Received: 30 April 2022 | Revised: 5 March 2023 | Accepted: 9 March 2023

DOI: 10.1111/exsy.13292

ORIGINAL ARTICLE

Expert Systems WILEY

How interesting and coherent are the stories generated by a large-scale neural language model? Comparing human and automatic evaluations of machine-generated text

Dominic Callan | Jennifer Foster

School of Computing, Dublin City University, Dublin, Ireland

Correspondence
Dominic Callan and Jennifer Foster, School of Computing, Dublin City University, Dublin, Ireland.
Email: dominic.callan24@mail.dcu.ie and jennifer.foster@dcu.ie

Abstract

Evaluation of the narrative text generated by machines has traditionally been a challenge, particularly when attempting to evaluate subjective elements such as interest or believability. Recent improvements in narrative machine text generation have been largely driven by the emergence of transformer-based language models, trained on massive quantities of data, resulting in higher quality text generation. In this study, a corpus of stories is generated using the pre-trained GPT-Neo transformer model, with human-written prompts as inputs upon which to base the narrative text. The stories generated through this process are subsequently evaluated through both human evaluation and two automated metrics: BERTScore and BERT Next Sentence Prediction, with the aim of determining whether there is a correlation between the automatic scores and the human judgements. The results show variation in human evaluation results in comparison to modern automated metrics, suggesting further work is required to train automated metrics to identify text that is defined as interesting by humans.

KEYWORDS
evaluation, machine-generated text, natural language generation, transformers

1 | INTRODUCTION

Given a suitable prompt, autoregressive Transformer models (Brown et al., 2020; Radford et al., 2019; Vaswani et al., 2017; Zellers et al., 2019), trained to produce the most likely continuation of a sequence of tokens, can be used to generate text that is fluent and grammatical Clark et al. (2021). Such models can be used to generate news articles, answer questions, take part in a dialogue, produce summaries, and translate from one language to another. They can also be used to generate *narrative text* or *stories*. In this article, we focus on the task of evaluating the quality of machine generated stories or narrative text.

Many challenges exist in the evaluation of machine generated text. Automatic metrics which compare the system output to a gold standard or ground truth reference are commonly used. This comparison can take the form of simple word/character n-gram overlap metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) or metrics based on vector similarity such as BERTScore (Zhang et al., 2020a) and MoverScore (Zhao et al., 2019). The latter are better at accounting for paraphrases, that is when the output and the reference have a similar meaning but use

Abbreviations: BERT, bidirectional encoder representations from transformers; GPT, generative pre-trained transformer; NLG, natural language generation.

different words. All these metrics work best when more than one reference is available for comparison, because, for many generation tasks, there is more than one 'correct' system output. For story generation, the space of acceptable outputs becomes too large for reference-based evaluation to be feasible. Given a story prompt, there are innumerable story continuations which might be considered good stories, and evaluating in this manner does not allow for the possibility of correct but novel generation (Akoury et al., 2020; Purdy et al., 2018; Roemmele et al., 2017). Therefore, reference-free evaluation is needed.

Human evaluation is often superior to automatic evaluation because human language is designed for human communication and people are able to detect semantic differences and nuances that cannot be captured by sometimes crude and brittle automatic metrics. However, when asking people to judge the quality of an automatically generated story, we need to think about what constitutes quality. What makes a good story? In this article, we ask people to judge automatically generated stories in as simple a way as possible—does the story make sense (coherence) and do they want to keep reading (interest)? The human evaluation takes the form of a survey in which participants are presented with short narrative passages generated by an auto-regressive transformer language model (Black et al., 2021; Brown et al., 2020). They are asked to evaluate how interesting the prompt and story are, how coherent the story is and whether it is related to the prompt. The prompts are from a reddit.com 'writing prompts' dataset Fan et al. (2018).

Automatic evaluation, although difficult to get right for natural language generation (NLG) tasks, still has the benefit of being useful for testing intermediate versions of a system. For story generation, it would be useful to have an automatic metric that can tell us something about the quality of a story. Given the two criteria that we explore with the human evaluation, that is coherence and interest, we employ two automatic metrics in an attempt to measure the former. We investigate the correlation between these automatic metrics and the human judgements, and find only a low positive correlation.

The article is organized as follows: in Section 2 we discuss in more detail the problems of defining story interest and evaluating machine-generated stories; in Section 3, we describe our story generation model and our human and automatic methods for evaluating the generated stories; in Section 4, we present the findings of our study, before concluding and discussing potential future work in Section 5.

2 | BACKGROUND

Story generation differs significantly from other machine text generation challenges; rather than focusing on word overlap with an input or reference text as the metric for success, developing a believable narrative text requires composing coherent natural language texts that describe a sensible sequence of events (Yao et al., 2019). A 'good' or successfully generated story is a subjective idea. There are many criteria that should be considered, with the result that the evaluation of stories is a difficult problem that is relatively understudied (Lowe et al., 2017).

Challenges exist in evaluating machine-generated narrative text, both in human evaluation and through automatic metrics. The problems associated with one form of evaluation drives progress in the other. Chaganty et al. (2018) note that the many problems with automated evaluation metrics motivate the need for human evaluation, and Clark et al. (2021) comment that the absence of good ways of encoding aspects of what constitutes human quality output means that we must rely on human evaluation of our models. At the same time, the many challenges around reliable and scalable human evaluation have driven the development of automated evaluation systems. In this section, we first discuss the challenges associated with each type of evaluation (Sections 2.1 and 2.2), before going on to discuss the challenges associated with evaluating stories (Section 2.3).

2.1 | Human evaluation of machine generated text

NLG has long been identified as a difficult and complex area to evaluate accurately (Howcroft et al., 2020). Human evaluation is still considered to be the benchmark for evaluating machine generated outputs, noted as being crucial for developing successful NLG systems (van der Lee et al., 2021) and as being the gold-standard for evaluation (Hashimoto et al., 2019). A goal of NLG is to produce fluent outputs that can be read by laypeople (Yao et al., 2019); it is suitable therefore that this same group of 'laypeople' review the output where possible.

However, undertaking human evaluation of machine generated text systems also involves many challenges. Human interaction can be slow, is expensive and often hard to scale up (Chaganty et al., 2018; Lowe et al., 2017); in particular, the high cost is an influence on progress towards building reliable automated evaluators. Purdy et al. (2018) observe that the cost of human evaluation represents a bottleneck to AI research on story generation. Training of evaluators on what to expect and setting context and expectations can help them to focus on specific features of the text, which can be necessary given a tendency for humans to focus on form and fluency ahead of content (Clark et al., 2021).[1]

Human evaluators also have no way of evaluating diversity; they can assign a score to the quality of a text that has been produced without knowing whether it has simply been reproduced from the training data (Hashimoto et al., 2019). An automated evaluation system that could evaluate true novelty would bring value in these instances; a machine-generated story that did not directly plagiarize the training data may achieve a higher score to reward this. Hashimoto et al. (2019) describe this evaluation of diversity as crucial for creative, open-ended tasks.

