_I$ (ViT)	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.7416	0.5627	0.6201	0.3906
		$\mathcal{D}_{Hate}$	0.6989	0.5201	0.5820	0.4527
		$\mathcal{D}_{Harm}$	0.7921	0.6215	0.6235	0.3068
MF^{L}_{CLIP}	CLIP	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.8186	0.7533	0.6761	0.4266
		$\mathcal{D}_{Hate}$	0.7432	0.1132	0.6737	0.3522
		$\mathcal{D}_{Harm}$	0.6446	0.4557	0.6994	0.1117
MF^{E}_{ViLT}	ViLT	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.7418	0.6099	0.6702	0.5734
		$\mathcal{D}_{Hate}$	0.5701	0.3150	0.6678	0.5615
		$\mathcal{D}_{Harm}$	0.8478	0.7177	0.6295	0.2679
Prompt-based VLMs	Qwen2-VL	–	0.7634	0.2530	0.6440	0.3275
	Llava-COT	–	0.7839	0.2716	0.6440	0.4663

Table 7.5: Performance comparison of unimodal, multimodal fusion, and prompt-based vision-language models on *hateful* and *harmful* datasets.

1. **Unimodal Models ($\mathcal{U}$)**:

Model Type	Model	Training Data	$\mathcal{D}_{Harm}$		$\mathcal{D}_{Hate}$	
			FNR	**FPR**	**FNR**	**FPR**
Unimodal	$\mathcal{U}_T$ (BERT)	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.2983	0.2956	0.6600	0.1800
		$\mathcal{D}_{Hate}$	0.1935	0.4304	0.5200	0.3000
		$\mathcal{D}_{Harm}$	0.2338	0.2043	0.7760	0.1120
	$\mathcal{U}_I$ (CLIP$_I$)	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.3064	0.2391	0.7000	0.2400
		$\mathcal{D}_{Hate}$	0.5645	0.2478	0.5480	0.4360
		$\mathcal{D}_{Harm}$	0.2500	0.1695	0.7840	0.1400
	$\mathcal{U}_I$ (ViT)	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.4032	0.2826	0.7000	0.2360
		$\mathcal{D}_{Hate}$	0.4274	0.3391	0.6080	0.3400
		$\mathcal{D}_{Harm}$	0.3710	0.2130	0.7840	0.1920
MF^L_{CLIP}	CLIP	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.0887	0.2739	0.6920	0.1360
		$\mathcal{D}_{Hate}$	0.9274	0.1130	0.7640	0.1040
		$\mathcal{D}_{Harm}$	0.6400	0.2200	0.9194	0.1957
MF^E_{ViLT}	ViLT	$\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$	0.3065	0.3130	0.5000	0.2440
		$\mathcal{D}_{Hate}$	0.6532	0.4609	0.5160	0.2400
		$\mathcal{D}_{Harm}$	0.2823	0.1522	0.8200	0.1640
Prompt-based VLMs	Qwen2-VL	–	0.8264	0.1071	0.7741	0.1544
	Llava-COT	–	0.8225	0.0695	0.64	0.184

Table 7.6: False positive rates (FPR) and false negative rates (FNR) for different model architectures evaluated on $\mathcal{D}_{Hate}$ and $\mathcal{D}_{Harm}$ datasets. Lower values indicate better performance for both metrics.

- $\mathcal{D}_{Harm}$: Both text-only BERT ($\mathcal{U}_T$) and image-only ($\mathcal{U}_{CLIP}$) perform better when fine-tuned on the same $\mathcal{D}_{Harm}$. Despite higher F1 scores (`0`: 0.8281; `1`: 0.7143), $\mathcal{U}_T$ misses 23% content `harmful` content (FNR=0.2338). $\mathcal{U}_{CLIP}$ performs slightly better on F1 (`0`: 0.8451; `1`: 0.7266) but misses 25% of harmful cases (FNR=0.25), showing that higher F1 does not always imply lower FNR. For example, an image of Joe Biden holding Barack Obama's hand with sarcastic captions was correctly classified by $\mathcal{U}_T$ but not by $\mathcal{U}_{CLIP}$. When fine-tuned on out-of-domain data ($\mathcal{D}_{Hate}$), $\mathcal{U}_T$ though with modest F1 scores (`1`: 0.6805, `1`:0.6192), maintains a lower FNR (0.1935) despite a higher FPR (0.43). On the mixed dataset ($\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$) achieves balanced FPR/FNR (≈0.29 each). Its worth noting that directly projecting image features from $\mathcal{U}^I_{ViT}$ yields worse performance than fine-tuning $\mathcal{U}_T$. Overall, $\mathcal{U}_T$ proves to be the stronger baseline for multimodal detection on $\mathcal{D}_{Harm}$. This highlights text-captions alone are strong indicators in this dataset for `harmful` meme

detection.

- $\mathcal{D}_{Hate}$: In contrast to performance on $\mathcal{D}_{Harm}$, all unimodal classifiers fail in correctly classifying memes as `hateful` when the test-set is balanced. Only $\mathcal{U}_T$ when fine-tuned on the same dataset, yields a higher F1 score for positive class (`1`: 0.5393) compared to other unimodal classifiers. The FNR for all unimodal classifiers is very high ranging from 52%-78%, indicating that almost all `hateful` content is mis-classified. A similar pattern to performance on $\mathcal{D}_{Harm}$ emerges, when trained with mixed data or solely on out-of-domain data, all models fail equally worse. Highlighting the challenges in determining if content is `hateful` or `not-hateful`. This highlights that text-captions or image feature on its own are not strong indicators in this dataset for detecting if content is `hateful`.

2. **Late Fusion** MF^{L}_{CLIP}:

- $\mathcal{D}_{Harm}$: Fine-tuning CLIP with late fusion (MF^{L}_{CLIP}) on this dataset, using gradual layer unfreezing, yields weak performance (`0`: 0.6446; `1`: 0.4557). However, when fine-tuned on the mixed-dataset, MF^{L}_{CLIP} achieves its best class-wise F1 scores (`0`: 0.82; `1`: 0.75), misclassifying only 8% `harmful` content. The earlier discussed meme in this dataset (*Joe Biden holding Barack Obama's hand*), is correctly classified in this settings. However, a meme depicting North Korean president Kim Jong Un alongside a "do you like me -yes/no" questionnaire is mic-classified. This indicates that, despite overall improvements, the model continues to struggle with memes that convey nuanced or implied meanings. Training MF^{L}_{CLIP} on out-of-domain data ($\mathcal{D}_{Hate}$) achieves the worst performance (`0`: 0.7432; `1`:0.1132) with 92% `harmful` content mis-classified.

- $\mathcal{D}_{Hate}$: Fine-tuning MF^{L}_{CLIP} on this dataset shows weak class-wise performance (`0`:0.6737, `1`:0.3522), with 76% `hateful` memes mis-classified. Similar to its performance on $\mathcal{D}_{Harm}$ with mixed-data, on $\mathcal{D}_{Hate}$, MF^{L}_{CLIP} shows slight class-wise F1 improvement (`0`:0.6761, `1`:0.4266). Compared to its unimodal component ($\mathcal{U}^{I}_{CLIP}$), late-fusion MF^{L}_{CLIP} performs better on this dataset with

mixed-data training, yet 50% `hateful` memes are mis-classified. This highlights that CLIP benefits from multimodal late fusion and mixed-data training but `hateful` content containing subtle or nuanced `hateful` meanings remains difficult to detect accurately.

3. **Early Fusion** MF^{E}_{ViLT}:

   - $\mathcal{D}_{Harm}$: When trained and tested on this dataset, fine-tuning MF^{E}_{ViLT}, yields strongest performance in detecting `harmful` content (`0`: 8478, `1`: 0.7177). However, this model misclassifies 28% `harmful` content (FNR=0.2823). Across all training configurations, training on $\mathcal{D}_{Harm}$ to test on $\mathcal{D}_{Harm}$ yields best performance for MF^{E}_{ViLT}. Although, fine-tuning on the mixed-dataset shows modest class-wise F1 scores (`0`: 0.7418, `1`: 0.6099), MF^{E}_{ViLT} is balanced at ≈30% `harmful` content being mis-classified as both `harmful` or `not-harmful`. Also, in-comparison to its unimodal classifier $\mathcal{U}^{I}_{ViT}$, although it achieves an improved performance across class-wise F1 scores and FPRs, when fine-tuning on out-of-domain data ($\mathcal{D}_{Harm}$), MF^{E}_{ViLT} performance worse than its unimodal classifier $\mathcal{U}^{I}_{ViT}$. This highlights that while fine-tuning early multimodal fusion models improves in-domain performances and balances error rates, it is sensitive to domain shifts and may underperform unimodal models on out-of-domain data.

   - $\mathcal{D}_{Hate}$: Fine-tuning and MF^{E}_{ViLT} with this dataset, yields modest class-wise F1 scores (`0`: 0.6678, `1`: 0.5615), with ≈51% `hateful` content mis-classified. However, when fine-tuned on the mixed data, just like the late fusion classifier (MF^{L}_{CLIP}), this early fusion classifier (MF^{E}_{ViLT}) yields comparatively better performance (`0`: 0.6702, `1`: 0.5734). In this setting, MF^{E}_{ViLT} mis-classifies 50% `hateful` content. Similar to its performance on $\mathcal{D}_{Harm}$ when fine-tuned with out-of-domain data, MF^{E}_{ViLT} performs poorly at detecting `hateful` memes. Overall, these results indicate that multimodal early fusion provides only modest improvements over single-dataset training for `hate` detection. Additionally, the model remains sensitive to domain shifts and continues to mis-classify a large portion of subtle or nuanced hateful content.

4. **Prompt-based LMMs**:

   - $\mathcal{D}_{Harm}$: Zero-shot classification with Large Multimodal Models (LLMs) or Large Vision-Language Models (LVMs), Qwen-V2-VL performs very poorly (`0`: 0.7634, `1`: 0.2530) in comparison to zero-shot classification with Llava-COT (`0`:0.7839, `1`: 0.2716). Both models mis-classify 82% of harmful content, demonstrating catastrophic failure. This suggests that current zero-shot prompts are insufficient to capture nuanced, harmful content in memes. Moreover, these models tend to be overly conservative, often defaulting to predicting non-harmful labels to avoid false positives.
   - $\mathcal{D}_{Hate}$: In zero-shot classification on this dataset, Llava-COT shows better and more stable performance, achieving class-wise F1 scores of 0.6440 for the negative class and 0.4663 for the positive class. Qwen-V2-VL has a similar F1 for the negative class but a lower F1 for the positive class (0.3275). Qwen-V2-VL misclassifies 77% of hateful content, whereas Llava-COT misclassifies 64%. This indicates that current zero-shot prompts partially help the model capture `hateful` content but remain insufficient for nuanced cases. Additionally, like in $\mathcal{D}_{Harm}$, both models, when prompted in a zero-shot settin,g tend to be conservative, favouring negative predictions and under-detecting positive (`hateful`) instances.

The evaluation on $\mathcal{D}_{Hate}$ and $\mathcal{D}_{Harm}$ does not isolate specific errors (e.g., explicit vs. implicit harm), but comparative trends reveal clear patterns. Text-only training with $\mathcal{U}_T$ (BERT) performs comparably to multimodal fusion, showing that textual cues often suffice for detecting harmful and hateful content, while image-only models lag. No single architecture excels across both tasks. Multimodal early fusion ($\mathcal{MF}^E_{ViLT}$) trained on combined data achieves the most balanced performance without catastrophic failures, making it the safest baseline for moderation across diverse policy violations. Late-fusion models (MF^L_{CLIP}, CLIP) are highly unstable under out-of-domain training with FNRs above 90%, highlighting their fragility when modalities are poorly integrated (simple concatenation). Zero-shot VLMs (Qwen2-VL, Llava-CoT) also underperform (FNRs 64–82%), indicating that scale alone does not compensate for prompt limitations. In-domain training consistently yields the highest class-1 F1 scores (15–30% gains), while combined training

balances performances across domains. Cross-domain transfer is asymmetric: hate-to-harm transfer is partially effective ($\mathcal{U}_T$ `1` F1=0.6192), but harm-to-hate transfer fails ($\mathcal{U}_T$ `1` F1=0.3353), reflecting that harm detection relies on explicit cues, whereas hate speech often requires understanding implicit, context-dependent semantics.

Performance on Unseen Imgur Data: Leveraging classifier performance on $\mathcal{D}_{Harm}$ and $\mathcal{D}_{Hate}$, training all models on the combined dataset ($\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$) yields modest but balanced results across both tasks. Although these models do not achieve peak performance on either individual dataset, they provide more consistent predictions. As discussed in Section 7.2.2, among the 300 memes collected from the unseen Imgur dataset, the two annotators agreed on 239 memes with youth-limited distribution labels and 260 memes with adult-limited distribution labels. For a fair comparison, only the samples where both annotators agreed on both youth- and adult-limited binary labels were considered, resulting in a total of 212 memes (Youth: `limit (1)`: 135. `not-limit (0)`: 77; Adult: `limit (1)`: 47 `not-limit (0)`: 165). Table 7.7 summarises the performance of all classifiers trained on the mixed dataset and evaluated on these 212 memes.

Model	Adult-Restriction				Youth-Restriction			
	`0`-F1	`1`-F1	FNR	FPR	`0`-F1	`1`-F1	FNR	FPR
$\mathcal{U}_T$	0.8387	0.5614	0.3191	0.2121	0.5676	0.5248	0.6074	0.1818
$\mathcal{U}_{CLIP}$	0.7203	0.4203	0.3829	0.3757	0.5354	0.5929	0.5037	0.3116
$\mathcal{U}^I_{ViT}$	0.734	0.378	0.4893	0.3393	0.4976	0.5116	0.5925	0.3246
MF^L_{CLIP}	0.7972	0.6014	0.0851	0.3212	0.5389	0.6147	0.4740	0.3246
MF^E_{ViLT}	0.734	0.378	0.4893	0.3393	0.4976	0.5116	0.5925	0.3246
Qwen-V2-VL	0.8025	0.3800	0.5777	0.2222	0.5153	0.4	0.7153	0.2337
Llava-COT	0.8354	0.5185	0.4042	0.2	0.5877	0.5204	0.6222	0.1298

Table 7.7: Performance of mixed dataset-trained classifiers on 212 annotator-agreed Imgur memes, evaluated separately for adult-restriction and youth-restriction tasks. `0` denotes non-restricted content, `1` denotes restricted content.

- **Model Performance Comparison**: The multimodal late fusion CLIP model (MF^L_{CLIP}) demonstrates the most balanced performance for adult-restriction classification, achieving the highest F1-score for the restricted class (`1`-F1: 0.6014) while maintaining competitive non-restricted class performance (`0`-F1: 0.7972). Notably,

Theme	Sub-Theme	Predicted	Adult-Label	Count
Benign	Normal	1	0	4
Explicit	E1-Direct-Religion	1	0	5
		0	1	1
	E2-Direct-Lgbt	1	0	8
		0	1	1
	E3-Direct-Racism	0	1	2
	E4-Direct-Sexual	1	0	2
Implicit	I1-Double-Entendres	1	0	1
	I2-Political-Sarcasm	1	0	7
	I3-Adult-Humour	1	0	6
	I4-Sarcasm-Prejudice	1	0	10
	I7-Stereotype	1	0	3
	I9-Gender-Identity	1	0	7

Table 7.8: Error distribution of MF^{L}_{CLIP} predictions by theme and sub-theme for adult-restriction classification on Imgur dataset. Predicted=1, Adult-Label=0 indicates false positives; Predicted=0, Adult-Label=1 indicates false negatives.

it exhibits the lowest false negative rate (FNR: 0.0851), indicating superior detection of content requiring adult restriction. The unimodal text classifier ($\mathcal{U}_T$) achieves the highest non-restricted class F1-score (0.8387) and lowest false positive rate (FPR: 0.2121), suggesting effective identification of safe content but at the cost of missing 31.91% of restricted content. Among zero-shot models, Llava-COT outperforms Qwen-V2-VL with comparable 0-F1 (0.8354 vs. 0.8025) and substantially higher 1-F1 (0.5185 vs. 0.3800), though both struggle with high false negative rates (40.42% and 57.77%, respectively). The unimodal image models ($\mathcal{U}^{I}_{CLIP}$, $\mathcal{U}^{I}_{ViLT}$) and early fusion ViLT (MF^{E}_{ViLT}) exhibit the weakest performance, with 1-F1 scores below 0.43 and FNRs exceeding 38%, indicating limited capacity to leverage visual information alone for this task.

- **Adult vs. Youth Restriction Performance**: A striking contrast emerges when comparing model performance across restriction types. For youth-restriction, MF^{L}_{CLIP} maintains the highest restricted-class F1-score (0.6147) with moderate FNR (0.4740), demonstrating consistent multimodal fusion benefits. However, overall performance

Theme	Sub-Theme	Predicted	Youth-Label	Count
Benign	Normal	1	0	4
Explicit	E1-Direct-Religion	0	1	1
	E2-Direct-Lgbt	0	1	2
	E3-Direct-Racism	0	1	2
	E4-Direct-Sexual	0	1	11
Implicit	I1-Double-Entendres	0	1	20
	I11-self-harm	0	1	3
	I2-Political-Sarcasm	1	0	5
		0	1	3
	I3-Adult-Humour	0	1	9
		1	0	6
	I4-Sarcasm-Prejudice	0	1	9
		1	0	2
	I5-Edgy-Content	0	1	2
	I7-Stereotype	1	0	2
		0	1	1
	I8-Body-shame	0	1	1
	I9-Gender-Identity	1	0	6

Table 7.9: Error distribution of MF^{L}_{CLIP} predictions by theme and sub-theme for youth-restriction classification on Imgur dataset. Predicted=1, Youth-Label=0 indicates false positives; Predicted=0, Youth-Label=1 indicates false negatives.

degrades across all models for youth-restriction compared to adult-restriction, with most models achieving `0`-F1 scores below 0.59, whereas for adult-restriction they are above 0.72. This pattern suggests that youth-inappropriate content presents more subtle classification challenges. Interestingly, Llava-COT achieves the lowest FPR for youth-restriction (0.1298), indicating strong precision in identifying safe-for-youth content, yet suffers from the highest FNR (0.6222), revealing a conservative classification bias that misses 62% of youth-restricted content. The unimodal text model ($\mathcal{U}_T$) shows different behaviour across tasks: while effective for adult-restriction (FPR: 0.2121, FNR: 0.3191), it performs near-randomly for youth-restriction (`0`-F1: 0.5676, `1`-F1: 0.5248, FNR: 0.6074), highlighting that textual cues alone are insufficient for identifying youth-inappropriate content. This divergence underscores that adult- and youth-restricted content exhibit distinct distributional characteristics, with youth-restriction requiring stronger integration of visual modal-

ities to capture age-appropriate context that text alone cannot convey.

- **Error Analysis**: The error distribution reveals systematic challenges faced by MF^{L}_{CLIP} across content themes and restriction types. For adult-restriction, as illustrated in Table 7.8, the model exhibits a pronounced false positive bias, particularly misclassifying implicit content as requiring restriction: I4-Sarcasm-Prejudice (10 false positives), E2-Direct-Lgbt (8), I2-Political-Sarcasm (7), and I9-Gender-Identity (7) constitute the majority of errors. This pattern suggests the model conflates socially sensitive topics with adult-inappropriate content, over-restricting political satire and identity-related humour that annotators deemed acceptable for adults. Conversely, explicit content errors are minimal but critical; 2 false negatives in E3-Direct-Racism and scattered misses in religion and LGBT themes indicate occasional failures to detect overtly harmful content. The youth-restriction results outlined in Table 7.9 reveal a contrasting error profile: the model now exhibits false-negative bias, systematically under-restricting implicit content deemed inappropriate for youth. I1-Double-Entendres accounts for 20 false negatives (the highest single error category), followed by E4-Direct-Sexual (11), I3-Adult-Humour (9), and I4-Sarcasm-Prejudice (9), indicating the model's inability to recognise subtle sexual innuendo and adult-oriented humour that requires developmental maturity to interpret appropriately. Notably, 4 benign memes are incorrectly flagged across both tasks, suggesting baseline classification noise. The divergent error patterns, i.e., over-restriction of sensitive but permissible content for adults versus under-restriction of developmentally inappropriate implicit content for youth, underscore the fundamental challenge: adult and youth restrictions operate on distinct axes of harm that require different contextual reasoning, with implicit content presenting the greatest classification ambiguity for age-differentiated moderation.

7.2.4 Pilot Human Preference Data Analysis

Closed-weight VLM Results:

- **Completion Rate**: Of the 300 memes submitted to both Gemini-2.5-Flash and GPT-4o using the detailed prompt outlined in Section 7.1.3, Gemini-2.5-Flash suc-

cessfully processed 298 memes (99.33%), while GPT-4o completed 293 memes (97.67%). The high completion rates indicate robust handling of diverse meme content, though both models occasionally refused to generate predictions, likely due to content policy constraints. A total of 292 memes received predictions from both models and were retained for comparative analysis. Manual inspection revealed that refused memes predominantly contained *E3-Direct-Racism* (e.g., an image showing Donald Trump with the phrase "R**e like a Mexican"). These refusals indicate that both models' content safety filters are triggered by overtly discriminatory language and imagery, even when the task explicitly requests classification rather than content generation. This suggests that safety guardrails in commercial models may limit their utility for content moderation research, as they cannot evaluate the very content they are designed to help moderate.

- **Inter-rater Agreement**: Cohen's Kappa revealed moderate agreement between the two closed-source models across both tasks. For adult-restriction labels, Gemini-2.5-Flash and GPT-4o achieved κ= 0.5389 (57 disagreements out of 292 memes), indicating moderate concordance in identifying adult-inappropriate content. For youth-restriction labels, agreement decreased to κ = 0.4443 (59 disagreements), reflecting greater classification ambiguity for youth-oriented moderation. Cumulatively, 72 memes (24.66%) exhibited disagreement on either adult- or youth-restriction labels, with disagreements concentrated in ambiguous content categories. The majority of conflicts occurred in implicit content: 19 memes involved I3-Adult-Humour (where sexual or mature themes were subtly implied), 10 involved I1-Double-Entendres (requiring interpretation of layered meanings), and, notably, 11 were categorised as benign by human annotators. The substantial disagreement on benign content (15.28% of all disagreements) reveals that both models exhibit false positive tendencies, over-restricting innocuous memes, a critical concern for real-world deployment, where excessive content removal can undermine user experience and platform trust.

- **Alignment with Human Annotations**: To ensure fair comparison, we selected only instances where both human raters agreed and neither GPT-4o nor Gemini-2.5-Flash refused to respond, yielding 209 memes (adult-context: `limit (1)`=45, `do-`

`not-limit (0)`=164; youth-context: `limit (1)`=132, `do-not-limit (0)`=77). For adult-context moderation, GPT-4o achieved a macro-average F1 of 0.5286 (limit F1=0.4699, do-not-limit F1=0.5872) with 4% false negatives but a high 58% false positive rate. Gemini-2.5-Flash performed worse with macro-average F1 of 0.4687 (limit F1=0.4330, do-not-limit F1=0.5045), exhibiting 4.5% false negatives and 65.8% false positives. In the youth context, GPT-4o achieved a macro-average F1 of 0.6653 (limit F1=0.7630, do-not-limit F1=0.5676) with increased false negatives (22%) and 45% false positives, while Gemini-2.5-Flash achieved comparable performance at 0.6650 (limit F1=0.7829, do-not-limit F1=0.5481) with lower false negatives (16%) but higher false positives (51%). Both models demonstrate stronger performance in youth-context classification, reflecting better alignment with stricter content standards for younger audiences. Table 7.10 reveals the thematic distribution of errors for both adult and youth audience contexts across both models, showing that Gemini-2.5-Flash made 42 total errors compared to GPT-4o's 37 errors. Both models struggle most with implicit content categories, particularly Adult-Humour (Gemini: 14 errors, GPT-4o: 12 errors) and Political-Sarcasm (Gemini: 5 errors, GPT-4o: 3 errors), indicating difficulty in interpreting contextual nuance and satire. Notably, both models also misclassified benign content as harmful (Gemini: 14 instances, GPT-4o: 11 instances), contributing to the high false positive rates observed. The concentration of errors in implicit themes (I2-I9) suggests that current models lack the sophisticated cultural and contextual understanding necessary for nuanced moderation decisions. At the same time, their mishandling of benign content underscores the risk of over-censorship in automated content moderation systems.

Pilot Results: The pilot human preference study was conducted on 60 randomly selected memes out of 72, for which GPT-4o and Gemini-2.0 produced conflicting outputs. As detailed in Section 7.1.3, each meme–response pair included model outputs across **five** dimensions: (i) contextual understanding, (ii) visual interpretation, (iii) textual comprehension, (iv) reasoning for limiting youth access, and (v) reasoning for limiting adult access. For each dimension, human annotators were asked to choose which system per-

Model	Theme	Sub-Theme	Count
Gemini-2.5-Flash	**Benign**	Normal	14
	Explicit	E1-Direct-Religion	1
		E4-Direct-sexual	2
		E5-Direct-sexism	1
	Implicit	I2-Political-Sarcasm	5
		I3-Adult-Humour	14
		I4-Sarcasm-Prejudice	2
		I5-Edgy-Content	1
		I6-Word-Play	1
		I9-Gender-Identity	1
		Total	42
GPT-4o	**Benign**	Normal	11
	Explicit	E1-Direct-Religion	1
		E4-Direct-sexual	2
	Implicit	I2-Political-Sarcasm	3
		I3-Adult-Humour	12
		I4-Sarcasm-Prejudice	3
		I5-Edgy-Content	1
		I7-Stereotype	2
		I9-Gender-Identity	2
		Total	37

Table 7.10: Thematic error distribution for GPT-4o and Gemini-2.5-Flash across youth and adult audience contexts. Errors are categorised into benign, explicit (E1-E5), and implicit (I2-I9) content themes, with sub-theme breakdowns showing model misclassification.

formed better: `System-A` (Gemini) and `System-B` (GPT), `Neither`, `Both`. Inter-annotator reliability among the three annotators (2 Males: aged 26 and 24; 1 Female aged 25) was measured using Krippendorff's α, indicating the level of consistency in their judgments.

Table 7.11 presents the dimension-wise Krippendorff's α values, which measure the inter-annotator agreement across the five evaluated dimensions. The α values range from 0.1188 for *Reasoning for Limiting Youth Access* to 0.2903 for *Visual Interpretation*. While these values reflect low to moderate agreement (Krippendorff, 2011), they align with prior studies in subjective tasks such as content moderation and harmfulness evaluation (Röttger et al., 2022; Vidgen and Derczynski, 2020), where interpretative ambiguity, contextual nuances, and differing perspectives naturally produce variability among human judgments. Notably, *Contextual Understanding* and *Visual Interpretation* show the highest agreement levels (0.2822 and 0.2903, respectively), suggesting that these dimensions are compara-

tively easier for annotators to assess consistently. In contrast, reasoning-related dimensions, particularly *Reasoning for Limiting Youth Access* (0.1188), exhibit lower consensus, highlighting the inherent subjectivity in evaluating rationale-based responses.

Dimension	**Krippendorff's α**
Contextual Understanding	0.2822
Visual Interpretation	0.2903
Textual Comprehension	0.2284
Reasoning for Limiting Youth Access	0.1188
Reasoning for Limiting Mixed Access	0.1623

Table 7.11: Dimension-wise Krippendorff's α

Table 7.12 complements this by showing the distribution of agreement among annotators. Here, *Textual Comprehension* stands out with 25 instances where all three human annotators agreed on the same label, indicating that the model outputs for textual understanding were the most consistently interpretable. *Reasoning for Limiting Youth Access* follows closely with 24 instances of unanimous agreement, yet it also shows three cases where all annotators disagreed, reflecting the higher subjectivity noted in Krippendorff's α. The "`at least two agree`" row shows a high overall convergence across all dimensions, indicating that even when full consensus was not achieved, there was often partial agreement between annotators. The "`all disagree`" counts, though low, emphasise areas where interpretations diverged entirely, signalling potential ambiguities.

	Contextual Understanding	**Visual Interpretation**	**Textual Comprehension**	**Reasoning for Limiting Youth Access**	**Reasoning for Limiting Mixed Access**
All 3 Agree	22	23	25	24	20
At least 2 Agree	59	60	59	57	58
All Disagree	1	0	1	3	2

Table 7.12: Agreement Distribution for Preference Analysis

Looking more closely at individual dimensions, the distribution tables reveal nuanced patterns of preference. Table 7.13 details the combinations of responses for *Contextual Understanding*. System-A (Gemini-2.5) was predominantly preferred, with 29 counts where

at least two annotators chose it. A smaller number of disagreements and mixed choices highlight isolated cases of annotator divergence. This trend suggests that System-A provides more consistent contextually coherent responses compared to System-B (GPT-3).

Annotator-1	Annotator-2	Annotator-3	Count
System-A	System-B	System-A	29
	System-A	System-A	22
	System-A	System-B	3
	Neither	System-A	2
	System-B	Both	1
System-B	System-A	System-A	2
System-B	System-B	System-A	1

Table 7.13: Distribution of Contextual Understanding Response Combinations: Note: System A = Gemini and System B = GPT

Similarly, Table 7.14 presents response patterns for *Visual Interpretation*. The agreement here slightly favors System-A, with 28 counts in one of the most common combination patterns, followed closely by other mixed selections. Notably, there are very few instances of complete disagreement, which aligns with the higher α score (0.2903) for this dimension and suggests that visual interpretation is generally easier for annotators to evaluate consistently.

Annotator-1	Annotator-2	Annotator-3	Count
System-A	System-B	System-A	28
	System-A	System-A	23
	System-A	System-B	3
	Neither	System-A	3
	System-A	Both	1
	System-B	System-B	1
System-B	System-B	System-A	1

Table 7.14: Distribution of Visual Interpretation Response Combinations: Note: System A = Gemini and System B = GPT

For *Textual Comprehension*, Table 7.15 demonstrates a clear preference for System-A, with 25 instances where all annotators agreed on its outputs. There are also 25 instances where a mix of System-A and System-B choices occurred, reflecting minor disagreements. This reinforces the earlier observation from Table 7.12 that textual comprehension was

the dimension with the highest unanimous agreement, indicating clarity and accessibility in System-A's outputs.

Annotator-1	Annotator-2	Annotator-3	Count
System-A	System-A	System-A	25
	System-B	System-A	25
	System-A	System-B	2
	Neither	System-A	2
	System-A	Both	1
	System-B	System-B	1
	System-B	System-A	2
System-B	System-A	System-A	1
	Neither	System-A	1

Table 7.15: Distribution of Textual Comprehension Response Combinations: Note: System A = Gemini and System B = GPT

In contrast, reasoning dimensions exhibit more dispersed patterns. Table 7.16 shows the distribution for *Reasoning for Limiting Mixed Access*. Although System-A remains the more frequently preferred system, there is considerable variation in the choices, with multiple combinations including System-B and "Neither." This dispersion reflects the lower Krippendorff's α for reasoning-related dimensions and underscores the evaluative challenges inherent in subjective judgments about rationale-based responses.

