Accelerating Workflows in Video Game Translation: A Recommender System for Review and Post-Edit Assignments

TENDAI MUKANDE, Research Ireland ML-LABS, Dublin City University, Ireland
DIMITER DINKOV, Kantan AI, Keywords Studios, Dublin, Ireland
RICCARDO SUPERBO, Kantan AI, Keywords Studios, Dublin, Ireland
NOEL O'CONNOR, Insight Research Ireland Centre for Data Analytics, Dublin City University, Ireland

The advancement in Neural Machine Translation (NMT) has significantly improved the localisation of content across multiple languages, offering fluency and efficiency. However, in complex applications, such as the translation of video games, where there is a need to preserve original player experiences, NMT alone cannot capture subtleties such as humour. As a result, human oversight remains essential to ensure that audio and text translations retain their original intent and appeal. In batch workflows that involve millions of words, the assignment of post-editing/review jobs to the rightful personnel is tedious to complete manually. We present the progress and challenges encountered during the development of a recommender system (RS) designed to enhance video game translation workflows. The main goal of this work is to improve time and cost efficiency in real-time localisation workflows and ensure that unique aspects of game narratives are preserved while meeting the demands of global audiences. The results of the online A/B evaluation in more than 292,000 video game translation workflows demonstrate that the recommender system achieves more than 90% time savings and up to 76% cost reduction compared to manual assignment.

CCS Concepts: • **Information systems → Recommender systems**; **Recommender systems**; • **Computing methodologies → Machine translation**;

Additional Key Words and Phrases: Neural Machine Translation, Recommender Systems, Editorial Workflow, Machine Learning, Automated Decision Engines, Resource Allocation.

ACM Reference Format:
Tendai Mukande, Dimiter Dinkov, Riccardo Superbo, and Noel O'Connor. 2025. Accelerating Workflows in Video Game Translation: A Recommender System for Review and Post-Edit Assignments. In *Proceedings of ACM Transactions on Recommender Systems (ACM).* ACM, New York, NY, USA, 18 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The globalisation of commerce and communication creates an overwhelming demand for high-quality translation services, as businesses seek to expand their reach to international markets and business organisations strive to communicate across diverse language barriers. The need for translations that are not only accurate but culturally sensitive is necessary for effective global

Authors' Contact Information: Tendai Mukande, Research Ireland ML-LABS, and Dublin City University, Dublin, Ireland, tendai.mukande@mail.dcu.ie; Dimiter Dinkov, Kantan AI, and Keywords Studios, Dublin, Dublin, Ireland, dimiterd@kantanai.io; Riccardo Superbo, Kantan AI, and Keywords Studios, Dublin, Dublin, Ireland, riccardosuperbo22@gmail.com; Noel O'Connor, Insight Research Ireland Centre for Data Analytics, and Dublin City University, Dublin, Ireland, Noel.OConnor@dcu.ie.

ACM, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

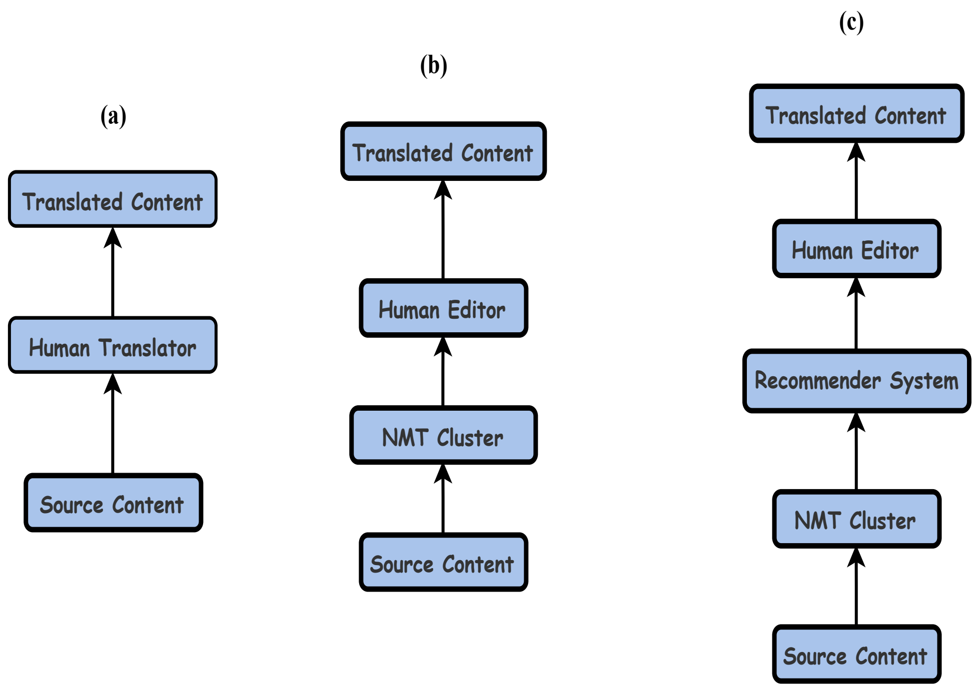

Fig. 1. Overview of translation workflows: (a) Traditional human translation. (b) NMT generated content reviewed by a human editor and (c) an advanced NMT workflow enhanced by a recommender system to address the (time and cost) inefficiencies of manual assignment in large-scale settings.

communication. Traditionally, human translators have been relied upon to meet these demands, as shown in Figure 1(a), but the process is costly, labour intensive, and time consuming [26]. This has driven the adoption of Neural Machine Translation (NMT), a breakthrough technology that uses deep learning algorithms and large data sets to produce fast and fluent translations [56], significantly speeding up the translation process compared to statistical machine translation [53]. Advanced NMT solutions such as KantanMT[1] have enabled organisations to reach global markets faster by accelerating the implementation of large-scale translation delivery.

Although NMT models have revolutionised translation services, they are still far from perfect [12]. These models excel in many contexts, but they often struggle with context-specific nuances such as idiomatic expressions, and specialised terminology, which are essential in niche use cases such as video game localisation.[2] For example, the creative and dynamic language of video games presents challenges: games require translations that not only convey meaning, but also preserve emotional tone. NMT, while effective in technical translations, often misses the mark on such creative elements, leading to inappropriate output [16, 49]. This limitation requires the refinement of NMT-generated content by human editors, as illustrated in Figure 1(b), to be more contextually appropriate and linguistically accurate [20]. This combination of NMT and human expertise is crucial in game localisation, where the integrity of original content needs to be maintained in

[1] KantanMT is an advanced NMT platform developed by KantanAI.
[2] Localisation in this context refers to the process of adapting a video game to suit the language, norms, legal requirements, and player expectations of the respective region.

various linguistic contexts [12]. For our use case, the real-time allocation of the right personnel in batch workflows involving more than 35 million[3] words translated per month is not trivial. The main challenge is to match the right editors with the right task and to ensure that the editors have the necessary expertise for the required task. Manual task assignment is time-consuming and can result in missed deadlines for ongoing projects. In addition, the high volume of job workflows generated by more than 7500[4] NMT engines monthly creates the need for a scalable and systematic assignment solution that is computationally efficient in terms of memory usage, latency and cloud bills. Moreover, the objectives of privacy, fairness and transparency are paramount as well as accuracy of the assignments. To address these issues, we propose an RS that balances these objectives by automatically assigning human editors to NMT-generated translations, as illustrated in Figure 1(c), based on various factors such as their labour cost, productivity and expertise.