Results of human evaluations are not always repeatable (Celikyilmaz et al., 2020) and there is little standardization across evaluations in the NLG space. Howcroft et al. (2020) suggest that there is need to standardize experimental design and terminology in the field to be able to compare human evaluations across different studies, as there is great inconsistency in how these human evaluations are run.

When crowdsourcing human evaluations, Celikyilmaz et al. (2020) highlight issues with using sources like Amazon Mechanical Turk, especially when the task is to evaluate longer text sequences. These workers are typically more used to evaluating microtasks and may be less experienced with evaluating stories. Strong clear guidelines and instructions need to be issued to maximize the effectiveness of these evaluations. Lowe et al. (2017) however warn that there must be a balance, as too much instruction can introduce bias. van der Lee et al. (2021) caution that there is a risk of inadvertently recruiting bots or participants who want to get paid for as little work as possible.

2.2 | Automatic evaluation of machine generated text

Automated evaluation has traditionally proven difficult in NLG. Text generation can go wrong in different ways while still receiving the same scores on automated metrics (van der Lee et al., 2021). Even modern automated systems do not have the same common-sense approach that good human evaluators have (Clark et al., 2021). Many automatic metrics exist currently. BLEU (Papineni et al., 2002) has traditionally been used in NLG systems to evaluate word overlap, however it is not a suitable metric for measuring the success of developing narrative text. Chaganty et al. (2018) note that while BLEU is cheap to run, it correlates poorly with human judgement. By rewarding word overlap, BLEU assigns a positive value to repetition, an element of machine text-generation that is to be avoided in story generation. Howcroft et al. (2020) reiterate this, however, noting that whilst it is true that BLEU correlates poorly, the lack of an alternative means that it is still widely used for many text-evaluation tasks. As a metric, BLEU breaks down when the space of allowable outputs is large, as in open-ended generation, as with prompts and stories (Yao et al., 2019). Other metrics have emerged. BLEURT (Sellam et al., 2020) is a bidirectional encoder representations from transformers (BERT)-based evaluation metric that is fine-tuned on synthetically generated sentence pairs using automatic evaluation scores such as BLEU. It is then further fine-tuned on machine generated texts and human written references using human evaluation scores and automatic metrics as labels. The Automatic Dialogue Evaluation Model (ADEM) proposed by Lowe et al. (2017) is a model-based evaluation that is learned from human judgements. It is mainly used for evaluating dialogue generation and is shown to correlate well with human judgement. Hashimoto et al. (2019) propose Human Unified with Statistical Evaluation (HUSE), focusing on open ended text generation tasks, including story completion. This model differs to ADEM by combining statistical evaluation and human judgements in a single model which is trained to distinguish machine text from human text in an attempt to evaluate both text quality and diversity. Like HUSE, MAUVE (Pillutla et al., 2021) is a statistical metric that attempts to capture the difference between human and machine generated text, but unlike HUSE, it is fully automated and not trained on human judgements.

2.3 | Evaluating stories

To evaluate the extent to which subjective attributes like interest, creativity or believability are applicable in machine generated text, certain criteria must be defined as metrics. In their in-depth study of human evaluations of automatically generated text, van der Lee et al. (2019) reported that the most used metrics in these types of studies were fluency, naturalness, quality, and meaning-preservation, but ultimately, they note that the criteria chosen should depend on the specific task. Gatt and Krahmer (2018) produce a similar list, also including 'Human-likeness' and 'Genre compatibility'.

Celikyilmaz et al. (2020) discuss certain attributes that should be present in machine generated text, including overall style, formality, or tone of the generated text. They add that there should be a 'typicality' to the generated text, meaning that it should be the type of text that we often see. They note that Lexical Cohesion[2] is one of the most used metrics to evaluate story generation. Words in the generated sentence should be semantically related to the words in the story context (Roemmele et al., 2017), and the extent of this can dictate the score that a generated story receives. Roemmele et al. (2017) also observe that 'style' matching, using a similarity comparison of the number of adverbs, adjectives, conjunctions, determiners, nouns, pronouns, prepositions, and punctuation tokens in the input and the output is an important evaluation metric, as a consistent writing style from the input through to the output improves the naturalness of the generated story.

Accuracy is of less concern for story generation, as their output cannot usually be judged by fidelity to an identifiable, external input (van der Lee et al., 2021). Grammaticality and fluency are not significant problems with modern transformer-based systems in comparison with older systems—the errors are instead often semantic or narrative (Yao et al., 2019). People can easily recognize non-sequitur sequences of events or conclusions, even when they are grammatical (Purdy et al., 2018).

The difference between well written, coherent text, and interesting text is difficult to define. Generating text that simply describes a sequence of events alone is not enough for it to be considered interesting and coherent (McIntyre & Lapata, 2009). A significant challenge is developing a narrative, theme or plot that will be accepted by humans as genuine and believable. McIntyre and Lapata (2009) give the example that if a character steals something and then runs away, a logical expectation from the listening audience might be that the character is caught;

the human listener can recognize immediately when something does not make sense sequentially in a narrative, even though they often do not realize they are doing this.

3 | METHODOLOGY

In this section, we describe how we generated the stories to be used in this study (Section 3.1) and how we evaluated these stories (Section 3.2).

3.1 | Generating the stories

3.1.1 | Text generation system

Vaswani et al. (2017) introduced the Transformer as a solution to certain problems that existed in using recurrent neural networks, including their inability to deal with long sequences in story generation. Transformer models dispense with a recurrent architecture, making use solely of an 'attention' function and parallelised processing. The attention function allows tokens to 'attend' to each other and helps to identify for each token how relevant other tokens in the input and/or output sequence are.

For this study, our text generator is a generative pre-trained transformer (GPT) (Brown et al., 2020) Transformer language model which is trained to predict the next token in a sequence, where every token can only attend to context to its left. Licencing costs prevented the use of the GPT-3 model. The GPT-Neo 2.7B parameter transformer model is used instead. Developed by EleutherAI, it is designed to be an open-source replication of Open-AI's GPT-3 architecture (Black et al., 2021). The GPT-Neo model is trained on 'The Pile' dataset, an 825GB diverse open-source English text corpus targeted at training large scale data models (Gao et al., 2020). The Pile is made up of 22 smaller datasets, including BookCorpus2, YouTube closed-captions, Project Gutenberg, and English Wikipedia.

3.1.2 | Input prompts

The dataset used as a source of input prompts to the text generation system is a set of prompts taken from the reddit.com 'writing prompts' dataset, introduced by Fan et al. (2018). The themes of the prompts vary, although they are often centred around genres such as fantasy or sci-fi. They are used as inputs to the transformer model which generates more text (the story) as output. The average prompt length is 147 characters or 27 words. The shortest is eight words and the longest is 56. In the event of there being more than one sentence within the prompt, it is treated as one sequence and a single input to the model instead of two separate sentences.