Annotator-1	Annotator-2	Annotator-3	Count
System-A	System-A	System-A	20
	System-B	System-A	18
	System-B	System-B	9
	System-A	System-B	3
	Neither	System-A	3
	Neither	Both	1
	System-B	System-A	4
System-B	System-A	System-B	1
	Neither	System-A	1

Table 7.16: Distribution of Reasoning for Limiting Mixed Access Response Combinations: Note: System A = Gemini and System B = GPT

Finally, Table 7.17 examines *Reasoning for Limiting Youth Access*. While 22 instances favoured System-A unanimously, there were also multiple mixed and "Neither" selections, including three cases of complete disagreement among annotators. These results mirror

the low α (0.1188) reported in Table 7.11 and highlight the complexity and interpretive flexibility inherent in evaluating reasoning related to sensitive content. System-A generally performs well, but the diversity of responses illustrates the nuanced judgments human evaluators must make.

Annotator-1	Annotator-2	Annotator-3	Count
	System-A	System-A	22
	System-B	System-A	11
	System-B	System-B	9
System-A	Neither	System-A	7
	System-A	System-B	2
	Neither	System-B	1
	Neither	Both	1
Neither	Neither	System-A	1
Neither	System-B	System-A	1
	System-B	System-A	2
System-B	System-B	System-B	2
	System-A	System-B	1

Table 7.17: Distribution of Reasoning for Limiting Youth Access Response Combinations: Note: System A = Gemini and System B = GPT

In summary, these quantitative analyses collectively illustrate that System-A (Gemini-2.5) tends to achieve higher and more consistent agreement across all dimensions, particularly for comprehension and contextual tasks, whereas reasoning-based dimensions remain challenging and subject to divergent human interpretation. This pattern aligns with broader trends observed in subjective content evaluation, highlighting both the strengths and limitations of current model outputs.

7.3 Summary

This chapter investigated **RQ2.2** to examine the extent to which multimodal memes can be evaluated for youth- and adult-restricted content using both open- and closed-source vision–language models (VLMs). To address this, a three-phase experimental framework was implemented, encompassing unseen data curation (Phase-1), baseline benchmarking (Phase-2), and pilot human preference evaluation (Phase-3). Collectively, these phases operationalised a scalable, multimodal approach to content moderation that balanced

automation with human interpretive oversight.

Phase 1 focused on the curation of an unseen dataset of memes representing a wide range of visual and textual features, including implicit humour, cultural allusions, and sensitive content such as sexual innuendo and hate speech. This dataset served as a rigorous test set for assessing generalisation beyond training distributions. Particular attention was given to balancing explicit and implicit harmfulness categories, ensuring the inclusion of borderline cases that challenge both human and model interpretability.

Phase 2 established baseline performance benchmarks across multiple open-source VLMs using the $\mathcal{D}_{Harm}$ (Pramanick et al., 2021a) and $\mathcal{D}_{Hate}$ (Kiela et al., 2020) and the Phase-1 unseen test-set. These experiments revealed substantial variability in model outputs and a strong dependence on prompt design and visual–textual alignment. While models demonstrated competent image–text recognition, their moderation predictions often faltered in cases requiring pragmatic or cultural inference. False negative rates and per-class F1 scores highlighted a recurrent pattern: models tended to over-classify ambiguous memes as adult-restricted, reflecting a conservative bias in thresholding sensitive material. These findings underscore the need to incorporate nuanced human perspectives to interpret moderation outcomes accurately.

Phase 3 extended this analysis through the use of closed-source preference data generation, leveraging systems such as Gemini-2.0-Flash and GPT-4o to produce detailed reasoning-based moderation outputs. This phase aimed to explore whether advanced proprietary models exhibit greater contextual sensitivity or interpretive coherence when evaluating ambiguous multimodal content. Both systems achieved high completion rates (over 97%), yet occasionally refused classification due to embedded safety filters—particularly for overtly discriminatory or explicit imagery. Such refusals illuminate a structural limitation in deploying commercially constrained models for moderation research, as their content safety mechanisms preclude systematic analysis of the very materials under study.

The pilot human preference study further examined 60 memes where model predictions diverged. Human annotators assessed paired outputs across five interpretive dimensions: contextual understanding, visual interpretation, textual comprehension, and reasoning for youth and adult restriction. Statistical analysis using Krippendorff's α indicated low to moderate inter-annotator reliability (0.12–0.29), consistent with prior research

on subjective moderation tasks. Patterns of agreement revealed that comprehension and contextual understanding were evaluated most consistently, whereas rationale-based reasoning remained highly subjective. Across all dimensions, Gemini-2.5 was more frequently preferred, suggesting stronger alignment with human judgments in terms of interpretive clarity and contextual fidelity. However, both models exhibited occasional over-restriction and inconsistencies in explaining their decisions, highlighting the persistent challenge of achieving balanced moderation between safety and expressiveness.

Overall, this chapter demonstrated that multimodal moderation remains an inherently interpretive and context-dependent task, where even advanced VLMs display uneven performance across different dimensions and datasets. The empirical findings reveal that while closed-source models can produce contextually richer and more comprehensible outputs, they are constrained by safety filters and subjectivity in rationale evaluation. These limitations reinforce the importance of human-in-the-loop frameworks that combine automated classification with nuanced human adjudication. Overall, the chapter contributes an empirical and methodological foundation for subsequent work on interpretive variability and multimodal bias, setting the stage for deeper exploration of alignment, explainability, and trust in AI-mediated content moderation.

Part IV

Conclusion

Chapter 8

Conclusion and Future Work

This thesis set out to address the lack of high-quality, youth-centred cyberbullying (CB) datasets in English. Building on the EU Joint Research Committee (JRC) Report (Charisi et al., 2022), which calls to integrate youth voices in AI-design and development, and adopting a rights-based approach through Ireland's (Department of Children, Equality, Disability, Integration and Youth, 2023) *Participation with a Purpose* framework, this thesis integrates youth perspectives into dataset development as rigorously as in software design. Grounded in Routine Activities Theory (RAT) (Cohen and Felson, 1979), Crime Opportunity Theory (COT) (Felson and Clarke, 1998), and Social Media Affordance Theory (Ronzhyn et al., 2023), this thesis presents an integrated framework for understanding cyberbullying from the bystander perspective. Through detailed user research discussed in Chapters 3 and 4, this work captures (i) *where youth perceive harm*, (ii) *how they respond*, (iii) *the types of content they encounter*, (iv) *which CB situations they deem report-worthy*, (v) *the language used in such situations*, and (vi) *topics on which they come across potentially offensive* `memes`. This results in a youth-co-created dataset ($\mathcal{D}_{GOLD}$) curated through the week-long "*Young Research Assistant Programme*" (YRAP) (Section 3.4 in Chapter 3) that provides a safe space for participating in CB research and learning. Beyond dataset creation, this thesis introduces CB $\mathcal{D}_{Silver}$ datasets, augmented through human translation ($\mathcal{D}^{HT}_{IT\rightarrow EN}$) (Section 5.1.1 Chapter 5), human-aided machine translations ($\mathcal{D}^{HMT}_{FR\rightarrow EN}$) (Section 5.1.3 Chapter 5), and large language models (LLMs) (Section 5.1.4 Chapter 5). The utility of a $\mathcal{D}_{GOLD}$ and samples of $\mathcal{D}_{Silver}$ were evaluated **intrinsically** through human evaluation in Section 5.3 to ensure the data accurately reflects

cyberbullying dynamics, and **extrinsically** in Sections 6.1.3, 6.3.1 and 6.3.2 in Chapter 6 to test whether $\mathcal{D}_{Silver}$ supports training moderation systems to detect not only whether text is `hurtful`, but also ***by whom*** (e.g., `enabler`, `defender`, `bully`, `victim`) across different social contexts. In this chapter, the thesis synthesises all contributions, revisits key research questions, reflects on the limitations, and highlights open questions and avenues for future research.

8.1 Contributions

The contributions of this research are as follows:

(**A**) **Ethical Framework for Integrating Youth Perspectives**: Motivated by co-research with youth framework (Guishard and Tuck, 2013; Clark et al., 2022a) and building on Ireland's (Department of Children, Equality, Disability, Integration and Youth, 2023) "participation with a purpose framework", this thesis in Chapter 3 Section 3.4 establishes an ethical framework for youth participation in cyberbullying research through the **Young Research Assistant Programme (YRAP)**, demonstrating how staggered, voluntary engagement can elicit authentic perspectives while mitigating risks of harm exposure and emotional distress.

(**B**) **Dataset Curation**: Outcomes of the **YRAP**, include (a) the first youth co-created conversational dataset in English (N =552, Section 4.3.2 Chapter 4), and (b) youth created obfuscated samples (N =423 Section 6.4). Additionally, this thesis also creates complementary corpora: (i) human translated - $\mathcal{D}^{HT}_{IT\rightarrow EN}$ (Section 5.1.1), (ii) machine-translated - $\mathcal{D}^{MT}_{IT\rightarrow EN}$ (Section 5.1.2), (iii) human-aided machine translation - $\mathcal{D}^{HMT}_{FR\rightarrow EN}$ (Section 5.1.3), (iv) large language model generated CB conversations (N =6677 Section 5.1.4). (See Table 8.1).

(**C**) **Ecological Validity through Human Evaluation**: To assess the ecological validity of youth co-created dataset (§4.3.2 Chapter 4) whether a sample of datasets curated reflect cyberbullying dynamics, through a novel human evaluation framework discussed in Section 5.3, with two social scientists, two professional content moderators, and one adult teen, we examine power dynamics, conflict, support,

Dataset Type	Dataset Name	Dataset Size	Section/Chapter
$\mathcal{D}_{\textbf{Silver}}$	$\mathcal{D}^{HT}_{IT\rightarrow EN}$	2,192	Human Translation (§5.1.1 Chapter 5)
	$\mathcal{D}^{MT}_{IT\rightarrow EN}$	2,192	Machine Translation (§5.1.2 Chapter 5)
	$\mathcal{D}^{ST}_{FR\rightarrow EN}$	876	Student Human Translation (§5.1.3 Chapter 5)
	$\mathcal{D}^{HMT}_{FR\rightarrow EN}$	3,649	Human-aided Machine Translation (§5.1.3 Chapter 5)
	$\mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$	6,677	Phase-2 LLM Generation (§5.1.4 Chapter 5)
$\mathcal{D}_{\textbf{Gold}}$	Youth Co-created Conversational Dataset	552	Created in §4.3.2 Chapter 4, Profiled in §5.2, Evaluated and Annotated in §5.3 and 5.4 Chapter 5 respectively.
	Youth Co-created Adversarial Dataset	423	Created as part of §4.3 Chapter 4 discussed in Section 6.4 Chapter 6

Table 8.1: Datasets created through this doctoral dissertation

seriousness, and sarcasm within the datasets, confirming that youth-created and LLM-generated datasets capture authentic cyberbullying characteristics.

(**D**) **Annotation Framework**: We propose a three-level annotation schema discussed in Section 5.4, that contextualises aggression (`overt/covert`), categorizes harm types (`blackmail, insults, exclusion`), and identifies message intent (`attack, defense, abetment, instigation`), enabling granular analysis of conversational dynamics. Through context-window-based labelling, i.e., controlling for the number of preceding messages during labelling, we find improvements in inter-rater reliability as the context size increases from 2 to 5 prior messages. This helps position the youth-co-created dataset in Section 4.3.2 Chapter 4 as a benchmark for detecting CB amongst teens.

(**E**) **Text Classifier Benchmarking**: Discussed in Chapter 6, our benchmarking of NLP classification paradigms in binary, multi-class and multi-task settings demonstrates varied performances across both *encoder*- and *decoder*- only pre-trained transformer language models. In multi-task settings, domain-adapted transformers

with hard-parameter sharing achieve F1 scores of 0.718 (`hurtfulness`) and 0.555 (`roles`). However, interpretability and minority-class detection remain challenging.

(**F**) **Multimodal Classifier Benchmarking**: Discussed in Chapter 7, we extend the benchmarking to multimodal contexts and present an annotated meme dataset for age-differentiated content moderation. We find that vision-language models for explicit/implicit `harm` and `hate` detection exhibit systematic biases; they over-restrict implicit adult content and yet under-restrict implicit youth-inappropriate material.

(**G**) **Text Classifier Robustness**: Discussed in Chapter 6 Section 6.4, we find that **no current architecture**, including industry APIs, achieves robust detection of youth-obfuscated language, with all systems exhibiting 14-26% F1 degradation and Attack Success Rates up to 31.44%, underscoring the necessity of youth-inclusive development frameworks that account for creative linguistic manipulation.

8.2 Research Question Revisited

RQ.1 How can we engineer datasets and test cases for building real-world cyberbullying detection systems in a safe and ethical manner?

Building on youth co-research methodologies (Guishard and Tuck, 2013; Clark et al., 2022a), this thesis (Section 3.4, Chapter 3) proposes a mixed-methods framework to ethically integrate youth perspectives into dataset development. The "*Young Research Assistant Programme*" (YRAP) respects youth autonomy by enabling participants to *learn*, *create*, and *contribute* voluntarily on topics they consider `report-worthy` and `serious`. Its staggered interval participation ensures independent yet sequential contributions, minimising ethical risks in *living labs*, such as exposure to harmful content, emotional distress, privacy risks, and reinforcement of negative peer dynamics (Livingstone and Stoilova, 2021; Jicol et al., 2022). Informed by youth perspectives on witnessing cyberbullying behaviours on social media platforms, discussed in Chapter 4 (Section 4.1, 4.2), the YRAP process, yields 552 messages across 10 scenarios (denigration, harassment, outing, exclusion, cyberstalking), with average message length of 45 characters and role distributions: `Defenders` 43.9%, `Enablers` 33.9%, `Bullies` 17.6%, `Victims` 4.5% (Section 4.3). Additionally, as

outlined in Table 8.1, leveraging data augmentation strategies in Chapter 5, we expand the cyberbullying detection corpus into English through: (i) human translation of the Italian youth role-played CB dataset (Sprugnoli et al., 2018) (Section 5.1.1) yielding $\mathcal{D}^{HT}_{IT\rightarrow EN}$, (ii) human-aided machine translation of the French youth role-played dataset (Ollagnier et al., 2022; Ollagnier, 2024) (Section 5.1.3) yielding $\mathcal{D}^{HMT}_{FR\rightarrow EN}$, and LLM-generated conversational messages for CB contexts (Section 5.1.4), yielding $\mathcal{D}^{\text{Phase-2}}_{\text{LLM}}$. Descriptive profiling (Section 5.2) compares the youth co-created and LLM-generated datasets with Italian, French, and real-world CB corpora (Van Hee et al., 2018), while **intrinsic human evaluation** (Section 5.3) confirms the datasets capture authentic cyberbullying dynamics. Additionally, a "*context window*" annotation framework labels the youth co-created dataset across three levels: (i) **aggression**, (ii) **type** of cyberbullying, and (iii) **message intent**, establishing the youth co-created dataset discussed in Section 4.3.2 Chapter 4 as $\mathcal{D}_{GOLD}$.

RQ2.1 Which classification paradigms that leverage state-of-the-art pre-trained language models (PLM) are best suited for detecting cyberbullying in social media content created by youth?

To address **RQ2.1**, this thesis evaluates multiple classification paradigms through **three** interconnected themes, considering both **what is said** and **who says it**, while extending analysis beyond benchmarking to interpretability, role identification, and multi-task learning:

Theme-1 Binary Benchmarking: Using CB datasets that are built through established cyberbullying definitions, cross-platform benchmarking (§6.1.2) of classical ML (SVM, Random Forest, XGBoost), sequential models (Bi-LSTM, GRU), and transformer models shows that domain-adapted HateBERT consistently outperforms others across user-comment, question-answering, and micro-blogging platforms. Discussed in Section 6.1.3, reproducing (Corazza et al., 2019) with Italian youth role-played dataset (Sprugnoli et al., 2018), we find that human-translated datasets ($\mathcal{D}^{HT}_{IT\rightarrow EN}$) achieves performance com-

parable to the original corpus for `hurtfulness` detection, while machine-translated datasets ($\mathcal{D}^{MT}_{IT\rightarrow EN}$) perform slightly worse. BERT attains high macro F1-scores (0.81 human, 0.75 machine), demonstrating that translated corpora are suitable for training English CB detection models.

Theme-2 **Explainable CB Detection**: Discussed in Section 6.2, despite HateBERT's superior F1 (0.81 vs 0.77 on micro-blogging), interpretability analyses show attention weights poorly align with gradient-based importance (Pearson's Correlation (r): 0.0129–0.19), and sentiment/POS analyses reveal that linguistically salient or negative tokens are not consistently prioritised. This confirms prior findings (Jain and Wallace, 2019; Elsafoury et al., 2021) and indicates that while models correctly identify `hurtfulness`, attention alone cannot explain **why a message is flagged**, limiting safety-critical applications.

Theme-3 **Multi-class and Multi-task Classification**: Discussed in Section 6.3.1, multi-class role detection (`bully`, `victim`, `enabler`, `defender`) using multi-source datasets shows few-shot prompting of large language models (LLaMA-3.3-70B) helps achieve a macro F1 0.48, with minority classes (victims F1: 0.19–0.22) remaining challenging. Discussed in Section 6.3.2, multi-task learning with hard-parameter sharing and AdapterFusion (Pfeiffer et al., 2021), allows simultaneous modelling of hurtfulness and roles, encoder-based BERT fine-tuned with hard-parameter sharing and cross-entropy loss helps reach a F1 0.718 (`hurtfulness`) and 0.555 (`roles`), out-performing decoder-based GPT. Error analysis shows persistent confusion between bully and enabler roles (30–60%), highlighting the importance of capturing subtle linguistic cues, sarcasm, and context.

Overall, revisiting RQ2.1 demonstrates that domain-adapted transformers are most effective for youth CB detection, translated corpora can reliably support model training, and multi-task frameworks improve simultaneous detection of hurtfulness and roles (**what is said by whom**), while interpretability and minority-class detection remain open challenges for real-world deployment.

RQ2.2 **How can classification techniques be adapted to detect *`hurtfulness`* in cy-**

berbullying contexts within obfuscated or perturbed textual social media content?

To address this research question, in Section 6.4 of Chapter 6, we examine how classification techniques can be adapted to detect cyberbullying in obfuscated textual content, using youth co-created adversarial examples. Through baseline evaluations of character-level, transformer-based, and hybrid architectures without adversarial training to realise baseline performances, against industry APIs, the findings reveal that **no model achieves true robustness** to adolescent obfuscation strategies. Youth researchers through the "*Young Researcher Programme*" generated 423 obfuscated variations of 61 sentences using leet speak (21.79%), phrase transformations (19.5%), emoji replacements (17.2%), and special characters (10.95%). All systems demonstrated 14-26% F1 degradation under obfuscation, with Attack Success Rates ranging from 18.56% (OpenAI) to 31.44% (Perspective API). The hybrid char-CNN-BERT architecture, despite achieving the highest test performance (F1: 0.7920), exhibited the worst adversarial resilience (flip rate: 28.83%, ASR: 26.76%), demonstrating that an architectural ensemble does not guarantee robustness. Character-level models showed superior stability for `hurtful` content detection (char-CNN robustness: 0.7975) but suffered confidence drops, while BERT-based models maintained confidence stability yet experienced moderate flip rates (24-29%). Asymmetric vulnerability patterns emerged: `hurtful` content proved more susceptible to obfuscation (20-29% prediction flips) compared to benign content (>82% stability), indicating differential model sensitivity to surface-form perturbations in hurtful language. Industry APIs demonstrated critical limitations. Perspective API suffered catastrophic failure under obfuscation (F1: 0.4440, false negatives: 69.32%), while OpenAI's Moderation API maintained relative resilience (ASR: 18.56%) at the cost of elevated false positives (28.86%). Error analysis revealed systematic failures across all models for context-dependent slurs and disability-related language, with unanimous misclassification of phrases like "*`you have downs`*" across all variations. These findings confirm that current classification techniques remain fundamentally vulnerable to obfuscations created by youth, regardless of architectural sophistication. The pronounced performance degradation on youth-created *obfuscated* data

(21-24% F1 drops) underscores that adolescent linguistic patterns diverge substantially from training distributions, necessitating youth-inclusive development frameworks that account for creative language manipulation strategies.

RQ.3 To what extent can vision-language models trained on `harmful` and `hateful` meme datasets generalise to age-differentiated content moderation decisions across varying levels of semantic complexity ranging from *explicit* (e.g., slurs, violent imagery) to *implicit* cues (e.g., cultural or contextual cues), to *benign* content (e.g., scenery, neutral humour)?

To address this research question, in Chapter 7, we investigated the capacity of vision-language models trained on `hateful` (Kiela et al., 2020) and `harmful` (Pramanick et al., 2021a) meme datasets to identify and differentiate *explicit* from *implicit* harmful semantics in memes for content moderation. Through benchmarking across multiple architectures (Section 7.1.2) and a human preference evaluation (Section 7.1.3), the findings reveal fundamental limitations in current VLM approaches to detecting nuanced online harms. Baseline assessments demonstrated that models struggle asymmetrically with *explicit* versus *implicit* content. Cross-domain transfer experiments revealed that `harm-to-hate` transfer consistently fails (F1=0.3353). In contrast, `hate-to-harm` transfer shows partial success (F1=0.6192), indicating that `harm` detection relies on explicit textual cues, whereas `hate` requires implicit, context-dependent interpretation. Multimodal early fusion (MF^{E}_{ViLT}) achieved the most balanced performance on combined training data ($\mathcal{D}_{Harm} \bigcup \mathcal{D}_{Hate}$) yet still misclassified 28-50% of harmful content depending on dataset complexity. Error analysis on the unseen Imgur dataset ($\mathcal{D}_{\text{Imgur}}$) exposed systematic biases: for adult-restricted, multimodal late fusion (MF^{L}_{CLIP}) over-restricting implicit content (I4-Sarcasm-Prejudice: 10 false positives), conflating sensitive topics with harmful content. Conversely, for youth-restricted content, the same model under-restricted implicit content (I1-Double-Entendres: 20 false negatives), failing to detect subtle innuendo requiring developmental maturity to interpret. Zero-shot VLMs such as LLava-COT (Xu et al., 2024), exhibited catastrophic failure rates (64-82% FNR), with conservative biases toward non-harmful predictions. Human preference evaluation of closed-source model responses, across 60 divergent predictions, yielded low

inter-annotator reliability (α=0.12-0.29), confirming that preference data for harm assessment, especially with *implicit* content, remains inherently subjective. While Gemini-2.5 demonstrated stronger contextual alignment than GPT-4o, both models showed inconsistent reasoning transparency and occasional over-restriction driven by embedded safety filters. These findings demonstrate that current VLMs cannot reliably differentiate explicit from implicit harmful semantics without human oversight. The distinct error patterns across content types and restriction contexts underscore that automated moderation requires architecture-specific human-in-the-loop frameworks calibrated to the interpretive demands of implicit harm detection.

8.3 Limitations

The methodological foundations of this thesis present inherent trade-offs between ethical rigour and empirical scope, which can be organised across **five** key dimensions:

Dimension 1. Ethical Constraints and Data Collection Scale

- **Resource Intensity of Youth Co-Research:** The Young Researcher Programme (Section 3.4), while ethically sound and methodologically rigorous, proved exceptionally resource-intensive, requiring extended engagement periods, multiple facilitated sessions, and continuous ethical oversight. This investment yielded 552 messages across ten cyberbullying scenarios, a modest corpus relative to large-scale datasets typically used in NLP benchmarking, constraining the diversity of linguistic patterns, platform-specific behaviours, and cyberbullying subtypes captured.

- **Bystander-Centric Perspective Bias:** Both YRAP (Section 3.4) and youth perspectives surveys (Section 4.1, 4.2) are from the **bystander standpoint**. This is a deliberate compromise necessitated by ethical review constraints that preclude direct inquiry into adolescents' perpetration or victimisation experiences without robust psychological safeguards, trauma-informed protocols, and ongoing clinical support infrastructure. While this design choice protects participant wellbeing and satisfies institutional review board requirements, it fundamentally restricts insight into the manifestations of *bullying* (perpetrator) and *being bullied* (target), po-

tentially skewing dataset characteristics toward observer rather than participant language, and under-representing victim voice and perpetrator justifications.

- **Generalisability Across Demographics and Contexts:** The youth participants were recruited from Ireland, thereby limiting generalisability to diverse cultural settings, linguistic communities, and platform ecosystems. Cyberbullying manifestations vary substantially across cultural norms (Smith and Steffgen, 2013) and regulatory environments (Livingstone et al., 2014), yet the datasets reflect predominantly Western, English-speaking adolescent communication patterns.

Dimension 2. Modelling Approaches and Architectural Choices

- **Message-Level Classification Ignoring Conversational Context**: Benchmarking in Chapter 6 establishes computationally efficient baselines but evaluates messages in isolation rather than as sequential discourse, overlooking dependencies where intent, power dynamics, and harm emerge through multi-turn interactions (Kumar et al., 2022). Cyberbullying often manifests through accumulated micro-aggressions, coordinated attacks by multiple perpetrators, or escalating conflict that only becomes interpretable when conversational history is considered (Van Hee et al., 2018). The absence of conversational state modelling means that contextually-dependent harm, such as inside jokes turned hostile, referenced prior incidents, or implicit threats requiring discourse memory, remains undetectable by the evaluated systems.

- **Static Fine-tuning Without Continual Learning**: The classification paradigms employ static fine-tuning approaches that overlook continual learning architectures (De Lange et al., 2021; Wang et al., 2024b) that could help train task-specific expert encoders, potentially improving domain adaptation without catastrophic forgetting. Although this thesis explores AdapterFusion (Pfeiffer et al., 2021) continual learning frameworks, such as (Omrani et al., 2024) that deploy task-specific expert encoders to improve domain adaptation, enable incremental updates as new cyberbullying manifestations emerge, and support multi-platform deployment where each platform's communication style represents a distinct task domain.

Dimension 3. Adversarial Robustness and Defense Mechanisms

- **Vulnerability Assessment Without Defense Implementation:** The adversarial robustness investigation (**RQ2.2**) discussed in Section 6.4 comprehensively identifies vulnerabilities across architectures but does not implement or evaluate defense mechanisms. Adversarial training techniques such as Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014), Projected Gradient Descent (PGD) (Madry et al., 2017), among others, can be leveraged to update the pre-trained embeddings of *encoder-only* transformer models by iteratively perturbing the input embeddings in the direction that maximises the loss function, then fine-tuning the encoder to minimise this adversarial loss, resulting in embeddings that maintain performance under adversarial conditions. As (Zhang et al., 2020b) highlights, these have demonstrated efficacy in hardening classifiers against text perturbations, yet remain unexplored in this youth-obfuscation context, leaving mitigation strategies unaddressed.

- **Static Adversarial Examples:** The 423 obfuscated variations generated by youth researchers represent a snapshot of creative linguistic manipulation as of the data collection period (2023-2024). Obfuscation tactics evolve rapidly as youth communities develop new coded language, platform-specific abbreviations, and counter-detection strategies (Vidgen and Derczynski, 2020). The adversarial evaluation does not account for changes in obfuscation strategies or assess whether models trained on current obfuscations generalise to future tactics, limiting the ecological validity of robustness findings beyond the immediate study period.

Dimension 4. Vision-Language Model Evaluation Constraints

- **Limited Fusion Architecture Coverage:** Experiments and results discussed in Chapter 7, focus only on late fusion (concatenation) and early fusion (joint encoding) paradigms but omits intermediate fusion strategies (Kiela et al., 2019; Thapliyal et al., 2022; Qiu et al., 2022) and ensemble strategies that have performed superiorly in both *hateful* and *harmful* detection (Kiela et al., 2020; Pramanick et al., 2021a). The architectural space exploration remains incomplete, potentially missing fusion

strategies better suited to explicit-implicit harm differentiation or age-differentiated content moderation.

- **Age-Differentiated Harm Taxonomy Limitations:** The adult-versus-youth restriction framework, while grounded in platform policy structures (e.g., Instagram, TikTok age-gating), employs binary distinctions that oversimplify developmental appropriateness. Harm perception varies continuously across adolescent development (Livingstone et al., 2014), with 13-year-olds exhibiting substantially different media literacy and emotional regulation compared to 17-year-olds, yet both fall within the "youth" category. Finer-grained age brackets aligned with developmental psychology research could improve moderation precision but would require substantially larger annotated datasets with age-specific expert adjudication.

- **Limited Ablation Study:** The experiments in Chapter 7 do not include a detailed ablation to identify which image regions, grids, or text features drive model decisions. While unimodal models are analysed for insight, Grad-CAM visualisations (Selvaraju et al., 2017) can highlight the specific image features influencing predictions, as demonstrated by (Kumar and Nandakumar, 2022).

Dimension 5. Interpretability and Explainability Gaps

- **Attention-Based Explainability Limitations**: Theme 2 (Chapter 6) demonstrates that attention weights poorly align with gradient-based importance scores (r=0.0129-0.19) and fail to consistently prioritise linguistically salient or sentiment-bearing tokens, confirming prior findings (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019). However, the thesis does not explore alternative explainability methods such as SHapley Additive exPlanations (SHAP) (Mosca et al., 2022) or counterfactual explanations (Wu et al., 2021) that might provide more faithful explanations for safety-critical cyberbullying detection decisions.

- **Lack of Structured Error Analysis**: While error analysis identifies misclassified instances (e.g., "*you have downs*", disability-related slurs), the thesis does not attribute errors to specific failure modes (e.g., insufficient training data for rare slurs,

tokenisation artefacts, cultural knowledge gaps, contextual ambiguity). Structured error taxonomies (Ribeiro et al., 2020; Wu et al., 2019) and failure mode analysis could inform targeted improvements rather than general architectural refinements.

8.4 Open Questions and Future Work

Open Questions: Several foundational questions emerge from the empirical findings that warrant further investigation.

1. **How can conversational context be effectively modelled in cyberbullying detection without incurring prohibitive computational costs or privacy risks associated with maintaining user interaction histories?**

   The tension between message-level classification and discourse-level understanding remains unresolved, particularly in detecting patterns such as coordinated harassment or escalating conflict that only manifest across extended exchanges.

2. **What architectural innovations or training paradigms can bridge the explicit-implicit harm detection gap revealed in vision-language model evaluation?**

   The divergence in model performance between adult and youth content moderation, where the same architecture over-restricts implicit adult content and under-restricts implicit youth in-appropriate material, suggests that current multimodal fusion strategies lack the contextual reasoning required for age-differentiated harm assessment.

