# (Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts

**Minghao Wu**[1] **Jiahao Xu**[2] **Yulin Yuan**[3] **Gholamreza Haffari**[1]
**Longyue Wang**[4†] **Weihua Luo**[4] **Kaifu Zhang**[4]
[1]Monash University, Australia [2]Nanyang Technological University, China
[3]University of Macau, Macau SAR, China [4]Alibaba International Digital Commerce, China
{minghao.wu, gholamreza.haffari}@monash.edu
Jiahao004@e.ntu.edu.sg, yulinyuan@um.edu.mo
{wanglongyue.wly, weihua.luowh, kaifu.zkf}@alibaba-inc.com

## Abstract

Literary translation remains one of the most challenging frontiers in machine translation due to the complexity of capturing figurative language, cultural nuances, and unique stylistic elements. In this work, we introduce TRANSAGENTS, a novel multi-agent framework that simulates the roles and collaborative practices of a human translation company, including a CEO, Senior Editor, Junior Editor, Translator, Localization Specialist, and Proofreader. The translation process is divided into two stages: a preparation stage where the team is assembled and comprehensive translation guidelines are drafted, and an execution stage that involves sequential translation, localization, proofreading, and a final quality check. Furthermore, we propose two innovative evaluation strategies: Monolingual Human Preference (MHP), which evaluates translations based solely on target language quality and cultural appropriateness, and BLP, which leverages large language models like GPT-4 for direct text comparison. Although TRANSAGENTS achieves lower $d$-BLEU scores, due to the limited diversity of references, its translations are significantly better than those of other baselines and are preferred by both human evaluators and LLMs over traditional human references and GPT-4 translations. Our findings highlight the potential of multi-agent collaboration in enhancing translation quality, particularly for longer texts.[1]

## 1 Introduction

Machine translation (MT) has achieved significant advancements recently, driven by breakthroughs in deep learning and neural networks (Cho et al., 2014; Sutskever et al., 2014; Vaswani et al., 2017; Gu et al., 2019; Liu et al., 2020; Fan et al., 2021). Despite these technological strides, literary translation remains a challenging area for MT systems. Literary texts, with their complex language, figurative expressions, cultural nuances, and unique stylistic elements, pose significant hurdles that are difficult for machines to overcome (Voigt and Jurafsky, 2012). This complexity makes literary translation one of the most challenging areas within machine translation, referred to as ''the last frontier of machine translation'' (Klemin, 2024).

Recent research in multi-agent systems, especially those using large language models (LLMs), shows significant promise (Yao et al., 2023; Wang et al., 2023c; Dong et al., 2023). These systems harness the collective intelligence of multiple agents, outperforming individual models in dynamic environments that require complex problem-solving and collaboration. In this work, we leverage LLMs to simulate the various roles in a human translation company and propose TRANSAGENTS, a multi-agent framework designed to mimic the best practices of human translation. The translation process in TRANSAGENTS is divided into two main stages: preparation and execution, each comprising several sub-stages. In the preparation stage, the designated CEO agent selects a Senior Editor based on the specific needs of the client. The Senior Editor then assembles a team from a roster of roles, including Junior Editor, Translator, Localization Specialist, and Proofreader. Together, the Senior Editor and Junior Editor draft a comprehensive translation guideline to ensure consistency and quality throughout the process. During the execution stage, the Translator agent translates the text chapter by chapter,

†Longyue Wang is the corresponding author.

[1]GitHub Repository: https://github.com/minghao-wu/transagents.



Transactions of the Association for Computational Linguistics, vol. 13, pp. 901–922, 2025. https://doi.org/10.1162/tacl.a.25
Action Editor: Daniel Gildea. Submission batch: 2/2025; Revision batch: 8/2024; Published 8/2025.

adhering to the guidelines. The Localization Specialist revises the translated text to ensure cultural and contextual alignment with the target audience. Subsequently, the Proofreader reviews the translation to eliminate errors and verify adherence to the guideline. Finally, the Senior Editor conducts a thorough quality check to ensure the final translation meets the highest standards. By dividing the workflow into distinct stages and assigning specialized roles, TRANSAGENTS effectively emulates the structured approach of human translation companies, leveraging the capabilities of LLMs to enhance translation quality and efficiency.

Furthermore, evaluating the accuracy and quality of literary translations is challenging due to the subjective nature of literature and the potential imperfections in reference translations (Thai et al., 2022; Freitag et al., 2023). To address these challenges, in addition to conventional MT evaluation metrics, we propose two innovative evaluation strategies: *Monolingual Human Preference* (MHP) and *Bilingual LLM Preference* (BLP). Monolingual Human Preference involves human evaluators from the target audience assessing translations without referring to the original text. This approach focuses on fluidity, readability, and cultural appropriateness, simulating the real-world consumption of literature. BLP leverages advanced LLMs, such as GPT-4, which are provided with the original texts to facilitate direct comparison. This method aims to utilize the superior translation capabilities of LLMs, mitigating the impact of imperfect reference translations.

Our empirical findings reveal that, while TRANSAGENTS achieves the lowest $d$-BLEU scores, it significantly outperforms the state-of-the-art machine translation (MT) method in terms of GEMBA-DA. Additionally, both human evaluators and a LLM evaluator prefer the translations produced by TRANSAGENTS over human-written references and GPT-4 translations. Our analysis suggests that the significant decline in $d$-BLEU scores is primarily due to the limited diversity of the reference translations. In contrast, the translations generated by TRANSAGENTS exhibit significantly greater diversity compared to other approaches. Furthermore, we demonstrate the effectiveness of our design in terms of agent profiling and collaboration strategies. Interestingly, while TRANSAGENTS excels in translating long texts and significantly outperforms other baselines in this area, it struggles with translating shorter texts effectively.

Our contributions are summarized as follows:

- We present TRANSAGENTS, a novel multi-agent system for literary translation that emulates the traditional translation publication process. This multi-agent approach effectively addresses the complex nuances inherent in literary works.
- We propose two novel evaluation strategies, *Monolingual Human Preference* (MHP) and *Bilingual LLM Preference* (BLP) to assess the quality of translations. These methods address the limitations of traditional machine translation evaluation metrics.
- Despite achieving lower $d$-BLEU scores, our empirical results demonstrate that translations produced by TRANSAGENTS are significantly better than other approaches in terms of GEMBA-DA and preferred by both human evaluators and language models over human references and GPT-4 translations. Furthermore, we offer a comprehensive analysis of the strengths and weaknesses of TRANSAGENTS.