In this paper, we present our efforts in developing an RS that dynamically assigns profiled human candidate editors based on their language proficiency, productivity, cost and domain experience in video game localisation. The RS ensures that NMT-generated translations are matched with the **most valuable**[5] candidate editors, preserving both the quality of translation (audio and text) and the creative aspects of the original game content, whereby the complexity of language and thematic relevance amplifies the challenges of manual task allocation. Unlike standard translations, video game localisation involves dialogue and in-game mechanics, which require a deeper understanding of both the language of the target audience, hence, the ability to match editors with the right expertise is crucial. Our proposed solution reduces time bottlenecks in video game translation, producing content that resonates with global audiences while maintaining the creative integrity of the original work in real time. We address challenges encountered in complex editorial workflows large-scale settings, to optimise assignment of the right editors to the right task and ensure high video game translation quality, ultimately improving the end-user experience. Our proposed model accelerates the localisation process in a cost-effective manner. The results of the online A/B evaluation of more than 292,000 large-scale video game translation projects demonstrate that the RS achieves significant time and cost savings compared to manual assignment.

2 Related Work

Research on the development of automated decision engines to improve NMT workflows is a promising area [13, 34]. As demand for high-quality translations grows, such as in technical domains, the synergy between human expertise and machine efficiency is needed to improve NMT output at scale. In this section, we discuss the related research to this work. We give an overview of the advances in NMT, highlight the need for human oversight in NMT systems, and then discuss the application of Recommender Systems (RSs) in automated workflows.

2.1 Advances in Neural Machine Translation

NMT has seen advances over the past decade, driven by breakthroughs in deep learning and the availability of large datasets [10]. Early models, such as recurrent neural networks (RNNs) and long-short-term memory networks (LSTMs), laid the foundation for modern architectures, but faced challenges such as vanishing gradients and difficulty modelling long-range dependencies [19]. The introduction of the transformer self-attention mechanism [51] marked a paradigm shift in NMT, enabling better context handling and parallel processing of data, leading to substantial improvements in translation fluency and accuracy [9]. Transformer-based methods have been

[3]Source: https://www.kantanai.io/

[4]Source: https://www.kantanai.io/

[5]In this context, the most valuable candidate depends on the task requirements. For example, a candidate with lower productivity can be prioritised if their cost is lower and the project is not urgent.

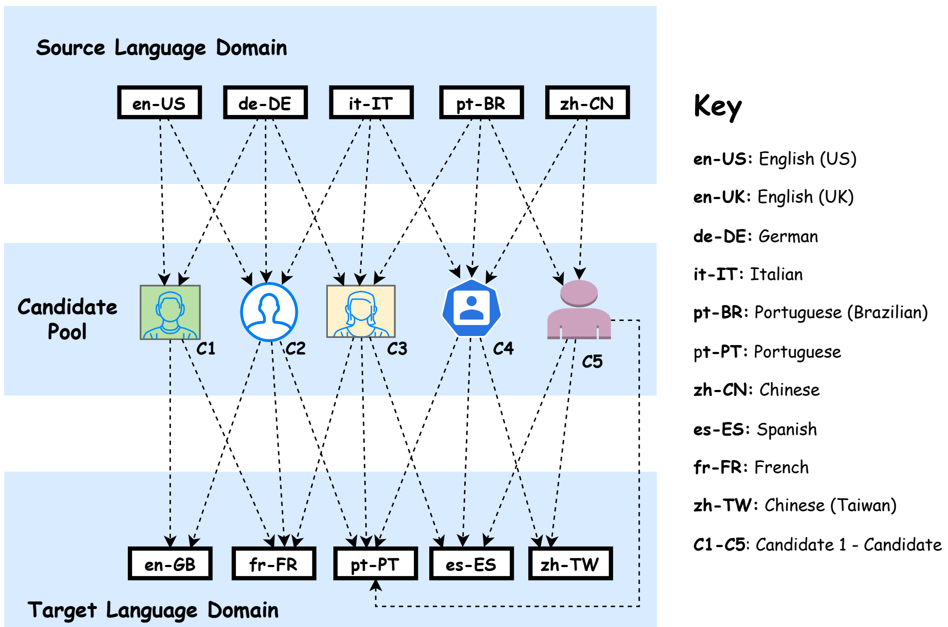

Fig. 2. Overview of the review and post-editing task showing source and target domain languages as well as a candidate pool. Review tasks are assigned to the most valuable candidates in the pool based on their language competencies, cost, productivity, quality scores and game domain experience.

shown to be effective in various scenarios, including multilingual and multi-source translations [40, 44], making NMT more viable for large-scale and real-time applications. Despite their potential, NMT models face limitations in contextually sensitive or creative content, such as idiomatic expressions or poetry [52]. They often struggle to capture the cultural nuances in translations [22] of such content. In addressing these limitations, human editors have participated in the refinement of the NMT output to maintain narrative fidelity [10]. Approaches such as human-in-the-loop not only improve translation quality but also provide valuable feedback for fine-tuning models [20], ensuring that human expertise complements the efficiency of NMT.

2.2 Human oversight in Neural Machine Translation

Despite advances in NMT, recent research highlights the important role of human oversight in the production of high-quality translations [12, 20, 59]. Although most NMT methods excel at speeding up translations, they often fail to capture the subtleties of language. This limitation becomes evident especially in context-sensitive scenarios such as legal documentation, medical communication, or creative content such as poetry, comedy and video games [30]. Human post-editing, as shown in Figure 2, is needed to bridge these gaps, especially in specialised domains where precision is paramount. Post-editors address a wide range of issues, including grammatical inaccuracies, mistranslated idioms, tonal inconsistencies and contextual errors. Moreover, human editors have an innate ability to resolve ambiguities, adhere to stylistic guidelines, and refine translations to preserve the intent and integrity of the original text [18, 59]. An example is the audio and text translation of video games, which demands not only linguistic accuracy but also an appreciation of regional preferences. Although NMT systems are effective for straightforward translation tasks, they often struggle with more complex contexts, which can compromise the player experience and reduce the authenticity of the game in localised markets [12]. Human oversight through post-editing and proofreading ensures that translations align with cultural norms, thus preserving the essence of the original content [38].

2.3 Recommender systems in automated workflows

Research on RSs has evolved significantly over the past decades, progressing from their early use in suggesting products or content [58] to tackling more complex and dynamic applications [17, 45, 54]. The integration of RSs into automated decision workflows presents an exciting avenue for innovation in industries such as gaming, e-commerce and global communications. As demand for high-quality translations and other specialised outputs increases, the role of RSs in automating and optimising these processes becomes increasingly necessary [2]. Despite these advances in RSs, their application in specialised workflows, such as game localisation, remains underexplored [35]. NMT often fails to capture content such as idiomatic expressions [39]. These shortcomings require human oversight, where editors with the right linguistic expertise refine machine-generated translations to meet the demands of localised markets. However, manual assignment of tasks to editors is inefficient, especially in large-scale workflows. Traditional methods rely heavily on static rules and generally do not meet the dynamic and multifaceted requirements of localisation processes [20]. For example, game localisation projects often involve strict deadlines, evolving linguistic standards, and the need to adapt translations to various contexts.