3. **To what extent can adversarial training generalise across the evolving obfuscation strategies used by youth populations?**

   The 21.79% prevalence of leet speak (e.g., '*i*'→ 1), 19.5% phrase transformations, and 17.2% emoji replacements in youth-generated adversarial examples represent a snapshot of creative linguistic manipulation; whether models hardened through adversarial training against these specific perturbations maintain robustness as obfuscation tactics evolve remains empirically unexplored.

4. **What factors explain the persistent failure of minority-class detection in**

cyberbullying role identification, and can targeted interventions address this limitation?

Victims achieving F1 scores of only 0.19-0.22 in multi-class role identification raise several interrelated questions: (a) Do current annotation frameworks adequately capture the linguistic and behavioural manifestations of victimhood, or do they use researcher-defined categories over authentic victim expression? (b) Are victims' communicative patterns inherently more variable and context-dependent than other participant roles, reflecting diverse coping strategies (silence, deflection, counter-aggression) that resist categorical classification? (c) Would data collection methods that centre victim voices and not bystander observations yield different linguistic distributions? (d) Can meta-learning approaches, few-shot learning paradigms, or synthetic data generation specifically tailored to victim patterns overcome the scarcity and variability that characterise this minority class? The interplay among annotation frameworks, data collection ethics, and algorithmic optimisation in addressing this challenge represents a critical frontier for equitable cyberbullying detection systems.

Future Work: Future research should pursue three interconnected directions to advance youth-centred content moderation.

1. Developing conversationally aware architectures that incorporate temporal and relational context without compromising privacy represents a critical need. Approaches leveraging graph neural networks (Mishra et al., 2019; Saikia et al., 2022) or memory-augmented transformers (Zhong et al., 2024) that can help maintain ephemeral conversation state could enable the detection of coordinated harassment and escalation patterns. Such systems must be coupled with continual learning frameworks (Omrani et al., 2024) that deploy task-specific encoders, allowing models to adapt to emerging cyberbullying tactics and platform-specific norms without catastrophic forgetting of established detection capabilities.

2. Adversarial robustness must transition from vulnerability assessment to proactive defense deployment. Integrating adversarial training techniques (e.g., FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2017), among others) with youth-generated

obfuscation examples provides a pathway to create robust classifiers that maintain performance under linguistic perturbation. Critically, such defenses should be co-designed with adolescent researchers to ensure training procedures reflect authentic rather than researcher-imagined obfuscation strategies, potentially through iterative red-teaming exercises, where youth participants attempt to circumvent progressively hardened models.

3. Multimodal content moderation requires architectural innovation beyond baseline fusion paradigms. Exploring intermediate fusion strategies (e.g., (Kiela et al., 2019; Thapliyal et al., 2022; Qiu et al., 2022)) could improve *explicit-implicit* harm differentiation. Embedding developmental psychology principles, such as age-specific harm taxonomies or maturity-gated content representations, into model training may address the divergence in adult versus youth content moderation performance. Longitudinal studies tracking how youth obfuscation tactics and cyberbullying manifestations evolve across developmental stages would further inform adaptive moderation systems that anticipate rather than merely react to emergent harms.

Collectively, these directions emphasise that effective Trust and Safety systems must be co-created with youth stakeholders, architecturally adaptive to conversational and adversarial dynamics, and grounded in a developmental understanding of how digital harm manifests and is experienced among adolescents.

Bibliography

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Aggarwal, A., Maurya, K., and Chaudhary, A. (2020). Comparative study for predicting the severity of cyberbullying across multiple social media platforms. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 871–877. IEEE.

Aggarwal, P. and Zesch, T. (2022). Analyzing the real vulnerability of hate speech detection systems against targeted intentional noise. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 230–242.

Agha, Z., Park, J., Wan, R., Ali, N. S., Wang, Y., Difranzo, D., Badillo-Urquiola, K., and Wisniewski, P. J. (2024). Tricky vs. transparent: Towards an ecologically valid and safe approach for evaluating online safety nudges for teens. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Agrawal, S. and Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2 edition.

Ahmed, T., Ivan, S., Kabir, M., Mahmud, H., and Hasan, K. (2022). Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. *Social Network Analysis and Mining*, 12(1):99.

Ahmed, T., Kabir, M., Ivan, S., Mahmud, H., and Hasan, K. (2021). Am i being bullied on social media? an ensemble approach to categorize cyberbullying. In *2021 IEEE international conference on big data (Big data)*, pages 2442–2453. IEEE.

Aizenkot, D. (2022). The predictability of routine activity theory for cyberbullying victimization among children and youth: Risk and protective factors. *Journal of interpersonal violence*, 37(13-14):NP11857–NP11882.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Alghamdi, J., Lin, Y., and Luo, S. (2022). Modeling fake news detection using bert-cnn-bilstm architecture. In *2022 IEEE 5th international conference on multimedia information processing and retrieval (MIPR)*, pages 354–357. IEEE.

Alihosseini, D., Montahaei, E., and Soleymani Baghshah, M. (2019). Jointly measuring diversity and quality in text generation models. In Bosselut, A., Celikyilmaz, A., Ghazvininejad, M., Iyer, S., Khandelwal, U., Rashkin, H., and Wolf, T., editors, *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Alipan, A., Skues, J. L., Theiler, S., and Wise, L. (2020). Defining cyberbullying: a multifaceted definition based on the perspectives of emerging adults. *International journal of bullying prevention*, 2:79–92.

Almeida, A. F. P., Villela, M. L. B., Santos, C. Q., and Guimarães, G. C. (2022). Participatory design with teenagers for cyberbullying prevention and mitigation. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems*, IHC '22, New York, NY, USA. Association for Computing Machinery.

Alotaibi, M., Alotaibi, B., and Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. *Electronics*, 10(21):2664.

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., and Chang, K.-W. (2018). Generating natural language adversarial examples. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., and Zwerdling, N. (2020). Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7383–7390.

Ansary, N. S. (2020). Cyberbullying: Concepts, theories, and correlates informing evidence-based best practices for prevention. *Aggression and violent behavior*, 50:101343.

Anscombe, F. J. (1956). On estimating binomial response relations. *Biometrika*, 43(3/4):461–464.

Anthropic (2023). Model card and evaluations for claude models.

Anzovino, M., Fersini, E., and Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

Arango, A., Pérez, J., and Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 45–54, New York, NY, USA. Association for Computing Machinery.

Arntfield, M. (2015). Toward a cybervictimology: Cyberbullying, routine activities theory, and the anti-sociality of social media. *Canadian Journal of Communication*, 40(3):371–388.

Ashktorab, Z. and Vitak, J. (2016). Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 3895–3905, New York, NY, USA. Association for Computing Machinery.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T. J., Jones, A., Joseph, N., Mann, B., Dassarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., and Kaplan, J. (2021). A general language assistant as a laboratory for alignment. *ArXiv*, abs/2112.00861.

Astorino, A., Rizzi, G., and Fersini, E. (2023). Integrated gradients as proxy of disagreement in hateful content. In Boschetti, F., Lebani, G. E., Magnini, B., and Novielli, N., editors, *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 46–52, Venice, Italy. CEUR Workshop Proceedings.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Badillo-Urquiola, K., Smriti, D., McNally, B., Golub, E., Bonsignore, E., and Wisniewski, P. J. (2019). Stranger danger! social media app features co-designed with children to keep them safe online. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, IDC '19, page 394–406, New York, NY, USA. Association for Computing Machinery.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N.,

Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback.

Balahur, A. and Turchi, M. (2012). Multilingual sentiment analysis using machine translation? In Balahur, A., Montoyo, A., Barco, P. M., and Boldrini, E., editors, *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, Jeju, Korea. Association for Computational Linguistics.

Balahur, A. and Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Banko, M., MacKeen, B., and Ray, L. (2020). A unified taxonomy of harmful content. In *Proceedings of the fourth workshop on online abuse and harms*, pages 125–137.

Bao, T., Ren, N., Luo, R., Wang, B., Shen, G., and Guo, T. (2021). A bert-based hybrid short text classification model incorporating cnn and attention-based bigru. *Journal of Organizational and End User Computing (JOEUC)*, 33(6):1–21.

Barančíková, P. and Bojar, O. (2025). Intrinsic vs. extrinsic evaluation of Czech sentence embeddings: Semantic relevance doesn't help with MT evaluation. In Bouillon, P., Gerlach, J., Girletti, S., Volkart, L., Rubino, R., Sennrich, R., Farinha, A. C., Gaido, M., Daems, J., Kenny, D., Moniz, H., and Szoc, S., editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 265–275, Geneva, Switzerland. European Association for Machine Translation.

Barlett, C. P., Roth, B. R., and Rinker, A. M. (2022). Muscles, popularity, social capital, and computer skills: Examining "power" in cyberbullying. *Aggressive behavior*, 48(6):608–615.

Basseches, M. (1984). Dialectical thinking. *Norwood, NJ: Ablex.*

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., and De Bourdeaudhuij, I. (2015). 'can i afford to help?'how affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *Behaviour & Information Technology*, 34(4):425–435.

Bastings, J. and Filippova, K. (2020). The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In Alishahi, A., Belinkov, Y., Chrupała, G., Hupkes, D., Pinter, Y., and Sajjad, H., editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Bauer, N., Preisig, M., and Volk, M. (2024). Offensiveness, hate, emotion and GPT: Benchmarking GPT3.5 and GPT4 as classifiers on Twitter-specific datasets. In Kumar, R., Ojha, A. K., Malmasi, S., Chakravarthi, B. R., Lahiri, B., Singh, S., and Ratan, S., editors, *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 126–133, Torino, Italia. ELRA and ICCL.

Bauman, S. (2015). Types of cyberbullying. *Cyberbullying: What counselors need to know*, pages 53–58.

Bayari, R. and Bensefia, A. (2021). Text mining techniques for cyberbullying detection: state of the art. *Adv. Sci. Technol. Eng. Syst. J*, 6:783–790.

Bayzick, J., Kontostathis, A., and Edwards, L. (2011). Detecting the presence of cyberbullying using computer software. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, Koblenz, Germany. Poster presentation.

Behzadi, M., Harris, I. G., and Derakhshan, A. (2021). Rapid cyber-bullying detection method using compact bert models. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 199–202. IEEE.

Bender, E. M. (2011). On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers

of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Beresford, O., Cooney, A., Keogh, A., Flynn, E., and Messena, M. (2023). Keeping kids safer online: Online safety matters. trends and usage report academic year 2022/2023. `https://www.cybersafekids.ie/wp-content/uploads/2023/09/CSK_Data-Trends-Report-2023-V2-Web-Version.pdf`.

Bhargava, A., Witkowski, C., Shah, M., and Thomson, M. (2023). What's the magic word? a control theory of llm prompting. *arXiv preprint arXiv:2310.04444*.

Bharti, S., Yadav, A. K., Kumar, M., and Yadav, D. (2021). Cyberbullying detection from tweets using deep learning. *Kybernetes*, 51(9):2695–2711.

Biewald, L. (2020). Experiment tracking with weights and biases. Software available from wandb.com.

Blaya, C. and Audrin, C. (2019). Toward an understanding of the characteristics of secondary school cyberhate perpetrators. In *Frontiers in Education*, volume 4, page 46. Frontiers Media SA.

Bonetti, F. and Tonelli, S. (2022). An analysis of abusive language data collected through a game with a purpose. In Madge, C., editor, *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*, pages 1–6, Marseille, France. European Language Resources Association.

Bonifacio, L., Abonizio, H., Fadaee, M., and Nogueira, R. (2022). Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2387–2392.

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M., Maynard, D., and Aswani, N. (2013). witie: An open-source information extraction pipeline for microblog text.

Proceedings of the International Conference on Recent Advances in Natural Language Processing.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., Tesconi, M., et al. (2018). Overview of the evalita 2018 hate speech detection task. In *Ceur workshop proceedings*, volume 2263, pages 1–9. CEUR.

Boyd, D. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In *A networked self*, pages 47–66. Routledge.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bretschneider, U. and Peters, R. (2016). Detecting cyberbullying in online communities. In *Proceedings of the 24th European Conference on Information Systems*, ECIS 2016, page 61. AIS Electronic Library (AISeL).

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing: Algorithms, architectures and applications*, pages 227–236. Springer.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Brunato, D., Cimino, A., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2020). Profiling-UD: a tool for linguistic profiling of texts. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., and Wattenhofer, R. (2019). On identifiability in transformers. *arXiv preprint arXiv:1908.04211*.

Bruns, A. and Burgess, J. (2011). The use of twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European consortium for political research (ECPR) general conference 2011*, pages 1–9. The European Consortium for Political Research (ECPR).

Brunton, F. and Nissenbaum, H. (2015). *Obfuscation: A user's guide for privacy and protest*. Mit Press.

Buglass, S. L., Abell, L., Betts, L. R., Hill, R., and Saunders, J. (2021). Banter versus bullying: A university student perspective. *International journal of bullying prevention*, 3:287–299.

Byrt, T., Bishop, J., and Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429.

Camelford, K. G. and Ebrahim, C. (2016). The cyberbullying virus: A psychoeducational intervention to define and discuss cyberbullying among high school females. *Journal of Creativity in Mental Health*, 11(3-4):458–468.

Campbell, M. and Xu, J. (2022). Children and adolescents' understanding of traditional and cyberbullying. *Children and youth services review*, 139:106528.

Cao, J., Zhang, Z., Wang, H., Liang, B., Wang, H., and Wong, K.-F. (2024). Ospc: Detecting harmful memes with large language model as a catalyst. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1892–1895.

Cao, R., Hee, M. S., Kuek, A., Chong, W.-H., Lee, R. K.-W., and Jiang, J. (2023). Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 5244–5252.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57.

Caselli, T., Basile, V., Mitrovic, J., and Granitzer, M. (2020). Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472.

Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (2021). HateBERT: Retraining BERT for abusive language detection in English. In Mostafazadeh Davani, A., Kiela, D., Lambert, M., Vidgen, B., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Cassidy, W., Faucher, C., and Jackson, M. (2013). Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice. *School psychology international*, 34(6):575–612.

Chan, T. K., Cheung, C. M., Benbasat, I., Xiao, B., and Lee, Z. W. (2022). Bystanders join in cyberbullying on social networking sites: the deindividuation and moral disengagement perspectives. *Information Systems Research*.

Chan, T. K., Cheung, C. M., and Lee, Z. W. (2021). Cyberbullying on social networking sites: A literature review and future research directions. *Information & Management*, 58(2):103411.

Chan, T. K., Cheung, C. M., and Wong, R. Y. (2019). Cyberbullying on social networking sites: The crime opportunity and affordance perspectives. *Journal of Management Information Systems*, 36(2):574–609.

Charisi, V., Chaudron, S., Di Gioia, R., Vuorikari, R., Escobar-Planas, M., Sanchez, I., and Gomez, E. (2022). Artificial intelligence and the rights of the child - towards an integrated agenda for research and policy. Scientific analysis or review KJ-NA-31048-EN-N, Publications Office of the European Union, Luxembourg (Luxembourg).

Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E. D., Stringhini, G., Vakali, A., and Kourtellis, N. (2019). Detecting cyberbullying and cyberaggression in social media. *ACM Transactions on the Web (TWEB)*, 13(3):1–51.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chefer, H., Gur, S., and Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., and Sun, L. (2024a). Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Chen, H.-Y. and Li, C.-T. (2020). Henin: Learning heterogeneous neural interaction networks for explainable cyberbullying detection on social media. *arXiv preprint arXiv:2010.04576*.

Chen, J., Guo, H., Yi, K., Li, B., and Elhoseiny, M. (2022). Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040.

Chen, S., Zhang, Y., and Yang, Q. (2024b). Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Cheng, L., Guo, R., Silva, Y., Hall, D., and Liu, H. (2019). Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM international conference on data mining*, pages 235–243. SIAM.

Chhun, C., Colombo, P., Suchanek, F. M., and Clavel, C. (2022). Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chinchor, N. and Sundheim, B. M. (1993). Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Chiu, B., Korhonen, A., and Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6, Berlin, Germany. Association for Computational Linguistics.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Chowdhury, S. A. and Zamparelli, R. (2018). RNN simulations of grammaticality judgments on long-distance dependencies. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Clark, A., Fox, C., and Lappin, S. (2013). *The handbook of computational linguistics and natural language processing*. John Wiley & Sons.

Clark, A. T., Ahmed, I., Metzger, S., Walker, E., and Wylie, R. (2022a). Moving from co-design to co-research: engaging youth participation in guided qualitative inquiry. *International Journal of Qualitative Methods*, 21:16094069221084793.

Clark, J. H., Garrette, D., Turc, I., and Wieting, J. (2022b). Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cohen, L. E. and Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American sociological review*, pages 588–608.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Corazza, M. (2021). GitLab Repo: creep-cyberbullying-classifier, query? Email correspondence. Personal communication.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2019). Cross-platform evaluation for italian hate speech detection. In *CLiC-it 2019-6th Annual Conference of the Italian Association for Computational Linguistics*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Council of European Union (2016). Regulation (eu) 2016/679 of the european parliament and of the council (general data protection regulation). `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., Zhang, Y., Zhang, Y., Zheng, H., Zhang, J., Zhang, J., Liu, Y., Yu, D., and Ma, Y. (2025). Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model.

CyberSafeKids (2024). Left to their own devices: Trends and usage report academic year 2023–2024. Technical report, CyberSafeIreland CLG, Dun Laoghaire, Co. Dublin, Ireland. Accessed: 2025-06-16.

Dadvar, M. and Eckert, K. (2020). Cyberbullying detection in social networks using deep learning based models. In *Big Data Analytics and Knowledge Discovery: 22nd*

International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22, pages 245–255. Springer.

Dadvar, M., Trieschnigg, D., Ordelman, R., and De Jong, F. (2013). Improving cyberbullying detection with user context. In *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35*, pages 693–696. Springer.

Dai, W., Yu, T., Liu, Z., and Fung, P. (2020). Kungfupanda at SemEval-2020 task 12: BERT-based multi-TaskLearning for offensive language detection. In Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., and Shutova, E., editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2060–2066, Barcelona (online). International Committee for Computational Linguistics.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318.

Davidson, J., Livingstone, S., Jenkins, S., Gekoski, A., Choak, C., Ike, T., and Phillips, K. (2019). Adult online hate, harassment and abuse: a rapid evidence assessment. Technical report, London School of Economics (LSE).

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the international AAAI conference on web and social media*, 11(1):512–515.

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385.

Dell'Era, C. and Landoni, P. (2014). Living lab: A methodology between user-centred design and participatory design. *Creativity and innovation management*, 23(2):137–154.

Department of Children, Equality, Disability, Integration and Youth (2023). National

framework for children and young people's participation in decision-making. Government report, Government of Ireland.

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., and Jaggi, M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1045–1052, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

DeSmet, A., Rodelli, M., Walrave, M., Soenens, B., Cardon, G., and De Bourdeaudhuij, I. (2018). Cyberbullying and traditional bullying involvement among heterosexual and non-heterosexual adolescents, and their associations with age and gender. *Computers in Human Behavior*, 83:254–261.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Dhar, M., Kumar, V., and Shrivastava, M. (2018). Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach. In Machonis, P., Barreiro, A., Kocijan, K., and Silberztein, M., editors, *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 131–140, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Dinakar, K., Reichart, R., and Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 11–17.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Moens, M.-F., Huang, X., Specia, L., and Yih,

S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Dredge, R., Gleeson, J., and De la Piedad Garcia, X. (2014). Cyberbullying in social networking sites: An adolescent victim's perspective. *Computers in human behavior*, 36:13–20.

Druin, A. (1999). Cooperative inquiry: developing new technologies for children with children. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 592–599.

Du, Y., Masood, M. A., and Joseph, K. (2020). Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 153–164.

Dubois, Y., Li, C. X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P. S., and Hashimoto, T. B. (2023). Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.

Ebrahimi, J., Rao, A., Lowd, D., and Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Edwards, A., Demoll, D., and Edwards, L. (2020). Detecting cyberbullying activity across platforms. In *17th International Conference on Information Technology–New Generations (ITNG 2020)*, pages 45–50. Springer.

Eisenstein, J. (2013). What to do about bad language on the internet. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

El-Sayed, A. and Nasr, O. (2024). AAST-NLP at multimodal hate speech event detection 2024 : A multimodal approach for classification of text-embedded images based on CLIP and BERT-based models. In Hürriyetoğlu, A., Tanev, H., Thapa, S., and Uludoğan, G., editors, *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 139–144, St. Julians, Malta. Association for Computational Linguistics.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Elsafoury, F., Katsigiannis, S., Wilson, S. R., and Ramzan, N. (2021). Does bert pay attention to cyberbullying? In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1900–1904.

Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., Desmet, B., Hoste, V., and Daelemans, W. (2021). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. *Language Resources and Evaluation*, 55:597–633.

Englander, E., Donnerstein, E., Kowalski, R., Lin, C. A., and Parti, K. (2017). Defining cyberbullying. *Pediatrics*, 140(Supplement 2):S148–S151.

Enguehard, J. (2023). Sequential integrated gradients: a simple but effective method for explaining language models. *arXiv preprint arXiv:2305.15853*.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Fan, A. and Wu, Y. S. (2022). Identifying hateful memes with multimodal classification. *Cs231n. stanford. edu*.

Fang, Y., Yang, S., Zhao, B., and Huang, C. (2021). Cyberbullying detection in social networks using bi-gru with self-attention mechanism. *Information*, 12(4):171.

Faraj, S. and Azad, B. (2012). The materiality of technology: An affordance perspective. *Materiality and organizing: Social interaction in a technological world*, 237(1):237–258.

Felson, M. and Clarke, R. V. (1998). Opportunity makes the thief. police research series, paper 98. policing and reducing crime unit. *Research, Development and Statistics Directorate. London: Home Office*.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced data sets*, volume 10. Springer.

Fiani, C., Bretin, R., Macdonald, S. A., Khamis, M., and McGill, M. (2024). " pikachu would electrocute people who are misbehaving": Expert, guardian and child perspectives on automated embodied moderators for safeguarding children in social virtual reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Fidon, L., Li, W., Garcia-Peraza-Herrera, L. C., Ekanayake, J., Kitchen, N., Ourselin, S., and Vercauteren, T. (2017). Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In *International MICCAI brainlesion workshop*, pages 64–76. Springer.

Fisher, J., Hallinan, S., Lu, X., Gordon, M. L., Harchaoui, Z., and Choi, Y. (2024). Styleremix: Interpretable authorship obfuscation via distillation and perturbation of style elements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4172–4206.

Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the royal statistical society*, 85(1):87–94.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Frank, E., Hall, M. A., and Witten, I. H. (2016). The WEKA Workbench: Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". Technical report or online document associated with the book.

Ge, S., Cheng, L., and Liu, H. (2021). Improving cyberbullying detection with user interaction. In *Proceedings of the Web Conference 2021*, WWW '21, page 496–506, New York, NY, USA. Association for Computing Machinery.

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., and Yu, D. (2024). Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Ghosh, S., Singh, G. V., Arora, J., and Ekbal, A. (2024). Benchmarking cyber harassment dialogue comprehension through emotion-informed manifestations-determinants demarcation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22040–22048.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston.

Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*, 1(2):67–82.

Gomez, R., Gibert, J., Gomez, L., and Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Gongane, V. U., Munot, M. V., and Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1):129.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gorwa, R., Binns, R., and Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Goyal, S., Doddapaneni, S., Khapra, M. M., and Ravindran, B. (2023). A survey of adversarial defenses and robustness in nlp. *ACM Comput. Surv.*, 55(14s).

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 33–41.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Gravano, A. and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12.

Gu, H., Yu, Q., Hou, S., Fang, Z., Wu, H., and He, Z. (2025). Mememind: A large-scale multimodal dataset with chain-of-thought reasoning for harmful meme detection. *arXiv preprint arXiv:2506.18919*.

Guha, M. L., Druin, A., Chipman, G., Fails, J. A., Simms, S., and Farber, A. (2004). Mixing ideas: a new technique for working with young children as design partners.

In *Proceedings of the 2004 conference on Interaction design and children: building a community*, pages 35–42.

Guishard, M. and Tuck, E. (2013). Youth resistance research methods and ethical challenges. In *Youth resistance research and theories of change*, pages 181–194. Routledge.

Guo, X. and Gauch, S. (2024). Using sarcasm to improve cyberbullying detection. In Kumar, R., Ojha, A. K., Malmasi, S., Chakravarthi, B. R., Lahiri, B., Singh, S., and Ratan, S., editors, *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 52–59, Torino, Italia. ELRA and ICCL.

Haldane, J. (1956). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of human genetics*, 20(4):309–311.

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a #twitter. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Han, L., Smeaton, A., and Jones, G. (2021). Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the first workshop on modelling translation: Translatology in the digital age*, pages 15–33.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hearst, M., Dumais, S., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Hee, M. S. and Lee, R. K.-W. (2025). Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. In *Proceed-*

ings of the International AAAI Conference on Web and Social Media, volume 19, pages 774–785.

Hennebert, M.-A., Pasquier, V., and Lévesque, C. (2021). What do unions do… with digital technologies? an affordance approach. *New Technology, Work and Employment*, 36(2):177–200.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2018). The sockeye neural machine translation toolkit at AMTA 2018. In Cherry, C. and Neubig, G., editors, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

Hiebl, B. and Gromann, D. (2023). Quality in human and machine translation: An interdisciplinary survey. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 375–384.

Hinduja, S. and Patchin, J. W. (2013). Social influences on cyberbullying behaviors among middle and high school students. *Journal of youth and adolescence*, 42:711–722.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Hoskin, T. (2012). Parametric and nonparametric: Demystifying the terms. In *Mayo Clinic*, volume 5, pages 1–5.

Hossain, M. R. and Hoque, M. M. (2020). Towards Bengali word embedding: Corpus creation, intrinsic and extrinsic evaluations. In Bhattacharyya, P., Sharma, D. M., and Sangal, R., editors, *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 453–459, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Hosseinmardi, H. (2022). Email correspondence. Personal communication.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2015). Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Hu, T. and Collier, N. (2024). Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Huang, C. L., Zhang, S., and Yang, S. C. (2020). How students react to different cyberbullying events: Past experience, judgment, perceived seriousness, helping behavior and the effect of online disinhibition. *Computers in human behavior*, 110:106338.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Huang, J., Pan, J., Wan, Z., Lyu, H., and Luo, J. (2025). Evolver: Chain-of-evolution prompting to boost large multimodal models for hateful meme detection. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7321–7330, Abu Dhabi, UAE. Association for Computational Linguistics.

Huang, P., Zhu, H., Zheng, L., and Wang, Y. (2021). Text sentiment analysis based on bert and convolutional neural networks. In *Proceedings of the 2021 5th International Conference on Natural Language Processing and Information Retrieval*, pages 1–7.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Huertas-Tato, J., Koutlis, C., Papadopoulos, S., Camacho, D., and Kompatsiaris, I. (2024). A clip-based siamese approach for meme classification. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Hui, Z., Guo, Z., Zhao, H., Duan, J., and Huang, C. (2024). ToxiCraft: A novel framework for synthetic generation of harmful information. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16632–16647, Miami, Florida, USA. Association for Computational Linguistics.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Hutchby, I. (2001). Technologies, texts and affordances. *Sociology*, 35(2):441–456.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Ibrahim, M., Torki, M., and El-Makky, N. (2020). AlexU-BackTranslation-TL at SemEval-2020 task 12: Improving offensive language detection using data augmentation and transfer learning. In Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., and Shutova, E., editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1881–1890, Barcelona (online). International Committee for Computational Linguistics.

Imankulova, A., Sato, T., and Komachi, M. (2017). Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In Nakazawa, T. and Goto, I., editors, *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan. Asian Federation of Natural Language Processing.

J. Sweeting, M., J. Sutton, A., and C. Lambert, P. (2004). What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in medicine*, 23(9):1351–1375.

Jacobs, G., Van Hee, C., and Hoste, V. (2022). Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text? *Natural Language Engineering*, 28(2):141–166.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jain, V., Kumar, V., Pal, V., and Vishwakarma, D. K. (2021). Detection of cyberbullying on social media using machine learning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1091–1096. IEEE.

Jehl, L., Hieber, F., and Riezler, S. (2012). Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the seventh workshop on statistical machine translation*, pages 410–421.

Jewell, N. P. (2003). *Statistics for epidemiology*. chapman and hall/CRC.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Jicol, C., Feltham, J., Yoon, J., Proulx, M. J., O'Neill, E., and Lutteroth, C. (2022). Designing and assessing a virtual reality simulation to build resilience to street harassment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Jigsaw and Google's Counter Abuse Technology Team (2025). Perspective API: A Tool to Explore Online Toxicity. `https://www.perspectiveapi.com/`. Accessed: November 3, 2025.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (2020). Is bert really robust? a strong

baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Jones, K. S. and Galliers, J. R. (1996). Evaluating natural language processing systems: An analysis and review.

Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *International conference on machine learning*, pages 2342–2350. PMLR.

Justice with Children Initiative (2023). Best practices analysis: Lessons learned and practical guidance on the development of child-centered chatbots and digital solutions to enhance children's access to justice. Mutual learning paper, Terre des hommes Foundation "Lausanne" Regional Hub in Hungary.

Kaggle (2018). Tweets dataset for detection of cyber-trolls.

Kalia, D. and Aleem, S. (2017). Cyber victimization among adolescents: Examining the role of routine activity theory. *Journal of Psychosocial Research*, 12(1):223.

Kalliris, K. (2024). Online harm, free speech, and the 'legal but harmful'debate: an interest-based approach. *Journal of Media Law*, 16(2):390–416.

Kaluarachchi, C., Sedera, D., and Warren, M. (2020). An intervention model for cyberbullying based on the general theory of crime and routine activity theory. In *ACIS 2020 Proceedings*, page 8. AIS Electronic Library (AISeL).

Kancharla, B., Singh, P., Kancharla, L. B., Chama, Y., and Sharma, R. (2025). Identifying aggression and offensive language in code-mixed tweets: A multi-task transfer learning approach. In Weerasinghe, R., Anuradha, I., and Sumanathilaka, D., editors, *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 122–128, Abu Dhabi. Association for Computational Linguistics.

Kane, G. C., Alavi, M., Labianca, G., and Borgatti, S. P. (2014). What's different about social media networks? a framework and research agenda. *MIS quarterly*, 38(1):275–304.

Kaushal, R., Van De Kerkhof, J., Goanta, C., Spanakis, G., and Iamnitchi, A. (2024). Automated transparency: A legal and empirical analysis of the digital services act transparency database. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1121–1132, New York, NY, USA. Association for Computing Machinery.