## 2 Related Work

**Machine Translation** Machine translation (MT) has seen significant advancements recently. However, these improvements are mainly at the sentence level. Recent efforts focus on integrating contextual information to enhance translation quality beyond individual sentences (Wang et al., 2017; Wu et al., 2023; Herold and Ney, 2023; Wu et al., 2024b). LLMs have also shown superior capabilities in MT (Xu et al., 2023a; Robinson et al., 2023; Wang et al., 2023a; Wu et al., 2024a). Despite progress, MT performance in the general domain is saturating, shifting interest towards literary translation, which demands accuracy and cultural nuance. Karpinska and Iyyer (2023) show that while LLMs effectively leverage document-level context for literary translation, they still make critical errors. Wang et al. (2024a) evaluate LLMs' ability to translate long texts and propose context extrapolation to improve translation quality. Additionally, recent research explores using LLMs for evaluating literary translations (Yan et al., 2024; Zhang

et al., 2024). In this work, we introduce a novel multi-agent virtual company for literary translation and propose two evaluation strategies for assessing translation quality.

**Multi-Agent Systems** Intelligent agents are designed to understand their environments, make informed decisions, and respond appropriately (Wooldridge and Jennings, 1995). Compared to single-agent setups, multi-agent systems leverage collaboration among multiple agents based on LLMs to tackle complex problems or simulate real-world environments effectively (Guo et al., 2024). Recent studies have shown promising outcomes in areas such as software development (Qian et al., 2023; Hong et al., 2023), multi-robot collaboration (Mandi et al., 2023; Zhang et al., 2023), evaluation (Chan et al., 2023; Verga et al., 2024), and fact-checking (Du et al., 2023a). Additionally, extensive research explores using multiple agents to simulate societal, economic, and gaming environments (Park et al., 2022, 2023; Xu et al., 2023b; Li et al., 2023; Mukobi et al., 2023). Liang et al. (2023) proposes leveraging multi-agent debate for machine translation. However, their approach is limited to the sentence level.

**Ours** In this work, we utilize LLMs to replicate the traditional translation workflow employed by human translation companies, streamlining the process step by step and present a demo (Wu et al., 2024c). Recently, Briakou et al. (2024) introduce a step-by-step approach for translating long texts. Additionally, Wang et al. (2024c) proposed a multi-agent framework designed to preserve long-term memory in document-level machine translation. These studies are concurrent with our work and share similar goals of enhancing translation quality through structured methodologies.

## 3 TRANSAGENTS: A Multi-Agent Company for Literary Translation

### 3.1 Company Organization

To simulate the entire book translation process, TRANSAGENTS involves a diverse range of roles besides the CEO, including senior editors, junior editors, translators, localization specialists, and proofreaders, each with distinct responsibilities:

- **Senior Editors**: Senior editors are responsible for overseeing the content production process. Their primary duties encompass setting editorial standards, guiding other team members, and making decisions regarding publication schedules and content direction.
- **Junior Editors**: Junior editors work under senior editors, managing day-to-day editorial workflows, editing content, and assisting in content planning, and handling communications with various roles within the team.
- **Translators**: Translators convert written material from one language to another, preserving the original text's tone, style, and context.
- **Localization Specialists**: Localization specialists adapt content for specific regions or markets, translating and adjusting cultural references to resonate with local audiences.
- **Proofreaders**: Proofreaders perform final checks for grammar, spelling, punctuation, and other possible errors.

```
Name: Sofia Chang
Languages: English, Mandarin, Spanish, French
Nationality: Canadian
Gender: Female
Age: 47
Education: Ph.D. in Comparative Literature
Personality: meticulous, introverted, perfectionist,
↪  critical, thoughtful
Hobbies: gardening, chess, watercolor painting
Rate per word: 0.12
Years of working: 22
Profession: Senior Editor
Role prompt: You are Sofia Chang, a highly esteemed Senior
↪  Editor [TRUNCATED]
```

Figure 1: An example profile of **Senior Editor**.

**Agentization** To enhance the realism and efficacy of our translation process simulation, we use GPT-4-TURBO to generate 30 diverse virtual agent profiles for each role. As shown in Figure 1, these profiles encompass a wide range of attributes beyond language skills, including gender, nationality, rate per word, education, experience, and more. This detailed approach enriches the simulation's authenticity and reflects the complexity and diversity of real-world translation settings, thereby supporting and inspiring future research.

### 3.2 Translation Workflow

As visualized in Figure 2, we introduce the book translation workflow in our company TRANSAGENTS, including two main stages—preparation (Section 3.2.1) and execution (Section 3.2.2)—in this section.

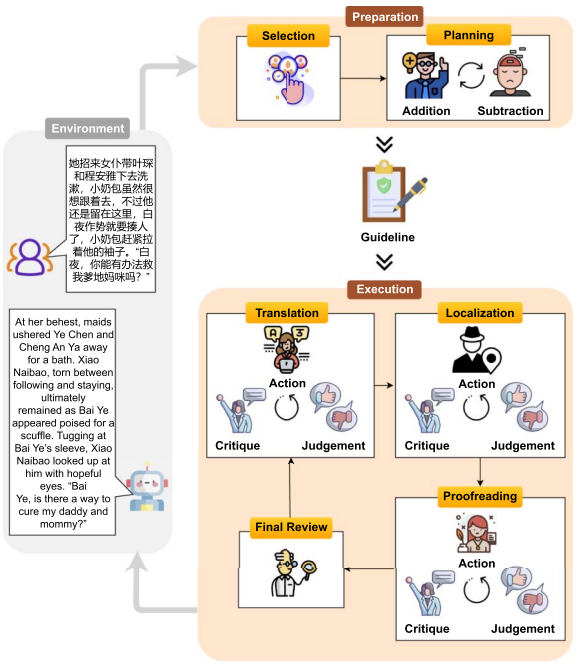


Figure 2: The workflow of TRANSAGENTS.

### 3.2.1 Preparation

**Project Member Selection** In our company, we create 30 agent profiles for each role, each with a unique role assignment prompt (see Figure 1). These prompts help assign specific roles to agents before dialogues begin. Initially, the CEO selects a Senior Editor for the book translation project, considering the client's requirements and the candidates' qualifications. Once chosen, the Senior Editor collaborates with the CEO to form the project team, taking into account the candidates' skills and backgrounds. Additionally, we employ a *self-reflection* strategy (Yao et al., 2023; Shinn et al., 2023; Qian et al., 2023) by using a ''ghost agent'' to prompt the CEO to reconsider their decision, as they sometimes struggle to select a Senior Editor with the necessary language skills.