3.1.3 | Output stories

800 stories were generated by GPT-Neo, given 800 input prompts. Given that the focus of this analysis is on narrative style text, when either the prompts or the stories were of a non-narrative nature[3] these were excluded. From the remaining corpus of narrative-style stories, 100 prompt-story pairs were chosen at random for evaluation. The average story length is 77 words, the longest has 96 words and the shortest has 54. A cap of 400 characters was used and the average character count is 381 characters. This cap was implemented both as a method of maintaining coherence but also as a consideration to the survey participants who would be reviewing each story.

Models trained on such huge datasets will almost inevitably have some undesirable content in their training sets (Gao et al., 2020). For the purposes of undertaking human evaluation, a word search was first completed on the data for common offensive terms before the human-survey stage commenced. When offensive terms were found, this prompt-story set was removed from the data and replaced with another at random.

3.2 | Evaluating the stories

3.2.1 | Human evaluation by survey

Human evaluation was undertaken using anonymous surveying where participants[4] were first advised that the stories were written by machine. For each pair, the participants were shown the prompt and the subsequent story generated by the GPT-Neo model and asked to answer the following four questions, using a 7 point-Likert scale.[5] The wording of these questions is designed to ask in plain English terms about the coherence and interestingness of the machine-generated stories.

1. **Q1** *How related do you think the story is to the prompt?* It was important to record the perceived semantic connection between the prompt and story; an interesting story could be produced by the system, however if it did not relate to the prompt then the objective of the task has not been achieved.
2. **Q2** *How much sense does the story make to you?* A question on the story making sense to the evaluator was a simple, colloquial way of asking about story coherence. This was introduced to observe whether a story needs to be coherent to be interesting to a reader, or conversely if an incoherent story was likely to be deemed uninteresting.
3. **Q3** *How interesting is the PROMPT to you?* Separate to the interest-level of the story, evaluators were asked if they found the prompt interesting, as their level of interest in the prompt may impact their interest in the resulting story generated.
4. **Q4** *How interesting is the STORY to you? (Would you read more?)* This question directly asks the evaluators whether they found the story interesting, but also includes attempts to pin down the subjective notion of 'interesting' by paraphrasing it as a desire to read more of the story.

There was also a further optional free-text question at the end of each survey, for evaluators to leave general comments or impressions. Two sample prompt/story pairs were included in the instructions of the survey, to provide context on the type of text that the evaluator would be reading in the survey and to set their expectations—see Figure 1. The 100 prompt/story pairs were split into five sets of 20 pairs to reduce the chances of evaluators tiring or growing bored and abandoning the survey. Each prompt/story was reviewed by a minimum of six unique reviewers, although the majority were reviewed by seven or more.

3.2.2 | Automatic evaluation metrics

Upon review of the available automatic metrics, BERTScore and BERT Next Sentence Prediction (NSP) were chosen as the metrics that suit the challenges of this study best. BERTScore attempts to focus on semantic coherence, an essential element for a strong, interesting narrative, whilst BERT-NSP aims to identify whether a sequence within a narrative is likely based on what has preceded it, also important for a story to flow coherently. Both automatic evaluation metrics are based on BERT (Devlin et al., 2018) which, like GPT, is a Transformer language model. However, BERT is not trained to predict the next most likely token in a sequence. Instead, it is trained to predict a masked token given the tokens on its left and right, and to predict whether two sequences follow on from each other. It uses bidirectional self-attention, whereas GPT uses constrained self-attention where every token can only attend to context to its left.

BERTScore

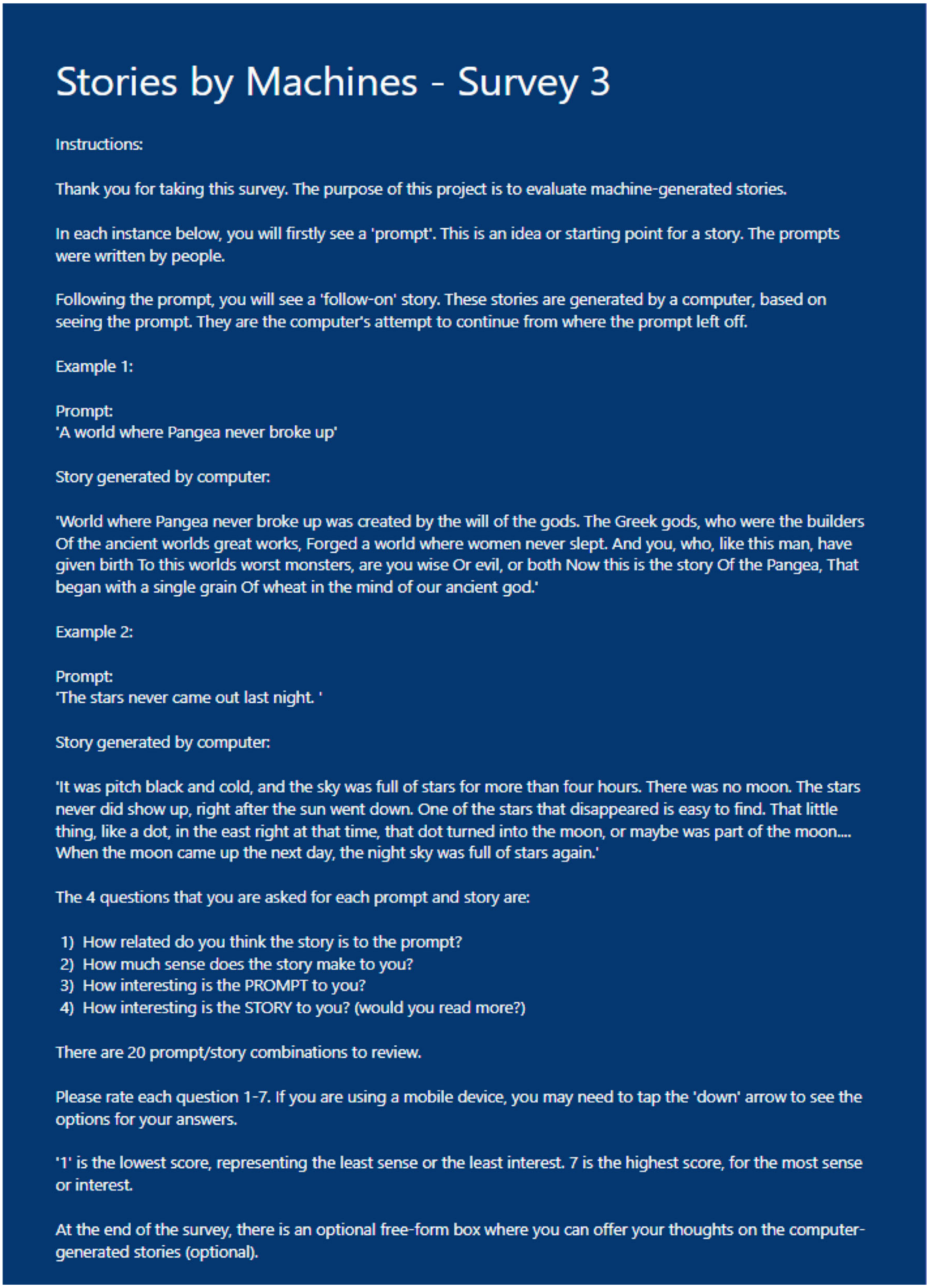
BertScore is a language generation evaluation metric based on BERT (Zhang et al., 2020a). It calculates a similarity score for two sentences, as a sum of cosine similarities between the contextualized word embeddings produced by a pretrained BERT model for each word in each sentence. BERTScore was originally developed to be used as a reference-based evaluation metric for machine translation and image captioning. BERTScore outperforms the previous automatic evaluation approaches in this area because of its capacity to use the contextual embeddings generated by a BERT model for evaluation. By assigning different embeddings to words depending on their surrounding context, it attempts to reward semantic relationships between an input and an output. Whilst it was developed for image captioning and machine translation tasks, BERTScore is designed to be task-agnostic. We attempt to use BERTScore as a reference-free evaluation metric by using it to measure semantic similarity between the prompt and the story, and between the story sentences themselves—core elements of coherent story generation.