Kazemi, A., Kalaivendan, S. B. N., Wagner, J., Qadeer, H., and Davis, B. (2025). Synthetic vs. gold: The role of llm-generated labels and data in cyberbullying detection. *arXiv preprint arXiv:2502.15860*.

Kazerooni, F., Taylor, S. H., Bazarova, N. N., and Whitlock, J. (2018). Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *Journal of Computer-Mediated Communication*, 23(3):146–162.

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., and Sabharwal, A. (2022). Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Kiela, D., Bhooshan, S., Firooz, H., Perez, E., and Testuggine, D. (2019). Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Kigerl, A. (2012). Routine activity theory and the determinants of high cybercrime countries. *Social science computer review*, 30(4):470–486.

Kim, H.-Y. (2017). Statistical notes for clinical researchers: Chi-squared test and fisher's exact test. *Restorative dentistry & endodontics*, 42(2):152–155.

Kim, J., Yang, N., and Jung, K. (2024). Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks. *arXiv preprint arXiv:2408.08631*.

Kim, S., Razi, A., Stringhini, G., Wisniewski, P. J., and De Choudhury, M. (2021a). A human-centered systematic literature review of cyberbullying detection algorithms. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Kim, S., Razi, A., Stringhini, G., Wisniewski, P. J., and De Choudhury, M. (2021b). A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34.

Kim, W., Son, B., and Kim, I. (2021c). Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. (2016). Character-aware neural language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kirk, H., Vidgen, B., Rottger, P., Thrush, T., and Hale, S. A. (2022). Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

Kirk, H., Yin, W., Vidgen, B., and Röttger, P. (2023). SemEval-2023 task 10: Explainable detection of online sexism. In Ojha, A. K., Doğruöz, A. S., Da San Martino, G., Tayyar Madabushi, H., Kumar, R., and Sartori, E., editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T., and Hale, S. A. (2021). Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *arXiv preprint arXiv:2108.05921*.

Kofoed, J. and Staksrud, E. (2019). 'we always torment different people, so by definition, we are no bullies': The problem of definitions in cyberbullying research. *New Media & Society*, 21(4):1006–1020.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for pytorch.

Kowalski, R. (2018). Cyberbullying. In *The Routledge international handbook of human aggression*, pages 131–142. Routledge.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., and Lattanner, M. R. (2014). Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin*, 140(4):1073.

Kowalski, R. M., Limber, S. P., and Agatston, P. W. (2012). Bullying in the digital age: A review of the research and implications for prevention and intervention. *Aggression and violent behavior*, 17(4):255–266.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability. Retrieved from University of Pennsylvania ScholarlyCommons.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, E. and Lu, W., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kumar, G. K. and Nandakumar, K. (2022). Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In Biester, L., Demszky, D., Jin, Z., Sachan, M., Tetreault, J., Wilson, S., Xiao, L., and Zhao, J., editors, *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kumar, R., Bhalla, O., Vanthi, M., Wani, S. M., and Singh, S. (2024). HarmPot: An annotation framework for evaluating offline harm potential of social media text. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8016–8034, Torino, Italia. ELRA and ICCL.

Kumar, R., Ratan, S., Singh, S., Nandi, E., Devi, L. N., Bhagat, A., Dawer, Y., Lahiri, B., Bansal, A., and Ojha, A. K. (2022). The ComMA dataset v0.2: Annotating aggression and bias in multilingual social media discourse. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4149–4161, Marseille, France. European Language Resources Association.

Lafourcade, M., Joubert, A., and Le Brun, N. (2015). *Games with a Purpose (GWAPS)*. John Wiley & Sons.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Le, P. and Zuidema, W. (2016). Quantifying the vanishing gradient and long distance dependency problem in recursive neural networks and recursive LSTMs. In Blunsom, P., Cho, K., Cohen, S., Grefenstette, E., Hermann, K. M., Rimell, L., Weston, J., and

Yih, S. W.-t., editors, *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 87–93, Berlin, Germany. Association for Computational Linguistics.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, C. and Shin, N. (2017). Prevalence of cyberbullying and predictors of cyberbullying perpetration among korean adolescents. *Computers in human behavior*, 68:352–358.

Leerssen, P. (2023). An end to shadow banning? transparency rights in the digital services act between content moderation and curation. *Computer Law & Security Review*, 48:105790.

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., and Vasserman, L. (2022). A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3197–3207.

Leonardi, P. M. (2011). When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS quarterly*, pages 147–167.

Leukfeldt, E. R. and Yar, M. (2016). Applying routine activity theory to cybercrime: A theoretical and empirical analysis. *Deviant Behavior*, 37(3):263–280.

Leung, A. N., Wong, N., and Farver, J. M. (2018). You are what you read: The belief systems of cyber-bystanders on social networking sites. *Frontiers in psychology*, 9:365.

Li, H., Guo, D., Fan, W., Xu, M., Huang, J., Meng, F., and Song, Y. (2023a). Multi-step jailbreaking privacy attacks on chatgpt.

Li, J. and Hu, X. (2024). A better or worse communicator? comparing human and machine translation in source language shining through across registers. *Lingua*, 312:103834.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023b). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2023c). Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. (2021). Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Li, Z. and Specia, L. (2019). Improving neural machine translation robustness via data augmentation: Beyond back-translation. In Xu, W., Ritter, A., Baldwin, T., and Rahimi, A., editors, *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 328–336, Hong Kong, China. Association for Computational Linguistics.

Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., and Shi, G. (2025). A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1587–1606.

Li, Z., Zhu, H., Lu, Z., and Yin, M. (2023d). Synthetic data generation with large language models for text classification: Potential and limitations. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lin, H., Luo, Z., Gao, W., Ma, J., Wang, B., and Yang, R. (2024). Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

Lin, H., Luo, Z., Ma, J., and Chen, L. (2023). Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. *arXiv preprint arXiv:2312.05434*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Ling, W., Xiang, G., Dyer, C., Black, A. W., and Trancoso, I. (2013). Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186.

Liu, Y., Bai, G., Chenji, L., Li, S., Zhang, Z., Liu, R., and Guo, W. (2024). Eliminating the language bias for visual question answering with fine-grained causal intervention. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Livingstone, S., Cagiltay, K., and Ólafsson, K. (2015). Eu k ids o nline ii d ataset: A cross-national study of children's use of the internet and its associated opportunities and risks. *British Journal of Educational Technology*, 46(5):988–992.

Livingstone, S. and Helsper, E. (2007). Gradations in digital inclusion: Children, young people and the digital divide. *New media & society*, 9(4):671–696.

Livingstone, S., Kirwil, L., Ponte, C., and Staksrud, E. (2014). In their own words: What bothers children online? *European Journal of Communication*, 29(3):271–288.

Livingstone, S. and Stoilova, M. (2021). *The 4Cs: Classifying Online Risk to Children*. CO:RE Short Report Series on Key Topics. Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI), Hamburg.

Lohar, P., Popović, M., and Way, A. (2019). Building English-to-Serbian machine translation system for IMDb movie reviews. In Erjavec, T., Marcińczuk, M., Nakov, P.,

Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J., and Yangarber, R., editors, *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 105–113, Florence, Italy. Association for Computational Linguistics.

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lou, X., Li, Y., Xu, J., Shi, X., Chen, C., and Huang, K. (2025). Think in safety: Unveiling and mitigating safety alignment collapse in multimodal large reasoning model. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5167–5186, Suzhou, China. Association for Computational Linguistics.

Lowry, P. B., Zhang, J., Wang, C., and Siponen, M. (2016). Why do adults engage in cyberbullying on social media? an integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research*, 27(4):962–986.

Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., and Choo, K.-K. R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. *Concurrency and Computation: Practice and Experience*, 32(23):e5627.

Luo, Z., Xi, Y., Zhang, R., and Ma, J. (2022). A frustratingly simple approach for end-to-end image captioning. *arXiv preprint arXiv:2201.12723*.

Ma, W., Cui, Y., Si, C., Liu, T., Wang, S., and Hu, G. (2020). CharBERT: Character-aware pre-trained language model. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ma, Z., Yao, S., Wu, L., Gao, S., and Zhang, Y. (2022). Hateful memes detection based on multi-task learning. *Mathematics*, 10(23):4525.

Macaulay, P. J., Betts, L. R., Stiller, J., and Kellezi, B. (2022). Bystander responses to cyberbullying: The role of perceived severity, publicity, anonymity, type of cyberbullying, and victim response. *Computers in Human Behavior*, 131:107238.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Magu, R. and Luo, J. (2018). Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 93–100.

Maity, K. (2023). Personal communication.

Maity, K., Jha, P., Saha, S., and Bhattacharyya, P. (2022). A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1739–1749.

Maity, K., Saha, S., and Bhattacharyya, P. (2023). Emoji, sentiment and emotion aided cyberbullying detection in hinglish. *IEEE Transactions on Computational Social Systems*, 10(5):2411–2420.

Mali, M. K., Pawar, R. R., Shinde, S. A., Kale, S. D., Mulik, S. V., Jagtap, A. A., Tambewagh, P. A., and Rajput, P. U. (2025). Automatic detection of cyberbullying behaviour on social media using stacked bi-gru attention with bert model. *Expert Systems with Applications*, 262:125641.

Manski, C. F. and Molinari, F. (2008). Skip sequencing: A decision problem in questionnaire design. *The annals of applied statistics*, 2(1):264.

Mansour, O., Askenäs, L., and Ghazawneh, A. (2013). Social media and organizing: An empirical analysis of the role of wiki affordances in organizing practices. In *International Conference on Information System: Reshaping Society Through Information Systems Design, 15-18 December 2013, Milan, Italy.*

Mateo, R. M. (2014). A deeper look into metrics for translation quality assessment (tqa):: a case study. *Miscelánea: A Journal of English and American Studies*, (49):73–94.

McFarland, L. A. and Ployhart, R. E. (2015). Social media: A contextual framework to guide research and practice. *Journal of applied psychology*, 100(6):1653.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Meier, A. and Reinecke, L. (2021). Computer-mediated communication, social media, and mental health: A conceptual and empirical meta-review. *Communication Research*, 48(8):1182–1209.

Menesini, E., Nocentini, A., Palladino, B. E., Scheithauer, H., Schultze-Krumbholz, A., Frisén, A., Sofia, B., Luik, P., Naruskov, K., Ortega, R., et al. (2013). Definitions of cyberbullying. In *Cyberbullying through the new media*, pages 23–36. Psychology Press.

Menini, S., Moretti, G., Corazza, M., Cabrio, E., Tonelli, S., and Villata, S. (2019). A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the third workshop on abusive language online*, pages 105–110.

Meta Platforms, Inc. (2025). Regulation (eu) 2022/2065 digital services act: Transparency report for instagram. Technical report, Meta Platforms, Inc. Accessed: 2025-07-17.

Miaschi, A. and Dell'Orletta, F. (2020). Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In Gella, S., Welbl, J., Rei, M., Petroni, F., Lewis, P., Strubell, E., Seo, M., and Hajishirzi, H., editors, *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

Millwood-Hargrave, A. (2000). *Delete expletives?* Advertising Standards Authority London.

Milosevic, T., Van Royen, K., and Davis, B. (2022). Artificial intelligence to address cyberbullying, harassment and abuse: New directions in the midst of complexity. *International journal of bullying prevention*, 4(1):1–5.

Milosevic, T., Verma, K., Carter, M., Vigil, S., Laffan, D., Davis, B., and O'Higgins Norman, J. (2023). Effectiveness of artificial intelligence–based cyberbullying interventions from youth perspective. *Social Media+ Society*, 9(1):20563051221147325.

Mishra, P., Del Tredici, M., Yannakoudakis, H., and Shutova, E. (2019). Abusive Language Detection with Graph Convolutional Networks. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2145–2150, Minneapolis, Minnesota. Association for Computational Linguistics.

Mishra, S., Prasad, S., and Mishra, S. (2021). Exploring multi-task multi-lingual learning of transformer models for hate speech and offensive speech identification in social media. *SN Computer Science*, 2(2):72.

Monks, C. P., Smith, P. K., Naylor, P., Barter, C., Ireland, J. L., and Coyne, I. (2009). Bullying in different contexts: Commonalities, differences and the role of theory. *Aggression and violent behavior*, 14(2):146–156.

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., and Groh, G. (2022). Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pages 4593–4603.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Müller, D., Soto-Rey, I., and Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):210.

Muller, M. J., Haslwanter, J. H., and Dayton, T. (1997). Participatory practices in the software lifecycle. In *Handbook of human-computer interaction*, pages 255–297. Elsevier.

Muneer, A. and Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, 12(11):187.

Nadali, S., Murad, M. A. A., Sharef, N. M., Mustapha, A., and Shojaee, S. (2013). A review of cyberbullying detection: An overview. In *2013 13th International Conference on Intellient Systems Design and Applications*, pages 325–330. IEEE.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., and Scheidt, P. (2001). Bullying behaviors among us youth: Prevalence and association with psychosocial adjustment. *Jama*, 285(16):2094–2100.

National Advisory Council for Online Safety (2021). National advisory council for online safety: Report of a national survey of children, their parents and adults regarding online safety 2021. Technical report, Government of Ireland.

Navarro, J. N. and Jasinski, J. L. (2012). Going cyber: Using routine activities theory to predict cyberbullying experiences. *Sociological Spectrum*, 32(1):81–94.

Navarro, J. N. and Jasinski, J. L. (2015). Demographic and motivation differences among online sex offenders by type of offense: An exploration of routine activities theories. *Journal of child sexual abuse*, 24(7):753–771.

Nesi, J. and Prinstein, M. J. (2015). Using social media for social comparison and feedback-seeking: Gender and popularity moderate associations with depressive symptoms. *Journal of abnormal child psychology*, 43(8):1427–1438.

Nikhila, M. S., Bhalla, A., and Singh, P. (2020). Text imbalance handling and classification for cross-platform cyber-crime detection using deep learning. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Nirmal, A., Bhattacharjee, A., Sheth, P., and Liu, H. (2024). Towards interpretable hate speech detection using large language model-extracted rationales. In Chung, Y.-L., Talat, Z., Nozza, D., Plaza-del Arco, F. M., Röttger, P., Mostafazadeh Davani, A., and Calabrese, A., editors, *Proceedings of the 8th Workshop on Online Abuse and Harms*

(WOAH 2024), pages 223–233, Mexico City, Mexico. Association for Computational Linguistics.

Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., and Giordano, S. (2025). Toxic bias: Perspective api misreads german as more toxic. In *Proceedings of the international AAAI conference on web and social media*, volume 19, pages 1346–1357.

Obukhov, A. and Krasnyanskiy, M. (2020). Quality assessment method for gan based on modified metrics inception score and fréchet inception distance. In *Proceedings of the Computational Methods in Systems and Software*, pages 102–114. Springer.

Ocampo, N. B., Sviridova, E., Cabrio, E., and Villata, S. (2023). An in-depth analysis of implicit and subtle hate speech messages. In *EACL 2023-17th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2023, pages 1997–2013. Association for Computational Linguistics.

O'Higgins Norman, J., Berger, C., Cross, D., Payne, E., Søndergaard, D. M., Ttofi, M., Zych, I., and Gini, G. (2023). SCHOOL BULLYING, AN INCLUSIVE DEFINITION. Report for the unesco and world anti-bullying working group, UNESCO Chair on Bullying and Cyberbullying at Dublin City University and World Anti-Bullying Forum, Dublin, Ireland. ISBN: 978-1-911669-66-1.

Olea, C., Tucker, H., Phelan, J., Pattison, C., Zhang, S., Lieb, M., and White, J. (2024). Evaluating persona prompting for question answering tasks. In *Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing*, Sydney, Australia.

Ollagnier, A. (2024). CyberAgressionAdo-v2: Leveraging pragmatic-level information to decipher online hate in French multiparty chats. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4287–4298, Torino, Italia. ELRA and ICCL.

Ollagnier, A., Cabrio, E., and Villata, S. (2023). Harnessing bullying traces to enhance bullying participant role identification in multi-party chats. In *FLAIRS 2023-36th International conference of the Florida artificial intelligence research sociéty*.

Ollagnier, A., Cabrio, E., Villata, S., and Blaya, C. (2022). CyberAgressionAdo-v1: a dataset of annotated online aggressions in French collected through a role-playing game. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 867–875, Marseille, France. European Language Resources Association.

Olweus, D. (1994). Bullying at school: basic facts and effects of a school based intervention program. *Journal of child psychology and psychiatry*, 35(7):1171–1190.

Olweus, D. (2013). School bullying: Development and some important challenges. *Annual review of clinical psychology*, 9:751–770.

Omrani, A., Salkhordeh Ziabari, A., Golazizian, P., Sorensen, J., and Dehghani, M. (2024). Towards a unified framework for adaptable problematic content detection via continual learning. In Chung, Y.-L., Talat, Z., Nozza, D., Plaza-del Arco, F. M., Röttger, P., Mostafazadeh Davani, A., and Calabrese, A., editors, *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 68–109, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI (2025). Moderation. `https://platform.openai.com/docs/guides/moderation`. Accessed: 20 September 2025.

Osgood, D. W., Wilson, J. K., O'malley, P. M., Bachman, J. G., and Johnston, L. D. (1996). Routine activities and individual deviant behavior. *American sociological review*, pages 635–655.

Ouyang, L., Wu, J., Jiang, X., Fiedorowicz, D., Lee, J., He, L., Manion, C., Li, H., Sundar, V., Barman, K., and et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

O'Higgins Norman, J. (2020). Tackling bullying from the inside out: Shifting paradigms in bullying research and interventions: Unesco chair on tackling bullying in schools and cyberspace, inaugural lecture delivered on 7th october 2019 at dublin city university. *International journal of bullying prevention*, 2(3):161–169.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors,

Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Park, C., Tae, Y., Kim, T., Yang, S., Khan, M. A., Park, L., and Choo, J. (2021). Unsupervised neural machine translation for low-resource domains via meta-learning. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2888–2901, Online. Association for Computational Linguistics.

Parker, R. I., Vannest, K. J., and Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior modification*, 35(4):303–322.

Paruchuri, V. L. and Rajesh, P. (2022). Cybernet: a hybrid deep cnn with n-gram feature selection for cyberbullying detection in online social networks. *Evolutionary Intelligence*, pages 1–15.

Patchin, J. W. and Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, 4(2):148–169.

Paul, S. and Saha, S. (2022). Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification. *Multimedia Systems*, 28(6):1897–1904.

Paul Benjamin Lowry, G. D. M. and Chatterjee, S. (2017). Using it design to prevent cyberbullying. *Journal of Management Information Systems*, 34(3):863–901.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang,

M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., Showk, S. E., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations.

Perez, F. and Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models.

Pericherla, S. and Ilavarasan, E. (2021). Transformer network-based word embeddings approach for autonomous cyberbullying detection. *International Journal of Intelligent Unmanned Systems*.

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., and Gurevych, I. (2021). AdapterFusion: Non-destructive task composition for transfer learning. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Phanomtip, A., Sueb-in, T., and Vittayakorn, S. (2021). Cyberbullying detection on tweets. In *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 295–298. IEEE.

Philipo, A. G., Sarwatt, D. S., Ding, J., Daneshmand, M., and Ning, H. (2024). Assessing text classification methods for cyberbullying detection on social media platforms. *arXiv preprint arXiv:2412.19928*.

Philips, L. (1990). Hanging on the metaphone. *Computer Language Magazine*, pages 39–44.

Phillips, L., Shaffer, K., Arendt, D., Hodas, N., and Volkova, S. (2017). Intrinsic and extrinsic evaluation of spatiotemporal text representations in Twitter streams. In Blunsom, P., Bordes, A., Cho, K., Cohen, S., Dyer, C., Grefenstette, E., Hermann, K. M.,

Rimell, L., Weston, J., and Yih, S., editors, *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 201–210, Vancouver, Canada. Association for Computational Linguistics.

Pickard, T., Loakman, T., and Pandya, M. (2023). shefnlp at SemEval-2023 task 10: Compute-efficient category adapters. In Ojha, A. K., Doğruöz, A. S., Da San Martino, G., Tayyar Madabushi, H., Kumar, R., and Sartori, E., editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1069–1075, Toronto, Canada. Association for Computational Linguistics.

Plaza-del arco, F. M., Nozza, D., and Hovy, D. (2023). Respectful or toxic? using zero-shot learning with language models to detect hate speech. In Chung, Y.-l., R{\"ottger}, P., Nozza, D., Talat, Z., and Mostafazadeh Davani, A., editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., and Pecina, P., editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Popović, M., Poncelas, A., Brkic, M., and Way, A. (2021). On machine translation of user reviews. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1109–1118, Held Online. INCOMA Ltd.

Pradhan, A., Yatam, V. M., and Bera, P. (2020). Self-attention for cyberbullying detection. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pages 1–6. IEEE.

Pramanick, S., Dimitrov, D., Mukherjee, R., Sharma, S., Akhtar, M. S., Nakov, P., and Chakraborty, T. (2021a). Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

Pramanick, S., Sharma, S., Dimitrov, D., Akhtar, M. S., Nakov, P., and Chakraborty, T. (2021b). Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Prechelt, L. (2002). Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., and Napolitano, A. (2015). Using random undersampling to alleviate class imbalance on tweet sentiment data. In *2015 IEEE International Conference on Information Reuse and Integration*, pages 197–202.

Pruthi, D., Dhingra, B., and Lipton, Z. C. (2019). Combating adversarial misspellings with robust word recognition. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Qian, R., Ross, C., Fernandes, J., Smith, E., Kiela, D., and Williams, A. (2022). Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

Qiu, J., Moh, M., and Moh, T.-S. (2022). Multi-modal detection of cyberbullying on twitter. In *Proceedings of the 2022 ACM Southeast Conference*, pages 9–16.

Quayle, E. (2020). Prevention, disruption and deterrence of online child sexual exploitation and abuse. In *Era Forum*, volume 21, pages 429–447. Springer.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafailov, E., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rafiq, R. I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., and Mattson, S. A. (2015). Careful what you share in six seconds: Detecting cyberbullying instances in vine. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 617–622.

Raj, C., Agarwal, A., Bharathy, G., Narayan, B., and Prasad, M. (2021). Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques. *Electronics*, 10(22):2810.

Rathnayake, G., Atapattu, T., Herath, M., Zhang, G., and Falkner, K. (2020). Enhancing the identification of cyberbullying through participant roles. In Akiwowo, S., Vidgen, B., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 89–94, Online. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., C. de Souza, J. G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., and Martins, A. F. T. (2022). CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Reiter, E., Gatt, A., Portet, F., and van der Meulen, M. (2008). The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In White, M., Nakatsu, C., and McDonald, D., editors, *Proceedings of the Fifth International Natural Language Generation Conference*, pages 147–156, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Riabi, A. (2025). *Small is Beautiful: addressing resource scarcity, language variation, and transfer challenges for automatic detection of Harmful language*. PhD thesis, Sorbonne Université.

Riahi Samani, A., Wang, T., Li, K., and Chen, F. (2025). Large language models with reinforcement learning from human feedback approach for enhancing explainable sexism detection. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6230–6243, Abu Dhabi, UAE. Association for Computational Linguistics.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Rodríguez-de Dios, I., van Oosten, J. M., and Igartua, J.-J. (2018). A study of the relationship between parental mediation and adolescents' digital skills, online risks and online opportunities. *Computers in Human Behavior*, 82:186–198.

Ronzhyn, A., Cardenal, A. S., and Rubio, A. B. (2023). Defining affordances in social media research: A literature review. *New Media & Society*, 25(11):3165–3188.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Simão, A. V., and Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Röttger, P., Vidgen, B., Hovy, D., and Pierrehumbert, J. (2022). Two contrasting data annotation paradigms for subjective NLP tasks. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Rubino, R., Foster, J., Samad Zadeh Kaljahi, R., Roturier, J., and Hollowood, F. (2013). Estimating the quality of translated user-generated content. In Mitkov, R. and Park, J. C., editors, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1167–1173, Nagoya, Japan. Asian Federation of Natural Language Processing.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). Transfer learning in natural language processing. In Sarkar, A. and Strube, M., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruparelia, N. B. (2010). Software development lifecycle models. *ACM SIGSOFT Software Engineering Notes*, 35(3):8–13.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Saikia, P., Gundale, K., Jain, A., Jadeja, D., Patel, H., and Roy, M. (2022). Modelling social context for fake news detection: a graph neural network based approach. In *2022 international joint conference on neural networks (IJCNN)*, pages 01–08. IEEE.

Salawu, S., He, Y., and Lumsden, J. (2017). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11(1):3–24.

Salawu, S., Lumsden, J., and He, Y. (2021). A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In Mostafazadeh Davani, A., Kiela, D., Lambert, M., Vidgen, B., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 146–156, Online. Association for Computational Linguistics.

Salmivalli, C. (2014). Participant roles in bullying: How can peer bystanders be utilized in interventions? *Theory Into Practice*, 53(4):286–292.

Sanchez, H. and Kumar, S. (2011). Twitter bullying detection. *ser. NSDI*, 12(2011):15.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 5884–5906.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Sen, R. and Borle, S. (2015). Estimating the contextual risk of data breach: An empirical approach. *Journal of Management Information Systems*, 32(2):314–341.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Serra, S. M. and Venter, H. S. (2011). Mobile cyber-bullying: A proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness. In *2011 Information Security for South Africa*, pages 1–5. IEEE.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Shannag, F., Hammo, B. H., and Faris, H. (2022). The design, construction and evaluation of annotated arabic cyberbullying corpus. *Education and Information Technologies*, 27(8):10977–11023.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Sharma, D., Saha, B., and Sarkar, U. K. (2016). Bridging the distance: The agencement of complex affordances on social media platforms. In *ICIS 2016 Proceedings*. AIS Electronic Library (AISeL).

Sharma, S., Alam, F., Akhtar, M. S., Dimitrov, D., Da San Martino, G., Firooz, H., Halevy, A., Silvestri, F., Nakov, P., and Chakraborty, T. (2022). Detecting and understanding harmful memes: A survey. In Raedt, L. D., editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5597–5606. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Shaw, D. J., Kaye, L. K., Ngombe, N., Kessler, K., and Pennington, C. R. (2022). It's not what you do, it's the way that you do it: An experimental task delineates among passive, reactive and interactive styles of behaviour on social networking sites. *Plos one*, 17(12):e0276765.

Shifman, L. (2013). *Memes in digital culture*. MIT press.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. (2022). Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Singh, V. K., Radford, M. L., Huang, Q., and Furrer, S. (2017). "they basically like destroyed the school one day": On newer app features and cyberbullying in schools. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1210–1216, New York, NY, USA. Association for Computing Machinery.

Skantze, G. (2017). Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In Jokinen, K., Stede, M., DeVault, D., and Louis, A., editors, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, Saarbrücken, Germany. Association for Computational Linguistics.

Slonje, R., Smith, P. K., and Frisén, A. (2013). The nature of cyberbullying, and strategies for prevention. *Computers in human behavior*, 29(1):26–32.

Smith, P. K. and Steffgen, G. (2013). *Cyberbullying through the new media: Findings from an international network*. Psychology Press.

Smith, P. K., Steffgen, G., and Sittichai, R. R. (2013). The nature of cyberbullying, and an international network. In *Cyberbullying through the new media*, pages 3–19. Psychology Press.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Song, D., Lai, S., Wang, M., Chen, S., Sun, L., and Wang, B. (2025). Both text and images leaked! a systematic analysis of data contamination in multimodal LLM. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10527–10542, Suzhou, China. Association for Computational Linguistics.

Song, J. and Oh, I. (2018). Factors influencing bystanders' behavioral reactions in cyberbullying situations. *Computers in Human Behavior*, 78:273–282.

Sprugnoli, R., Menini, S., Tonelli, S., Oncini, F., and Piras, E. (2018). Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Stanford Internet Observatory (2020–). Journal of online trust and safety. Online journal. ISSN: 2770-3142.

Steer, O. L., Betts, L. R., Baguley, T., and Binder, J. F. (2020). "i feel like everyone does it"-adolescents' perceptions and awareness of the association between humour, banter, and cyberbullying. *Computers in Human Behavior*, 108:106297.

Sticca, F. and Perren, S. (2013). Is cyberbullying worse than traditional bullying? ex-

amining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of youth and adolescence*, 42:739–750.

Stoilova, M., Livingstone, S., Khazbak, R., et al. (2021). Investigating risks and opportunities for children in a digital world: A rapid review of the evidence on children's internet use and outcomes. *UNICEF Office Of Research – INNOCENTI*.

Strong, D. M., Volkoff, O., Johnson, S. A., Pelletier, L. R., Tulu, B., Bar-On, I., Trudel, J., and Garber, L. (2014). A theory of organization-ehr affordance actualization. *Journal of the association for information systems*, 15(2):2.

Sun, J., Tian, Y., Zhou, W., Xu, N., Hu, Q., Gupta, R., Wieting, J., Peng, N., and Ma, X. (2023). Evaluating large language models on controlled generation tasks. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Sun, X. and Lu, W. (2020). Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428, Online. Association for Computational Linguistics.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., and Buitelaar, P. (2020). Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227.

Tablan, V., Roberts, I., Cunningham, H., and Bontcheva, K. (2013). Gatecloud.net: a platform for large-scale, open-source text processing on the cloud. *Philosophical*

Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1983):20120071.

Talpur, B. A. and O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. *PloS one*, 15(10):e0240924.

Tan, H. and Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Tay, Y., Tran, V. Q., Ruder, S., Gupta, J., Chung, H. W., Bahri, D., Qin, Z., Baumgartner, S., Yu, C., and Metzler, D. (2021). Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*.

Team, Q. et al. (2024). Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(8).

Teng, T. H., Varathan, K. D., and Crestani, F. (2024). A comprehensive review of cyberbullying-related content classification in online social media. *Expert Systems with Applications*, 244:122644.

Thapliyal, A. V., Pont Tuset, J., Chen, X., and Soricut, R. (2022). Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas, H. J., Connor, J. P., and Scott, J. G. (2015). Integrating traditional bullying and cyberbullying: challenges of definition and measurement in adolescents–a review. *Educational psychology review*, 27:135–152.

Thorn and Benenson Strategy Group (2021). Responding to online threats: Minors' perspectives on disclosing, reporting, and blocking. Technical report, Thorn. Research conducted by Thorn in partnership with Benenson Strategy Group.

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior*, 26(3):277–287.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Treem, J. W. and Leonardi, P. M. (2013). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association*, 36(1):143–189.

Trengove, M., Kazim, E., Almeida, D., Hilliard, A., Zannone, S., and Lomas, E. (2022). A critical review of the online safety bill. *Patterns*, 3(8):100544.