3 System Design

We develop an recommender system to address task assignment bottlenecks in video game localisation workflows. We integrate data from multiple sources, such as editors' performance metrics and project requirements, to allocate tasks, such as assigning a translation task to editors with demonstrated expertise in the target culture and familiarity with the terminology of gaming. Similarly, editors who are experienced in handling creative narratives are prioritised for tasks involving emotionally resonant content. This approach not only improves efficiency, but also ensures that the quality of localisation meets the expectations of players in different regions [37]. Our model is scalable, making it suitable for both small- and large-scale projects. Using modular Application Programming Interface (API) endpoints, we streamline tasks such as data processing, training and real-time inference. Real-time inference APIs allow for low-latency responses by dynamically scaling resources to match traffic demands, and ensuring that the system remains responsive under high load. Additionally, caching and batch endpoints reduce latency and computational overhead by optimising the flow of data and reducing unnecessary computations.

3.1 Workflow Architecture Overview

The RS is integrated with NMT clusters and data clusters to ensure seamless task allocation. NMT clusters generate incoming job tasks, and candidate editor profiles are continuously updated and stored in the database, as illustrated in Figure 3. The overall workflow architecture illustrated in Figure 3 includes candidate profile, recommendation, task assignment and feedback loop. When a recommendation task is received from NMT clusters, the RS recommends the top N candidates from the pool and assigns the most valuable candidate. We **hypothesize** that RS improves the efficiency of the workflow while ensuring higher quality in translations, maintaining the cultural integrity of the translated content. .

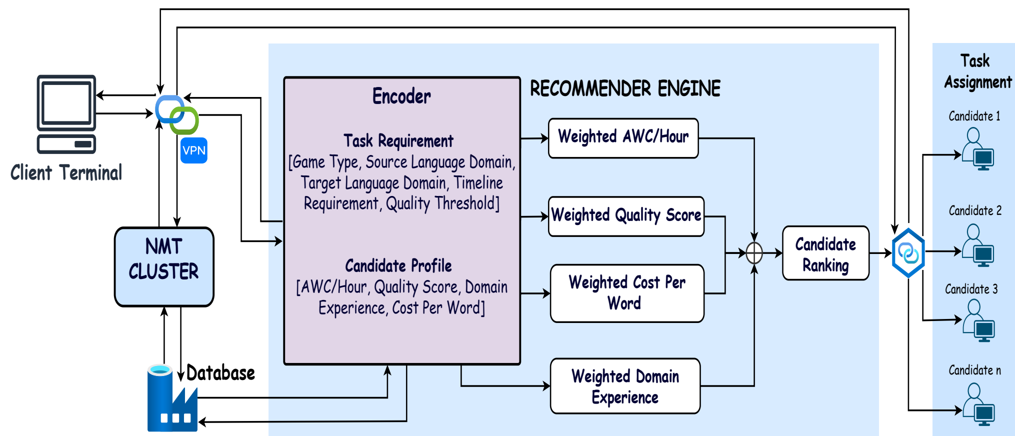

Fig. 3. Architecture overview of the localisation workflow. The RS is integrated to NMT clusters and data clusters to ensure seamless task allocation. NMT clusters generate incoming job tasks, and candidate editor profiles are continuously updated and stored in the database.

Ranking Task. Candidate editor profile attributes such as **language proficiency,**[6] **domain experience**[7] and **previous performance metrics.**[8] Profiling ensures that evolving skills and experience are continuously updated in candidate profiles to facilitate precise and effective task matching. The RS evaluates the editor profiles against the specific requirements of each job task and generates a candidate ranking based on a combination of weighted metrics. The metrics considered in the ranking score calculation include **weighted Average Word Count Per Hour (AWC/hour)**, which is used to determine the productivity of the candidates, and the **weighted quality score**, which is used to evaluate the quality of previous work done by the candidates. **Weighted Cost per Word** balances cost-efficiency, crucial for projects constrained by tighter budgets, and **weighted domain experience** calculates the expertise of editor in specific domains, ensuring that culturally sensitive adaptations are assigned to the most experienced candidates for the particular game domain. Human reviewers with domain experience improve NMT performance by applying their deep understanding of the context of the game, the tone of the narrative, and the expectations of the players. Their familiarity with in-game terminology, character voices, and genre conventions allows them to identify and correct ambiguous or inappropriate translations that would otherwise go unnoticed by less experienced candidates. **Language proficiency** determines the level of editor proficiency in the source and target languages, ensuring that editors with higher proficiency are prioritised for tasks requiring high linguistic precision. These metrics are aggregated into a final suitability score for each editor, ranking them by relevance to the task at hand.

Task assignment: Queued jobs in the workflow are assigned to the top N-ranked candidates according to priority. The highest ranked candidate is first notified and given the option to accept or decline the task. If the candidate declines or fails to respond within the designated notification period, the task is passed to the next candidate in the ranking. This process continues until the task is accepted. Once a candidate accepts, they complete the task, and their profile is updated to

[6]The candidate's expertise in specific source and target languages.
[7]Previous experience of the candidate in executing translation projects of a particular video game domain, such as Athena, Minecraft, Steelhead, Woodstock, etc.
[8]Previous performance metrics of the candidate, such as AWC/hour, quality rating, cost per word, etc.

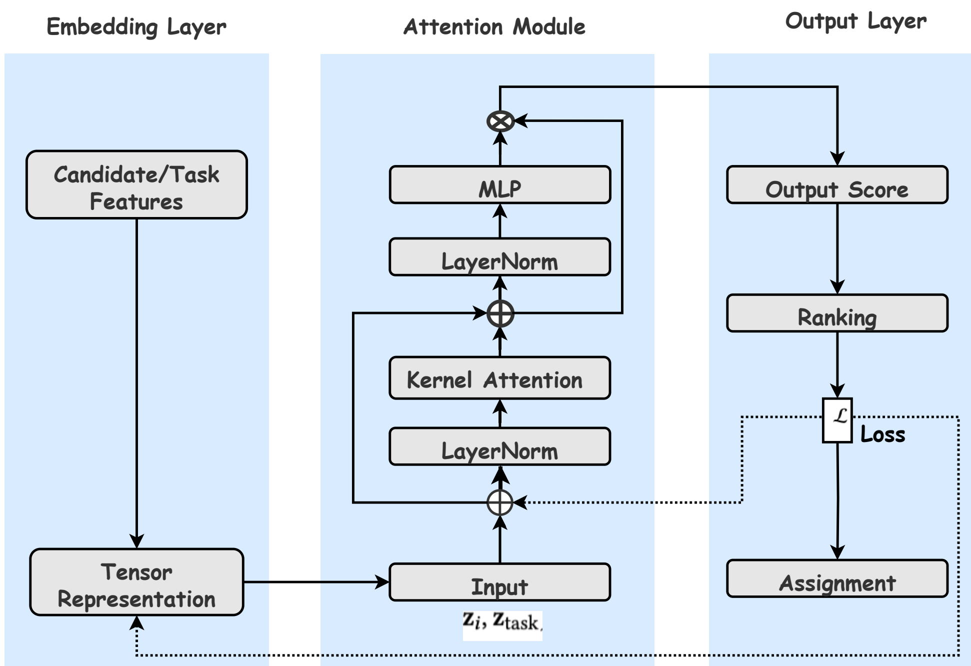

Fig. 4. Recommendation model overview. Task and candidate feature vectors are embedded to obtain lower-dimensional representations and passed to the attention module. The attention layer allows the model to capture relevant aspects of the candidates feature set relative to the task, weighting them accordingly. The final ranking score is used to determine candidate suitability for specific tasks. The loss function combines the ranking loss with fairness and bias losses.

inform future assignment decisions. Our architecture features a feedback loop that continuously refines the RS. After each completed task, feedback such as quality ratings, word count per hour, and quality ratings by end users is collected to update editor profiles and adjust the weighting of evaluation metrics. This adaptive mechanism improves the predictive accuracy of the system over time, aligning the recommendations more closely with the evolving project requirements.