**Addition-by-Subtraction Collaboration** In our framework, we introduce the *Addition-by-Subtraction Collaboration* between two agents. Unlike the debate-style strategy, which involves multiple agents proposing answers and a third-party agent concluding, our method uses only two agents. One is the *Addition agent*, who extracts extracting as much relevant information as possible, and the other is the *Subtraction agent*, who reviews and eliminates redundant details, providing feedback to the Addition agent. We present the details of our collaboration strategy in

**Algorithm 1:** Addition-by-Subtraction Collaboration

**Input :** $\mathbf{C}$, the context; $\mathbf{I}$, the instruction; $\mathbf{M}$, the maximum number of iterations; $m$, the index of current iteration; $\mathbf{A}$, the Addition agent; $\mathbf{S}$, the Subtraction agent; $\mathbf{F}$, the feedback from the Subtraction agent; $\mathbf{R}'$, the temporary response;

**Output:** $\mathbf{R}$, the final response;

$\mathbf{H} \leftarrow [\mathbf{C}; \mathbf{I}]$ ▷ Initialize the conversation;
$\mathbf{R} \leftarrow \emptyset$ ▷ Initialize the response;
$m \leftarrow 0$ ▷ Current round;
**while** $m < \mathbf{M}$ **do**
  $m \leftarrow m + 1$;
  $\mathbf{R}' \leftarrow \mathbf{A}(\mathbf{H})$ ▷ Generate detailed response;
  $\mathbf{F} \leftarrow \mathbf{S}(\mathbf{H}, \mathbf{R}')$ ▷ Review and remove redundant information;
  $\mathbf{H} \leftarrow \mathbf{H} + [\mathbf{R}'; \mathbf{F}]$ ▷ Append $\mathbf{R}'$ and $\mathbf{F}$ to the conversation history $\mathbf{H}$;
  **if** $\mathbf{R} = \mathbf{R}'$ **then**
    Break ▷ Stop iterating as no further revisions are needed;
  $\mathbf{R} \leftarrow \mathbf{R}'$;
Return the final response $\mathbf{R}$;

Algorithm 1. The Addition agent $\mathbf{A}$ generates an initial response with comprehensive content. The Subtraction agent $\mathbf{S}$ then reviews this response, removes redundancies, and the agents iterate this process until no further revisions are necessary. The iteration process can terminate before reaching the maximum number of iterations if the ''stop'' condition is met. This approach is referred to as the ''early exit'' mechanism. Note that the maximum number of iterations used in Algorithm 1 is 2.

**Long-Term Memory Management** One of the principal challenges in literary translation is to maintain the coherence, cohesion, and consistency throughout the translated text. In this work, we classify memory into two types: *short-term memory*, which pertains to the context within the current conversation, and *long-term memory*, which relates to the context spanning the entire book. Recent efforts allow LLMs to effectively process the sequence containing several thousands of tokens (Xiong et al., 2023; Gu and Dao, 2023;

Botev et al., 2024), but they remain inadequate for handling the broader context required for entire books (Karpinska et al., 2024). Addressing this limitation, we employ multi-agent collaboration to develop a comprehensive set of translation guidelines, guiding all the agents involved in the translation process. In TRANSAGENTS, there are five components in the translation guidelines: the glossary, the book summary, the tone, the style, and the target audience. We design different strategies to process them:

- **Glossary and Book Summary**: The glossary of the book compiles essential terms from the source language and provides their corresponding translations in the target language, while the book summary provide a comprehensive overview of the whole book. Both components are facilitated by the collaboration between the Junior Editor (Addition Agent $\mathbf{A}$) and the Senior Editor (Subtraction Agent $\mathbf{S}$), employing the *Addition-by-Subtraction Collaboration* as depicted in Algorithm 1. For example, when constructing the glossary, the Junior Editor extracts as many terms as possible, while the Senior Editor attempts to remove generic terms.
- **Tone, Style, and Target Audience**: The translation of a book is more than just a word-for-word conversion; it's a delicate process of adapting tone, style, and content to resonate with the target audience while staying true to the original text's essence. In TRANSAGENTS, the Senior Editor defines the tone, the style, and the target audience of the translations based on a random chapter.

Overall, the glossary, book summary, tone, style, and target audience together form the comprehensive translation guidelines. These guidelines are prefixed as an essential part of the prompts for all roles involved in the translation process. We present an example prompt for the Translator in Figure 3. More prompts are in Appendix D.

### 3.2.2 Execution

In the execution phase, the process is divided into four sub-stages: translation, localization, proofreading, and final review. The first three sub-stages employ a collaborative strategy as detailed in Algorithm 2. Here, the Translator, Localization Specialist, and Proofreader act as

```
# Translation Guidelines

## Glossary
罗德: Rhode
虚空之龙: Void Dragon
星月佣兵团: Star Moon Mercenary Corps
[TRUNCATED]

## Book Summary
The book centers on Rhode Alante, initially a Summoner
↪  Swordsman in the game 'Dragon Soul Continent,'
↪  [TRUNCATED]

## Tone
The tone of the book is adventurous and immersive with
↪  elements of fantasy and suspense. [TRUNCATED]

## Style
The book is a gripping blend of fantasy and litRPG,
↪  characterized by its immersive world-building, dynamic
↪  combat scenes, and a clear progression system.
↪  [TRUNCATED]

## Target Audience
The target audience for this book includes young adults
↪  and adults who enjoy fantasy and adventure genres,
↪  particularly those who are fans of MMORPG [TRUNCATED]

# Chapter Text
序章 传奇落幕
乌云笼罩着天空，昏暗无光的地面上四周都是一片狼籍。
[TRUNCATED]

# Instruction
Translate the chapter text from Chinese into English.
↪  Ensure that your translation closely adheres to the
↪  provided translation guidelines [TRUNCATED]
```

Figure 3: An example prompt for the Translator, including the translation guidelines, the chapter text in the source language, and the instruction.

Action agents $\mathbf{P}$. The Junior Editor serves as the Critique agent $\mathbf{Q}$, and the Senior Editor functions as the Judgment agent $\mathbf{J}$. The Senior Editor also conducts the final checks before publication.

**Trilateral Collaboration** We divide the collaboration in TRANSAGENTS into three branches, referred to as *Trilateral Collaboration*:

- **Action**: Executes instructions and implements required actions.
- **Critique**: Reviews the response and provides constructive feedback to the Action branch.
- **Judgment**: Makes the final decision on whether the response is satisfactory or needs further revision.