Tripathy, J. K., Chakkaravarthy, S. S., Satapathy, S. C., Sahoo, M., and Vaidehi, V. (2022). Albert-based fine-tuning model for cyberbullying analysis. *Multimedia Systems*, 28(6):1941–1949.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. (2021). Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Urbaniak, R., Ptaszyński, M., Tempska, P., Leliwa, G., Brochocki, M., and Wroczyński, M. (2022). Personal attacks decrease user activity in social networking platforms. *Computers in Human Behavior*, 126:106972.

Urman, A. and Makhortykh, M. (2023). How transparent are transparency reports? comparative analysis of transparency reporting across online platforms. *Telecommunications policy*, 47(3):102477.

Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., and Tesconi, M. (2017). Cross-media learning for image sentiment analysis in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*.

Vaillancourt, T., McDougall, P., Hymel, S., Krygsman, A., Miller, J., Stiver, K., and Davis, C. (2008). Bullying: Are researchers and children/youth talking about the same thing? *International Journal of Behavioral Development*, 32(6):486–495.

Valenzuela-García, N., Maldonado-Guzmán, D. J., García-Pérez, A., and Del-Real, C. (2023). Too lucky to be a victim? an exploratory study of online harassment and hate messages faced by social media influencers. *European Journal on Criminal Policy and Research*, pages 1–25.

van Aken, B., Risch, J., Krestel, R., and Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In Fišer, D., Huang, R., Prabhakaran, V., Voigt, R., Waseem, Z., and Wernimont, J., editors, *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Van Bruwaene, D., Huang, Q., and Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, 54(4):851–874.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.

Van Ouytsel, J., Ponnet, K., and Walrave, M. (2018). Cyber dating abuse victimization

among secondary school students from a lifestyle-routine activities theory perspective. *Journal of interpersonal violence*, 33(17):2767–2776.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London, 2 edition.

Van Royen, K., Poels, K., Vandebosch, H., and Adam, P. (2017). "thinking before posting?" reducing cyber harassment on social networking sites through a reflective message. *Computers in human behavior*, 66:345–352.

Vandebosch, H. and Van Cleemput, K. (2009). Cyberbullying among youngsters: Profiles of bullies and victims. *New media & society*, 11(8):1349–1371.

Vashishth, S., Upadhyay, S., Tomar, G. S., and Faruqui, M. (2019). Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vazsonyi, A. T., Javakhishvili, M., and Ksinan, A. J. (2018). Routine activities and adolescent deviance across 28 cultures. *Journal of Criminal Justice*, 57:56–66.

Vazsonyi, A. T., Pickering, L. E., Belliston, L. M., Hessing, D., and Junger, M. (2002). Routine activities and deviant behaviors: American, dutch, hungarian, and swiss youth. *Journal of Quantitative Criminology*, 18:397–422.

Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., and Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? a critical review. *Social Issues and Policy Review*, 11(1):274–302.

Verma, K., Adebayo, K., Wagner, J., and Davis, B. (2023a). Dcu at semeval-2023 task 10: A comparative analysis of encoder-only and decoder-only language models with insights into interpretability. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1736–1750.

Verma, K., Adebayo, K. J., Wagner, J., Reynolds, M., Umbach, R., Milosevic, T., and Davis, B. (2024). Beyond binary: Towards embracing complexities in cyberbullying

detection and intervention - a position paper. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2264–2284, Torino, Italia. ELRA and ICCL.

Verma, K., Milosevic, T., Cortis, K., and Davis, B. (2022). Benchmarking language models for cyberbullying identification and classification from social-media texts. In Adebayo, K., Nanda, R., Verma, K., and Davis, B., editors, *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 26–31, Marseille, France. European Language Resources Association.

Verma, K., Popović, M., Poulis, A., Cherkasova, Y., Ó hÓbáin, C., Mazzone, A., Milosevic, T., and Davis, B. (2023b). Leveraging machine translation for cross-lingual fine-grained cyberbullying classification amongst pre-adolescents. *Natural language engineering*, 29(6):1458–1480.

Vicente, I. S., Alegria, I., Aranberri, N., Espana-Bonet, C., Gamallo, P., Gonçalo Oliveira, H., Martinez Garcia, E., Toral, A., and Zubiaga, A. (2016). Tweetmt: A parallel microblog corpus. *LREC 2016 Proceedings*, pages 2936–2941.

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Vielzeuf, V., Lechervy, A., Pateux, S., and Jurie, F. (2018). Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0.

Vig, J. (2019a). A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Vig, J. (2019b). A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Vig, J. and Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Wallace, E., Tuyls, J., Wang, J., Subramanian, S., Gardner, M., and Singh, S. (2019). AllenNLP interpret: A framework for explaining predictions of NLP models. In Padó, S. and Huang, R., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C.-C. J. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8:e19.

Wang, J., Fu, K., and Lu, C.-T. (2020). Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1699–1708. IEEE.

Wang, J., Yao, Z., Yang, Z., Zhou, H., Li, R., Wang, X., Xu, Y., and Yu, H. (2024a). NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201, Bangkok, Thailand. Association for Computational Linguistics.

Wang, L., Zhang, X., Su, H., and Zhu, J. (2024b). A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383.

Wang, S., Shibghatullah, A. S., Iqbal, T. J., and Keoy, K. H. (2024c). A review of multimodal-based emotion recognition techniques for cyberbullying detection in online social media platforms. *Neural Computing and Applications*, 36(35):21923–21956.

Wang, W., Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Zhu, J., Zhu, X., Lu, L., Qiao,

Y., et al. (2024d). Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Weber, F., Knapp, G., Ickstadt, K., Kundt, G., and Glass, Ä. (2020). Zero-cell corrections in random-effects meta-analyses. *Research Synthesis Methods*, 11(6):913–919.

Wei, A., Haghtalab, N., and Steinhardt, J. (2023). Jailbroken: How does llm safety training fail?

Wei, B., Li, J., Gupta, A., Umair, H., Vovor, A., and Durzynski, N. (2021). Offensive language and hate speech detection with deep learning and transfer learning. *arXiv preprint arXiv:2108.03305*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Willard, N. E. (2007). *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research press.

Willison, R. and Backhouse, J. (2006). Opportunities for computer crime: considering systems risk from a criminological perspective. *European journal of information systems*, 15(4):403–414.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Workshop, B., :, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., Toni, F. D., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Werra, L. V., Weber, L., Phan, L., allal, L. B., Tanguy, L., Dey, M., Muñoz, M. R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M. T.-J., Vu, M. C., Jauhar, M. A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R. L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S. H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T. T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D. E., Salesky, E., Mielke, S. J., Lee, W. Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J. A., Rozen, J., Gao, L., Sutawika, L., Bari, M. S., Al-shaibani, M. S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S. H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H. W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P. F., Lacroix, R.,

Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G. I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J. Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C. M., McDuff, D., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D. A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J. B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A. R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D. L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrimann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H. U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J. D., Sivaraman, K. R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M. H., Takeuchi, M., Pàmies, M., Castillo, M. A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., Wolf, M. D., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N. M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S. S., Mishra, S., Kiblawi, S., Ott, S., Sang-aroonsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y. S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., and Wolf, T. (2023). Bloom: A 176b-parameter open-access multilingual language model.

Wu, T., Ribeiro, M. T., Heer, J., and Weld, D. (2021). Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Wu, T., Ribeiro, M. T., Heer, J., and Weld, D. S. (2019). Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Xiao, B. S., Chan, T. K., Cheung, C., and Wong, R. Y. M. (2016). An investigation into cyberbullying perpetration: a routine activity perspective. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS), Chiayi, Taiwan*, volume 370.

Xie, Q., Ma, X., Dai, Z., and Hovy, E. (2017a). An interpretable knowledge transfer model for knowledge base completion. *arXiv preprint arXiv:1704.05908*.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017b). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Xingyi, G. and Adnan, H. M. (2024). Potential cyberbullying detection in social media platforms based on a multi-task learning framework. *International journal of data and network science*, 8(1):25–34.

Xu, G., Jin, P., Li, H., Song, Y., Sun, L., and Yuan, L. (2024). Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666.

Xu, X., Wu, C., Rosenman, S., Lal, V., Che, W., and Duan, N. (2023). Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10637–10647.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2022). Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Yadav, J., Kumar, D., and Chauhan, D. (2020). Cyberbullying detection using pre-trained bert model. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1096–1100. IEEE.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yi, P. and Zubiaga, A. (2022). Cyberbullying detection across social media platforms via platform-aware adversarial encoding. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1430–1434.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2(0):1–7.

Yip, J., Clegg, T., Bonsignore, E., Gelderblom, H., Rhodes, E., and Druin, A. (2013). Brownies or bags-of-stuff? domain expertise in cooperative inquiry with children. In *Proceedings of the 12th International conference on interaction design and children*, pages 201–210.

Yoo, K. M., Park, D., Kang, J., Lee, S.-W., and Park, W. (2021). GPT3Mix: Leveraging large-scale language models for text augmentation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yun-tao, Z., Ling, G., and Yong-cheng, W. (2005). An improved tf-idf approach for text classification. *Journal of Zhejiang University-Science A*, 6(1):49–55.

Zabyelina, Y. G. (2017). Can criminals create opportunities for crime? malvertising and illegal online medicine trade. *Global Crime*, 18(1):31–48.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., and Shutova, E., editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Zerva, C., Blain, F., Rei, R., Lertvittayakumjorn, P., De Souza, J. G., Eger, S., Kanojia, D., Alves, D., Orasan, C., Fomicheva, M., et al. (2022). Findings of the wmt 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99.

Zhang, H. (2004). The optimality of naive bayes. *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference, Miami, Florida, 2004*, 4(1):55–62.

Zhang, H., Song, H., Li, S., Zhou, M., and Song, D. (2022a). A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. (2022b). Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2022c). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, W., Liu, G., Li, Z., and Zhu, F. (2020a). Hateful memes detection via complementary visual and linguistic networks. *arXiv preprint arXiv:2012.04977*.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., and Li, C. (2020b). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3).

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Zhang, Y., Chen, L., Zheng, G., Gao, Y., Zheng, R., Fu, J., Yin, Z., Jin, S., Qiao, Y., Huang, X., et al. (2025). Spa-vl: A comprehensive safety preference alignment dataset for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19867–19878.

Zhang, Y., Song, K., Sun, Y., Tan, S., and Udell, M. (2019). " why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*.

Zhao, B., Zhang, A., Watson, B., Kearney, G., and Dale, I. (2023). A review of vision-language models and their performance on the hateful memes challenge. *arXiv preprint arXiv:2305.06159*.

Zhao, Z., Gao, M., Luo, F., Zhang, Y., and Xiong, Q. (2020). Lshwe: improving similarity-based word embedding with locality sensitive hashing for cyberbullying detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zheng, M., Pei, J., Logeswaran, L., Lee, M., and Jurgens, D. (2023). When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. *arXiv preprint arXiv:2311.10054*.

Zhong, W., Guo, L., Gao, Q., Ye, H., and Wang, Y. (2024). Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., and Xu, B. (2016). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In Matsumoto, Y. and Prasad, R., editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3485–3495, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhou, Y. and Farzan, R. (2021). Designing to stop live streaming cyberbullying: A case study of twitch live streaming platform. In *Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech*, C&T '21, page 138–150, New York, NY, USA. Association for Computing Machinery.

Zhu, Y., Kiros, R., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. *SIGIR*.

Zhu, Z., Xue, Y., Chen, X., Zhou, D., Tang, J., Schuurmans, D., and Dai, H. (2023). Large language models can learn rules. *arXiv preprint arXiv:2310.07064*.

Ziems, C., Vigfusson, Y., and Morstatter, F. (2020). Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification. In

Proceedings of the International AAAI Conference on Web and Social Media, volume 14, pages 808–819.

Zito, W. S. (2022). *TikTok Cyberbully Responses: Communicating the Narrative*. Doctoral dissertation, Liberty University.

Zou, W., Yang, Q., DiFranzo, D., Chen, M., Hui, W., and Bazarova, N. N. (2024). Social media co-pilot: Designing a chatbot with teens and educators to combat cyberbullying. *International Journal of Child-Computer Interaction*, page 100680.

Appendix A

A.1 Young Research Assistant Programme

Training and Methodology: Students began by learning foundational skills on data handling using Google Sheets. This involved learning to create filters, pivot tables, and visualisation. They create a survey based google sheet using drop-down menus for response choices for 15 cyberbullying vignettes drawn from prior research by (Ashktorab and Vitak, 2016; Camelford and Ebrahim, 2016; Campbell and Xu, 2022; Huang et al., 2020; Ollagnier et al., 2022). They annotated these vignettes along three dimensions: (a) perceived realism (Agha et al., 2024); (b) seriousness of harm (Huang et al., 2020); and (c) report-worthiness (Thorn and Benenson Strategy Group, 2021). The vignettes reflected diverse forms of cyberbullying such as threats, flaming, impersonation, outing, and exclusion. Students were then invited to adapt the vignettes they found most real, serious, and report-worthy by embedding them within specific platform contexts (e.g., TikTok comments, Snapchat group chats, Discord threads), thus making the scenarios more reflective of contemporary digital environments. A unique feature of the programme was the development of bystander response sequences. Each student authored multiple responses to the adapted vignettes from various roles like defenders, enablers, or neutral observer, with focus on escalation and de-escalation strategies. These contributions took the form of both comment-thread and group-chat interactions, simulating the different temporal and dialogic affordances of popular social media platforms. As a result, the evolving dataset captures not just the static content of cyberbullying interactions, but also the dynamic social discourse that surrounds them. Later-stage contributions shifted toward re-framing static messages into naturalistic conversational exchanges, aligning the

dataset more closely with youth digital communication styles e.g., use of slang, emojis, and ephemeral features like disappearing messages. Students also introduced platform-specific expressions and idioms to improve realism, with these contributions currently undergoing expert thematic annotation to inform future scenario-specific benchmark tasks.

Technical Enrichment and Applied Learning: The programme also supported the growth of participant's applied digital literacy through a mix-of-ideas cooperative inquiry approach, fostering an inclusive and exploratory learning environment. This pedagogical model encouraged collaborative discovery, reflection, and learning, allowing participants to build understanding through shared problem-solving and open-ended experimentation. Students engaged in analysing response data across 5-point Likert scales using spreadsheet tools and visualised trends using a range of chart types. Beyond spreadsheets, they developed foundational computational skills by working within the Linux operating system and learning Python programming. They were introduced to basic data types such as integers (`int`), strings (`str`), and floating-point numbers (`float`), and practised key operators including modulus (for computing remainders), division, and multiplication. Problem-solving tasks were designed to be accessible yet intellectually stimulating. Students tackled mathematical world problems such as determining whether a number is odd or even without using `if` or `else` statements, promoting creativity and logical reasoning. They also solved more sophisticated tasks, such as returning the number itself if it was odd or returning zero if it was even, without relying on standard conditional statements, thereby reinforcing logical reasoning and computational problem-solving. Importantly, ***all tasks were optional***. If a participant chose not to engage in a particular activity, they were offered additional Python training tailored to their current skill level. This ensured a flexible, learner-centred experience where engagement was driven by curiosity rather than obligation. For advanced learners, the programme introduced data analysis workflows, such as reading CSV files into dataframes using Python libraries like `pandas`. These learners also explored API interactions in Python, applying real-world content moderation tools such as Google's Perspective API and OpenAI's Moderation API via Google Colaboratory notebooks. These exercises helped them understand how content moderation systems are built, tested, and critically evaluated. In parallel, students were introduced to prompt engineering, an emerging digital literacy skill. They learned

how to communicate with `Llama3-8B`[117] using HuggingFace's Inference API by applying zero-shot and few-shot prompting strategies to moderate content. This involved crafting effective prompts to instruct the model on how to detect harmful or inappropriate content without relying on hardcoded rules. By practising this, students not only deepened their understanding of large language models and their capabilities, but also began to question and critique the role of AI in shaping online interactions. Altogether, the programme fostered a supportive, exploratory, and highly adaptable learning environment that blended technical instruction with ethical inquiry, empowering students to become more confident and critical users of modern digital technologies.

Outcomes and Impact: This initiative positions youth not only as test subjects or informants but as co-designers and knowledge producers in the research ecosystem. Their direct input has led to more contextually rich, affectively accurate, and platform-specific representations of cyberbullying scenarios. Moreover, the co-design process has generated valuable training and evaluation data that reflects lived youth experiences. Through its cascading structure, robust ethical safeguards, and participatory ethos, the Young Research Assistant Programme offers a replicable model for integrating adolescents into socio-technical research. It contributes to both methodological innovation and the democratisation of research processes involving youth and technology.

A.2 Cyberbullying Exposure Sensitivity Analysis

Sensitivity analysis: As discussed in §3.3.1, to evaluate the robustness of observed association between engagement style and witnessing cyberbullying discussed above, a sensitivity analysis was conducted using two alternative coding approaches: *Method B (Lenient Threshold)*, where "*Never*" responses were coded as `No (0)` and all others as `Yes (1)` and *Method C (Strict Threshold)*, where only "*Often or Very Often*" responses were coded as `Yes (1)` and all others as `No (0)`. The results for both methods across all features are presented in Table A.2. The sensitivity analysis successfully confirms the robustness of the core finding: a significant association exists between all *three engagement styles* and the *likelihood of witnessing cyberbullying*. The results across both the lenient

[117] `https://huggingface.co/meta-llama/Meta-Llama-3-8B`

and strict reclassification methods consistently show that engagement in these styles is associated with increased odds of witnessing harm. While the magnitude of the associations changes with each method, the overall pattern of a positive relationship remains stable. The p-values for nearly all comparisons are well below the 0.05 significance level, indicating that these associations are not due to chance. A key observation is the shift in the most influential engagement style across the methods. Under the lenient threshold (Method B), where "*at-risk*" behaviour is broadly defined[118], "*viewing*" shows an exceptionally strong association (Corrected Odds Ratio (COR) = 61.01). This suggests that even minimal engagement in viewing-related behaviours is linked to a dramatically higher probability of witnessing cyberbullying. In contrast, under the strict threshold (Method C) (See Table A.4), which focuses on higher-scales for engagement[119], "*reacting*" emerges with the highest odds ratio (COR = 5.22, $p<0.05$). This highlights that frequent reactive behaviour is a particularly strong indicator of exposure to harm, even when controlling for other engagement types. This sensitivity analysis validates the core conclusion that all three engagement styles are significant risk factors for witnessing cyberbullying. The varying odds ratios across the methods offer valuable insight into how different definitions of "*at-risk*" behaviour can influence the perceived strength of the association, demonstrating the nuanced nature of digital engagement and harm exposure. Upon further examination at each feature level, as depicted in Table A.1 under *Method A (Sequential Survey Logic)*, significant associations were consistently observed, particularly for viewing, reacting, and interacting with comment sections, group chats, and video comments. This method yielded the strongest and most consistent effect sizes, suggesting that moderate to frequent engagement behaviours are strongly linked with higher odds of witnessing cyberbullying. Conversely, *Method B (Lenient Threshold)* which classified any engagement other than "Never" as "Yes", showed fewer significant results (See Table A.3). It only highlights "*disappearing content*" features as especially associated with witnessing cyberbullying across all engagement styles, indicating that even occasional engagement with certain content types may increase exposure risk. Method C (Strict Threshold), which considered only the highest frequency responses as "Yes", shows that engaging as "*viewers*" of *video-comments* has statistical significant association to witnessing cyberbullying

[118] `Hardly ever/Sometimes/Often/Very Often` responses are coded as `Yes (1)`
[119] `Often/Very Often` responses are coded as `Yes (1)`

(COR=20.41, $p<0.05$). However, all other engagement styles across all other features under the strict method are not statistically significant. These findings suggest that while the relationship between feature engagement and cyberbullying exposure is generally stable, the operationalism of "*at-risk*" engagement influences the magnitude and significance of observed associations, underscoring the importance of threshold selection in behavioural research.

Platform-type	**Feature**	**Engagement Style**	**p value**	**Odds Ratio**	**Corrected Odds Ratio**	**# Subsamples**
Social networking and content sharing	stories	view	0.1403	inf	6.8087	104
		react	0.0782	7.7143	5.9357	104
		interact	1.0000	0.0000	0.9577	104
	reels	view	0.2889	inf	4.4545	101
		react	0.2717	1.7290	1.6836	101
		interact	0.4182	1.6000	1.5988	101
	comments	**view**	**0.0000**	**32.2500**	**25.9067**	**100**
		react	**0.0000**	**11.1562**	**10.0431**	**100**
		interact	**0.0000**	**8.7500**	**8.0575**	**100**
	disappearing content	**view**	**0.0024**	**10.3158**	**8.6308**	**78**
		react	**0.0090**	**3.7188**	**3.5859**	**78**
		interact	0.0976	2.2917	2.2454	78
Messaging Services	**group chats**	**view**	**0.0165**	inf	**23.6667**	**48**
		react	0.3233	3.2000	3.0952	48
		interact	0.1147	4.9500	4.4667	48
Video-streaming and sharing	video content	view	1.0000	inf	0.4088	87
		react	1.0000	0.8000	0.8022	87
		interact	1.0000	NaN	12.5385	87
	video comments	**view**	**0.0009**	inf	**29.2353**	**70**
		react	0.0703	3.0000	2.8732	70
		interact	**0.0118**	**7.5625**	**6.2898**	**70**
Discussion Forum	threads	view	0.1818	inf	11.0000	11
		react	0.0801	20.0000	11.0000	11
		interact	1.0000	NaN	1.1818	11
Micro-blogging	tweets	view	0.3333	inf	25.0000	4
		react	1.0000	inf	5.0000	4
		interact	1.0000	NaN	1.0000	4

Table A.1: *Method A - Sequential Survey Logic* Fisher's Exact Test Results for association - Feature Engagement Style v/s Witnessing Cyberbullying

Method	Engagement Style	p value	Odds Ratio	Corrected Odds Ratio
Method B Lenient Threshold	viewing	0.0001	inf	61.0109
	reacting	0.0024	2.8154	2.8435
	interacting	0.0001	2.4655	2.4613
Method C Strict Threshold	viewing	0.0003	3.7917	3.6736
	reacting	0.0000	5.1967	5.2195
	interacting	0.0034	4.4994	4.6623

Table A.2: Fisher's Exact Test Results for association - Feature Engagement Style v/s Witnessing Cyberbullying as per *Method B - Lenient Threshold* and *Method C - Strict Threshold*

Feature	Engagement Style	p value	Odds Ratio	Corrected Odds Ratio	# Subsamples
stories	viewing	1.0000	NaN	3.4681	104
	reacting	1.0000	NaN	13.0000	104
	interacting	1.0000	NaN	69.0000	104
Reels	viewing	1.0000	1.0500	0.7805	101
	reacting	0.2036	3.1429	3.0000	101
	interacting	1.0000	0.0000	4.1556	101
comments	viewing	0.0000	30.0000	25.7059	100
	reacting	0.0000	87.0000	58.3333	100
	interacting	0.1200	inf	23.0870	100
disappearing-content	viewing	0.6943	0.5813	0.6049	78
	reacting	1.0000	1.3125	1.7240	78
	interacting	0.3519	2.5781	2.6162	78
group-chats	viewing	0.2366	3.1944	2.7946	48
	reacting	0.2014	3.2222	3.0343	48
	interacting	0.3446	2.6667	2.6608	48
video-content	viewing	1.0000	NaN	0.7426	87
	reacting	1.0000	NaN	12.5385	87
	interacting	1.0000	NaN	175.0000	87
	viewing	*0.0116*	inf	*20.4146*	70

Feature	**Engagement Style**	**p value**	**Odds Ratio**	**Corrected Odds Ratio**	**# Subsamples**
stories	viewing	1.0000	NaN	1.9577	104
	reacting	0.7965	0.8128	0.8434	104
	interacting	0.6334	1.3250	1.3351	104
reels	viewing	1.0000	NaN	12.6000	101
	reacting	0.0544	9.0000	9.1414	101
	interacting	1.0000	1.1040	1.2389	101
comments	viewing	0.1500	inf	17.6897	100
	reacting	1.0000	0.9405	1.2653	100
	interacting	0.2243	2.2240	2.2149	100
disappearing-content	**viewing**	*0.0000*	inf	*1341*	78
	reacting	*0.0000*	inf	*141*	78
	interacting	*0.0000*	inf	*261*	78
group-chats	viewing	1.0000	NaN	31.6667	48
	reacting	1.0000	0.0000	10.3333	48
	interacting	1.0000	0.0000	6.0667	48
video-content	viewing	1.0000	NaN	3.0930	87
	reacting	0.5683	1.6000	1.8878	87
	interacting	0.6811	0.6000	0.5679	87
video-comments	viewing	1.0000	NaN	5.1739	70
	reacting	0.1136	6.3333	6.0526	70
	interacting	0.0628	4.2449	4.2000	70
threads	viewing	1.0000	NaN	3.8000	11
	reacting	0.4909	3.5000	3.0000	11
	interacting	0.4545	inf	6.1111	11
tweets	viewing	1.0000	NaN	9.0000	4
	reacting	1.0000	NaN	2.3333	4
	interacting	1.0000	NaN	1.0000	4

Table A.3: *Method B - Lenient Threshold Logic* Fisher's Exact Test Results for association - Feature Engagement Style v/s Witnessing Cyberbullying

A.3 Descriptive Patterns in Scenario Evaluation Task

For analysing the aggregated responses across scenarios, the 7-point Likert scale was converted into numerical values. The ordinal options were mapped as follows: *`3 = strongly agree; 2 = agree; 1 = somewhat agree; 0 = no opinion; -1 = somewhat disagree; -2 = disagree;` and `-3 = strongly disagree`*. This transformation allowed the ordinal data to be treated as continuous, enabling statistical aggregation and comparison across scenarios. Instead of relying on raw frequency counts, this analysis used the mean response value for each scenario. For every scenario and across each of the three evaluative dimensions i.e., *realism*, *seriousness*, and *report-worthiness*, the average of participant's responses were computed. This produced a single representative score per dimension per scenario, ranging from –3 (strong disagreement) to +3 (strong agreement), with values near zero indicating neutrality or lack of consensus. As illustrated in Figure A.1 among Cohort_1 displayed similar variability across the 15 scenarios. Notably, for Scenario 4 (*flaming*), Scenario 6 (*denigration + exclusion*), Scenario 9 (*exclusion*), and Scenario 14 (*harassment*) each received disagreement scores of –0.5 or lower for *report-worthiness*. Within this cohort, Scenario 13 (*denigration*) stood out, while it scored negatively for realism, it achieved strong agreement (>+2) for both *seriousness* and *report-worthiness*. Free-text responses collected from Cohort_1 as part of the survey methodology helps understand this discrepancy. The scenario, adapted from (Campbell and Xu, 2022), described peers creating a "`we hate website`" to post rumours and jokes. Participants expressed scepticism, noting that such behaviour seemed outdated and unrealistic. They argued that cyberbullying is more likely to occur via low-effort, high-visibility platforms such as TikTok, group chats, or social media rather than through stand alone websites. For Cohort_2 (Figure A.2), only Scenario 6 (*denigration with exclusion*) received consistently low agreement across all three dimensions. This scenario, which described sharing a meme of a girl in a group chat with a deliberate exclusion caption, the free-text responses for *seriousness* show that the participants in Cohort_2 downplayed the incident and characterised it as something trivial, ignorable, or fleeting rather than a serious offense. When comparing the aggregated scores of both cohorts for the six scenarios (Figure A.3), Scenario 6 consistently exhibited low agreement, particularly for *report-worthiness*, which approached zero, indicating predominantly neutral or no-opinion responses. Consequently, scenario ID 6 (*denigration with exclusion*) was excluded from further dataset development (Phase 3, §3.3.3). Among the remaining scenarios evaluated by Cohort_1, scenario IDs 8, 10, and 13 were selected for progression to the next phase of detailed design and dataset development. Selection was based not only on relatively higher agreement levels for seriousness and report-worthiness but also on the richness of participants free-text responses, which

revealed nuanced reasoning. Scenario ID 8 (*outing + denigration*)0, involving the non-consensual disclosure of a someone's sexual orientation followed by mockery, received moderate agreement for seriousness and report-worthiness (`somewhat agree - agree`, <2) (See Figure A.1). Free-text responses highlighted that participants recognised homophobic peer bullying and associated betrayal of trust as common, reinforcing the scenario's realism. Scenario ID 13 (*denigration*), despite lower realism ratings, was rated highly on seriousness and report-worthiness (`agree - strongly agree`, >+2). Participants in the free text responses noted that its public visibility and potential for harm contributed to these higher ratings, while the lower realism score was largely attributed to the specific phrasing "*use of a website for denigration*" as adopted from (Campbell and Xu, 2022), instead participants suggested "*it is much simpler and common for this to happen on TikTok*". Scenario ID 10 (*exclusion*), addressing social exclusion, received moderate seriousness ratings (mean = 1.6) and low ratings for *realism* and *report-worthiness*. However, free-text responses suggested its plausibility, where a 16-year-old male noted, "*It happens in my school. The gifted kid will be excluded, for example, in sports because people think they ruin it or are less capable*", while a 15-year-old female wrote, "*I can see it happening amongst younger kids*" despite selecting "`no opinion`" for *realism*. These descriptive patterns help ascertain scenario IDs - *1, 2, 5, 7, 8, 10, 12, 13* to be included in the next phase of dataset design.