3.2 Recommendation Model

The recommendation model retrieves data from a large database of candidate profiles to match them with job task requirements, including source and target languages, project timeline, and budget constraints. The RS ranks potential editors based on profile metrics such as language proficiency, domain expertise, and historical task performance. It balances subject-matter expertise while adhering to budget and deadline constraints. The goal is to improve task execution efficiency while reducing the overall computational cost of the workflow [27, 41]. This approach ensures the selection of editors who meet both quality and efficiency standards, which contributes to the overall success of the localisation project. The architecture of the recommendation model is shown in Figure 4.

3.2.1 ***Embedding Layer***. The RS ranks candidates for translation jobs by incorporating various contextual factors. To enable context-aware decision making, we use higher-order tensor representations that capture both candidate features and task requirements. This approach allows the model to express complex relationships between multiple factors in a compact and expressive way. For example, a highly experienced candidate who charges less and has a lower AWC/hour may be preferred in cost-sensitive scenarios. In contrast, a candidate with a high AWC/hour but limited experience might still be suitable if the task is urgent. By jointly modelling attributes such as candidate skills, quality, and task complexity, the model can make more informed ranking decisions. We define feature vectors for the task and candidates as tensors:

$$\mathbf{x}_{\text{task}} = [\text{source language domain, target language domain, project timeline, budget, quality threshold}] \tag{1}$$

where $\mathbf{x}_{\text{task}}$ is a vector that contains the features of the task, such as the required skills and project requirements. Similarly, the **candidate feature vector $\mathbf{x}_i$:**

$$\mathbf{x}_i = [\text{Language Competencies}_i, \text{AWC}_i, \text{Cost}_i, \text{Game Type Experience}_i, \text{Quality Score}_i] \tag{2}$$

Both task and candidate feature are transformed to obtain lower-dimensional vector space representations, which help capture non-linear relationships between the features. The candidate and task feature vectors are transformed as follows: **Candidate vector transformation $\mathbf{z}_i$:**

$$\mathbf{z}_i = \text{MLP}(\mathbf{x}_i) = \sigma(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x}_i + \mathbf{b}_1) + \mathbf{b}_2) \tag{3}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are weight matrices for MLP layers, $\mathbf{b}_1$ and $\mathbf{b}_2$ are biases for MLP layers and σ is the ReLU activation function. **Task vector transformation $\mathbf{z}_{\text{task}}$:**

$$\mathbf{z}_{\text{task}} = \text{MLP}(\mathbf{x}_{\text{task}}) = \sigma(\mathbf{W}'_2\sigma(\mathbf{W}'_1\mathbf{x}_{\text{task}} + \mathbf{b}'_1) + \mathbf{b}'_2) \tag{4}$$

where: $\mathbf{W}'_1$ and $\mathbf{W}'_2$ are task-specific weight matrices, and $\mathbf{b}'_1$ and $\mathbf{b}'_2$ are task-specific biases. After the candidate and task feature vectors are transformed, the next step is to determine the similarity between the transformed candidate representation and the task representation to measure the similarity between the transformed task and candidate vectors. The result is a learnable weight matrix that allows the model to determine how well the candidates match the task requirements.

$$\text{Sim}(\mathbf{z}_i, \mathbf{z}_{\text{task}}) = \mathbf{z}_i^T\mathbf{W}_{\text{sim}}\mathbf{z}_{\text{task}} \tag{5}$$

where $\mathbf{W}_{\text{sim}}$ is a learned weight matrix used to scale the similarity.

3.2.2 ***Attention Layer***. The attention layer refines the candidate's representation according to the needs of the task. The attention mechanism allows the model to dynamically adjust the importance of each feature based on specific task requirements. For instance, if the task requires high language domain proficiency, low cost, and quick turnaround, the attention layer enables the model to learn and prioritise this combination of features. This dynamic weighting allows the model to learn more complex relationships between the task and candidate features when ranking the candidate profiles. Each candidate embedding $\mathbf{z}_i$ is transformed into query, key and value vectors as follows:

$$q_i = \mathbf{W}_q\mathbf{z}_i, \quad k_i = \mathbf{W}_k\mathbf{z}_i, \quad v_i = \mathbf{W}_v\mathbf{z}_i \tag{6}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are learnt weight matrices; and q_i, k_i, v_i represent the query, key, and value vectors for candidate i. In standard self-attention [51], the weights are computed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{7}$$

where Q, K, V are matrices formulated by stacking q_i, k_i, v_i for all candidates within a particular language domain and d is the dimensionality of the query and the key vectors. To address the high computational cost of standard self-attention, which is quadratic in terms of memory [48], we apply a kernel function to approximate standard self-attention. This reduces the computational complexity of the attention mechanism from $O(N^2)$ to $O(N)$, where N is the number of candidates. In this way, we efficiently refine candidate representations while maintaining scalability. In our model, we approximate the softmax function as follows:

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \approx \phi(Q)\phi(K)^T \tag{8}$$

where $\phi(\cdot)$ is a kernel-induced feature map. With this approximation, we calculate the attention output as:

$$\text{Attention}(Q, K, V) = \phi(Q)\left(\phi(K)^T V\right) \tag{9}$$

The attention output for the candidate i, h_i, is computed as:

$$h_i = \sum_{j,k} \alpha_{ijk}\left(v_j \odot v_k\right) \tag{10}$$

where the indices j and k iterate over all candidates, representing pairwise interactions between candidates, v_j and v_k are the value vectors for candidates j and k, respectively, α_{ijk} is the attention weight derived from the kernelized feature representations of queries and keys. and $\odot$ denotes the element-wise multiplication of vectors v_j and v_k.

3.2.3 ***Output Layer***. The final candidate score i is calculated as a combination of the feature-based representation $\mathbf{z}_i$ and the attention-refined representation h_i:

$$\text{Score}_i = \mathbf{w}_{\text{features}}^T \mathbf{z}_i + \mathbf{w}_{\text{attn}}^T h_i \tag{11}$$

where $\mathbf{w}_{\text{features}}$ is the weight vector for the feature-based representation, $\mathbf{w}_{\text{attn}}$ is the weight vector for the attention output h_i.Once the scores for all candidates have been calculated, they are sorted in descending order to generate the ranked list.