The following example shows a prompt/story sentence pair from the data that achieves a high re-scaled BERTScore[6] of 0.287 and low 1-gram BLEU score[7] of 0.1226:

> ***Prompt:*** *You are born with the ability to stop time, but 1 day you see something else is moving when you have already stopped time.*
> *Story sentence*: Your brain takes over and tells you to move, but you cannot.

The BERTScore metric tokenises two selections of text that are to be compared, and using contextual embeddings produced by BERT, derives a semantic similarity metric by calculating cosine similarities between the embeddings. Two approaches were undertaken to obtain two BERTScore metrics for each story. In the first approach, the BERTScore is calculated between *each sentence in the story and the prompt*, and the resulting scores are averaged.

1. BERTScore is calculated comparing the Prompt to Sentence 1 in the story.
2. BERTScore is calculated comparing the Prompt to Sentence 2 in the story … and so on



Stories by Machines - Survey 3

Instructions:

Thank you for taking this survey. The purpose of this project is to evaluate machine-generated stories.

In each instance below, you will firstly see a 'prompt'. This is an idea or starting point for a story. The prompts were written by people.

Following the prompt, you will see a 'follow-on' story. These stories are generated by a computer, based on seeing the prompt. They are the computer's attempt to continue from where the prompt left off.

Example 1:

Prompt:
'A world where Pangea never broke up'

Story generated by computer:

'World where Pangea never broke up was created by the will of the gods. The Greek gods, who were the builders Of the ancient worlds great works, Forged a world where women never slept. And you, who, like this man, have given birth To this worlds worst monsters, are you wise Or evil, or both Now this is the story Of the Pangea, That began with a single grain Of wheat in the mind of our ancient god.'

Example 2:

Prompt:
'The stars never came out last night. '

Story generated by computer:

'It was pitch black and cold, and the sky was full of stars for more than four hours. There was no moon. The stars never did show up, right after the sun went down. One of the stars that disappeared is easy to find. That little thing, like a dot, in the east right at that time, that dot turned into the moon, or maybe was part of the moon.... When the moon came up the next day, the night sky was full of stars again.'

The 4 questions that you are asked for each prompt and story are:

1) How related do you think the story is to the prompt?
2) How much sense does the story make to you?
3) How interesting is the PROMPT to you?
4) How interesting is the STORY to you? (would you read more?)

There are 20 prompt/story combinations to review.

Please rate each question 1-7. If you are using a mobile device, you may need to tap the 'down' arrow to see the options for your answers.

'1' is the lowest score, representing the least sense or the least interest. 7 is the highest score, for the most sense or interest.

At the end of the survey, there is an optional free-form box where you can offer your thoughts on the computer-generated stories (optional).

FIGURE 1 Instructions to survey participants.

This score is identified as `BERTScore-1` in the results. The second approach compares *each sentence to the previous sentence*, rather than comparing each sentence to the prompt.

1. BERTScore is calculated comparing the Prompt to Sentence 1 in the story.

2. BERTScore is calculated comparing Sentence 1 to Sentence 2 in the story … and so on.

These scores were again aggregated and are captured as `BERTScore-2`. By taking this dual approach, it can be observed first if individual segments of the story are semantically linked back to the prompt, but also if each segment is semantically linked to the previous segment.

BERT next sentence prediction

BERT is trained on two tasks, Masked Language Modelling and NSP (Devlin et al., 2018). NSP is the task of predicting whether a sentence logically succeeds the previous sentence and is designed to learn the relationships between sentences. The model is trained on a large, labelled set of sentence pairs, where 50% of the time the next sentence is a random sentence, and 50% of the time it is the actual next sentence (Devlin et al., 2018). For this study, this BERT-NSP model is implemented as the second automated metric to evaluate stories. Each sentence pair is tokenised, and the BERT model processes the sentences and outputs 0 to indicate that Sentence-Two does follow Sentence-One, and 1 when it believes it does not.

Similarly to BERTScore, BERT-NSP evaluates the prompt/story pairs in two different ways. `BERT-NSP-1` looks at predicting whether each sentence in the story logically follows on from the prompt, whereas `BERT-NSP-2` compares the prompt to the first sentence of the story, and then each subsequent sentence to the previous sentence.

4 | RESULTS

4.1 | Human evaluation

The average scores received for each question are listed in Table 1. Prompt interest (Q3) received the highest average rating of 4.45 out of 7. In comparison, story interest (Q4) received the lowest average rating of 3.26 out of 7. This suggests that, in general the model was unable to produce stories of greater interest than the prompts.

There was no preference observed for either longer prompts or longer stories. Focusing on average ratings across the four questions, the highest rated prompt/story pair is the following:

> ***Prompt:*** *In the future, Disney purchases an entire planet and makes the whole thing into one giant amusement park.*
> ***Story:*** *The problem is that a large portion of the population hates the park, hates its corporate overlords, hates its theme park, and is starting to revolt. At the park itself, the people are divided. Some people are happy, some people are not. One thing that makes this story different from movies like the Matrix is that the people who live at the park have no control over the things that are happening there.*

The story remains coherent and relevant to the prompt throughout, not deviating, even at its latter stages. Introducing the term 'corporate overlords' puts a somewhat dystopian slant to the narrative that wasn't necessarily part of the prompt, showing progression of the story in a defined and novel direction, which the evaluators seemed to have enjoyed.