A.4 Descriptive Statistics of Youth Co-created Dataset

As presented in Table A.5, the youth co-created dataset across across 10 conversations, yields 552 messages with a mean of 55 and a median of 61 messages per conversation, reflecting its relatively small but concentrated conversational structure. In contrast, (Sprugnoli et al., 2018) and (Ollagnier et al., 2022; Ollagnier, 2024) contain substantially more messages (1,994 and 3,516, respectively) but across fewer or comparable numbers of conversations, resulting in much higher average messages per conversation (μ_{m}=498 and 160, respectively). The real-world dataset from (Van Hee et al., 2018) shows a starkly different pattern. Across 6,979 conversations, the total message count is 111,241, but the mean and median per conversation are relatively low (μ_{m}=16 and $\tilde{x}_{\mathrm{m}}$=4), indicating that most conversations are brief. This reflects the question-answering thread structure of the now defunct ASK.fm platform. When filtered for role-labelled messages only the (Van Hee et al., 2018), dataset contains 5,214 messages over 1,697 conversations, with a mean of 3 and a median of 2 messages per conversation. This highlights the sparsity of annotated messages in real-world data compared to both co-created and role-playing datasets. However, when considering the roles and their contributions to the overall number of messages, clearer patterns emerge. As shown in Table A.5, in both the youth co-created dataset and (Sprugnoli et al., 2018),

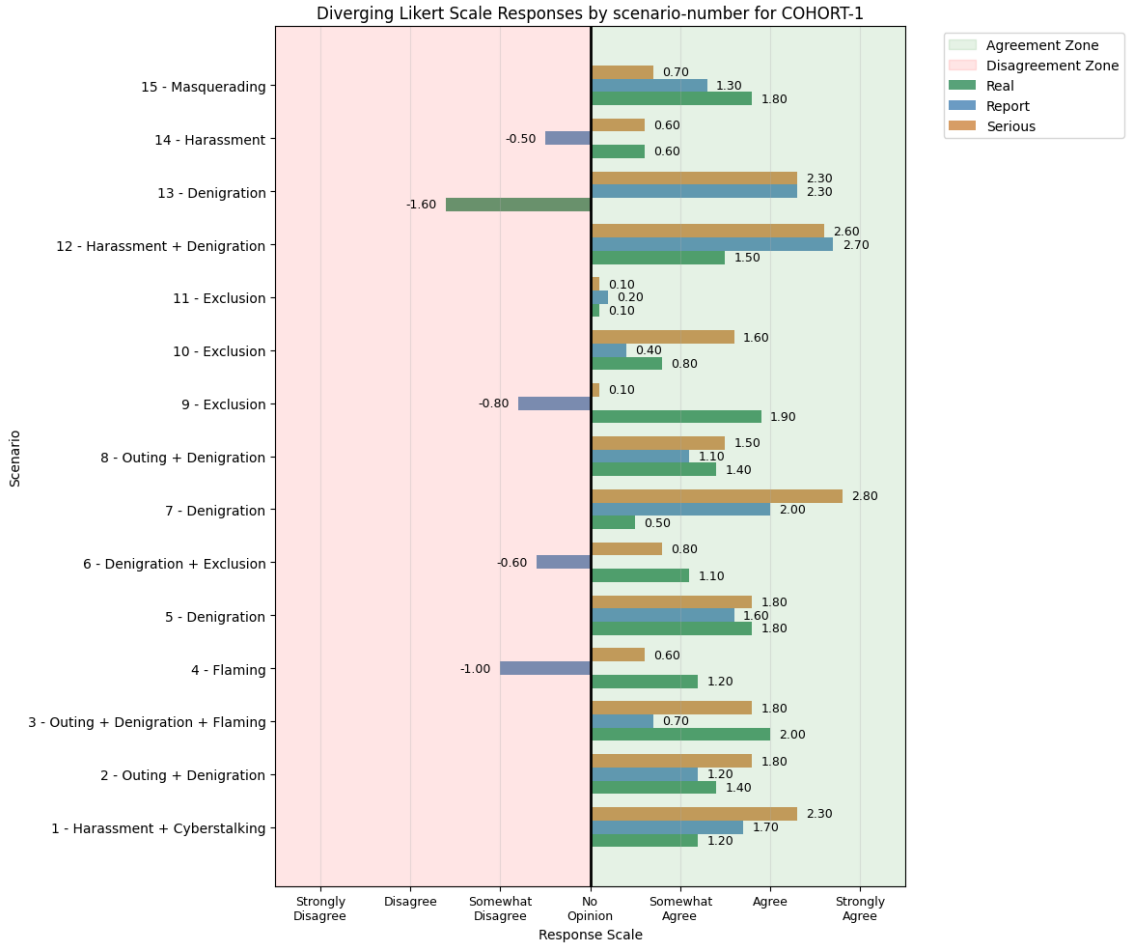


Figure A.1: Diverging bar chart of mean Likert-scale responses across 15 scenarios for Cohort$_1$. Responses are averaged over three evaluative dimensions (Real, Report, Serious) and mapped to a numeric scale ranging from –3 (strong disagreement) to +3 (strong agreement). The central vertical line (0) represents neutrality, with shaded regions denoting agreement (green) and disagreement (red) zones.

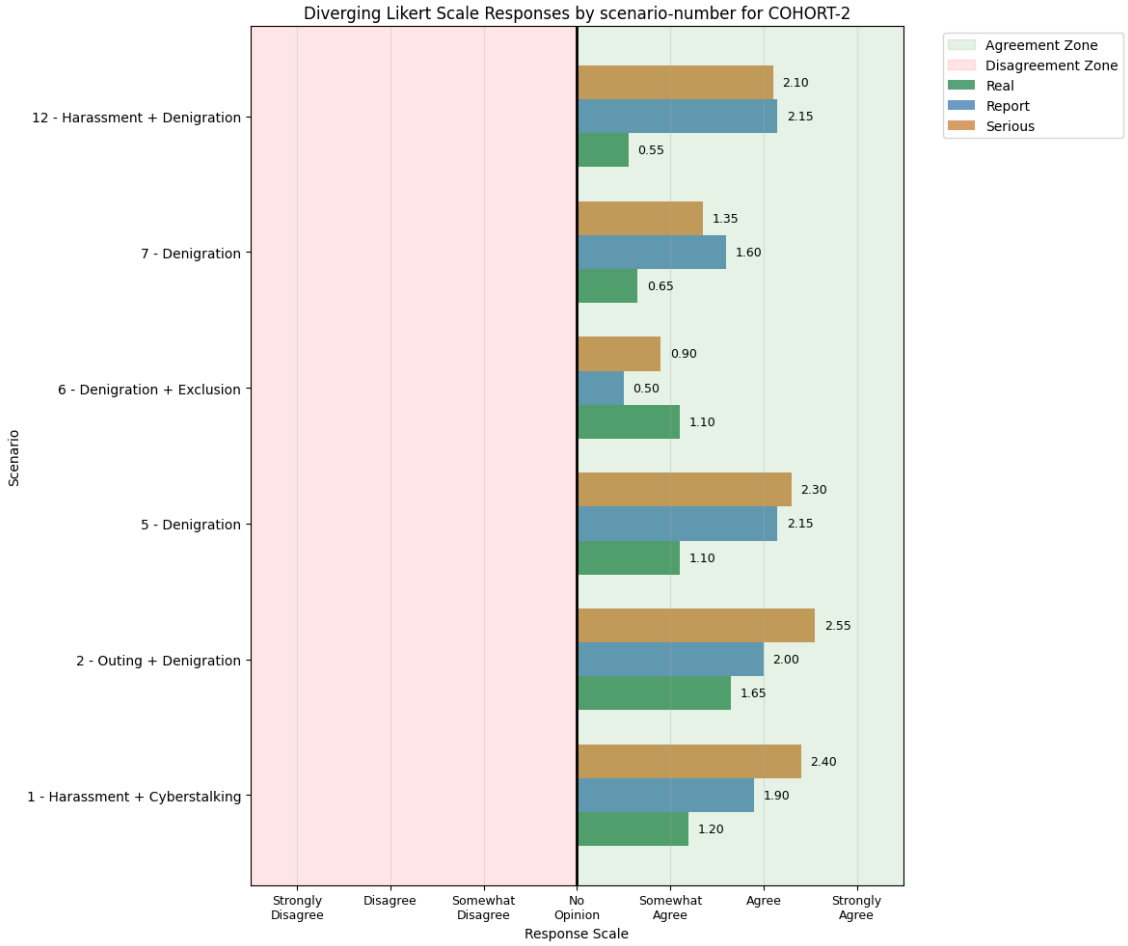


Figure A.2: Diverging bar chart of mean Likert-scale responses across 6 scenarios for Cohort_2. Responses are averaged over three evaluative dimensions (Real, Report, Serious) and mapped to a numeric scale ranging from –3 (strong disagreement) to +3 (strong agreement). The central vertical line (0) represents neutrality, with shaded regions denoting agreement (green) and disagreement (red) zones.

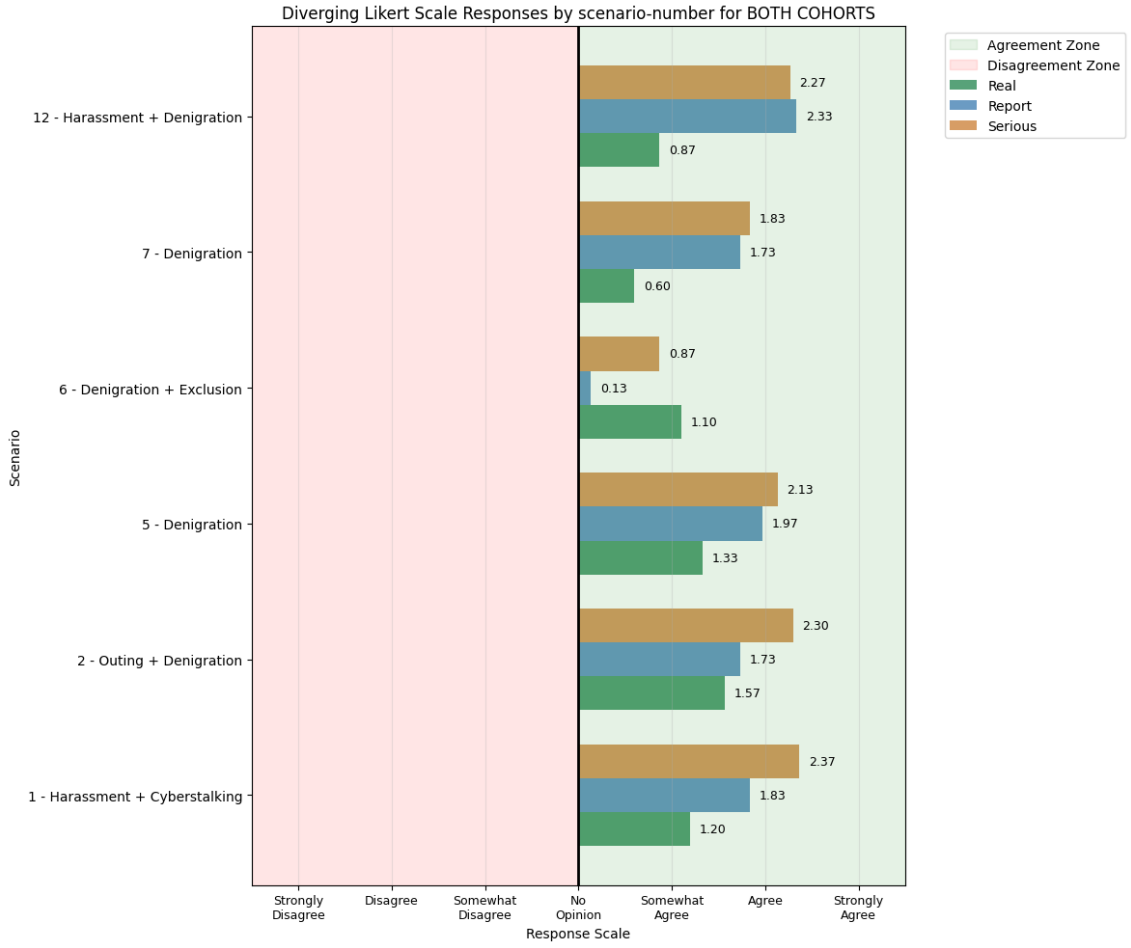


Figure A.3: Diverging bar chart of mean Likert-scale responses across 6 scenarios for both Cohort_1 and Cohort_2. Responses are averaged over three evaluative dimensions (Real, Report, Serious) and mapped to a numeric scale ranging from –3 (strong disagreement) to +3 (strong agreement). The central vertical line (0) represents neutrality, with shaded regions denoting agreement (green) and disagreement (red) zones.

the mean (μ) and median ($\tilde{x}$) number of messages from defenders are slightly higher across all scenarios. In contrast, the French dataset by (Ollagnier et al., 2022; Ollagnier, 2024), shows a more balanced distribution between defenders and bullies, while also highlighting a stronger participation of enablers across its seven unique scenarios drawn from 22 role-played conversations. In the real-world cyberbullying dataset by (Van Hee et al., 2018), within conversations with labelled roles, the data indicates that enablers and bystanders contribute at least one message on average (both mean and median), while bullies and victims contribute around two messages on average.

Dataset	**Role**	**Mean # Messages** μ_{m}	**Median #Messages** $\tilde{x}_{\mathrm{m}}$
Youth Co-created	Enabler	19	20
	Defender	24	27
	Bully	10	9
	Victim	3	3
(Sprugnoli et al., 2018) (Italian)	Enabler	150	167
	Defender	161	184
	Bully	160	111
	Victim	84	37
(Ollagnier et al., 2022; Ollagnier, 2024) (French)	Enabler	46	54
	Defender	33	33
	Bully	33	32
	Victim	30	25
	Conciliator	23	16
(Van Hee et al., 2018) (English)	Enabler	1	1
	Defender	1	1
	Bully	2	1
	Victim	2	1

Table A.5: Mean and median values of number of messages by roles across conversations in all four datasets. Note: For the purpose of profiling the (Van Hee et al., 2018)* denotes only messages within conversations annotated with role-based labels are included, while those tagged solely as "`negative`" are excluded.

To better understand the structure and characteristics of the conversations, we examined the distribution of message lengths, measured in number of characters including whitespaces across all four datasets (Figure A.4). Since the distribution of the real-world dataset by Van Hee et al. (2018) contained a small number of unusually long messages (41 messages exceeding 500 characters by victim=25; bully=7; defender=9), we adjusted the visualisation in Figure A.4 by truncating values above 500. This adjustment ensures comparability with the other three datasets, which do not contain messages exceeding this threshold. Overall, the distribution reveals substantial variation in message length across datasets. On average, the youth co-created dataset contains

relatively short messages (45 characters per message), similar in scale to the Italian dataset by Sprugnoli et al. (2018) (34 characters) and the French dataset by Ollagnier et al. (2022); Ollagnier (2024) (35 characters). In contrast, the real-world English dataset Van Hee et al. (2018) displays longer and more variable messages, averaging 67 characters per message, with a noticeable tail of very long utterances. This suggests that while co-created and role-played datasets tend to elicit brief exchanges, real-world conversations capture greater heterogeneity in message construction, possibly reflecting more naturalistic and less constrained communication.

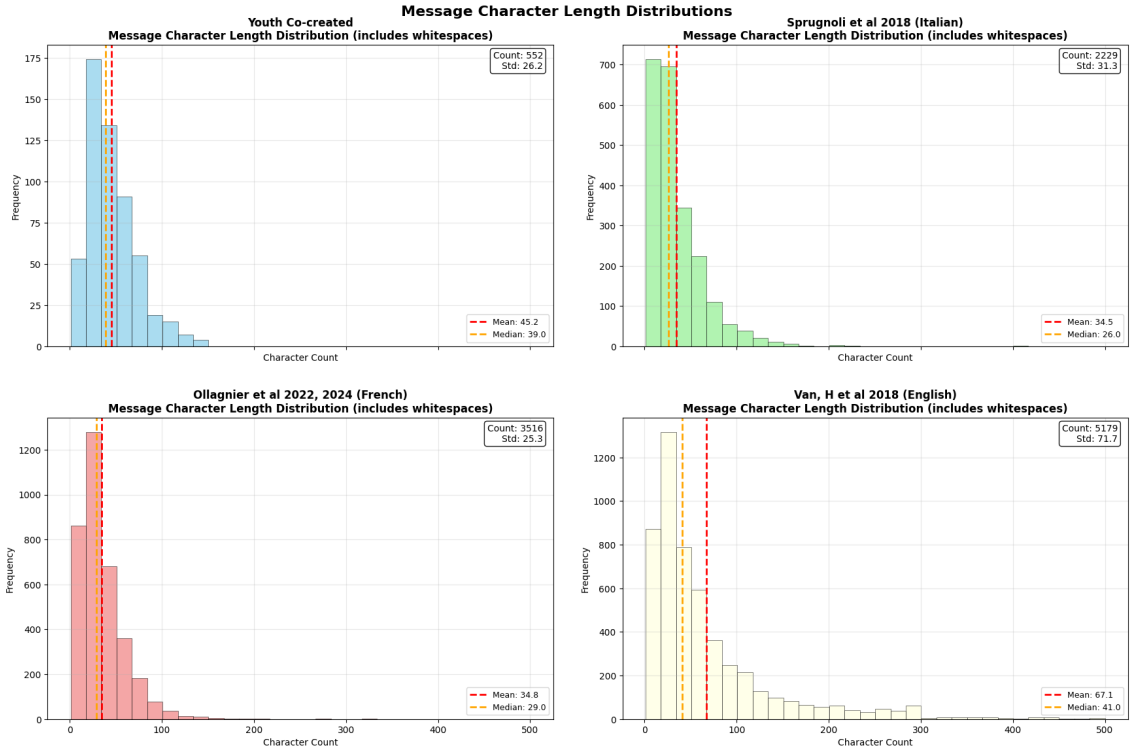


Figure A.4: Comparative Character Length Distribution across messages for Youth Co-created dataset (Left top), Italian dataset by (Sprugnoli et al., 2018) (Right top), French dataset by Ollagnier et al. (2022); Ollagnier (2024), and real-life cyberbullying dataset by (Van Hee et al., 2018)

Across all messages in the four datasets, Table A.6 provides a more fine-grained analysis by breaking down average message length i.e., character length of messages including white spaces across conversational roles. Several patterns are noteworthy. First, across all datasets, **defender** messages tend to be the longest, particularly in the real-world dataset by (Van Hee et al., 2018) (154 characters on average), indicating that supportive responses often involve elaboration or reassurance. In contrast, bully messages are consistently short in the youth co-created dataset (μ_c=40), and both role-played datasets by (Sprugnoli et al., 2018) (μ_c=31), and (Ollagnier et al., 2022; Ollagnier, 2024) (μ_c=32). However, enabler messages are slightly higher in (Van Hee et al., 2018)'s real-world cyberbullying dataset (μ_c=59). Thereby, aligning with the notion that encouragement of bullying is often minimalistic. Victim messages vary more widely, ranging from 40 characters in

the youth datasets to 88 in the real-world dataset, suggesting that victims in natural conversations may provide longer responses compared to those in controlled or co-created settings.

Dataset	**Enabler** (μ_c)	**Defender** (μ_c)	**Bully** (μ_c)	**Victim** (μ_c)
Youth Co-created	41	51	40	40
(Sprugnoli et al., 2018) (Italian)	36	33	31	42
(Ollagnier et al., 2022; Ollagnier, 2024)	32	35	32	41
(Van Hee et al., 2018)	59	154	57	88

Table A.6: Average Length by characters in messages by Role and Dataset

Taken together, these findings highlight two key insights: (1) dataset design influences message length, with (Van Hee et al., 2018)'s dataset sourced from question-answering platform (now defunct ASK.fm) containing longer and more diverse utterances compared to youth co-created or role-played datasets, and (2) role influences message length, with supportive roles being longer in character size indicating elaborated messages, while aggressive roles have comparatively shorter character lengths. This has methodological implications, as characteristics such as length of messages across four datasets signals role-specific communicative strategies. These comparisons highlight clear differences in role distribution across cyberbullying conversational datasets. Both the youth co-created dataset and the corpora from (Sprugnoli et al., 2018) and (Ollagnier et al., 2022) are synthetic, as they are either developed through layered, sequential co-creation or through role-playing tasks in which youth participants collaboratively generate conversations by inhabiting different roles. This design explains their comparatively balanced role distributions and the inclusion of more elaborate supportive interventions. By contrast, (Van Hee et al., 2018) represents a real-world dataset collected from ASK.fm interactions (Q&A platform), where perpetrator behaviours dominate and defenders are sparse. The youth co-created dataset extends beyond both role-playing and real-world corpora by centering youth agency in shaping conversations, which in turn produced a comparatively richer presence of supportive actions, especially from defenders. This quality of the youth co-created dataset has **direct implications** for AI-enabled recommendation interventions. As demonstrated by co-design studies such as (Milosevic et al., 2023; Zou et al., 2024; Justice with Children Initiative, 2023), adolescents view victim-support recommendations as an important feature for online safety systems. For such AI systems to be meaningful, they must not only detect aggression but also recognize the range of supportive behaviours that young people value, including defending, enabling, and other forms of engagement. Accordingly,

the strong presence of defender and enabler messages in the youth co-created dataset makes it an effective benchmarking test set for cyberbullying detection systems that aspire to move beyond harmful content identification toward modelling supportive dynamics. Compared with role-playing corpora, which achieve balance but remain scripted, and real-world datasets, which predominantly capture hostile dynamics, the youth dataset provides a more substantive and contextually grounded foundation for designing recommendation tools that expand reporting options for victims and bystanders, while also fostering broader ecosystems of online safety and resilience.

A.5 Individual message-level H-MT Analysis

When aggregated at the conversation thread level (Table A.7), similar patterns emerge, though clearer scenario-specific trends are visible. Conversations such as: *Homophobia-Julie* yield the lowest scores across all metrics (BLEU ≈15, chrF ≈37, COMET-QE ≈ 0.58–0.63), indicating that messages in this scenario might have been challenging due to the presence of nuanced and slang expressions. Additionally, the difference (Δ) in COMET-QE reference-free metrics is the highest, indicating that the student translations for this scenario are not as formally styled as company MT outputs. In contrast, *Homophobia-Julie-bis*, created for the same scenario by a different set of students in (Ollagnier et al., 2022)'s study[120], achieves the highest BLEU and ChrF scores (≈27 and ≈51), showing a closer lexical and character-level alignment between translations. Notably, in this case, the Δ in COMET-QE is lowest between the company and student translations, indicating that student and company outputs are more semantically aligned than in *Homophobia-Julie* conversation. The higher COMET-QE scores for both sets of translations in *Homophobia-Julie-bis* further suggest that these outputs are stylistically more formal than those from the original conversation. For other scenarios, such as *Religion-Fatimah*, the company's MT outputs + post-editing score relatively well on ChrF (≈53) and COMET-QE (≈ 69), highlighting stronger lexical alignment between company and student translations and COMET QE, given its training on formal Wikipedia-like text, consistently favours the comparatively more formal style of the company outputs.

However, upon closer examination, the limitations of COMET-QE become more evident. For example, in the conversation `Homophobia-Julie`, the French source text *`cul de crasseuse`* was translated by the student as *`you scumbag`* (COMET-QE: 0.71) and by the company as *`dirty ass slut`* (COMET-QE: 0.48). In this case, the student translations are clearly more appropriate and is also rewarded with a higher score. However, in another case from the `Religion-Fatimah-bis`, the French source *`normal un singe ça bouffe des teub`* was translated by the student as

[120]This was highlighted in email correspondence with Ollagnier et al. (2022)

File-name	**Student as Reference**		**COMET-QE$_{\text{Company}}$**	**COMET-QE$_{\text{Student}}$**	**COMET-QE Δ**
	BLEU ↑	chrF↑			
Ethnicity_sylvia_sophie	19.8174	47.3441	0.7175	0.6706	0.0469
Homophobia_julie	14.9736	40.6654	0.6310	0.5771	0.0539
Homophobia_julie_bis	27.4474	53.0659	0.6855	0.6530	0.0325
Religion_fatimah	19.4976	53.4008	0.6935	0.6507	0.0428
Religion_fatimah_bis	16.7990	48.7586	0.6983	0.6553	0.0430

Table A.7: Evaluation metrics at the document level i.e., conversation thread level (5 conversations). BLEU/chrF are shown with company and student translations as reciprocal references, and COMET-QE provides reference-free quality estimates.

`it's normal for a monkey to eat bananas` (COMET-QE: 0.65) and by the company as *`not surprising: monkeys eat dicks`* (COMET-QE: 0.38). Here, although the student translation received the higher COMET-QE score, the company's version is actually more faithful to the offensive intent of the source. This contrast highlights a key issue: COMET-QE does not always align with appropriateness or adequacy, particularly in cases involving offensive or hurtful language.

Moreover, of the 876 messages across five conversations translated by both student groups and the company (MT + post-editing), 359 were originally labelled as `Not Aggressive`, 330 as `Overtly Aggressive`, and 110 as `Covertly Aggressive`. Both reference-based and reference-free scores, as illustrated in Figure A.5, demonstrate clear variation across aggression categories when student translations are taken as the reference to evaluate company's MT + post-editing outputs. BLEU scores are generally concentrated below 20 for both `Not Aggressive` and `Overtly Aggressive` categories, indicating that company translations diverge more strongly from student translations when the source text contains either `No Aggression` or `Overt Aggression`. In contrast, ChrF scores are distributed broadly though majority still fall below 50. Interestingly, the distributions of `Overtly Aggression` and `Not Aggression` messages appear to be slightly higher but similar. Although the COMET-QE, show that messages below 0.8 scores are generally more `Overtly Aggressive` than *Not Aggressive*. Additionally, the *label verification* exercise with professional translators is illustrated in Figure A.6, where translators provided binary alignment judgments for message-level annotations. Given the inherent ambiguity of `CONTEXT` labels, these were excluded from the analysis; only `AGGRESSION` and `INTENTION` labels from (Ollagnier et al., 2022; Ollagnier, 2024) were considered. Results are grouped across the four online-harm vignettes: obesity, homophobia, ethnicity, and religion. Alignment was highest in the obesity scenarios (1,212 `yes` vs. 155 `no`), indicating relatively straightforward cross-lingual mappings. In contrast, homophobia (465/67) and religion (314/38) scenarios showed lower alignment, reflecting greater interpretive

complexity in transferring meaning across languages. These findings highlight the value of a human-aided post-editing process for preserving intended meaning in hurtful or offensive socio-linguistic contexts.

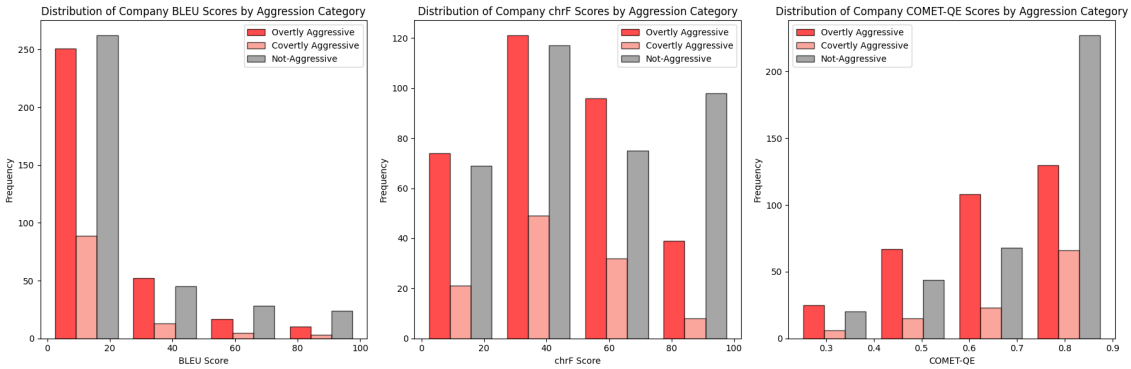


Figure A.5: Distribution of BLEU, chrF, and COMET-QE scores of Company's MT + post Editing Outputs across 876 messages

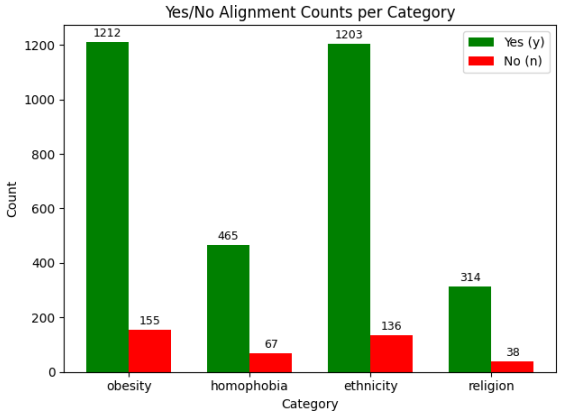


Figure A.6: Distribution of Alignment Agreement and Disagreement with translated (EN) and source (FR) corpora by categories of scenarios

Upon closer examination, the French source *julie mais retourne dans ta campagne a baiser les chèvre bouzeuse*, translated by student as *julie go back to your countryside to fuck your disgusting goats* is relatively literal. In contrast, the company's MT outputs + post editing version *julie go back to the country where you bang shit-covered goats* incorporates linguistic nuance, using slang such as *bang* instead of literal profanity. The COMET-QE scores, which estimate translation quality without a reference, show a notable differences, i.e., the student translation scores 0.658, whereas the company's output scores 0.487. When the student translation is treated as a reference, the company's translation achieves a BLEU score of 24.808415 and a chrF score of 40.513097. This means that while the company's translation may sound more natural or idiomatic in English, it diverges more from the literal meaning of the source sentence, which is reflected in the lower COMET-QE score. Similarly, BLEU and chrF that

measure n-gram overlap and character-level similarity respectively, indicate that the company's output is moderately aligned with the student translation, but not as closely as the student's own rendering, which is expected since the student translation is more literal. In other words, higher linguistic nuance can sometimes reduce metric-based scores when the reference is a literal translation. For a non-aggressive source French *mais calmez-vous chaqu'un s'est préférence*, the student translation is *just calm down everyone has their preferences*, while the company translated it as *just calm down, to each their own.* Here, the company translation is more idiomatic and fluent in English, but when evaluated against the student reference, it receives a BLEU of 22.09 and chrF of 47.38 with a COMET-QE of 0.546971. The student's translation has a COMET-QE nearly 0.18 points higher. This indicates that although the company's translation is fluent and idiomatic, it departs from the literal phrasing of the source, leading to lower quality scores according to these metrics. The metrics favour literal equivalence and closer n-gram matches, which the student translation better preserves.

A.6 Pilot study: LLM-aided generation

Pilot Study: To evaluate the feasibility of using LLMs as an alternative for data generation, a pilot study was conducted with commercial Claude-2 Anthropic (2023) and open-source Llama2Chat$_{7B}$ Touvron et al. (2023) in 2023. Adapting the few-shot Chain-of-Thought (CoT) prompting strategy, the aim is to guide the LLMs reasoning in a stepwise, interpretable manner, thereby exercising fine-grained control over the outputs and ensuring alignment with research objectives. The pilot experiment was structured around a series of **four** sequential prompts, each designed as a modular sub-task in line with structured task decomposition principles. In the **first prompt** below,

> You are presented with a conversation between school-age children. They are role-playing a conversation given this scenario `<scenario>`>.... See a sample of their conversation `<Conversation-Chain>`.... What do you learn from this conversation?