3.2.4 ***Loss Function***. The training objective is to optimise candidate selection according to the needs of the task, minimise bias, and improve fairness. Bias may arise when recommendation scores are influenced by specific attributes, such as workload history, which could inadvertently disadvantage newly added but equally qualified candidates. In our context, the ranking of recommended candidates is not solely determined by the highest performance metrics, such as the lowest cost or the best quality, but rather by the specific requirements of each project. This approach ensures that all candidates receive fair consideration based on contextual relevance rather than historical or systemic biases. The loss function incorporates ranking loss, fairness loss, and bias mitigation loss, weighted by hyperparameters $\lambda_{\text{fairness}}$ and λ_{bias}:

$$L_{\text{total}} = \sum_{\text{batch}} \left(L_{\text{ranking}} + \lambda_{\text{fairness}} \cdot L_{\text{fairness}} + \lambda_{\text{bias}} \cdot L_{\text{bias}}\right) \tag{12}$$

where L_{ranking} is the ranking loss term, L_{fairness} is the fairness loss term, and L_{bias} is the bias mitigation loss term, designed to reduce bias in rankings.

3.3 Implementation

The RS is implemented using a scalable technology stack that handles dynamic requirements efficiently, thus enhancing real-time workflow assignments and maintaining consistent performance across large-scale projects. We develop our system framework using PyTorch. We use FastAPI [25] in the back-end to provide asynchronous endpoints for real-time management of task data and editor profiles, with the goal of minimising inference latency, and Apache Server as a reverse proxy to manage secure communication, load balancing, SSL termination, and caching between front-end and back-end, optimising performance and reliability. We adopt a continuous learning approach [28] rather than retraining the model from scratch. Specifically, new updates to the candidate profile (for example, changes in quality scores, productivity, or domain experience) are incrementally incorporated through periodic fine-tuning of the model using recent task–candidate interaction data. This allows the RS to adapt to evolving editor performance and new content domains while preserving previously learned knowledge. Full retraining from scratch is performed only at scheduled intervals (for example, quarterly or after significant data changes) to mitigate concept drift [15].

4 Evaluation

To evaluate the effectiveness of the RS, we conducted extensive A/B testing within real-world localisation workflows, benchmarking its performance against traditional manual task assignment. The assessment focused on assignment accuracy and overall operational efficiency. The evaluation encompassed the review of more than 292,000 NMT projects by more than 221,000 candidate profiles, covering 40+ language pairs and 62 video game genres[9]. The ranking order used as reference (ground truth) is derived from historical assignments that had been validated by human experts. This established a human-expert ranking baseline against which the RS's top-N predictions were compared. Thus, "correct order" refers to the empirically observed ranking of editors according to their historical suitability scores. Therefore, the top-N accuracy is computed by checking whether the ranked list by the RS matches this reference order of human-validated assignments.

Data sources. Neural machine-translated projects, mainly referred to as the tasks on which candidates are assigned to review, are generated from **NMT engine clusters** hosted on KantanStream[10]. The RS retrieves the profiles of all possible candidates for each particular task from a **database** of globally crowd-sourced[11] human translators, ranks the most valuable candidates and assigns tasks based on project requirements, as illustrated in Figure 3. Crowdsourcing offers a blend of cost efficiency, cultural authenticity, and community involvement [5]. This model also enables on-demand scaling based on project needs without hiring more in-house staff. In addition, crowd-sourcing fosters community building; as users become contributors, they develop a sense of ownership and a deeper connection to the final product.

Data preprocessing. The data pre-processing steps are as follows. Incomplete or duplicate records are removed, and outliers in numerical attributes, such as **AWC per Hour** and **Cost per Word**, are filtered using interquartile range thresholds. Continuous variables are normalised using min–max scaling. Categorical features, including **source language domain**, **target language domain**, and **game type experience**, are represented using learnable embeddings, as described in Section 3.2.1.

[9] Source: KantanAI
[10] KantanStream is a cloud-based translation management system developed by KantanAI, designed to streamline and automate localisation workflows. Source: kantanai.io/kantanstream/
[11] Source: https://www.memoq.com/pressroom/kantanstream-offers-seamless-connectivity-with-memoq-server/

Historical task performance metrics are then aggregated to generate stable candidate performance vectors, capturing average productivity and quality trends over time.

4.1 A/B Test Setup and Results

The A/B tests were carried out over a five-week period, during which tasks and candidates were divided between the control group (manual assignment) and the treatment group (RS-based allocation). Candidate profiles updates reflect changes in language proficiency[12], domain experience[13], and performance metrics including AWC/hour, quality scores and cost per word. To evaluate the performance of the recommendation model, we assess its ranking accuracy and compare its task assignment efficacy against manual allocation.

Ranking Performance: We evaluated the RS's ranking performance using rank-sensitive metrics. In particular, **top-N accuracy** is defined as 100% only when the top-N list includes all the most valuable candidates in the correct order to ensure that every ranking maximises translation quality, domain relevance, and cost efficiency. Hit Ratio (HR@N)[14] and Normalised Discounted Cumulative Gain (NDCG@N)[15] metrics were applied to determine the ranking performance. The RS achieves high accuracy in terms of ranking tasks based on language skills, domain expertise, and past project performance. The metrics shown in **Table 1** indicate that the RS not only identifies the right candidates, but also ranks them effectively, ensuring that the most suitable candidates are prioritised for each localisation task. The system continuously improves by updating candidate profiles and adapting to new task types and content genres. In addition, the RS adapts dynamically by updating candidate profiles and task descriptors after each assignment. This feedback loop enables continuous improvement, particularly in domains with evolving linguistic or gameplay requirements.

Batch Size	NDCG@5	NDCG@10	NDCG@20	HR@5	HR@10	HR@20
5,000 tasks	0.916	0.931	0.952	0.942	0.953	0.967
25,000 tasks	0.914	0.927	0.945	0.946	0.945	0.958
50,000 tasks	0.911	0.923	0.938	0.938	0.941	0.936
100,000 tasks	0.908	0.919	0.931	0.928	0.934	0.926
292,000 tasks	0.901	0.908	0.914	0.921	0.927	0.918

Table 1. Ranking performance of the recommender system across different batch sizes using NDCG@K and Hit Ratio@K.

Assignment Efficiency: The RS significantly reduces the task assignment time from **7 to 75 hours** (manual) to just **5–65 minutes**, for batch sizes ranging from 5,000 to 292,000 jobs, as shown in Table 2. This results in consistent time savings of more **90%**, effectively eliminating major workflow bottlenecks. For example, manually assigning 25,000 tasks could take up to 15 hours, while the RS completes the same workload in 15-20 minutes.

[12]For example, the candidate's expertise in specific language pairs such as English-to-Korean translation.
[13]For example, prior experience translating within a specific game domain, such as Chinese-to-Japanese translation for *Minecraft*.
[14]HR measures whether at least one relevant (i.e., high-value) candidate appears in the top-N list.
[15]NDCG measures ranking quality by rewarding higher placement of relevant candidates.

Batch Size	Manual Assignment Time	RS Assignment Time	Time Saved by RS
5,000 tasks	7–8 hours	5–7 minutes	90%+
25,000 tasks	10–15 hours	15–20 minutes	90%+
50,000 tasks	18–24 hours	25–32 minutes	90%+
100,000 tasks	25–32 hours	40–48 minutes	90%+
292,000 tasks	60–75 hours	55–65 minutes	90%+

Table 2. Overall time cost (including candidate profile updates, ranking and task assignment) for manual vs. RS-based methods across varying batch sizes. The indicated manual assignment time reflects cumulative person-hours (for example, 2 people × 2 hours = 4 hours).