Each branch is managed by a dedicated agent as detailed in Algorithm 2. The *Action agent* $\mathbf{P}$ generates a response $\mathbf{R}$ based on the context $\mathbf{C}$ and instruction $\mathbf{I}$. The *Critique agent* $\mathbf{Q}$ then critiques the response $\mathbf{R}$. The *Action agent* $\mathbf{P}$ can either accept the critiques and update the response or maintain the original response. Finally, the *Judgment agent* $\mathbf{J}$ evaluates the response $\mathbf{R}$ to decide if the discussion can be concluded, *without requiring the conversation history*, due to the agents' limited

**Algorithm 2:** Trilateral Collaboration

**Input :** $\mathbf{C}$, the context; $\mathbf{I}$, the instruction; $\mathbf{M}$, the maximum number of iterations; $m$, the index of current iteration; $\mathbf{P}$, the Action agent; $\mathbf{Q}$, the Critique agent; $\mathbf{J}$, the Judgment agent; $\mathbf{F}$, the feedback from the Critique agent; $\mathbf{D}$, the judgment decision;

**Output:** $\mathbf{R}$, the final response;

$\mathbf{H} \leftarrow [\mathbf{C}; \mathbf{I}]$ ▷ Initialize the conversation;
$m \leftarrow 0$ ▷ Current round;
**while** $m < \mathbf{M}$ **do**
  $m \leftarrow m + 1$;
  $\mathbf{R} \leftarrow \mathbf{P}(\mathbf{H})$ ▷ Generate response;
  $\mathbf{F} \leftarrow \mathbf{Q}(\mathbf{H}, \mathbf{R})$ ▷ Generate critiques;
  $\mathbf{H} \leftarrow \mathbf{H} + [\mathbf{R}; \mathbf{F}]$ ▷ Append $\mathbf{R}'$ and $\mathbf{F}$ to the conversation history $\mathbf{H}$;
  **if** $m > 1$ **then**
    $\mathbf{D} \leftarrow \mathbf{J}(\mathbf{C}, \mathbf{I}, \mathbf{R})$ ▷ The Judgment agent $\mathbf{J}$ evaluate the response quality;
    **if** $\mathbf{D} = \textit{TRUE}$ **then**
      Break ▷ Stop iterating if the Judgment agent $\mathbf{J}$ thinks the response is of high quality;
Return the final response $\mathbf{R}$;

capability of processing long-range context. Similar to Algorithm 1, we also introduce the ''early exit'' mechanism if *Judgment agent* $\mathbf{J}$ thinks the iteration process can terminate. Note that the maximum number of iterations in Algorithm 2 is 2.

**Translation, Localization, and Proofreading** The translation stage involves three key roles: the Translator, the Junior Editor, and the Senior Editor, who collaborate to translate the book from the source language to the target language on a chapter-by-chapter basis. The process starts with the Translator (Action agent $\mathbf{P}$) translating the chapter content. The Junior Editor (Critique agent $\mathbf{Q}$) then reviews the translation, ensuring it adheres to guidelines and identifying any potential errors or areas for improvement. Lastly, the Senior Editor (Judgment agent $\mathbf{J}$) evaluates the translation and decides if further revision is needed. Following translation, the cultural adaptation process begins. The Localization Specialist adapts the content to fit the cultural context of the target audience, ensuring it resonates well and maintains the intended meaning. Next, the Proofreader then checks for language errors. Throughout cultural adaptation and proofreading, both the Junior Editor and the Senior Editor continue to critique and evaluate the content for further refinement.

**Final Review** The final review is the concluding step in the editorial process. Here, the Senior Editor evaluates the translation quality and checks the flow between adjacent chapters. The review ensures each chapter is coherent and meets quality standards while maintaining smooth transitions for narrative consistency. If the translation does not meet the required standards during the final review, the steps in the execution stage are repeated to address any issues, as shown in Figure 2.

### 3.3 Discussion

**System Design** Our multi-agent framework aligns with the commonly recognized translation project management process (Landers, 2001; Pérez, 2002; Terhaar et al., 2019; Walker, 2022). Besides the preparation stage, there are three core steps in the execution stage—translation, localization, and proofreading—each with its own specific aims. The translation phase leverages multi-agent collaboration to draft a coherent, context-aware translation from the source text. Although the translation from this step is already highly accurate thanks to the detailed translation guidelines, its output can still contain inaccuracies or lack cultural nuances, which is why the subsequent phases are indispensable. Next, we employ the localization phase to adapt the text more deeply to cultural, stylistic, or domain-specific norms, which is a distinct and critical step for achieving a fluent and natural translation of the content. Furthermore, the proofreading step ensures the translated text meets quality assurance standards by resolving residual errors such as typos, omissions, or inconsistencies. Together, these steps enable TRANSAGENTS to produce high-quality translations, like those from human translation agencies.

**Collaboration Strategies Comparison** In this work, we design two collaboration strategies, *Addition-by-Subtraction Collaboration* and *Trilateral Collaboration*. These two strategies are designed based on the tasks the agents are dealing with and the capability of LLM backbone. They are not directly comparable. When constructing

the translation glossary, we observe that the *Addition agent* is very likely to include some generic words into the list, so we introduce the *Subtraction agent* to remove these generic words. The iteration stops until the translation glossary does not update. However, this strategy is not suitable for translation, localization, and proofreading, as adding and removing content can significant degrade the translation quality. Conversely, *Trilateral Collaboration* cannot replace *Addition-by-Subtraction Collaboration* either. In our preliminary study, we observe that the *Critique agent* and the *Judgment agent* often fail to exclude those generic terms if we leverage *Trilateral Collaboration* to construct the translation glossary, which make translation guideline overly lengthy. Long context modeling is one of the key challenges of current LLMs. An overly lengthy translation guideline can degrade the translation quality.

## 4 Experimental Setup

In this work, our experimental setup primarily follows the WMT2023 shared task on discourse-level literary translation (DLLT) (Wang et al., 2023b, 2024b). The following sections introduce the baselines and datasets (Section 4.1), and evaluation approaches (Section 4.2) used in our study.

### 4.1 Baselines and Datasets

We leverage the state-of-the-art LLM GPT-4-TURBO as the backbone of our agents,[2] and compare our approach with the unconstrained systems in WMT2023 shared task on DLLT, including LLAMA-MT (Du et al., 2023b), GPT-4 (OpenAI, 2023), GOOGLE TRANSLATE, DUT (Zhao et al., 2023), and HW-TSC (Xie et al., 2023). We also involve GPT-4-TURBO, GPT-4O-MINI,[3] and GPT-4O[4] as our baselines.

In this work, we only leverage the official test set of the WMT2023 shared task on DLLT for evaluation. The official test set is collected from 20 web novels, each of which consists 20 consecutive chapters, totaling 240 chapters. Each chapter contains approximately 1,404 English words on average. The test set contains two references: REFERENCE 1 is translated by human translators and REFERENCE 2 is built by manually aligning bilingual text in web pages.

[2] `gpt-4-1106-preview`.

[3] `gpt-4o-mini-2024-07-18`.

[4] `gpt-4o-2024-08-06`.

### 4.2 Evaluation

In this work, we employ two evaluation approaches: Standard Evaluation (Section 4.2.1) and Preference Evaluation (Section 4.2.2).