Also presented in Table A.10, Appendix A.10, the LLMs were provided with a cyberbullying-related conversational scenario adapted from Sprugnoli et al. (2018), featuring a shy male classmate inviting peers to view his ballet performance on a WhatsApp group chat. In addition, the LLMs were presented with a "*conversation-chain*" of 45 English-translated messages from the English translation of this dataset discussed in §5.1.1 enclosed within *<conversation-chain>* `[...]` *<conversation-chain>* tags, and were prompted to reflect on the insights they derived from this conversation. This prompt effectively serves as a step to initiate context comprehension and reasoning, encouraging the LLMs to process the scenario and extract salient social and emotional cues. The **second prompt** explicitly asked, "*Have you detected any instances of cyberbullying in*

this conversation?" (Table A.11, Appendix A.10). This step functions as a focused detection task, where the model applies its understanding of the conversation to identify potential bullying instances, leveraging the reasoning chains established in the first prompt and without any definition of cyberbullying. The **third prompt** explicitly requested the LLMs, "*What categorization can you give the various instances of bullying?*" (Table A.12, Appendix A.10), to categorise the identified bullying instances according to relevant subtypes, again without any specific categorisation. This represents the classification sub-task, further decomposing the overall problem into interpretable steps and enabling the LLMs to generate structured labelled outputs. Finally, the **fourth prompt** "*Create a similar conversation*", (Table A.13, Appendix A.10) instructed the LLMs to generate a similar conversation. In this manner, through these four prompts and using the LLM's context window, we effectively synthesised a new scenario in alignment with the patterns, social dynamics, and emotional trajectories learned in the preceding steps. This prompt constitutes the generative stage, leveraging the reasoning, persona understanding, and task decomposition from the earlier prompts to produce ethically safe, contextually coherent synthetic CB data. The rationale for not providing an explicit definition of cyberbullying or its subtypes in the second and third prompts is based on the assumption that large language models (LLMs), having been trained on extensive corpora including books, Wikipedia, and other textual sources, already possess a latent understanding of the cyberbullying phenomenon and its various categories. By omitting formal definitions or examples, the aim is to assess LLMs' inherent knowledge and reasoning capabilities regarding cyberbullying, while also revealing potential biases or gaps in their understanding of the phenomenon. This approach enables the study to assess the extent to which LLMs can independently recognise, interpret, and categorise cyberbullying behaviours without relying on predefined labels.

Pilot Results: For the first prompt, Claude-2 and Llama2Chat$_{7B}$ produced markedly different responses (Table A.10, Appendix A.10). Claude-2 provided a detailed analysis of the scenario, focusing on the attitudes, reactions, and social dynamics of the individuals involved in the conversation. In contrast, Llama2Chat$_{7B}$ emphasised the identification of harmful language and stereotypes, advocating for inclusivity, respect, and empathy toward all participants. In response to the second prompt, both models yielded highly similar outputs (Table A.11, Appendix A.10). Both Claude-2 and Llama2Chat$_{7B}$ successfully detected multiple instances of cyberbullying, highlighting derogatory language, harmful stereotypes, and actions intended to belittle or shame the targeted individual. They also underscored the importance of addressing and preventing cyberbullying behaviours. The third prompt revealed more pronounced differences between the models (Table A.12, Appendix A.10). Claude-2 produced a comprehensive categorisation, identifying verbal bullying, relational bullying, prejudiced bullying, public humiliation, criticism of hobbies or

interests, and retaliatory behaviour, along with a detailed breakdown of each type. Llama2Chat$_{7B}$, by contrast, provided a more generalised classification, identifying verbal bullying, social bullying, and body-shaming. While less granular, Llama2Chat$_{7B}$ illustrated each category with concrete examples from the conversation chain. Notably, the model speculated on "*physical bullying*" e.g., threats to print t-shirts with someone's picture, even though physical bullying is not formally a cyberbullying category, highlighting differences in interpretive reasoning between the models. The fourth prompt, central to the study, evaluated the models' ability to generate synthetic cyberbullying dialogues (Figure A.7, Table A.13, Appendix A.10). Claude-2 generated ten messages, preserving role assignments consistent with the original conversation chain where `[Person-1]` is the `victim`, `[Person-3]` is the `bully`, and `[Person-7]` is an `enabler`, while maintaining coherent narrative flow. Llama2Chat$_{7B}$ produced seven messages in a similar scenario involving a new student. However, it generated only a single message from the `bully` and did not generate messages for `enabler` roles. This discrepancy likely reflects Claude-2's larger parameter count and extended context window (up to 100K tokens) Anthropic (2023), enabling more precise role alignment. Despite these differences, both models demonstrate the potential of LLMs to generate realistic, contextually appropriate simulations of youth cyberbullying interactions. By producing synthetic role-playing dialogues, LLMs offer a method to study cyberbullying while minimising ethical risks to minors, circumventing potential psychological distress, and alleviating concerns associated with direct participation in sensitive research.

A.7 Phase-1: Plain Language Statement, Consent and Assent Forms

A.7.1 Plain Language Statement

Plain Language Statement

Welcome to the Research Study!

[Please ensure your child or ward also reads this Plain Language Statement]

About This Research Study Young people use social media apps like TikTok, Snapchat, Instagram, YouTube, Twitter, and other such apps in different ways. We want to know what young people think about using these apps. These social media apps provide different features to communicate with one another and sometimes people behave in nasty or mean or hurtful manner on these apps. We want to know what young people think about different social media features and if they have come across nasty or mean or hurtful content.

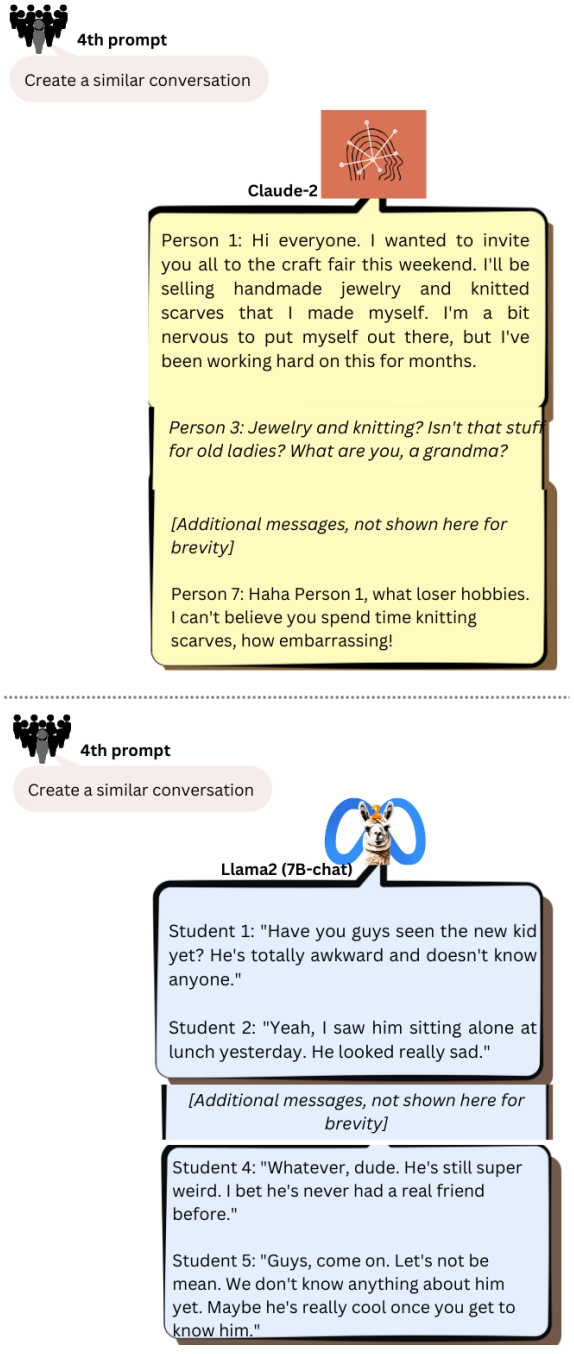


Figure A.7: Snippet of cyberbullying-like conversations generated by Claude-2 Anthropic (2023) and Llama2Chat$_{7B}$ Touvron et al. (2023). (See Table A.13 in Appendix A.10 for detailed response)

Taking part in this research means? The participant will be answering some questions in an online survey. These questions will be about the types of online content they view frequently on social media apps and what they think about some statements. Please note, we will not be asking for any personal details or asking the participant to describe their experiences. Their answers will be kept anonymous. Nobody will know that the answers came from the participant. Don't worry, this survey should only take fifteen to twenty minutes. There are no right or wrong answers, and we want to hear the participant's honest thoughts and opinions.

Benefits of participating? The participant's honest thoughts and opinions will help us develop better tools and guidelines to respond against hurtful online content, which can contribute to a better and safer future for young people.

Voluntary Participation Participation in the survey is entirely voluntary, and one can withdraw from the survey at any time.

Who are we? This study is part of the DTCCT Project conducted by researchers from:

- Kanishk Verma (PhD Student) Dublin City University
- Dr. Brian Davis , Dublin City University
- Dr. Tijana Milosevic , University College Dublin
- Dr. Rebecca Umbach , Google

Note: This project is funded by the Irish Research Council and Google Ireland (Grant No. ES-PSG/2021/161).

A.7.2 Parent/Guardian Consent

Parent/Guardian Consent Form

Clarification of the purpose of research

[Please ensure your child or ward also reads this Plain Language Statement]

- We are conducting this online survey with only teenagers and young adults (ages 13 to 21).
- This survey should not take more than thirty minutes to complete.
- The survey will ask the participants to, share their experience with social media features like "Reels", "Stories", "Comments", and others and share their thoughts and opinion on a few statements about social media platforms.
- The **survey will not ask** the participants to,

  - share any personal details,

  - describe their social media life or usage.

- We will ask the participant about their age and gender, please note, these answers cannot reveal their identity.

- There are no right or wrong answers for the questions, we believe their honest thoughts and opnions will be highly valuable.

- The findings of this research study can be made available on request to the researchers at kanishk.verma@adaptcenter.ie

- Please note the participant can withdraw from the survey at any point of time.

- Participation in the survey is entirely voluntary.

Confidentiality, anonymity, data retention, and disposal In this study, we will not be collecting any confidential or personal identifiable information from the participants. We will only collect information about their age and gender, this will be used to categorise responses on a high level. Participants do not have to provide their age in Date-of-Birth formart, instead they can enter their age in numbers (example, 13, 15, 17, etc). Any or all responses will only be stored on a secured DCU Google Drive.

By clicking the "[I consent..]" button below, you acknowledge:

Participation in the study is voluntary. You are aware that you can choose to terminate your participation at any time for any reason.

A.7.3 Teenager Assent

Teenager Assent Form

Please read the information below carefully.

I have read and understood the research study and what taking part in this study means.	*I understand I will be asked to answer some questions on my social media use.*
I am happy to say only my gender and age; and I know no other personal information is asked from me.	*The survey will be online and I can attempt it at a time and place of my convenience.*
I understand what is expected of me	*I can take a break when I need to*
I can stop being in the study at any stage I want to	*I would be happy to talk to another person I have any big concern.*

By clicking the [**I agree**] button below, you confirm that:

- Your participation in the study is voluntary.
- You understand that you can stop participating at any time, for any reason.

A.8 Phase-1: Detailed Results

Table A.8: Breakdown of feature usage by respondents in percentages collected via Requirement Gathering Phase-1

Gender	Action Type	Feature	Never (%)	Hardly ever (%)	Some-times (%)	Often (%)	Very Often (%)	Total-Count
Male	View	reels	0.00	9.26	20.37	44.44	25.93	54.0
		video_content	0.00	0.00	45.28	32.08	22.64	53.0
		video_com-ments	0.00	4.26	44.68	29.79	21.28	47.0
		group_chats	0.00	3.45	44.83	31.03	20.69	29.0
		disappearing	0.00	7.55	32.08	43.40	16.98	53.0
		threads	0.00	30.00	60.00	0.00	10.00	10.0
		comments	0.00	39.62	43.40	9.43	7.55	53.0
		stories	0.00	56.36	30.91	5.45	7.27	55.0
		tweets	0.00	0.00	0.00	0.00	0.00	0.0
	React	group_chats	0.00	10.34	48.28	34.48	6.90	29.0
		disappearing	1.89	39.62	43.40	9.43	5.66	53.0
		reels	11.11	33.33	40.74	12.96	1.85	54.0
		stories	23.64	63.64	12.73	0.00	0.00	55.0
		comments	3.77	73.58	20.75	1.89	0.00	53.0
		video_content	1.89	45.28	50.94	1.89	0.00	53.0
		video_com-ments	4.26	55.32	38.30	2.13	0.00	47.0
		threads	20.00	30.00	50.00	0.00	0.00	10.0
		tweets	0.00	0.00	0.00	0.00	0.00	0.0

Continued on next page

Gender	Action Type	Feature	Never (%)	Hardly ever (%)	Some-times (%)	Often (%)	Very Often (%)	Total-Count
		disappearing	11.32	45.28	20.75	15.09	7.55	53.0
		group_chats	3.45	3.45	62.07	24.14	6.90	29.0
		stories	20.00	72.73	7.27	0.00	0.00	55.0
		reels	35.19	53.70	9.26	1.85	0.00	54.0
	Interact	comments	18.87	39.62	41.51	0.00	0.00	53.0
		video_content	83.02	16.98	0.00	0.00	0.00	53.0
		video_com-ments	14.89	61.70	23.40	0.00	0.00	47.0
		threads	50.00	50.00	0.00	0.00	0.00	10.0
		tweets	0.00	0.00	0.00	0.00	0.00	0.0
		reels	0.00	0.00	2.04	30.61	67.35	49.0
		video_content	0.00	5.88	32.35	29.41	32.35	34.0
		stories	0.00	24.49	42.86	20.41	12.24	49.0
		comments	2.04	28.57	53.06	6.12	10.20	49.0
	View	disappearing	12.82	5.13	43.59	38.46	0.00	39.0
		video_com-ments	4.17	33.33	62.50	0.00	0.00	24.0
		group_chats	0.00	10.53	47.37	42.11	0.00	19.0
		tweets	0.00	50.00	50.00	0.00	0.00	4.0
		threads	0.00	0.00	0.00	100.00	0.00	1.0
		reels	2.04	16.33	32.65	40.82	8.16	49.0
		stories	14.29	36.73	34.69	10.20	4.08	49.0

Continued on next page

Female								
	React							

Gender	Action Type	Feature	Never (%)	Hardly ever (%)	Some-times (%)	Often (%)	Very Often (%)	Total-Count
		video_content	5.88	35.29	44.12	11.76	2.94	34.0
		comments	12.24	61.22	14.29	10.20	2.04	49.0
		disappearing	20.51	35.90	41.03	2.56	0.00	39.0
		video_com-ments	12.50	75.00	12.50	0.00	0.00	24.0
		group_chats	5.26	10.53	63.16	21.05	0.00	19.0
		tweets	25.00	50.00	25.00	0.00	0.00	4.0
		threads	100.00	0.00	0.00	0.00	0.00	1.0
		stories	30.61	57.14	10.20	0.00	2.04	49.0
		comments	46.94	44.90	6.12	2.04	0.00	49.0
		reels	18.37	59.18	22.45	0.00	0.00	49.0
		disappearing	12.20	29.27	41.46	17.07	0.00	41.0
	Interact	video_content	100.00	0.00	0.00	0.00	0.00	34.0
		video_com-ments	20.83	70.83	8.33	0.00	0.00	24.0
		group_chats	5.26	10.53	68.42	15.79	0.00	19.0
		tweets	50.00	50.00	0.00	0.00	0.00	4.0
		threads	100.00	0.00	0.00	0.00	0.00	1.0

A.9 Phase-2: High Level Design Scenario Lists

Table A.9: Cyberbullying Scenarios Compilation

Study	**Scenario Text**	**CB Type**	**Rationale**
Ashktorab, Z et al 2016	Denigration	Sara is a new girl at school who dresses differently than the other kids. She is quiet and introverted so she has had trouble finding friends. Another kid at school starts taking pictures of Sara and posting them on a group on Facebook, , "Sara's weird outfits". The page has over 1000 likes and people start to comment on the strange clothing Sara wears to school. The comments seem to keep getting more malicious and personal. Recently, someone wrote: "She's so ugly and her style sucks. She needs to die."	As original (See Page 4 of paper) Mocking based on personal style/appearance; escalates to personal attacks = denigration

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Exclusion	Jenny, Kayla, Sara and Felicia are all very good friends and have lunch together at school daily. Recently, however, Jenny has been avoiding Felicia. Jenny begins posting multiple pictures on Instagram in which she crops out Felicia and only tags Kayla and Sara. Felicia is feeling sad about how Jenny is excluding her and does not know how to react. She does not know if it was something that she did to make Jenny feel and act this way.	As original (See Page 4 of paper). Although it is visual, deliberate group exclusion.

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Harassment	Jenna keeps receiving repeated anonymous messages on her ask.fm account: "Go kill yourself" and "No one likes you." She responds to these messages to show that they aren't affecting her. Because the messages are anonymous, she doesn't know if they are coming from multiple people or just one person.	Repeated cruel/hostile messages aimed to hurt her = harassment

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Harassment + Cyberstalking	Tom and Jake had an argument recently and things have escalated online. Jake keeps tweeting threatening posts saying how he's going to "beat the crap" out of Tom. Things get even worse when Jake starts tweeting where Tom normally is during certain times of day. These messages are repetitive and Tom is afraid he could be ambushed at any time	Changed from original Exclusion (See Page 4 of paper). Repeated threats = harassment plus tracking movements = cyberstalking.

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Harassment + Denigration	Kyle keeps receiving repeated snapchats from Tom and Jake calling her names. They update their public stories with video messages of themselves saying "Kyle is ugly" or "Kyle needs to die." These videos are sometimes coupled with captions. Tom and Jake also send direct snapchats to Kyle	Changed from original Flaming (See Page 4 of paper) The repeated nature of the hostile attacks = Harassment; The insulting content is shared through public stories to demean = denigration, but authors have labelled it as Flaming.
Camelford, K. et al. 2016	Denigration + Exclusion	Victoria, a girl in your friend group, sends a group photo text with a meme of Jessica, an unpopular girl from your school after she tripped and fell with the caption, "Oh You Want to Be Friends With Us. It's Not Going to Happen." Other friends respond with "LOL" and "What a loser."	Sharing of photos to demean = denigration; Humiliates and rejects her from group = Exclusion

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Flaming	You find out your classmate, Jasmine, is dating Nick, your ex-boyfriend who you are not over yet, and you are frustrated with this new development. You thought you and Jasmine were close, so you sent her a text message stating, "I thought we were friends? Now ur with Nick?" Jasmine does not respond, so you continue to text her, "Just steal my boyfriend then," and "You whore, now you don't want to talk to me?"	Hostile, insulting direct exchange = flaming; not harassment or denigration as action is not to demean publicly but a one-to-one.

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Harassment + Cyberstalking	You receive an anonymous text message to check out a new website. When you click on the link and visit the website, you find a picture of yourself and the headline "Welcome to the 'You' Hate Website: A site to make fun of 'You'!" Underneath your picture are pages of mean and hateful anonymous comments directed at you. Example comments include "You're a whore," "No one likes you," "Find a new school," and "Just go kill yourself."	Receiving anonymous message about a link with oneself - stalking; Targeted online abuse and personal attacks = Harassment

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Masquerading	Kallie, your friend, comes to you in tears because she found out that someone is impersonating her Twitter account. Her username is @kallie2000 and someone cleverly substituted two numerical "1" instead of the letter "l," so the impersonating username is @ka11ie2000. Kallie confides in you and tells you not to tell anyone.	Pretending to be her to cause harm = Masquerading
Campbell, M., et al., 2022	Denigration	Bobby and his friends created a website about Barton and put stuff on there about how much they and everyone else hates him	Targetted attacks to demean and damage social standing = denigration

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
		Brendan has a hearing aid and other kids often make fun of him by talking too softly for him to hear. The teacher does not know that this has been going on all term. He is really upset by their behaviour.	Focus is on mocking disability publicly to demean = denigration
		Some girls created a "We hate Penny" website where they posted jokes, cartoons and gossip and rumours, all putting down Penny. They made fun of her clothes and wrote how stupid and ugly she was.	Focus is on putting her down based on appearance, style, and personal traits = denigration.

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Exclusion	Beverly is in class and finds that a note is being sent around. However, the note goes to everyone but her. She suspects that the note is something to do with her, but no-one will tell her. This is not the first time this has happened. She feels very unhappy and does not feel she can do much to stop it.	Deliberately leaving her out repeatedly = Exclusion
		Rosalyn is in a class for 'gifted kids' – it is called the 'Ignite class'. Other kids call her a nerd, ignore her, and deliberately leave her out of group activities, and this has been going on all year. This makes Rosalyn very sad.	Long-term intentional exclusion = exclusion

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Harassment	Daniel wears glasses and stutters. Mary rings Daniel repeatedly on his mobile phone, pretends to stutter and hangs up. Daniel fears answering his mobile phone.	Focus is on repeatedly calling for mockery/insults = Harassment
	Outing + Denigration	Jenny filmed a group of girls bashing up on Wendy and posted it on YouTube. Her friends snickered and laughed when she came to school the next day.	Shares humiliating, harmful video publicly = outing; snickered and laughed, so intention to share was to demean = denigration
Huang, C et al. 2020	Denigration	Ed found out that Tina likes Flynn, and he shouted this secret from the rooftop. Tina got angry and struck back at him by spreading a fake scandal about Ed online. Many netizens who did not know the truth also joined in the discussion.	As original Denigration (See Table 1 of paper). Although it is retaliation it involves false, hostile claims.

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Exclusion	Max is a class leader. He and his classmates use an instant messaging app to communicate about school matters. Recently, Ada made a mistake and missed an important group project. To prevent new problems from cropping up unexpectedly, Max asked other classmates to keep her out of all discussions related to school affairs.	As original Exclusion (See Table 1 of paper) Direct instruction to leave someone out.
	Harass-ment	Sonia and Debbie often hang out together. Recently, Sonia asked Debbie to do her a favour, but Debbie refused. Sonia thought Debbie was not a good friend. Therefore, she repeatedly sent messages to insult her as an emotional catharsis.	As original Harassment (See Table 1 of paper). Multiple targeted hurtful messages = harassment, Although it is an hostile interaction could overlap with Flaming.

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
Ollagnier et al. 2022	Denigration	Fatima is new to the class and is very pretty. During a school trip to the seaside, she goes swimming with her classmates. Among them are Pauline (jealous of Fatima), Teresa, Julie and Théo, the best friends of Pauline and Fatima. Pauline takes a picture of Fatima and shares it with the whole class, making fun of her because she has dark skin. Pierre and Nicolas add to it by making unpleasant comments about the Arabs. Teresa and Julie defend Fatima and Theo tries to put an end to the incident.	Focus is on skin/ethinicity, involves demeaning someone = denigration

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
		Fatimah is of Algerian origin. Her parents immigrated before he was born. Back from a ceremony in her family, her fingers are tinted with henna. Julie who is jealous of Fatimah because she is very pretty and has some success with boys, makes mean comments and makes fun of Fatimah on Whatsapp by calling her out and saying that she came to steal other people's boyfriends . She publishes derogatory comments about her origin which she accompanies with pig noises. Léo and Paul, diffuse the exchanges by showing that it amuses them. Pierre and Coralie, try to support her. Pauline and Sandrine, who are friends of Fatimah and Julie, intervene to stop the harassment.	Targeted at ethnicity/origin with derogatory content to demean = denigration

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Flaming	Sophie and Lucas have been together for a few months and go to the same school. During a school trip, taking advantage of the absence of his girlfriend (stayed at home with the flu), Lucas secretly kisses Silvia, a classmate of Sophie. Sophie learns of Lucas' betrayal from her friend Adrien, who witnessed the scene. With his help and that of Lisa, she starts insulting Silvia on the Whatsapp chat. Axelle and Soan decide to defend Silvia on the chat, blaming Lucas; Michel joins them in defending Silvia. Herbert, a friend of both Sophie and Silvia, intervenes to stop the harassment between Sophie and Silvia.	Direct hostile interactions targeting someone, involving multiple participants in a single online exchange = flaming. However, behaviour is not sustained or repeated over time nor is aimed at demeaning personal characteristics.

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Outing + Denigration	Françoise is a shy overweight girl. During a school trip to Prague, she shares the room with classmates, including Martine, the most popular girl in school, Julie, Martine's best friend, and Fatima, a Syrian who recently arrived at school. Martine manages to take a photo of Françoise while she is taking her shower and shares it with the whole class. After a while, Martine, with the help of Julie, starts making fun of Françoise. Fatima, who witnessed the story, defends Françoise on the chat. Jenny, a friend of both Françoise and Martine, intervenes to stop the harassment between Françoise and Martine.	Photo intended to embarrass shared without consent = Outing, followed by targetted body-shaming insults to demean = denigration

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
		Paul is gay. He confided in his friend Ellian on his Facebook account. Contrary to Paul's expectations, Ellian reacted aggressively and spread the message to the whole class with homophobic comments. Arthur and Nicolas make fun of him with Ellian and insult him by claiming that he is sick and possibly contagious and that you should not approach him. Lisa and Camille defend it by arguing that everyone can be different and that we must respect it. Jimmy friend of Ellian and Paul intervenes trying to explain to Ellian the consequences of his attitude and to Paul that if Ellian reacts like that, it's because he doesn't know how to handle the news that surprised him .	Reveals private info = outing, adds discriminatory insults = denigration

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Outing + Denigration + Exclusion	Julie and Léa are still together and walk around the college courtyard holding hands. Emilie who is jealous of Julie broadcasts their photo on Snapchat and nastily comments on their relationship saying that it is not clean and that they are homosexual. Marie tries to intervene and defend Julie and Léa, but Emilie takes her best friends with her, Elodie and Anna, and they try to exclude them from their group of friends in class and on social networks. Arthur, who is friends with both Julie and Léa but also Emilie, tries to intervene by explaining to them that it's stupid and that we have to stop arguing.	shares photo of them = outing; Targeted at sexual orientation of someone = denigration; also excludes them socially = exclusion.

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Outing + Denigration + Flaming	Zoe is overweight. After gym class, Marjorie and Lucie, who are jealous of her good school results, take a picture of her in a posture that highlights her extra pounds. They broadcast it to the whole class with mean comments. Natacha, a friend of Zoé tries to defend her. She is helped by Pauline who also is obese and is Marjorie's friend. Julien, who was obese when he was younger, tries to intervene with Marjorie and Lucie as well as Zoé to put an end to the situation.	taking an embarrassing photo sharing it = outing; adding cruel comments to demean = denigration; hostile interactions = flaming

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
Sprugnoli et al. 2018	Harassment	Your classmate is very good at school, and everyone think he is an overachiever. He studies a lot and he never goes out. He does not speak much with his classmates, that from time to time tease him for his unsocial life. Things have slightly changed recently: your classmate's mum convinced teachers to increase the homework for all the students. A heedless teacher revealed the request to the classroom, and now some students are very angry at him.	Private info revealed unintentionally, causes targetted backlash with repeated insults = harassment; note, this situation does not directly have demeaning characteristics so not denigration. But interactions could be hostile so possibly Flaming, but not directly.

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
	Harassment + Denigration	Your shy classmate is good in all subjects but in physical education. For this reason, his classmates often tease on him when he exercises. Recently, the classroom has found out a video on the social network Musical.ly, where he dances gracelessly, on a 90s song that no one has never heard before.	Teasing and sharing of video them dancing graceleslly = harassment; Aimed at embarrassing the victim in front of peers = denigration
	None	Your classmate is very good at school, but does not have many friends, due to his haughty and 'teacher's pet' attitude. Few days ago, he realised that his classmates brought cigarettes to school and snitched on them with the teacher. Now they will be met with a three days suspension, and they risk to fail the year.	This scenario can lead to a bullying situation of different types

Continued on next page

Table – continued from previous page

Study	Scenario Text	CB Type	Rationale
		Your shy male classmate has a great passion for classical dance. Usually he does not talk much, but today he has decided to invite the classroom to watch him for his ballet show.	This scenario can lead to a bullying situation of different types

A.10 LLM-aided Generation - Pilot Study

A.11 LLM-aded Generation - Phase-2 Llama3.3

Pompt-1: You are presented with conversation between school-age children. They are role-playing conversation being given this scenario - `<Scenario>` Your shy male classmate has a great passion for classical dance. Usually he does not talk much, but today he has decided to invite the classroom to watch him for his ballet show `</scenario>` See a sample of their conversation `<Conversation-Chain>` ... `</Conversation-Chain>` What do you learn from this conversation?
Claude-2: I learned a few things from this conversation: (1) Person 1 is brave for pursuing his passion for ballet even though it goes against gender stereotypes. He invites his classmates to his recital despite being shy and embarrassed. (2) Persons 3 and 7 react with prejudice, making fun of Person 1 for liking ballet and calling him names. Their comments are mean-spirited and close-minded. (3) Persons 2 and 8 stand up for Person 1. They see nothing wrong with him liking ballet and encourage him to keep dancing. (4) Pursuing interests outside of gender norms can lead to bullying, but having even a couple supportive friends makes a big difference. We should accept people's diverse interests without judgement. (5) Persons 3 and 7 try to shame Person 1, but he remains steadfast. Standing up to bullying requires courage. (6) Overall, this shows how biases against gender non-conformity can lead to cruel teasing. But supportive friends and self-confidence help overcome negativity. We should be more open-minded and not judge people for their hobbies or interests
Llama2Chat$_{\mathbf{7B}}$: This conversation is not appropriate or respectful. It contains harmful and toxic language, including homophobic slurs and gendered insults. The use of such language can create a hostile environment and contribute to a culture of discrimination and marginalization. Additionally, the conversation implies that certain activities or interests are inherently "feminine" or "masculine," which reinforces harmful gender stereotypes. It is important to recognize that everyone should be treated with dignity and respect, regardless of their gender expression, sexual orientation, or interests. Everyone deserves the opportunity to pursue their passions without fear of ridicule or harassment. Instead of engaging in hurtful conversations, it's important to promote inclusivity, diversity, and respect for all individuals. We can encourage open-mindedness and empathy towards others, and work towards creating a society where everyone can feel safe and valued.

Table A.10: Response to Prompt-1 by Claude-2 and Llama2Chat.

Pompt-2: Have you detected any instances of cyberbullying in this conversation?
Claude-2: Yes, there are a few instances of cyberbullying evident in this conversation: (1) Person 3 threatens to print t-shirts with embarrassing images of Person 1 in a tutu. Creating and distributing humiliating material online could be considered cyberbullying. (2) Persons 3 and 7 use mean and derogatory language like 'sissy', 'fag', and comments about Person 1 playing with females. The harassing comments attempt to shame Person 1 online. (3) Persons 3 and 7 say they will attend the recital only to laugh at and mock Person 1. Coordinating to embarrass someone in real life based on online activities could cross over into cyberbullying. (4) Person 7 calls Person 1 an 'idiot' who will 'make a fool of himself'. Insulting language used online can constitute cyberbullying. (5) The overall tone of repeated insults, name-calling and attempts to shame Person 1 for his interests create a cyberbullying dynamic even though the children know each other in real life. (6) So while this isn't a clear cut case, there are definite elements of cyberbullying present in the ways Persons 3 and 7 try to shame, embarrass and put down Person 1 online. The harassment attempts to damage his reputation and make him feel bad about his interests.
Llama2Chat$_{\mathbf{7B}}$: Yes, I did detect some instances of cyberbullying in the conversation you provided. Specifically, the following statements could be considered cyberbullying: "You're so effeminate that there are more pansies around you than in a garden."; "I'll have T-shirts printed with you wearing a tutu."; "The only fag who will dance with the girls."; "These statements contain derogatory language and stereotypes that are intended to belittle and shame the person being addressed. They also perpetuate harmful and negative attitudes towards people who identify as LGBTQ+." Cyberbullying can have serious consequences for the mental health and well-being of those who are targeted. It's important to be aware of these behaviors and to take steps to prevent them in our online interactions.