Cost Efficiency of RS-Based Task Assignment. Table 3 presents a cost efficiency comparison between manual task assignment and RS-based workflows. The results show that the RS is significantly more cost-efficient than manual assignment in terms of average cost per assignment and throughput. The integration of the RS into the localisation workflow delivers a 76.78% reduction in cost, highlighting its effectiveness in scaling large-scale localisation workflows.

Setup	Cost per Assignment (USD)	Throughput/$	Cost Savings
NMT + Manual Reviewer Assignment	$0.149	6.71 tasks/$	—
NMT + Recommender system	$0.0346	28.90 tasks/$	76.78%

Table 3. Cost efficiency comparison between cost efficiency comparison between *NMT+ Manual* and the *NMT+ Recommender system* task-assignment workflows.

4.2 Evaluation Summary

Across multiple high-volume video game localisation projects, the RS demonstrated consistent improvements in task assignment quality, speed, and cost-efficiency. As highlighted in Tables 1 and 2, the system achieves **time savings** of more than 90% in task allocation, completing batch assignments of up to 292,000 jobs in 55-65 minutes. **NDCG** and **HR**scores of more than 0.90, indicating high ranking accuracy of the model. Each task involves multi-criteria matching across cost, domain expertise, and linguistic quality, which is not trivial. The high NDCG and HR values instead demonstrate the robustness of the proposed model in capturing these dependencies within a practical setting. Moreover, the benchmark used in our evaluation is based on ground-truth human assignments derived from real production workflows, representing highly optimised and expert-curated decisions. Thus, the high accuracy of the recommendation system reflects the strong alignment of the model with expert human judgement rather than the simplicity of the prediction task. The cost efficiency analysis (Table 3) shows a **76.78% reduction in assignment costs** compared to manual workflows,which shows the effectiveness of the RS. These results demonstrate that the proposed RS not only outperforms manual assignment in key performance indicators, but also enables cost-effective localisation for large-scale video game workflows.

5 Overall Insights and Future Research Directions

As RSs continue to shape the digital landscape, their effective deployment is crucial for businesses to maintain competitive advantage. Industrial RSs have become integral in driving customer growth, improving user engagement, and improving business metrics such as sales, revenue, and customer retention [55]. However, for these systems to operate optimally in real-world scenarios, they need to overcome significant challenges related to scalability, efficiency, and ethical issues, such as privacy, fairness, and transparency [11]. Most existing RS research focuses on improving model performance, but there is also need to address challenges that arise in practical scenarios to ensure sustainability in industrial settings [8]. In addition, addressing ethical considerations in RSs is essential. Approaches such as fairness-aware learning, explainability, and privacy-preserving techniques can improve recommendation quality while also addressing the increasing concerns of users and regulatory bodies [31].

5.1 The need for scalable solutions

A significant bottleneck in the implementation of industrial-scale RSs is the high computational cost of deep learning models [8]. As the size of the model increases, the computational and economic challenges of scaling increase due to the substantial costs of the compute and storage resources needed for training and deployment [23, 42]. To address the high computational cost in practical recommendation scenarios, one research direction is to improve the efficiency of the model [33, 57]. Efficient methods would reduce both the cost and the time associated with training complex models in large-scale practical settings [43]. However, these methods need to consider the trade-off between computational efficiency and model performance. Hardware-algorithm co-design is another interesting research direction towards improving model efficiency [7, 36]. Heterogeneous data, for example in graph neural networks, is irregular and difficult to scale on GPU hardware [47]. To address this issue, some works propose hardware-algorithm co-design to improve model efficiency. For example, Neo [32] enables memory-efficient embedding computations through techniques such as hybrid kernel fusion, software-managed caching, and quality-preserving compression. These techniques reduce the memory footprint of large models while maintaining high computational efficiency. Another innovation in this direction is the Sublinear Deep Learning Engine proposed by Chen et al. [7], which combines intelligent algorithms with multi-core parallelism and workload optimisation to significantly improve computational efficiency. Another solution, TorchRec [36], which aims to improve the scalability and efficiency of RS models, further highlights the importance of dedicated infrastructure in improving the performance of RSs at scale.

5.2 Feedback And Continuous Learning

In practical industrial scenarios, continuous feedback is essential to ensure that the RS remains effective in handling evolving workflow and user requirements [46]. User feedback facilitates timely responses to performance variations, enabling the system to adapt dynamically, which is crucial to maintaining trust, transparency and engagement. Continuous learning, where the model incrementally incorporates new data without requiring complete retraining, is an interesting direction for large-scale, production-level RS deployments [6]. An important insight from our study is that integrating real-time feedback mechanisms can substantially improve the adaptability and accuracy of industrial RSs. In our workflow, real-time feedback is collected directly from project managers and editors through the editorial dashboard used for managing localisation assignments. This process ensures that future recommendations reflect updated performance trends while maintaining fairness and cost efficiency. Unlike consumer-facing RSs, metrics such as click-through or conversion rates are not applicable in this context; instead, performance evaluation focuses on editor productivity, translation quality, and task completion efficiency. A potential future research direction on feedback and continuous learning involves the development of improved user interfaces that better facilitate feedback collection and model interpretability [4]. In our setting, the RS integrates with a web-based editorial dashboard that allows reviewers and project managers to accept, monitor, and assess tasks in real-time. This interface provides transparency in recommendation decisions and enables structured feedback loops. However, several challenges remain, such as compliance with regional data privacy regulations and the need for scalable API endpoints capable of synchronising multilingual editorial data streams. Future research could explore conversational interfaces powered by large language models (LLMs) to improve usability and human–machine collaboration in continuous learning scenarios [1].

5.3 Multi-Objective Learning

Although academic research on RSs is mostly focused on optimising single objectives such as accuracy, privacy and fairness, most industrial applications need to balance multiple objectives such as accuracy, revenue generation, fairness and computational efficiency [50]. For example, while improving the accuracy of recommendations might lead to higher user satisfaction, it could also increase computational costs or potentially introduce biases. Therefore, multi-objective RSs aim to simultaneously optimise various performance metrics, addressing these competing goals within a single framework. For example, optimising for both accuracy and fairness could require the introduction of fairness constraints into the objective function, adding another layer of complexity to the model design and training process. This would introduce challenges in implementing effective multi-objective optimisation strategies at scale [54]. In practical applications, where computational cost is a major concern, evaluating and managing trade-offs between performance metrics such as efficiency, scalability and cost is essential [14]. Methods such as Pareto Front Learning provide a foundation for handling these multi-objective problems in RSs [24]. These models attempt to optimise multiple objectives simultaneously by considering the "Pareto Front," a set of non-dominated solutions where one objective cannot be improved without worsening another [54]. Further research into multi-objective RSs could focus on developing new algorithms that better balance trade-offs between various objectives which include scalability and user satisfaction.

5.4 Ethical Considerations

Our work lays the foundation for aligning the industrial deployment of RS with ethical AI principles, balancing performance with fairness and privacy [31]. In industrial RSs, success cannot be measured solely through business key performance indicators (KPIs); ethical considerations such as fairness, transparency, privacy, and bias mitigation are equally important [45]. For example, human review of NMT content also helps filter culturally sensitive content. To support transparency, system logs are audited by both internal and external stakeholders, which is necessary to promote trust and engagement. In our system design, we anonymise sensitive candidate information to maintain privacy [21] and comply with regulations such as the General Data Protection Regulations[16]. In addition, the global pool of candidates enables diverse access to our system in diverse linguistic and cultural settings. This aims to promote a global view of ethical evaluations [3, 29].