The lowest rated prompt/story pair, displaying few of the above-mentioned positive characteristics, is shown below:

> ***Prompt:*** *Everyone has disappeared but you, you think that everyone is just holding surprise party for you.*
> ***Story:*** *Then, the story starts. The film follows the adventures of jinx, a former pro wrestling manager who is now a private chef. His wife. jenna leigh, had just found out that he was dating a woman named dana bjorn Stresemann. A few months later, jinx and dana get back together after a break-up. Now, when jinx is going away for work, dana contacts the police with a missing person notice.*

The story does not relate back to the prompt and the context of the story contradicts the prompt semantically.

TABLE 1 Human evaluation mean scores from survey results.

Question	Average score
Q1. Story/prompt relationship	3.74/7
Q2. Story coherence	3.42/7
Q3. Prompt interest	4.45/7
Q4. Story interest	3.26/7

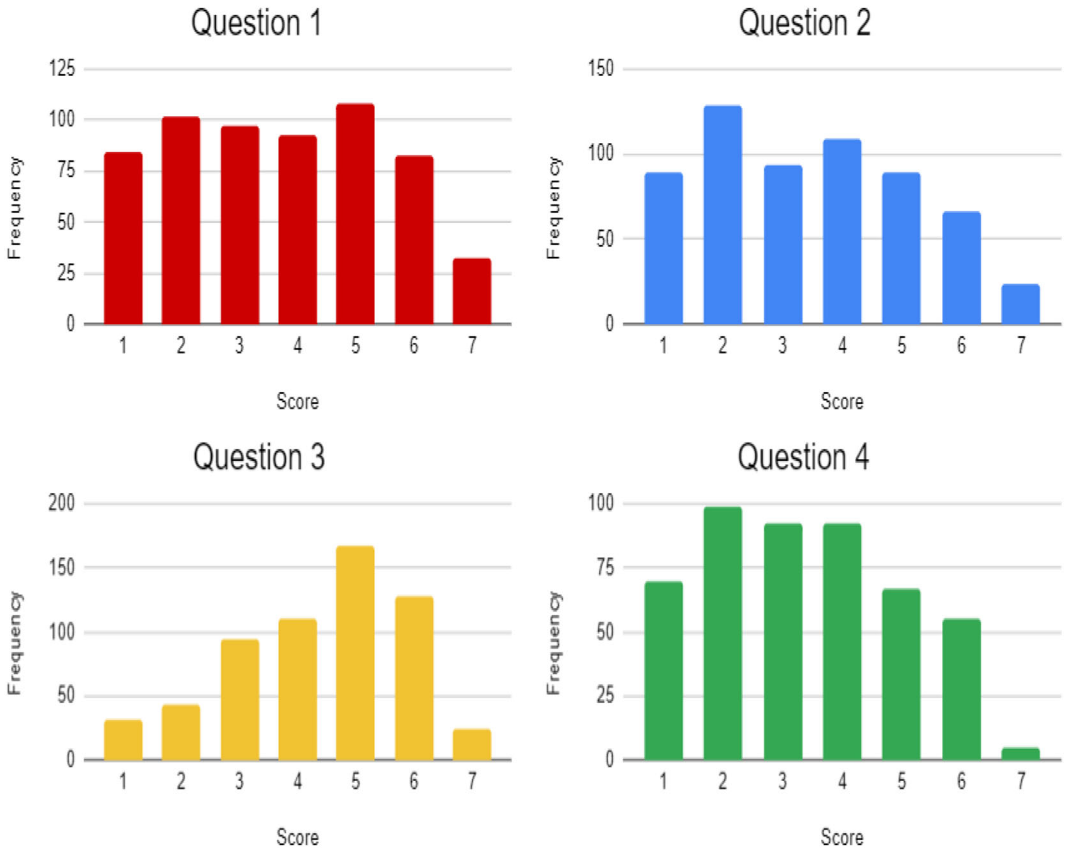


FIGURE 2 Histogram of survey responses per question type. Q1. Story/prompt relationship, Q2. Story coherence, Q3. Prompt interest, Q4. Story interest.

Figure 2 reports the number of times a particular score was awarded for each question type. The distribution of the scores is relatively evenly spread across the Likert-scale, although there is a noticeably lower number of high scores for Q4 (story interest) compared with the other question types. Each story was evaluated by at least six survey participants. Figure 3 reports, for each of the four questions, the distribution of the standard deviations in the scores for that question for a particular story. The standard deviations range from 0.4 to 2.4, with the majority of questions having a standard deviation over 1. This indicates a substantial level of disagreement amongst the survey participants in the scores they assigned to a particular story. As an example, the following story received story-interest scores (Q4) of 1,2,3,4,5,6, and 6 across the seven evaluators:

> ***Prompt:*** *You are immortal. The only time you will die is when you say your kill phrase that you created. You have serious anxiety over it.*
> ***Story:*** *You know that when you say it, you get a one in seven chance of dying, and when you say it, you never get a chance to kill anyone. So, you are living a normal life, and you even look happy. In the end, you get to experience an incredible feeling of happiness. You start feeling like everything is cool with you. The only problem is that you cannot actually touch your kill phrase.*

Evaluator disagreement like this suggests mixed interest in the topic, which could be categorized as being in the genre of sci-fi/thriller/crime. Broadly speaking, as many evaluators found it interesting as did not. Given the observation of inter-evaluator disagreement, it must be considered whether it is unrealistic to expect an automated metric to provide meaningful scores for subjective metrics, when humans often cannot agree upon the same metrics.

Table 2 and the heatmap in Figure 4 show a matrix illustrating correlation between the different survey questions and the four automated metrics implemented. Focusing for now on the correlation between the four survey questions, the strongest correlation of 0.80 is between story coherence (Q2) and story interest (Q4), suggesting that evaluators were most interested in the stories that they found to be the most coherent. A strong relationship (0.72) is also observed between the story coherence (Q2), and the story-prompt relationship (Q1), indicating potentially that evaluators factored in the connection between prompt and story when considering overall coherence; a story that was coherent, would be deemed less so if it did not follow on logically from the prompt. It should be noted that there was a low positive correlation of 0.34 reported