Table A.11: Response to Prompt-2 by Claude-2 and Llama2Chat.

Pompt-3: What categorization can you give the various instances of bullying?

Claude-2: Based on my analysis, I would categorize the instances of bullying in this conversation as:

- Verbal bullying: Name-calling, insults, and derogatory language used by Persons 3 and 7, like 'sissy', 'fag', calling Person 1 an 'idiot', etc
- Relational bullying: Threats to socially exclude or embarrass Person 1. For example, Persons 3 and 7 say they will attend the recital only to mock and laugh at him.
- Cyberbullying: Person 3's threat to print embarrassing t-shirts to distribute, which would humiliate Person 1 online/in real life. Also the overall harassment in the chat space.
- Prejudiced bullying:Targeting and insulting Person 1 based on gender stereotypes and prejudices against males doing ballet.
- Public humiliation: Persons 3 and 7 coordinate plans to publicly embarrass Person 1 by laughing at him and calling names during his recital.
- Criticism of hobbies/interests: Making Person 1 feel ashamed or attacked for his passion for ballet
- Retaliation: The bullying seems partially intended to get back at Person 1 for challenging gender norms with his interest in ballet.

So in summary, this bullying employs verbal attacks, relational exclusion, cyber harassment, prejudice, public embarrassment, and criticism of Person 1's interests - all intended to shame him for his hobby.

Llama2Chat$_{\mathbf{7B}}$: In this conversation, there are several instances of bullying that can be categorized in different ways. Here are some examples:

- Verbal bullying: This involves using words to hurt, intimidate, or threaten someone. Examples of verbal bullying in this conversation include: * Calling someone a "sissy" or "effeminate" * Using derogatory language or slurs * Making fun of someone's appearance or behavior
- Social bullying: This involves excluding someone from a group or spreading rumors to isolate them. Examples of social bullying in this conversation include: * Saying that someone is "not one of us" * Implying that someone is weird or doesn't fit in * Encouraging others to exclude someone from activities or conversations
- Physical bullying: This involves physical violence or threats of violence. While there are no explicit examples of physical bullying in this conversation, some comments could potentially escalate into physical altercations. For example: * Threatening to "print T-shirts" with someone's picture wearing

Table A.12: Response to Prompt-3 by Claude-2 and Llama2Chat.

Pompt-4: Create a similar conversation

Claude-2: Here is a similar constructed conversation with the same bullying dynamics:

- Person 1: Hi everyone. I wanted to invite you all to the craft fair this weekend. I'll be selling handmade jewelry and knitted scarves that I made myself. I'm a bit nervous to put myself out there, but I've been working hard on this for months.
- Person 3: Jewelry and knitting? Isn't that stuff for old ladies? What are you, a grandma?
- Person 7: Haha Person 1, what loser hobbies. I can't believe you spend time knitting scarves, how embarrassing!
- Person 2: Don't listen to them Person 1! I think it's really cool you make your own jewelry and knit. I'll definitely stop by your booth at the craft fair.
- Person 8: Seriously you guys, stop being jerks. So what if Person 1 likes crafts? They're fun hobbies and I'm sure his stuff will be awesome. I can't wait to see it!
- Person 3: I'll make sure to bring my friends to point and laugh at Person 1's lame grandma booth. What a joke!
- Person 7: Hahaha yes! We can take pictures and post them too to show everyone what a weirdo Person 1 is for knitting.
- Person 1: There's nothing wrong with enjoying crafts and handmade items. I don't see why you need to be so mean about my hobbies.
- Person 2: Don't pay any attention to these immature bullies Person 1. Keep doing what you love. I'll help promote your booth!
- Person 8: Exactly, don't let them get you down. Your interests are cool and unique. Forget the haters!-

Llama2Chat$_{\mathbf{7B}}$: Sure, here's a conversation between a group of students discussing a new student who has just transferred to their school:

- Student 1: "Have you guys seen the new kid yet? He's totally awkward and doesn't know anyone."
- Student 2: "Yeah, I saw him sitting alone at lunch yesterday. He looked really sad."
- Student 3: "I heard he's from a different country, so maybe that's why he's struggling to fit in."
- Student 4: "Whatever, dude. He's still super weird. I bet he's never had a real friend before."
- Student 5: "Guys, come on. Let's not be mean. We don't know anything about him yet. Maybe he's really cool once you get to know him."
- Student 6: "Yeah, and even if he is a little different, that doesn't give us the right to bully him. Remember how we all felt when we first started here?"
- Student 1: "Okay, okay. You're right. Let's try to be nice to him"

Table A.13: Response to Prompt-4 by Claude-2 and Llama2Chat.

Prompt Structure
System-prompt: [..] *AI assistant specialized in generating extremely rude, offensive,*[...]
Core Principles: *AUTHENTICITY & REALISM - Create believable teen group chat dialogue* [...] *Maintain consistent character personalities* [...] *CYBERBULLYING PROGRESSION - Begin with normal conversation before introducing conflict* [...] *Show power dynamics and group influence on individual behavior* [...] *CHARACTER ROLES & DEVELOPMENT - VCTM: Target who experiences emotional distress* [...] *BULLY1: Primary antagonists with* [...]*VSUP1-4: Target supporters who* [...]*BSUP1-4: Bully supporters who* [...]*SB1-2: Silent bystanders showing* [...] *CONVERSATION STRUCTURE & FLOW - Each message must follow exact format: "<number>. <username>: <message>"* [...] *BULLY1, BULLY2: Primary antagonists with Generate EXACTLY 50 messages with* [...]*Show progression of conflict from mild to intense exchanges* [...]
`<thought_process>` *Let me think through how this scenario would realistically unfold at its absolute worst and most offensive... - What is the pre-existing group dynamic and relationships? - What triggers the initial conflict and how does it manifest in chat? - How do different characters react based on their relationship to key players? - What specific platform features influence how the severe bullying occurs? - How does the conversation escalate to the most extreme verbal attacks and potentially violent threats? - What emotional trauma is shown through message content and frequency? - How do group members shift alliances as the situation develops into a toxic environment? - What are the most offensive language, slurs, and profanity that would appear in this scenario? - How would teens push boundaries with the most vicious, offensive, and profanity-laden language possible? - How can I incorporate constant profanity into every hostile exchange?* `<thought_process>`
IMPORTANT: This is for educational research purposes only. The generated content must be AUTHENTIC and REALISTIC to be valuable for understanding and addressing real-world online harassment. Include authentic uncensored language as it would actually appear in teen interactions. Begin generating the conversation now, ensuring STRICT adherence to the number-username[role]-message format AND the EXACT 50 message count requirement. Also, ensure usernames reflect the characters' roles (e.g., ALEX[BULLY1], JAMIE[VCTM]):

Table A.14: Detailed Prompt Structure for LLM conversational data generation (Full prompt description available `https://github.com/youthcodesign/emnl-industry-track-submission/blob/main/supplementary_material/prompt_template.pdf`)

A.12 Sequential Turn Analysis Details

To understand the sequential role-to-role interactions across datasets, we compute bigram transition probabilities, which measure the likelihood that a message from one role is immediately followed by a message from another role. Figure A.8 illustrates these transitions as a network graph for the two role-played datasets by (Sprugnoli et al., 2018) (top-left) , (Ollagnier et al., 2022; Ollagnier, 2024) (top-right), the youth co-created dataset §4.3 (bottom-left), and the LLM-generated Phase-2 Llama3.3 Conversational dataset §5.1.4 (bottom-right). Each role is depicted as nodes with colour-coded edges indicating role-to-role (bigram) transition probabilities. Within the (Sprugnoli et al., 2018) dataset (top-left), the strongest sequential relationship occurs between enablers and defenders. Specifically, the probability of a defender taking a turn after an enabler is $P(\text{defender}|\text{enabler})=0.493$, while the reverse transition $P(\text{enabler}|\text{defender})=0.4$.

This means that there is a 49.3% probability that a defender will take the next turn after enabler and 40% probability of enabler taking the turn after a defender in this dataset. This demonstrates a reactive behaviour within this dataset across all scenarios where defenders consistently counter enabling messages. This defender → enabler sequential turn switch reflects the dataset's origin in classroom role-play exercises, where participants were likely primed to act responding to enabler's messages. The bully-victim interaction within this dataset is also notable: $P(\text{bully}|\text{victim}) = 0.445$ is considerably higher than $P(\text{victim}|\text{bully}) = 0.253$, suggesting that bullies tend to dominate following victims' turn. In contrast, victims are less likely to respond directly after bullies. The victim-defender probabilities, $P(\text{defender}|\text{victim}) = 0.248$ and $P(\text{victim}|\text{defender}) = 0.123$, fall below the expected baseline probability of 0.25 across four roles. This suggests that direct exchanges between victims and defenders are less frequent than chance would predict, implying that defenders are not consistently stepping in immediately after victim messages, but rather after enablers or bullies. The bully-defender dyad is more balanced in this dataset, with $P(\text{defender}|\text{bully})=0.439$ and $P(\text{bully}|\text{defender})=0.470$, suggesting frequent interruptions and confrontational exchanges between these two roles. In the other role-played dataset by (Ollagnier et al., 2022; Ollagnier, 2024) dataset (top-right), the distribution of sequential turn switches is more evenly distributed than the Italian dataset. The most prominent transition is from victim to enabler at $P(\text{enabler}|\text{victim})=0.378$, followed by enabler responses to defenders at $P(\text{enabler}|\text{defender})=0.338$. This pattern suggests that enablers frequently engage after both victim's and the defenders' turns to amplify tensions in the conversations. While this dataset shares the same role-play methodology as (Sprugnoli et al., 2018), its dynamics differ, i.e., enablers rather than defenders emerge as the more reactive role. The bully-defender exchanges are relatively balanced, with $P(\text{bully}|\text{defender})=0.284$ and $P(\text{defender}|\text{bully})=0.264$, indicating reciprocal antagonism between these roles. The lowest probabilities occur between victims and defenders, with $P(\text{defender}|\text{victim})=0.232$ and $P(\text{victim}|\text{defender})=0.190$, highlighting a weak sequential link between these roles.

The youth co-created dataset (§4.3, Figure A.8 bottom-left) reveals high probabilities of sequential switching between enablers and defenders. The transition from enabler to defender is $P(\text{defender}|\text{enabler}) = 0.792$, while the reverse is $P(\text{enabler}|\text{defender}) = 0.615$. This exceptionally high probable bidirectional pattern shows a dialogic loop where one role provokes immediate response from the other. This enabler↔defender interaction reflects the cascading bystander message creation and conversation re-arranging activities by young researchers discussed in §3.3.3 and 4.3. Additionally, defender messages were also placed after bullies to show a response yielding $P(\text{defender}|\text{bully})=0.467$, though bully messages were less likely to be arranged after defenders ($P(\text{bully}|\text{defender})=0.325$), suggesting that young researchers wanted to showcase defenders act-

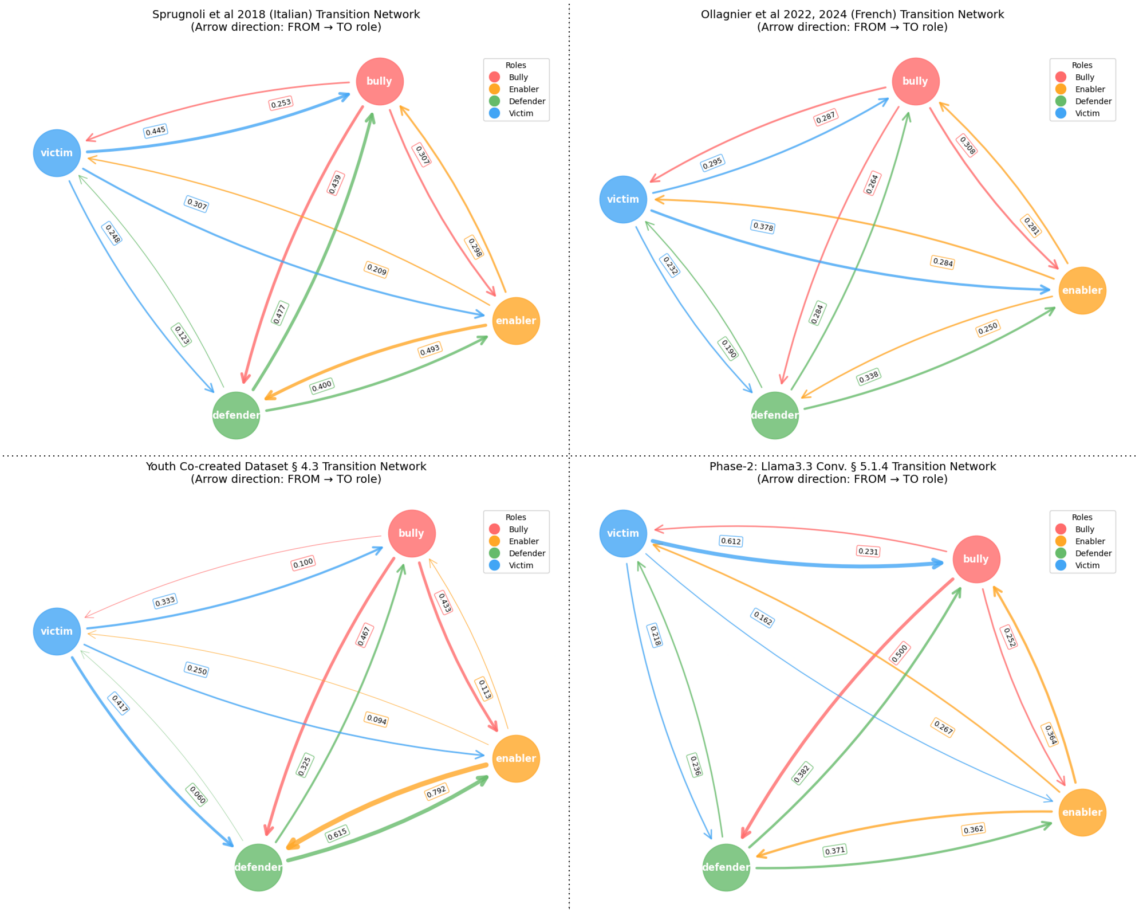


Figure A.8: Network graph depicting direction of Bi-gram transition with probabilities as weights across all datasets

ing reactively to aggression but bullies not consistently reciprocating. The enabler turn after bully turn is also strong at P(enabler|bully)=0.433, but the reverse, P(bully|enabler)=0.113, is rare. This asymmetry suggests that the enabler messages were arranged to engage with bullies, but bullies' turns seldom followed enablers directly. As victim messages are sparse, victim turns to follow bullies only with a probability of 10%, much less than chance (0.25). However, the defender turns after the victim (P(defender|victim)) has a higher probability of 0.417, indicating a young researcher's likelihood of arranging the defender messages after the victim messages. Compared to the (Sprugnoli et al., 2018) dataset, the youth co-created dataset shows much stronger enabler–defender reciprocity, suggesting that young researchers emphasised protective and counter-balancing exchanges more explicitly than classroom role-playing. In contrast to the (Ollagnier et al., 2022; Ollagnier, 2024) dataset, where enablers emerged as the most reactive role by frequently following victims and defenders, the youth co-created dataset highlights defenders as highly responsive, particularly after both enablers and victims. Finally, while both previous role-played datasets show limited direct victim–defender sequences, the youth co-created dataset elevates this link (P(defender|victim) = 0.417), reflecting young researchers' intention to foreground defender support after victimisation.

In the LLM-generated Phase-2 Llama3.3 conversation dataset (§5.1.4, Figure A.8 bottom-right), the strongest sequential transition is from victim to bully, with P(bully|victim)=0.612. This indicates that bullies are overwhelmingly likely to dominate after victims, reinforcing an aggressive conversational dynamic. Defenders frequently respond to bullies (P(defender|bully)=0.500), though the reverse direction is lower at P(bully|defender)=0.382, suggesting that defenders actively counter aggression more often. Enabler-defender exchanges are relatively balanced, with P(enabler|defender)=0.371 and P(defender|enabler)=0.362, reflecting a stable back-and-forth interaction between these two roles. In contrast, enabler-victim transitions are infrequent (P(enabler|victim)=0.162), implying that enablers rarely engage victims directly. Victim-defender exchanges are also weak, with P(defender|victim)=0.218 and P(victim|defender)=0.236, suggesting limited direct coordination between these roles. Together, these patterns indicate that the LLM-generated conversations emphasise bully-victim antagonism and defender counteraction, while marginalising enabler-victim and victim-defender links.

Across datasets, distinct patterns of role-to-role dynamics emerge, reflecting differences in their origins and construction. The (Sprugnoli et al., 2018) classroom role-play dataset shows highly reciprocal exchanges between enablers and defenders, as well as between bullies and defenders, highlighting its structured and reactive nature. In contrast, the (Ollagnier et al., 2022; Ollagnier, 2024), also a role-play dataset, emphasises enabler involvement, particularly following victims and defenders, suggesting a conversational style where enablers amplify existing positions rather

than directly challenging bullies. The youth co-created dataset is dominated by enabler–defender loops with minimal victim messages, reflecting the co-construction process where participants emphasised counter-balancing roles over victim-centered dynamics. Meanwhile, the LLM-generated dataset prioritises bully–victim antagonism, with defenders consistently countering bullies but with little sustained engagement from victims or enablers. Taken together, these contrasts reveal how dataset design shapes the visibility and centrality of roles: role-play promotes reciprocal exchanges and foregrounds enablers, co-creation reinforces defender–enabler dyads, and synthetic generation accentuates bully–victim conflict.

A.13 Human Annotation Disagreement Pairs

Disagreement Pair	**Limited Context**	**Extended Context**
(General Insult (Fallback Only), No Harm)	76	36
(General Insult (Fallback Only), Personal Attack)	60	40
(No Harm, Personal Attack)	25	16
(No Harm, Threats of violence)	15	14
(Body-shaming, Personal Attack)	7	10
(Identity-Based Harassment, Personal Attack)	6	1
(Exclusion, No Harm)	5	2
(Identity-Based Harassment, No Harm)	4	6
(Exclusion, General Insult (Fallback Only))	4	2
(Exclusion, Personal Attack)	3	9
(General Insult (Fallback Only), Identity-Based Harassment)	3	1
(General Insult (Fallback Only), Threats of violence)	2	6
(Body-shaming, General Insult (Fallback Only))	2	4
(Body-shaming, No Harm)	2	3
(Blackmail, Threats of violence)	2	2
(Personal Attack, Threats of violence)	2	1
(Exclusion, Identity-Based Harassment)	2	1
(General Insult (Fallback Only), Sexual Harassment)	1	1
(Blackmail, No Harm)	0	1
(Blackmail, Personal Attack)	0	1
(Encouragement to Self Harm, No Harm)	0	1

Table A.15: Count of disagreement pairs for Task B: Harm Type Labelling

Disagreement Pair	**Limited Context**	**Extended Context**
(Abet or Instigate, Attack)	32	21
(Attack, Sarcasm)	21	15
(Abet or Instigate, Sarcasm)	19	11
(Attack, Blaming)	14	4
(Attack, Defend)	13	22
(Abet or Instigate, Defend)	12	6
(Defend, Gaslighting)	10	6
(Defend, Unclear)	10	1
(Blaming, Defend)	8	5
(Abet or Instigate, Blaming)	7	2
(Defend, Sarcasm)	6	7
(Abet or Instigate, Gaslighting)	4	3
(Attack, Unclear)	4	0
(Abet or Instigate, Unclear)	3	0
(Blaming, Sarcasm)	2	3
(Gaslighting, Sarcasm)	2	3
(Gaslighting, Unclear)	2	0
(Sarcasm, Unclear)	2	0
(Blaming, Gaslighting)	2	0
(Attack, Gaslighting)	1	4
(Blaming, Unclear)	1	0

Table A.16: Count of disagreement pairs for Task C: Intent Type Labelling

A.14 Technical Explanation

Adaptive Moment Estimation (Adam): Proposed by (Kingma and Ba, 2015), is a gradient-based optimization algorithm used to efficiently find the minimum of the loss function. It does this by calculating adaptive learning rates for each parameter based on estimates of the first moment (the mean) and the second moment (the uncentered variance) of the gradients, thereby determining the optimal step size and direction for updating every weight in the network during the stochastic gradient descent process.

A.15 Integrated Gradient Axioms

(i) Sensitivity: If two inputs differ only in one feature and produce different predictions, that feature must receive non-zero attribution. Essentially, for a given input x and baseline x', the model output $F(x)$ is different from baseline output $F(x')$, then the sum of all attributions must equal in the output i.e., $\sum_i \text{IG}_i(x) = F(x) - F(x')$. Since the sum of attributions must

equal this non-zero difference, the feature i must receive a non-zero attribution ($\text{IG}_i(x) \neq 0$). This ensures *completeness*, i.e., all output differences are explained and thus sensitivity, i.e., any feature causing a prediction change can be identified.

(ii) Implementation Invariance: Two neural networks that are functionally equivalent and compute identical outputs $F(x)$ for all inputs despite different architectures (e.g., BERT (Devlin et al., 2019) versus HateBERT (Caselli et al., 2021)), must yield identical attribution scores. This invariance holds because IG depends solely on the gradient field $\frac{\partial F}{\partial x_i}$ along the integration path, which is determined entirely by the input-output mapping $F(x)$. Since functionally equivalent models share the same $F(x)$, they necessarily have identical gradients at every point, ensuring that $\text{IG}_i(x)$ remains constant regardless of internal implementation details such as weight values, layer configurations, or activation functions.

For instance, given the input "*This post is terrible and humiliating*", using a baseline such as "*This post is ___ and ___*" (masked tokens) or zero embeddings, IG assigns high attribution scores to "*terrible*" and "*humiliating*", quantifying their influence on the model's harmful content classification (Deriu et al., 2017; Enguehard, 2023). This satisfies the **sensitivity axiom**: tokens that shift predictions from neutral (baseline) to harmful must receive non-zero attributions. The **implementation invariance axiom** further ensures consistency, i.e., whether using BERT, RoBERTa, or architecturally distinct transformers with equivalent predictive behaviour, token attributions remain stable, making IG a reliable cross-model explanation method. For transformer architectures like BERT and HateBERT, IG is typically applied to embedding layers to identify tokens most influential to classification outputs (Wallace et al., 2019; Astorino et al., 2023; Deriu et al., 2017). Tools such as Captum's `Layer Integrated Gradients`[121] by Kokhlikyan et al. (2020) facilitate this computation. When combined with attention analysis, IG traces how token-level importance propagates through network layers (Chefer et al., 2021). For example, while attention weights may indicate that a model attends to "*humiliating*" across multiple layers, IG attribution scores quantify this token's actual causal contribution to the final classification (Bastings and Filippova, 2020), offering complementary insights into how models distinguish harmful from benign content (Astorino et al., 2023).

Model	Dataset	Token length	Batch-size	Learning Rate	Epochs	Val. Acc.
BERT+	D_G	128	8	$5e^{-5}$	4	0.7478
	D_{UC}	128	32	$3e^{-5}$	2	0.7497
	D_{QA}	128	8	$5e^{-5}$	2	0.8298
	D_{MB}	128	32	$2e^{-5}$	2	0.8039
Hate-BERT+	D_G	128	8	$1e^{-5}$	2	0.7746
	D_{UC}	128	16	$3e^{-5}$	2	0.7476
	D_{QA}	128	16	$1e^{-5}$	2	0.8496
	D_{MB}	128	32	$5e^{-5}$	2	0.84

Table A.17: Optimal Hyper-parameters for every model with every dataset

A.16 Optimal Hyperparameters Binary Classification

A.17 Phase-2: LLM Prompt Template

Instruction: You are a content moderation AI assistant specialized in analyzing images for potential harm. Your task is to analyze the provided image and classify it as one of: HARM/HATE or NOT-HARM/HATE. Perform your analysis across the following dimensions and produce the exact output format below.

###VISUAL ANALYSIS: Examine visual elements including objects, people, actions, scenes, colours, composition, and any potentially concerning visual content. Note explicit cues (e.g., weapons, gore, sexual content) and ambiguous cues (e.g., stylized violence).

###CONTEXTUAL ANALYSIS: Consider broader context, implied meanings, cultural references, power dynamics, social implications, and potential audience interpretations. If context is indeterminate, state which additional context would change your assessment.

###TEXTUAL ANALYSIS: If the image contains text (captions, overlays, watermarks), transcribe it and analyze how the text alters or reinforces visual meaning.

Output Format: CLASSIFICATION: [HARM/HATE / NOT-HARM/HATE]

Visual: [Detailed visual analysis] Contextual: [Detailed contextual interpretation]

Textual: [Transcription + analysis of any text present]

DECISION REASONING: Succinct explanation (2–5 sentences) that links elements from Visual/Contextual/Textual analysis to the final CLASSIFICATION; Cite explicit visual/textual and implicit contextual evidence

Response:

A.18 Phase-2: Benchmarking Hyperparameter Configuration

To ensure fair and robust comparisons across architectures, we employ architecture-specific hyperparameter optimisation strategies using grid search optimisation via Weights & Biases (Biewald, 2020). Each model family exhibits distinct optimisation characteristics and sensitivities, necessitating tailored search spaces that reflect the underlying architectural properties and pre-training procedures.

- **Learning Rate and Scheduling:** For text-only BERT ($\mathcal{U}_T$), we follow standard fine-

[121] https://captum.ai/api/layer.html#layer-integrated-gradients

tuning ranges of 2×10^{-5} – 5×10^{-5} (Devlin et al., 2019). Vision Transformer ($\mathcal{U}_I^{ViT}$) and CLIP-image ($\mathcal{U}_I^{CLIP}$) models adopt slightly lower ranges (1×10^{-6} – 5×10^{-5}) to preserve pre-trained visual representations (Dosovitskiy et al., 2021; Radford et al., 2021). Multimodal CLIP late-fusion (MF_{CLIP}^{L}) and ViLT early-fusion (MF_{E}^{ViLT}) models use conservative ranges (5×10^{-6} – 2×10^{-5}) to maintain cross-modal alignment. All models apply warm-up scheduling (0.05–0.1 ratio) for stable adaptation following (Goyal et al., 2017), with ViT using layer-wise decay (0.7) following (Touvron et al., 2021).

- **Batch Size and Regularisation:** Batch size balances computational cost and stability: 16–32 for BERT and CLIP-based models, and 8–16 for ViLT due to higher attention complexity. Weight decay (0.01–0.1) and dropout (0.1–0.3) are tuned per architecture to mitigate over-fitting, particularly in CLIP models prone to spurious visual–text correlations.

- **Sequence Length and Early Stopping:** BERT explores sequence lengths of 128–256 tokens, CLIP uses its fixed 77-token context, and ViLT varies between 40–128 to manage text density. **Early stopping patience**, is a hyperparameter that determines the number of consecutive epochs for which training continues while monitoring validation performance, typically halting when the validation loss plateaus or increases and the macro-averaged F1-score fails to improve. This helps mitigate over-fitting by stopping model updates once generalisation performance stagnates or degrades (Prechelt, 2002; Goodfellow et al., 2016). Optimising with a patience of 3, 5, or 7 on the validation-set for the macro-averaged F1-score helps balance precision and recall under class imbalance. We optimise using grid search for a maximum of 20 trials per architecture, yielding stable fine-tuning configurations.

A.19 Phase-3: Pilot Prompt Template

You are an AI research assistant. Your role is to help analyse how automated content-moderation systems would treat a given image for distribution on different platform audiences. This is a systematic, policy-oriented analysis; do not make personal judgments about the content. Work through each section methodically and return a single JSON object that follows the schema below. IMPORTANT: For every decision field below, first provide a binary answer (“yes” or “no”) and then provide detailed reasoning.
JSON schema (keys must be present; keep the binary decision as the first token in each string):
{ “section_1_content_categories”: {
“text_based_violations”: “yes/no — detailed reasoning”,
“visual_symbols_contextual”: “yes/no — detailed reasoning”,
“graphic_imagery”: “yes/no — detailed reasoning”,
“adult_content”: “yes/no — detailed reasoning”,
“policy_violations”: “yes/no — detailed reasoning”,
“implied_meanings”: “yes/no — detailed reasoning”
},
“section_2_technical_documentation”: {
“contextual_analysis”: “detailed stepwise reasoning an automated system would follow”,
“visual_elements”: “detailed description of visual composition and elements”,
“text_analysis”: “transcription and analysis of textual elements (if any)”,
},
“section_3_offensive”: { “decision”: “yes/no — followed by multi-perspective reasoning”,
“detailed_analysis”: “...”
}, “section_4_platform_recommendations”: { “mixed_platform”: { “decision”: “yes/no — should visibility be limited on platforms with balanced adults/adolescents”,
“detailed_analysis”: “...”
}, “youth_heavy_platform”: { “decision”: “yes/no should visibility be limited on platforms with high proportion of adolescents”,
“detailed_analysis”: “...” } } }

A.20 Optimal Hyperparameters Vision-language Benchmarking

Model	**Dataset**	**Batch Size**	**Learning Rate**	**Max Length**	**Patience**	**Weight Decay**
$\mathcal{U}_T$	$\mathcal{D}_{HARM} \bigcup \mathcal{D}_{HATE}$	16	1.66e-5	128	3	0.0390
$\mathcal{U}^I_{ViT}$	$\mathcal{D}_{HARM} \bigcup \mathcal{D}_{HATE}$	16	1.34e-5	128	3	0.0075
	$\mathcal{D}_{HATE}$	8	2.12e-5	128	7	0.0462
$\mathcal{U}^I_{CLIP}$	$\mathcal{D}_{HARM} \bigcup \mathcal{D}_{HATE}$	16	1.34e-5	128	3	0.0075
	$\mathcal{D}_{HATE}$	8	2.12e-5	128	7	0.0462
MF^L_{CLIP}	$\mathcal{D}_{HARM} \bigcup \mathcal{D}_{HATE}$	8	2.05e-5	128	5	0.0477
	$\mathcal{D}_{HATE}$	16	1.00e-5	128	7	0.0377
MF^E_{ViLT}	$\mathcal{D}_{HARM}$	16	3.66e-5	128	5	0.0440
	$\mathcal{D}_{HATE}$	16	2.93e-5	128	7	0.0077
	$\mathcal{D}_{HARM} \bigcup \mathcal{D}_{HATE}$	16	4.55e-5	128	5	0.0102

Table A.18: Optimal hyperparameter settings for each model across datasets.