#### 4.2.1 Standard Evaluation

Following Wang et al. (2023b), we use $d$-BLEU (Papineni et al., 2002; Post, 2018; Liu et al., 2020) to evaluate the translation quality,[5] as the translations may not strictly align with the source text on a sentence-by-sentence basis. To compute the $d$-BLEU score, we concatenate all the chapter translations into a single document for evaluation. Both references in the test set are used for computing $d$-BLEU. Furthermore, we use GEMBA-DA (Kocmi and Federmann, 2023b) with GPT-4O as the evaluator to assess translation quality chapter by chapter. It is important to note that recent neural MT metrics, such as COMET (Rei et al., 2020), are typically designed for sentence-level evaluation and have restricted context sizes. For instance, COMET has a context window of only 512 tokens, while DOCCOMET (Vernikos et al., 2022) supports slightly longer contexts but requires sentence-by-sentence alignment. Additionally, we also conduct the bootstrap resampling and significance testing (Koehn, 2004) to better understand the results.

#### 4.2.2 Preference Evaluation

Acknowledging the multifaceted nature of literary texts is essential, as they do not have a single, universal translation. Standard translation evaluations, which rely on comparisons to a standard reference, fall short in capturing this complexity. Following Thai et al. (2022), we segment each chapter into approximately 150-word sections based on the story development and then utilize both human evaluators and LLMs to assess translations based on their preferences. We detail our evaluation methods in this section.

**Monolingual Human Preference (MHP)** When reading a translated book, understanding the original language is unnecessary for the target audience. Thus, readers should prefer a better translation without referencing the original text. Human evaluators compare pairs of translation segments describing the same part of the story and select their preferred translation using the

[5] `nrefs:2|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`.

provided interface in Appendix A. To ensure context is considered, evaluators review all segments within a chapter in their original order, as segments may rely on preceding information.

**Implementation Details of MHP** In this work, we collect human preferences on translation segments through SurveyMonkey.[6] We only recruit evaluators from the United States to minimize potential impacts of demographics. Each translation pair is evaluated by at least 10 people and costs us 0.30 USD per annotation. During the annotation process, we randomly swap the positions of translation segments for each comparison to avoid the positional bias from human annotators. We filter out possible low-quality responses or human evaluators based on following criteria:

- Being labeled as low quality by SurveyMonkey's response quality model;
- Selecting the same option for all selections;
- Taking less than 20 seconds per annotation.

After filtering, we collect at least 5 responses per pair. Furthermore, we aggregate the human evaluations using majority voting, where the most selected option is considered the final result. If two translations receive the same number of votes, we record the result as ''No Preference'' (Tie).

**Bilingual LLM Preference (BLP)** Recent work demonstrates that the reference translations are likely to be of low quality (Freitag et al., 2023; Xu et al., 2024). Kocmi and Federmann (2023a) demonstrate that GPT-4 is capable of accurately estimating translation quality without the need for human reference translations. Hence, we require GPT-4-0125-PREVIEW to determine which translation segment is better, without providing the reference translations, as shown in Figure 4. Each segment pair is assessed in both forward and reversed directions.

**Evaluation Metrics** For both MHP and BLP, we use the winning rate (%), which is the percentage of instances where a model's generated chapter is preferred by either the human evaluators or the LLM, to measure the model performance.

**Comparison with Classical Human Evaluation** Classical human evaluation methods, such as the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2013), are typically designed for sentence-level MT. However, TRANSAGENTS translates the source text chapter by chapter. It is highly challenging for the annotators to perform MQM evaluations on ultra-long texts like our test examples, making the human evaluation process extraordinarily slow and expensive. In contrast, the MHP evaluation method offers several advantages. First, MHP does not require bilingual professional annotators. Instead, it relies on monolingual annotators from the target audience, simplifying the annotator recruitment process and enabling larger-scale human evaluations. Second, unlike classical human evaluation methods, MHP does not provide the source text to annotators. Instead, annotators are asked to select their preferred translations, allowing them to focus on text attributes beyond translation errors, such as naturalness, fluency, and other stylistic qualities. Furthermore, recent research has demonstrated the effectiveness of preference-based evaluation in estimating translation quality (Thai et al., 2022; He et al., 2024). These distinctions make MHP a more suitable evaluation method for our use case.

```
[The start of source]
[$src_lang]: $src
[The end of source]

[The start of assistant 1's translation]
[$tgt_lang]: $asst1
[The end of assistant 1's translation]

[The start of assistant 2's translation]
[$tgt_lang]: $asst2
[The end of assistant 2's translation]

We would like to request your feedback [TRUNCATED]
```

Figure 4: The prompt used for bilingual LLM preference evaluation.

## 5 Main Results

We present the main results on standard evaluation and preference evaluation in this section.

**Standard Evaluation Results** We present the automatic evaluation results in Table 1. Our approach performs poorly in terms of the $d$-BLEU metric, achieving the lowest scores among the compared methods. However, it is important to note that $d$-BLEU has limitations as an evaluation metric, as it primarily focuses on surface-level similarity and may not fully capture the quality and coherence of the generated text. Although the $d$-BLEU score of TRANSAGENTS is significantly lower than that of GPT-4, TRANSAGENTS achieves

[6] https://www.surveymonkey.com/.

| | $d$-BLEU ↑ | GEMBA-DA ↑ |
|---|---|---|
| REFERENCE 1 | — | $85.4_{\pm 0.5}$ |
| REFERENCE 2 | — | $82.6_{\pm 4.5}$ |
| LLAMA-MT | 43.1 | — |
| GPT-4-0613 | 43.7 | — |
| GOOGLE | 47.3 | — |
| DUT | 50.2 | — |
| HW-TSC | 52.2 | — |
| GPT-4-TURBO | $47.8_{\pm 2.6}$ | $85.8_{\pm 0.5}$ |
| GPT-4O-MINI | $46.1_{\pm 2.5}$ | $86.9_{\pm 0.4}$† |
| GPT-4O | $46.3_{\pm 2.6}$ | $87.1_{\pm 0.3}$† |
| TRANSAGENTS | $25.0_{\pm 2.4}$ | $87.7_{\pm 0.2}$† |

Table 1: $d$-BLEU and GEMBA-DA scores with standard deviation on WMT2023 DLLT test set. ↑ indicates higher is better. GEMBA-DA score is in the range from 0 to 100. † indicates the improvement is statistically significant against REFERENCE 1 at the significance level $\alpha = 0.05$ according to Koehn (2004).

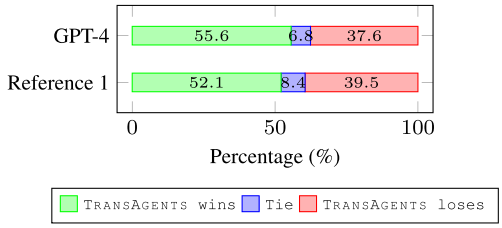


Figure 5: Monolingual Human Preference evaluation results. GPT-4 indicates GPT-4-1106-PREVIEW.

a higher GEMBA-DA score. The GEMBA-DA results suggest that TRANSAGENTS delivers better overall quality despite its lower performance on $d$-BLEU.