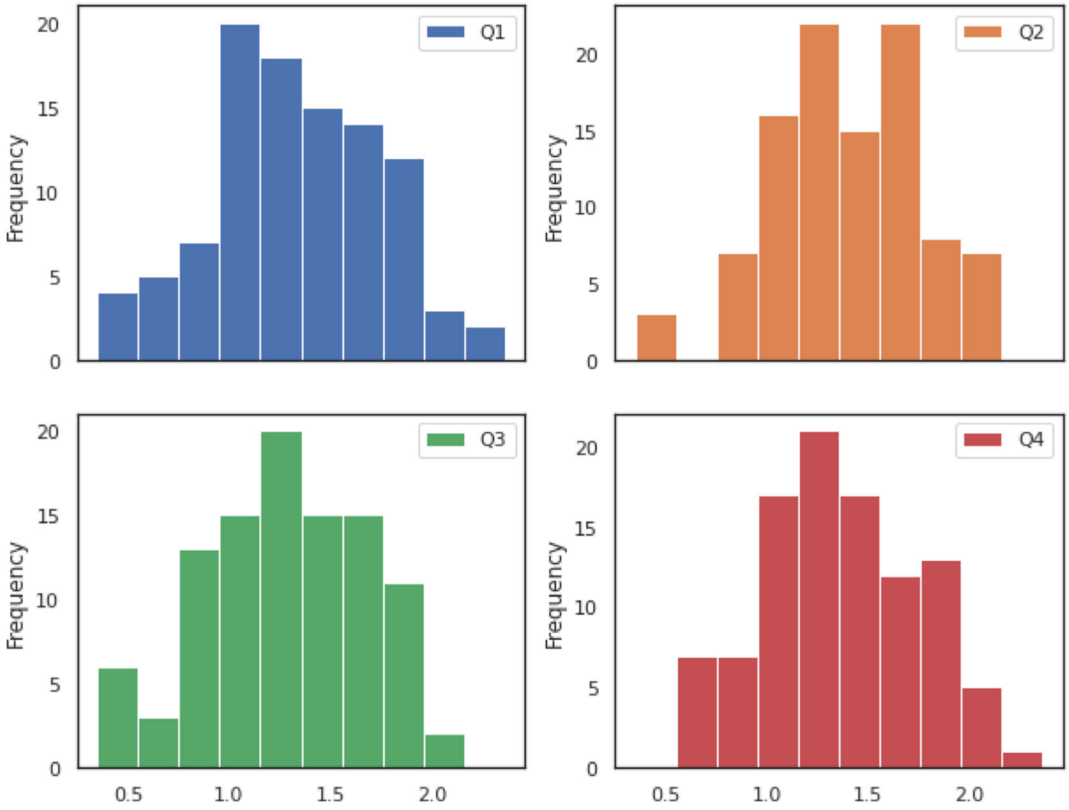


FIGURE 3 Distribution of standard deviation of survey participants' scores for each question. Q1. Story/prompt relationship, Q2. Story coherence, Q3. Prompt interest, Q4. Story interest.

TABLE 2 Correlation between human judgements and automated metrics.

Metric	Q1	Q2	Q3	Q4	BS1	BS2	NSP1	NSP2
Q1	1.00	0.72	0.24	0.62	0.41	0.29	0.25	0.25
Q2	0.72	1.00	0.19	0.80	0.25	0.21	0.12	0.15
Q3	0.24	0.19	1.00	0.34	0.09	0.12	0.02	0.06
Q4	0.62	0.80	0.34	1.00	0.28	0.23	0.10	0.25
BS1	0.41	0.25	0.09	0.28	1.00	0.66	0.32	0.26
BS2	0.29	0.21	0.12	0.23	0.66	1.00	0.24	0.27
NSP1	0.25	0.12	0.02	0.10	0.32	0.24	1.00	0.61
NSP2	0.25	0.15	0.06	0.25	0.26	0.27	0.61	1.00

Abbreviations: Q1, Story/prompt relationship; Q2, Story coherence; Q3, Prompt interest; Q4, Story interest.

between prompt interest (Q3) and story interest (Q4). Rating a prompt interesting did not necessarily mean that the story itself would be deemed interesting.

In general, the prompts were quite specific. They set a certain tone or introduced a theme that defined a direction that a story 'should' take. Whilst this was still an open-ended style task, vaguer, less specific prompts may provide more leeway for the model to produce stories that humans would deem relevant. While some stories generated could be considered related to the prompt, the story may have taken a secondary semantic element of the prompt to build upon, rather than use the predominant or primary theme or idea.

The last question in the survey invited the survey participants to comment on the stories. The following is a summary of their comments:

- They found the themes somewhat unsettling.
- Some stories did not make sense.
- Some stories came across as poetic, but also noted that this may have been a coincidence or fluke.

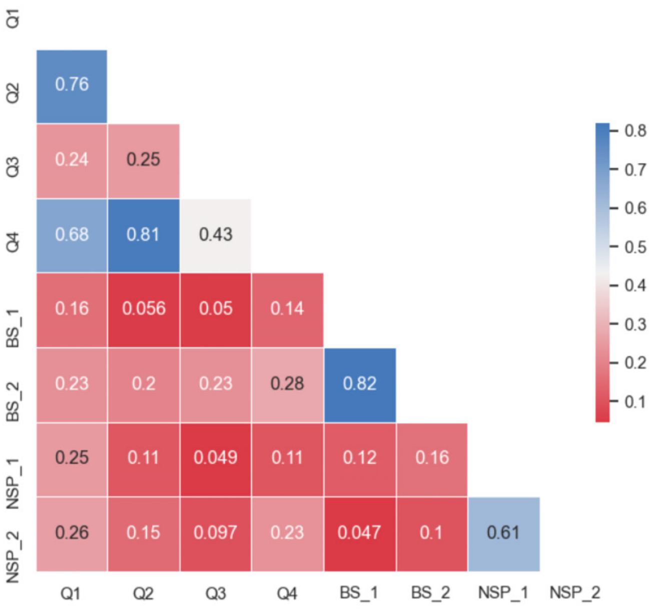


FIGURE 4 Correlation heat map.

- The stories came across like 'blurbs that would be seen on the back of a book cover'.
- The stories sometimes went in a direction that human stories would not, which generated interest.
- The stories fail to continue expanding on the most interesting part of the prompt.
- There are some funny generations, there is also a surreal aspect to some of them, and even some that are profound (e.g. 'I gave you all you gave me').
- The style seemed different to human writing, although this wasn't necessarily bad.
- The stories written by the computer were sometimes more abstract (than the prompts).

4.2 | Automatic evaluation

4.2.1 | BERTScore

For both BERTScore metrics, `BERTScore-1` where each sentence in the story is compared with the prompt and `BERTScore-2` where each sentence is compared with the previous sentence, cosine similarity was calculated for each sentence pair and then averaged for an overall score for a given prompt/story. We are interested in the correlation of the BERTScores with human judgements rather than raw scores, but it is worth noting that the BERTScore-2 figures tended to be slightly higher than the BERTScore-1 figures (a range of −0.168–0.439 with a mean of 0.138 versus a range of −0.187–0.365 with a mean of 0.074). See Figure 5 for the score distributions. Some of the highest BERTScore results were for stories that demonstrated a notable amount of repetition, for example the following which received the highest `BERTScore-1`:

> ***Prompt:*** *A dozen small alien ships enter the solar system, they ignore us. A few years later other ships show up, destroy the first visitors and leave. Ten years later two fleets arrive.*
> ***Story:*** *A decade later the aliens come again, this time with a fleet of ships, and destroy the visitors and leave. One thousand years later, a new alien ship arrives, a vessel similar to the first. One hundred years later the alien ships finally come again, this time with over 500 ships, destroy the 100 ships that came the previous year, then use the surviving alien vessels to create their base*

The re-use of the term 'years later' assisted in increasing the cosine similarity score. Despite the intention of maintaining a focus on rewarding semantic similarity, this shows that repetition is still also rewarded by this metric. This same prompt/story pair was the 15th highest rated by humans out of 100 stories evaluated.