6 Conclusion

In this work, we present a recommender system designed to accelerate large-scale video game localisation workflows. In complex tasks such as video game translation, where preserving the original player experience is critical, NMT alone often falls short, often struggling to accurately convey subtleties such as humour and emotional tone. To address this, human expertise remains essential. Our RS efficiently assigns the most suitable reviewers to post-edit NMT output, resulting in significant improvements in accuracy, processing speed, and cost-effectiveness across large-scale localisation pipelines. Empirical results from A/B testing on more than 292,000 tasks on more than

[16]https://gdpr.eu/

40 languages show that the RS significantly outperforms manual assignment, delivering up to 90% time savings and reducing costs by more than 76%. Future directions include advancing the proposed solution in other use cases that apply automated decision-making systems. Ethical design considerations, such as fairness and transparency, are integrated into the workflow, ensuring trust and global participation in industrial settings. The insights from this work pave the way for applying RSs in related domains, including poetry and comedy, where aligning human expression with machine intelligence remains necessary to preserve contextual, cultural and linguistic precision.

Acknowledgments

This research was carried out as part of the first author's placement at KantanAI, Keywords Studios Group[17]. This publication is supported by Science Foundation Ireland under Grant number 18/CRT/6183.

References

[1] Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. 2024. Experimental Interface for Multimodal and Large Language Model Based Explanations of Educational Recommender Systems. *arXiv preprint arXiv:2402.07910* (2024).

[2] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI We Trust? Perceptions About Automated Decision-making by Artificial Intelligence. *AI & Society* 35, 3 (2020), 611–623.

[3] Lex Beattie, Dan Taber, and Henriette Cramer. 2022. Challenges in Translating Research to Practice for Evaluating Fairness and Bias in Recommendation Systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 528–530.

[4] Joeran Beel and Haley Dixon. 2021. The 'Unreasonable' Effectiveness of Graphical User Interfaces for Recommender Systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 22–28.

[5] Maximiliana Behnke, Antonio Valerio Miceli Barone, Rico Sennrich, Vilelmini Sosoni, Thanasis Naskos, Eirini Takoulidou, Maria Stasimioti, Menno van Zaanan, Sheila Castilho, Federico Gaspari, et al. 2018. Improving Machine Translation of Educational Content via Crowdsourcing. In *11th Edition of the Language Resources and Evaluation Conference*. ELRA, Paris, France, 3343–3347.

[6] Guohao Cai, Jieming Zhu, Quanyu Dai, Zhenhua Dong, Xiuqiang He, Ruiming Tang, and Rui Zhang. 2022. Reloop: A Self-Correction Continual Learning Loop for Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 2692–2697.

[7] Beidi Chen, Tharun Medini, James Farwell, Charlie Tai, Anshumali Shrivastava, et al. 2020. Slide: In Defense of Smart Algorithms Over Hardware Acceleration for Large-Scale Deep Learning Systems. *Proceedings of Machine Learning and Systems* 2 (2020), 291–306.

[8] Derek Zhiyuan Cheng, Ruoxi Wang, Wang-Cheng Kang, Benjamin Coleman, Yin Zhang, Jianmo Ni, Jonathan Valverde, Lichan Hong, and Ed Chi. 2023. Efficient Data Representation Learning in Google-scale Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 267–271.

[9] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–38.

[10] Desakh Putu Setyalika Putri Dewayanti and Margana Margana. 2024. The impact of Contextual Understanding on Neural Machine Translation Accuracy: A Case Study of Indonesian Cultural Idioms in English Translation. *Englisia: Journal of Language, Education and Humanities* 12, 1 (2024), 223–236.

[11] Zhenhua Dong, Jieming Zhu, Weiwen Liu, and Ruiming Tang. 2023. Ten Challenges in Industrial Recommender Systems. *arXiv preprint arXiv:2310.04804* (2023).

[12] Marie Escribe. 2019. Human Evaluation of Neural Machine Translation: The Case of Deep Learning. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*. Incoma Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 36–46.

[13] Arno J Gingele, Hesam Amin, Kurt De Wit, Malte Jacobsen, Arjan Hageman, Kay van der Mierden, Julia Brandts, Jerremy Weerts, Matthew Barrett, Lana J Dixon, et al. 2023. Developing an AI-Based Decision Engine for Disease-Modifying Therapy in Heart Failure–A Pilot Study. *European Heart Journal-Digital Health* 6, 2 (2023), ztad075.

[14] Nyoman Gunantara. 2018. A Review of Multi-Objective Optimization: Methods and Its Applications. *Cogent Engineering* 5, 1 (2018), 1502242.

[17] https://www.keywordsstudios.com/

[15] Manzoor Ahmed Hashmani, Syed Muslim Jameel, Mobashar Rehman, and Atsushi Inoue. 2020. Concept drift Evolution in Machine Learning Approaches: A Systematic Literature Review. *International Journal on Smart Sensing and Intelligent Systems* 13, 1 (2020), 1.

[16] Hassan Mahill Abdallah Hassan, Abdelrahman Elyass Mohamed Abdelmajd, Aziz Abdulrab Saleh Al Salafi, et al. 2019. Investigating the Inadequacy of Machine Translation in Conveying the Sense and Sensibility Towards Arabic Texts Translated into English. *International Journal of Linguistics, Literature and Translation* 2, 1 (2019), 42–49.

[17] Talha Iqbal, Mehedi Masud, Bilal Amin, Conor Feely, Mary Faherty, Tim Jones, Michelle Tierney, Atif Shahzad, and Patricia Vazquez. 2024. Towards Integration of Artificial intelligence into medical Devices as a Real-Time Recommender System for personalised healthcare: State-of-the-art and Future Prospects. *Health Sciences Review* 10 (2024), 100150.

[18] Yanfang Jia, Michael Carl, and Xiangling Wang. 2019. How Does the Post-Editing of Neural Machine Translation Compare with From-Scratch Translation? A Product and Process Study. *The Journal of Specialized Translation* 31, 1 (2019), 60–86.

[19] Ahmed Khan and Aaliya Sarfaraz. 2019. RNN-LSTM-GRU Based Language Transformation. *Soft Computing* 23, 24 (2019), 13007–13024.

[20] Fatima Khan. 2024. Human-in-the-Loop Approaches to Improving Machine Translation. *Academic Journal of Science and Technology* 7, 1 (2024), 1–8.

[21] Kungang Li, Xiangyi Chen, Ling Leng, Jiajing Xu, Jiankai Sun, and Behnam Rezaei. 2024. Privacy Preserving Conversion Modeling in Data Clean Room. In *Proceedings of the 18th ACM Conference on Recommender Systems*. ACM, New York, USA, 819–822.

[22] Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate Meanings, Not Just Words: Idiomkb's Role in Optimising Idiomatic Translation with Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. AAAI, Palo Alto, CA, USA, 18554–18563.

[23] Yuening Li, Diego Uribe, Chuan He, Jiaxi Tang, Qingyun Liu, Junjie Shan, Ben Most, Kaushik Kalyan, Shuchao Bi, Xinyang Yi, et al. 2024. Short-form Video Needs Long-term Interests: An Industrial Solution for Serving Large User Sequence Models. In *Proceedings of the 18th ACM Conference on Recommender Systems*. ACM, New York, USA, 832–834.