**Preference Evaluation Results** We compare the performance of our TRANSAGENTS model with REFERENCE 1 and GPT-4-1106-PREVIEW using MHP and BLP evaluations. The results, presented as winning rates, are shown in Figure 5. The translations produced by TRANSAGENTS are marginally preferred by human evaluators compared to both REFERENCE 1 and GPT-4-1106-PREVIEW. The Cohen's $\kappa$ for the MHP between REFERENCE 1 and TRANSAGENTS is 0.17, while the Cohen's $\kappa$ for the MHP between GPT-4-1106-PREVIEW and TRANSAGENTS is 0.11. Additionally, the translations generated by TRANSAGENTS are also more favored by GPT-4-0125-PREVIEW compared to the other models, as shown in Figure 6. Both evaluation methods demonstrate that TRANSAGENTS can

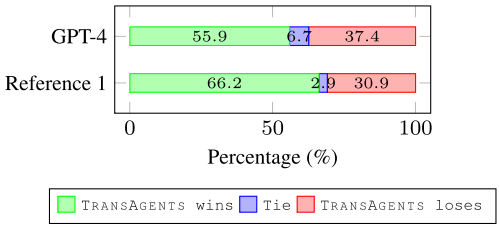


Figure 6: Bilingual LLM Preference evaluation results. GPT-4 indicates GPT-4-1106-PREVIEW.

| | $d$-BLEU | GEMBA-DA |
|---|---|---|
| GPT-4-TURBO | $47.8_{\pm 2.6}$ | $85.8_{\pm 0.5}$ |
| TRANSAGENTS | | |
| – Translation | $28.8_{\pm 2.3}$ | $86.4_{\pm 0.2}$ |
| – Localization | $25.5_{\pm 2.5}$ | $87.3_{\pm 0.3}$ |
| – Proofreading | $25.0_{\pm 2.4}$ | $87.7_{\pm 0.2}$ |

Table 2: $d$-BLEU and GEMBA-DA scores with standard deviation given by each stage in TRANSAGENTS on WMT2023 DLLT test set. Note that the ''proofreading'' translation is the final translation of TRANSAGENTS.

generate superior translations compared with the human translators and GPT-4-1106-PREVIEW.

## 6 Analysis

**What Causes TRANSAGENTS to ''Fail'' in Terms of $d$-BLEU?** As shown in Table 1, the translation produced by TRANSAGENTS achieves the lowest $d$-BLEU score among the compared methods. To investigate the reasons behind this, we evaluate the output of each stage in the TRANSAGENTS workflow using the official references from the WMT2023 DLLT test set. As shown in Table 2, although TRANSAGENTS achieves lower $d$-BLEU scores, the translation quality improves further in terms of GEMBA-DA after each step. This demonstrates the effectiveness of each step in TRANSAGENTS and highlights the limitations of BLEU. Furthermore, the translation guideline is likely to be the primary contributor to the final translation quality, as the largest decline in $d$-BLEU comes from the Translation step of TRANSAGENTS, compared to GPT-4-TURBO. Additionally, we conduct an experiment that rephrases the references of FLORES (Costa-jussà et al., 2022) and present the results in Appendix B.

**Strengths and Weaknesses of TRANSAGENTS Regarding Genres** The test examples span a variety of genres, including Video Games (VG),

| | Overall | VG | EF | SR | CR | F | SF | HT | FR |
|---|---|---|---|---|---|---|---|---|---|
| *Monolingual Human Preference* | | | | | | | | | |
| GPT-4-TURBO | 55.6 | 64.5 | 68.2 | 63.3 | 44.6 | 68.2 | 39.1 | 48.0 | **77.8** |
| REFERENCE 1 | 52.1 | **67.7** | 63.6 | 56.7 | 42.9 | 63.6 | 37.0 | 40.0 | 66.7 |
| *Bilingual LLM Preference* | | | | | | | | | |
| GPT-4-TURBO | 55.9 | **74.1** | 56.8 | 58.3 | 47.3 | 70.5 | 47.8 | 34.0 | 66.7 |
| REFERENCE 1 | 66.2 | **88.7** | 59.1 | 70.0 | 54.5 | 83.0 | 53.3 | 62.0 | 61.1 |

Table 3: The breakdown winning rate of TRANSAGENTS against GPT-4-TURBO and REFERENCE 1. Best results in each row are highlighted in **bold**. Worst results in each row are highlighted in underline.

| | MATTR ↑ | MTLD ↑ |
|---|---|---|
| REFERENCE 1 | 80.9 | 89.1 |
| GPT-4-1106-PREVIEW | 81.5 | 94.9 |
| TRANSAGENTS | | |
| – Translation | 83.5 | 117.0 |
| – Localization | 83.6 | 119.4 |
| – Proofreading | 83.6 | 119.4 |

Table 4: Linguistic diversity in terms of MATTR (up-scaled by $\times 100$) and MTLD. ↑ indicates higher is better.

Eastern Fantasy (EF), Sci-fi Romance (SR), Contemporary Romance (CR), Fantasy (F), Science Fiction (SF), Horror & Thriller (HT), and Fantasy Romance (FR). We present a detailed analysis of the performance of our model TRANSAGENTS, across these categories in Table 3. Our observations indicate that TRANSAGENTS excels in domains that demand extensive domain-specific knowledge, such as historical contexts and cultural nuances. These areas often pose significant challenges for translators. Meanwhile, TRANSAGENTS underperforms in contemporary domains, which do not require as much specialized knowledge. This trend underscores TRANSAGENTS's strengths and weaknesses.

**Linguistic Diversity** Linguistic diversity in literary texts is critical for enriching the reading experience. To quantify the linguistic diversity of the translation, we leverage two metrics: the Moving-Average Type-Token Ratio (MATTR) (Covington and McFall, 2010) and the Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010). As shown in Table 4, assisted by our translation guidelines, our initial translation significantly improves linguistic diversity compared to the reference text. Moreover, the localization step further enhances linguistic diversity, while the proofreading step does not affect it. These results demonstrate the effectiveness of our approach in preserving and enhancing the richness of language in the translated literary work.

**Agent Profile Design** Recent work demonstrates that diverse role-playing profiles can effectively enhance model performance (Chan et al., 2024). We design agent profiles with detailed and rich personas, as described in Section 3.1. To validate the effectiveness of our design choices on agent profiles, we categorize the agent profiles into four groups based on their level of detail:

- **None**: The role-playing profile is not used.
- **Minimum**: The profile includes only the job title of the agent, for example, `You are a Senior Editor.`
- **Language-Specified (LangSpec)**: The profile includes the job title and language skills of the agent, for example, `You are a Senior Editor, specializing in English and Chinese.`
- **Vivid**: The profile includes detailed and rich personas as described in Section 3.1.