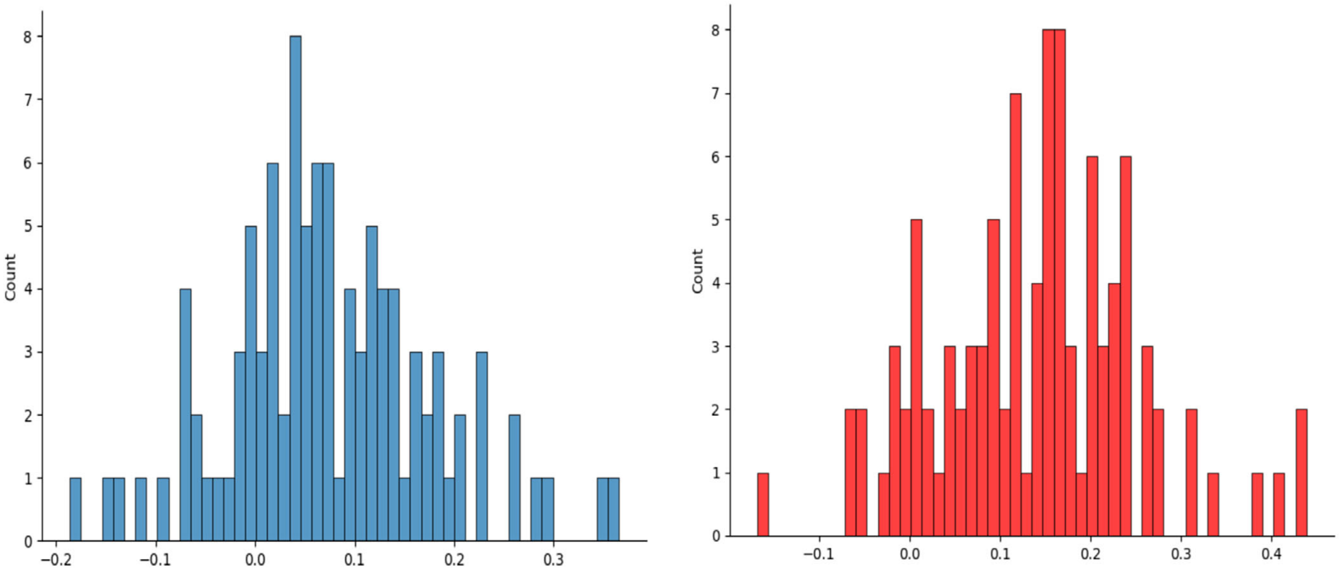


FIGURE 5 BERT score 1 (left) and BERT score 2 (right) distributions.

4.2.2 | BERT next sentence prediction

The BERT-NSP scores are relatively high; in most cases, both `BERT-NSP-1`, where each sentence relates back to the prompt, and `BERT-NSP-2`, where each sentence is compared with the previous sentence, predict that the next sentence does logically follow the preceding sentence. Within the 100 prompts/stories assessed, there were a total of 531 sentence pairs reviewed for next sentence prediction combinations. For `BERT-NSP-1`, 433 of the comparisons were deemed to be logical next sentences and only 98 were not. For `BERT-NSP-2`, an even higher number of sentences were predicted to follow on from the previous one; 497 of the 531 sentences were logical sequences, with only 34 not being predicted as logical. In hindsight, this is somewhat expected, given the similarities in training objectives (predicting the next token versus next sentence prediction) and training data between GPT-Neo and BERT, and indicates that BERT NSP prediction, used in this way, is not a useful metric for evaluating the coherence of narrative text.

4.2.3 | Correlation with human scores

The results in Table 2 and Figure 4 show weak correlation between human judgement scores and automated metric scores. The highest correlation between an automated and a human metric is 0.41 between story-prompt relatedness (Q1) and `BERTScore-1`. This aligns with what `BERTScore-1` is trying to achieve: semantic relatedness between the prompt and each story sentence. Regarding story interest as defined by humans, there was a very low correlation of 0.28 with `BERTScore-1` and 0.23 with `BERTScore-2`. There was almost no correlation with the BERT-NSP scores; this metric found for most cases that sentences logically followed each other, however it did not provide the more granular level of analysis that human surveying and BERTScore provided.

BERTScore and BERT-NSP scoring is undertaken at a sentence level and aggregated for each story, whereas the human evaluators were asked to judge the story in its entirety. This is relevant, as BERTScore scores may be impacted by one or two low results in a sentence-pair within a story, thereby lowering the overall average score for an otherwise strong story.

5 | CONCLUSION

Whilst it is established that modern transformer models generate significantly more fluent and grammatical text than their predecessors, evaluation of narrative elements of their output continues to be a challenge. Many standard automated evaluation metrics exist for text generation that reward repetition of the input; this is not a success metric in narrative text generation. We have presented an evaluation study which took the dual approach of (1) obtaining human judgements for a set of stories generated by the GPT-Neo Transformer model, focusing on criteria around story interest and coherence, and (2) carrying out automatic evaluations of the same texts. The automatic metrics implemented focus on semantic similarity estimation, rather than n-gram overlap.

Our survey results show a strong correlation between story coherence and story interest, with relatively low scores for both. There is a fine balance to suspending disbelief in storytelling and generating believable stories, and the GPT-NEO-generated text is shown in this study to often lack this level of nuance. Although the main goal of the study was to carry out a human evaluation of narrative text, a secondary goal was to evaluate the same stories using automatic metrics and to ascertain the usefulness of such metrics for this task by comparing the results they produce to the human evaluation results. We found low correlation between the automatic metrics, BERTScore and BERT NSP, and the human judgements, with BERTScore being more useful as a metric.

Future work in the area could involve identifying story prompts for a certain specific genre (crime for example) and generating stories consistent with this genre. If evaluators with an interest in this genre were recruited, this may reduce the occurrence of low scores due simply to evaluators' lack of interest in the topic, regardless of the quality of the output. Scores may become more reliable through this process by employing ‘expert evaluators’ for a given theme. There is also scope, given a sufficiently high volume of human judgements, to train a new evaluation system, allowing for the development of an automated metric fit for evaluating narrative text.