[24] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. 2019. A Pareto-Efficient Algorithm for Multiple Objective Optimization in E-commerce Recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, New York, USA, 20–28.

[25] Bill Lubanovic. 2023. *FastAPI: Modern Python Web Development* (1st ed.). O'Reilly Media, Inc., Sebastopol, CA, USA. 277 pages.

[26] Lieve Macken, Daniel Prou, and Arda Tezcan. 2020. Quantifying the Effect of Machine Translation in a High-Quality Human Translation Production Process. In *Informatics*, Vol. 7. MDPI, Basel, Switzerland, 12.

[27] Debabrata Mahapatra and Vaibhav Rajan. 2020. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *International Conference on Machine Learning*. PMLR, Cambridge, MA, USA, 6597–6607.

[28] Francesca Marzi, Giordano d'Aloisio, Antinisca Di Marco, and Giovanni Stilo. 2023. Towards a Prediction of Machine Learning Training Time to Support Continuous Learning Systems Development. In *European Conference on Software Architecture*. Springer, Cham, Switzerland, 169–184.

[29] Elio Masciari, Areeba Umair, and Muhammad Habib Ullah. 2024. A Systematic Literature Review on AI-Based Recommendation Systems and Their Ethical Considerations. *IEEE Access* 12 (2024), 121223–121241.

[30] Evgeny Matusov. 2019. The Challenges of Using Neural Machine Translation for Literature. In *Proceedings of the Qualities of Literary Machine Translation*. ACL, Stroudsburg, PA, USA, 10–19.

[31] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender Systems and Their Ethical Challenges. *Ai & Society* 35 (2020), 957–967.

[32] Dheevatsa Mudigere, Yuchen Hao, Jianyu Huang, Zhihao Jia, Andrew Tulloch, Srinivas Sridharan, Xing Liu, Mustafa Ozdal, Jade Nie, Jongsoo Park, et al. 2022. Software-Hardware Co-design for Fast and Scalable Training of Deep Learning Recommendation Models. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*. ACM and IEEE, New York, NY, USA, 993–1011.

[33] Tendai Mukande, Esraa Ali, Annalina Caputo, Ruihai Dong, and Noel E O'Connor. 2023. A Flash Attention Transformer for Multi-Behaviour Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 4210–4214.

[34] Bryan Ng. 2018. Self-Service Workflows for Recommendation Systems Using Online Machine Learning Services. In *2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*. IEEE, New York, NY, USA, 1–6.

[35] Minako O'Hagan. 2009. Towards a Cross-Cultural Game Design: An Explorative Study in Understanding the Player Experience of a Localised Japanese Video Game. *The Journal of Specialised Translation* 11, 1 (2009), 211–233.

[36] Intaik Park, Ehsan Ardestani, Damian Reeves, Sarunya Pumma, Henry Tsang, Levy Zhao, Jian He, Joshua Deng, Dennis Van der Staay, Yu Guo, et al. 2024. Toward 100TB Recommendation Models with Embedding Offloading. In *Proceedings of the 18th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 841–843.

[37] Silvia Pettini. 2021. *The Translation of Realia and Irrealia in Game Localization: Culture-Specificity Between Realism and Fictionality*. Routledge, Oxfordshire, UK.

[38] Marco Pirrone and Arianna D'Ulizia. 2024. The Localization of Software and Video Games: Current State and Future Perspectives. *Information* 15, 10 (2024), 648.

[39] Aung Pyae. 2018. Understanding the Role of Culture and Cultural Attributes in Digital Game Localization. *Entertainment Computing* 26 (2018), 105–116.

[40] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-Resource Languages: A Survey. *Comput. Surveys* 55, 11 (2023), 1–37.

[41] Michael Ruchte and Josif Grabocka. 2021. Scalable Pareto Front Approximation for Deep Multi-Objective Learning. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, New York, NY, USA, 1306–1311.

[42] Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, James Caverlee, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2024. Improving Data Efficiency for Recommenders and LLMs. In *Proceedings of the 18th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 790–792.

[43] Yuan Shao, Bibang Liu, Sourabh Bansod, Arnab Bhadury, Mingyan Gao, and Yaping Zhang. 2024. Optimizing for Participation in Recommender System. In *Proceedings of the 18th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 806–808.

[44] Felix Stahlberg. 2020. Neural Machine Translation: A Review. *Journal of Artificial Intelligence Research* 69 (2020), 343–418.

[45] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, et al. 2024. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *ACM Transactions on Recommender Systems* 2, 3 (2024), 1–57.

[46] Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. 2019. Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, New York, NY, USA, 645–651.

[47] Shyam Tailor. 2022. *Practical Processing and Acceleration of Graph Neural Networks*. Ph. D. Dissertation. University of Cambridge, Cambridge, UK.

[48] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *Comput. Surveys* 55, 6 (2022), 1–28.

[49] Kizito Tekwa and Jessica Liu Jiexiu. 2023. Neural Machine Translation Systems and Chinese Wuxia Movies: Moving into Uncharted Territory. In *Understanding and Translating Chinese Martial Arts*. Springer Nature Singapore, Singapore, 71–89.

[50] Ye Tian, Langchun Si, Xingyi Zhang, Ran Cheng, Cheng He, Kay Chen Tan, and Yaochu Jin. 2021. Evolutionary Large-Scale Multi-Objective Optimization: A Survey. *ACM Computing Surveys (CSUR)* 54, 8 (2021), 1–34.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS '17, Vol. 30)*. Curran Associates, Inc., Long Beach, CA, 5998–6008.

[52] Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang, and Boxing Chen. 2022. Challenges of Neural Machine Translation for Short Texts. *Computational Linguistics* 48, 2 (2022), 321–342.

[53] Xing Wang, Zhaopeng Tu, and Min Zhang. 2018. Incorporating Statistical Machine Translation Word Knowledge into Neural Machine Translation. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 26, 12 (2018), 2255–2266.

[54] Timo Wilm, Philipp Normann, and Felix Stepprath. 2024. Pareto Front Approximation for Multi-Objective Session-Based Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 809–812.

[55] Chuhan Wu, Qinglin Jia, Zhenhua Dong, and Ruiming Tang. 2023. Customer Lifetime Value Prediction: Towards the Paradigm Shift of Recommender System Objectives. In *Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 1293–1294.

[56] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation. *arXiv preprint arXiv:1609.08144* (2016).

[57] Rengan Xu, Junjie Yang, Yifan Xu, Hong Li, Xing Liu, Devashish Shankar, Haoci Zhang, Meng Liu, Boyang Li, Yuxi Hu, et al. 2024. Enhancing Performance and Scalability of Large-Scale Recommendation Systems with Jagged Flash Attention. In *Proceedings of the 18th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 778–780.

[58] Eva Zangerle and Christine Bauer. 2022. Evaluating Recommender Systems: Survey and Framework. *Comput. Surveys* 55, 8 (2022), 1–38.
[59] Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural Machine Translation Quality and Post-Editing Performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. ACL, Stroudsburg, PA, USA, 10204–10214.

Received 1 December 2024; revised 12 July 2025; accepted 11 November 2025