We present the results in Table 5 and demonstrate that incorporating a role-playing message can significantly enhance model performance. We observe that agent profile with more details typically lead to better performance. Our **Vivid** agent profile, which includes the most detailed and colorful information, achieves the best performance.

**Impact of Iterations** We conduct additional experiments to analyze the effect of varying $\mathbf{M}$ and present the results in Figure 7 for Algorithm 1 and Algorithm 2, respectively. Our findings indicate

| | MTLD ↑ | $BLP_{REF}$ ↑ | $BLP_{GPT}$ ↑ |
|---|---|---|---|
| TRANSAGENTS | | | |
| + None | 113.8 | 48.2 | 40.1 |
| + Minimum | 117.9 | 59.6 | 48.2 |
| + LangSpec | 118.8 | 64.4 | 52.7 |
| + Vivid | 119.4 | 66.2 | 55.9 |

Table 5: Analysis on agent profile design. MTLD measures the linguistic diversity. $BLP_{REF}$ and $BLP_{GPT}$ indicate the winning rate (%) of Bilingual LLM preference against REFERENCE 1 and GPT-4-TURBO. ↑ indicates higher is better.

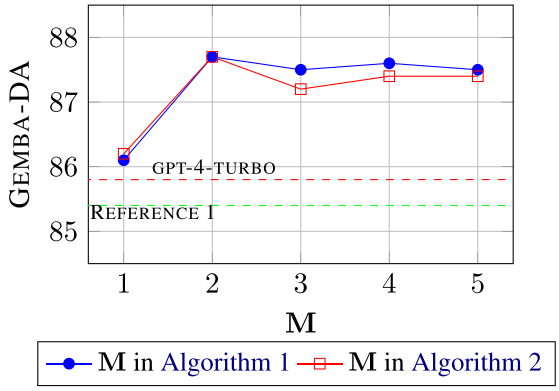


Figure 7: GEMBA-DA scores given by TRANSAGENTS with regard to M in Algorithm 1 and Algorithm 2. The green and red dashed lines indicate the GEMBA-DA scores of REFERENCE 1 and GPT-4-TURBO.

that for $\mathbf{M} > 2$, the improvements in translation quality become minimal. This behavior can be attributed to our ''early exit'' design, where the iteration terminates early if the glossary list and summary converge quickly (at line 9 of Algorithm 1) or if the Judgment agent **J** in Algorithm 2 determines that the output is of high quality (at line 10 of Algorithm 2). For ultra-long literary texts, the limitations of current LLMs in processing long contexts become more pronounced, and increasing **M** beyond 2 can sometimes degrade translation quality. Furthermore, thanks to our ''early exit'' design, the runtime of TRANSAGENTS increases by only about 10% when $\mathbf{M} > 2$.

**Short-Text Translation** We conduct additional experiments on English-Chinese and English-German translations using the WMT2024 General MT tasks. Due to budget constraints, we randomly selected 200 test examples from the entire test set, and the results are presented in Table 6. When translating shorter texts, TRANSAGENTS performs worse than GPT-4O-MINI and GPT-4O, as expected. Upon manual inspection of the translations and the conversations between agents, we found that the agents often fail to correctly determine the style, tone, and target audience when provided with only a short text. Consequently, the translations produced by TRANSAGENTS frequently misalign with the style and tone of the source and reference sentences and suffer from overtranslation issues. We present a case study for the short text translation in Section 7.

| | En-Zh | | En-De | |
|---|---|---|---|---|
| | *d*-BLEU | C.K. | *d*-BLEU | C.K. |
| GPT-4-TURBO | 42.7 | 79.3 | 25.5 | 70.5 |
| GPT-4O-MINI | 42.9 | 79.3 | 26.6 | 71.2 |
| GPT-4O | 45.6 | 80.5 | 28.3 | 72.4 |
| TRANSAGENTS | 29.3 | 75.5 | 24.0 | 72.4 |

Table 6: *d*-BLEU and COMETKIWI scores on English-Chinese and English-German translation tasks of WMT2024 General MT test set. C.K. indicates COMETKIWI.

| | G.D. | Runtime | Cost |
|---|---|---|---|
| REFERENCE 1 | 85.4 | a few weeks | est. 40K |
| GPT-4-TURBO | 85.8 | 1.1 | 42.1 |
| TRANSAGENTS | 87.7 | 13.1 | 487.3 |

Table 7: Comparison among REFERENCE 1, GPT-4-TURBO, and TRANSAGENTS. G.D. indicates GEMBA-DA. The runtime (in hours) and cost (in USD) are calculated based on the entire test set.

**Cost Analysis** We provide a more detailed evaluation of REFERENCE 1,[7] GPT-4-TURBO, and TRANSAGENTS in terms of translation quality, runtime, and financial costs, as shown in Table 7. Our analysis reveals that TRANSAGENTS introduces significant overhead in both runtime and costs compared to GPT-4-1106-PREVIEW. However, it remains substantially more cost-effective and efficient than professional human translators.

## 7 Case Study

In this section, we explore case studies on cultural adaptation and shorter texts. We also discuss a case on undertranslation and include comments from professional translators in Appendix C.

[7]The minimum rate for translation services is 0.12 USD per word. Source: `https://tinyurl.com/bdze92xr`.

| | |
|---|---|
| Source Text | 罗德抬起头来，正好看见一个中年男子推门走进来，他穿着冒险者的皮甲，一头鲜红的长发随意的向后梳理着，看见罗德，男子微微一笑。“你好，先生，我是星月佣兵团的团长，卡特。” |
| REFERENCE 1 | Rhode looked up and saw a middle-aged man pushing the door open. He wore an adventurer's leather armor and casually combed back his long red hair. The man smiled when he saw Rhode. "Hello, sir. I am the Guild Leader of the Star Moon Mercenaries, Carter." |
| GPT-4-TURBO | Rhode looked up and saw a middle-aged man entering through the door. He was dressed in an adventurer's leather armor, with a head of bright red hair casually combed back. Seeing Rhode, the man smiled slightly. "Hello, sir, I am the leader of the Star Moon Mercenary Group, Carter." |
| | TRANSAGENTS |
| TRANSLATION | Rhode looked up, only to behold a middle-aged man making his entrance. Clad in the leather armor of an adventurer, his fiery red hair swept back carelessly. Upon noticing Rhode, the man's lips curled into a modest smile. "Greetings, sir. I am the leader of the Star Moon Group, Carter." |
| LOCALIZATION | Rod looked up to see a middle-aged man entering. The man was dressed in the leather armor typical of adventurers, his fiery red hair casually swept back. Spotting Rod, the man offered a modest smile.. "Hello, sir. I am Carter, the leader of the Star Moon Group." |
| PROOFREADING | Rhode looked up to see a middle-aged man entering. The man was dressed in the leather armor typical of adventurers, his fiery red hair casually swept back. Spotting Rhode, the man offered a modest smile. "Hello, sir. I am Carter, the leader of the Star Moon Mercenary Corps." |

Table 8: Case study on cultural adaptation. The text in red indicates that the translation is accurate in meaning but not in cultural context. The text in blue indicate that the translation is accurate both in meaning and in cultural context.