An alternative mode of evaluation would be to run the model several times for each prompt, with the best outputs subsequently selected for human evaluation. Automated metrics could potentially be used to decide which of these outputs are best. This may result in more consistently coherent stories being evaluated, providing a higher quality starting point for more selective human evaluation.[8]

Story quality is a subjective notion, and one that is not easily pinned down. In this study, we have directly asked people whether a story is interesting to them, qualifying interesting as whether they would like to continue reading. This is based on the assumption that a good story is one that makes the reader want to keep reading. But which reader are we referring to? Someone may keep reading because they want to find out what happens next, someone else might keep reading because they relate to the characters, for someone else, it might be the descriptive language used. Future work could refine the notion of interestingness by carrying out a survey with longer stories and more specific questions related to plot, character, and style. This type of finer-grained evaluation would work better with a model capable of generating more coherent stories than the GPT-Neo model used in this study.

From a narrative perspective, the stories generated by GPT-Neo falls short in terms of consistently providing interest; their success is somewhat hit-and-miss. More immediate consistent success for these systems may be achieved through non-narrative-style text generation, or by employing a hybrid machine-human approach.

ACKNOWLEDGMENT

Open access funding provided by IReL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Dominic Callan https://orcid.org/0000-0002-9163-1777

ENDNOTES

[1] A recent attempt to evaluate machine generated text is the study of Clark et al. (2021) who used human evaluators to compare human authored and machine generated text across several domains, including story and news generation. They observe that human evaluators focused on form and structure rather than content in deciding whether a text was written by a machine or a human. This allowed for the conclusion that machines write fluently but did not address other narrative strengths of the text produced.

[2] Lexical Cohesion Halliday and Hasan (1976) refers to the use of related words to connect various parts of a text together in a coherent way.

[3] Occasionally the text outputs were not stories, but instead were sequences of text describing the third-party telling of a story.

[4] The participants were university graduates, aged between 25 and 40 and based in Dublin, Ireland. All participants were fluent English speakers. Google Forms (https://www.google.com/forms/about/) was used as the survey tool.

[5] van der Lee et al. (2019) reported that the Likert 5-point scale was the most used scale for Human evaluation in NLG. They conclude, however, that a 7-point scale is best, maximizing reliability, validity, and discriminative power.

[6] Zhang et al. (2020b) announced an optional improvement to BERTScore after the release of their original paper, to address the relatively small range observed between high and low scores. They suggest that the cosine similarity score be rescaled through a linear transformation, noting that this rescaling does not negatively impact correlation with human judgement. This rescaling is implemented in BERTScore calculations in this article.

[7] A variant of the BLEU evaluation metric calculated using word unigram overlap scores (Papineni et al., 2002).

[8] See, however, Ji et al. (2022) for a criticism of the use of automatic metrics to filter out items before human evaluation, in the context of open-domain dialogue systems.

REFERENCES

Akoury, N., Wang, S., Whiting, J., Hood, S., Peng, N., & Iyyer, M. (2020). *STORIUM: A dataset and evaluation platform for machine-in-the-loop story generation*. Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 6470–6484). Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.525

Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021). *GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow*. http://github.com/eleutherai/gpt-neo

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv, Abs/2005.14165*.

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. *arXiv Preprint arXiv:2006.14799*.

Chaganty, A. T., Mussman, S., & Liang, P. (2018). The price of debiasing automatic metrics in natural language evaluation. *arXiv Preprint arXiv:1807.02202*.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv Preprint arXiv:2107.00061*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv:1810.04805*.

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. *arXiv Preprint arXiv:1805.04833*.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv Preprint arXiv:2101.00027*.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Halliday, M., & Hasan, R. (1976). *Cohesion in english*. Routledge.

Hashimoto, T. B., Zhang, H., & Liang, P. (2019). Unifying human and statistical evaluation for natural language generation. *arXiv Preprint arXiv:1904.02792*.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., & Rieser, V. (2020). *Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions*. Proceedings of the 13th international conference on natural language generation (pp. 169–182).

Ji, T., Graham, Y., Jones, G., Lyu, C., & Liu, Q. (2022). *Achieving reliable human assessment of open-domain dialogue systems*. Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 6416–6437). Dublin, Ireland: Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.445

Lin, C.-Y. (2004). *ROUGE: A package for automatic evaluation of summaries*. Text summarization branches out (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics. https://aclanthology.org/W04-1013

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., & Pineau, J. (2017). Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv Preprint arXiv:1708.07149*.

McIntyre, N., & Lapata, M. (2009). *Learning to tell tales: A data-driven approach to story generation*. Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP (pp. 217–225).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: A method for automatic evaluation of machine translation*. Proceedings of the 40th annual meeting of the association for computational linguistics (pp. 311–318).

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., & Harchaoui, Z. (2021). Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, *34*, 4816–4828.

Purdy, C., Wang, X., He, L., & Riedl, M. (2018). *Predicting generated story quality with quantitative measures*. Fourteenth artificial intelligence and interactive digital entertainment conference, predicting generated story quality with quantitative measures.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Roemmele, M., Gordon, A. S., & Swanson, R. (2017). *Evaluating story generation systems using automated linguistic analyses*. SIGKDD 2017 workshop on machine learning for creativity (pp. 13–17).

Sellam, T., Das, D., & Parikh, A. P. (2020). *BLEURT: Learning robust metrics for text generation*. ACL.

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. J. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, *67*, 101151.

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. J. (2019). *Best practices for the human evaluation of automatically generated text*. INLG.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv, Abs/1706.03762*.

Yao, L., Peng, N., Weischedel, R. M., Knight, K., Zhao, D., & Yan, R. (2019). Plan-and-write: Towards better automatic storytelling. *arXiv, Abs/1811.05701*.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 9054–9065). Curran Associates, Inc. http://papers.nips.cc/paper/9106-defending-against-neural-fake-news.pdf

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020a). BERTSCORE: Evaluating text generation with bert. *arXiv, Abs/1904.09675*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020b). *Rescaling bertscore with baselines*. https://github.com/Tiiiger/bert_score/blob/master/journal/rescale_baseline.md. GitHub.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., & Eger, S. (2019). *MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance*. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 563–578). Hong Kong, China: Association for Computational Linguistics. https://aclanthology.org/D19-1053

AUTHOR BIOGRAPHIES

Dominic Callan recently graduated with a Master's degree in Artificial Intelligence from Dublin City University, where he focussed his research on large language models and machine generated language. He works as a Data Analyst for a start-up in the tech industry.

Dr. Jennifer Foster is a lecturer at the School of Computing in Dublin City University. Her expertise lies in the field of Natural Language Processing and she has authored 80+ peer-reviewed publications on topics including parsing, social media text analysis, sentiment analysis, question answering and Irish language resources. She is a regular programme committee member for the main conferences in her field, and from 2016 to 2019, served on the executive board of the Association for Computational Linguistics. She is currently the Chair of the Computing for Business programme at DCU.

How to cite this article: Callan, D., & Foster, J. (2023). How interesting and coherent are the stories generated by a large-scale neural language model? Comparing human and automatic evaluations of machine-generated text. *Expert Systems*, *40*(6), e13292. https://doi.org/10.1111/exsy.13292