**Cultural Adaptation** In Chinese, job titles are typically placed before a person's name, whereas in English, job titles usually come after the person's name. This order reflects differing linguistic and cultural conventions regarding the structuring of personal information in the two languages. As demonstrated in Table 8, both REFERENCE 1 and GPT-4-TURBO fail to correctly adjust the order of names and job titles, thus not adhering to the cultural norms expected in the target language. In contrast, TRANSAGENTS accurately reflects this cultural context in its translation after the translation-localization-proofreading workflow. In the initial translation phase, the output of TRANSAGENTS is of high quality but contains errors related to cultural adaptation and terminology usage. During the localization step, the cultural adaptation issues are resolved, but errors in terminology persist. Finally, the proofreading step addresses the remaining terminology issues. This example effectively demonstrates the utility of each step in the TRANSAGENTS process. This emphasizes the capability of TRANSAGENTS to provide translations that are culturally appropriate, ensuring an immersive reading experience for readers.

**Short-Text Translation** As mentioned in Section 6, the translations produced by TRANSAGENTS frequently misalign with the style and tone of the source and reference sentences due to the short context. For instance, as shown in Table 9, the translations generated by GPT-4-TURBO, GPT-4O-MINI, and GPT-4O are almost identical to the reference translation. In contrast, the translation produced by TRANSAGENTS are more polished and idiomatic in Chinese but diverge somewhat from the literal meaning of the original text.

| | |
|---|---|
| Source Text | Adapt the old, accommodate the new to solve issue |
| REFERENCE | 适应旧的，容纳新的，积极解决问题。(Adapt to the old, accommodate the new, and actively solve problems) |
| GPT-4-TURBO | 适应旧的，接纳新的以解决问题。(Adapt to the old and embrace the new to solve problems.) |
| GPT-4O-MINI | 适应旧的，容纳新的，以解决问题。(Adapt the old and accommodate the new to solve problems.) |
| GPT-4O | 适应旧事物，接纳新事物以解决问题。(Adapt to old things and accept new things to solve problems.) |
| TRANSAGENTS | 传承与创新并重，以应对当代挑战。(Inheritance and innovation are equally important to meet contemporary challenges) |

Table 9: Case study on short-text translation. The texts within parentheses are the back-translations provided by GOOGLE TRANSLATE.

## 8 Conclusion

We introduce TRANSAGENTS, a novel multi-agent framework for literary translation inspired by the structured workflows of human translation companies. The framework divides the process into preparation and execution stages, assigning distinct roles to agents. To evaluate its effectiveness, we propose innovative methods, including Monolingual Human Preference (MHP) and Bilingual LLM Preference (BLP). Our findings show that the framework excels in translating long texts and capturing the nuances of literary works, though challenges with shorter texts highlight areas for improvement. Insights from agent profiling and collaboration strategies offer guidance for refining future translation systems. In summary, our work paves the way for new directions in developing translation systems that combine LLMs with the best practices of human translation.

## Acknowledgments

We sincerely thank the action editor and reviewers for their valuable time, insightful feedback, and constructive suggestions. Their thoughtful comments have greatly contributed to improving the

quality and clarity of our work. We deeply appreciate their efforts in helping us refine this manuscript.

## References


Aleksandar Botev, Soham De, Samuel L. Smith, Anushan Fernando, George-Cristian Muraru, Ruba Haroun, Leonard Berrada, Razvan Pascanu, Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Sertan Girgin, Olivier Bachem, Alek Andreev, Kathleen Kenealy, Thomas Mesnard, Cassidy Hardin, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Armand Joulin, Noah Fiedel, Evan Senter, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, David Budden, Arnaud Doucet, Sharad Vikram, Adam Paszke, Trevor Gale, Sebastian Borgeaud, Charlie Chen, Andy Brock, Antonia Paterson, Jenny Brennan, Meg Risdal, Raj Gundluru, Nesh Devanathan, Paul Mooney, Nilay Chauhan, Phil Culliton, Luiz Gustavo Martins, Elisa Bandy, David Huntsperger, Glenn Cameron, Arthur Zucker, Tris Warkentin, Ludovic Peran, Minh Giang, Zoubin Ghahramani, Clément Farabet, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, Yee Whye Teh, and Nando de Frietas. 2024. Recurrentgemma: Moving past transformers for efficient open language models. *CoRR*, abs/2404.07839.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.123

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *CoRR*, abs/2308.07201.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100. https://doi.org/10.1080/09296171003643098

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration code generation via chatgpt. *CoRR*, abs/2304.07590.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023a. Improving factuality and reasoning in language models through multiagent debate. *CoRR*, abs/2305.14325.

Zefeng Du, Wenxiang Jiao, Longyue Wang, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. 2023b. On extrapolation of long-text translation with large language models. https://www.researchgate.net/profile/Longyue-Wang/publication

/373991070_On_Extrapolation_of_Long-Text_Translation_with_Large_Language_Models/links/6507a6929fdf0c69dfd42dc5/On-Extrapolation-of-Long-Text-Translation-with-Large-Language-Models.pdf

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22:107:1–107:48.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.5

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.51

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, pages 11179–11189.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *CoRR*, abs/2402.01680. https://doi.org/10.24963/ijcai.2024/890

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246. https://doi.org/10.1162/tacl_a_00642

Christian Herold and Hermann Ney. 2023. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.codi-1.15

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *CoRR*, abs/2308.00352.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.41

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A ''novel'' challenge for long-context language models. *CoRR*, abs/2406.16264. https://doi.org/10.18653/v1/2024.emnlp-main.948

Jeremy Klemin. 2024. The last frontier of machine translation. *The Atlantic*.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.64

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Clifford E. Landers. 2001. Literary translation: A practical guide. *Multilingual Matters*. https://doi.org/10.21832/9781853595639

Nian Li, Chen Gao, Yong Li, and Qingmin Liao. 2023. Large language model-empowered agents for simulating macroeconomic activities. *CoRR*, abs/2310.10436.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *CoRR*, abs/2305.19118.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. https://doi.org/10.1162/tacl_a_00343

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Zhao Mandi, Shreeya Jain, and Shuran Song. 2023. Roco: Dialectic multi-robot collaboration with large language models. *CoRR*, abs/2307.04738.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392. https://doi.org/10.3758/BRM.42.2.381, PubMed: 20479170

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. 2023. Welfare diplomacy: Benchmarking language model cooperation. *CoRR*, abs/2310.08901.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023–1 November 2023*, pages 2:1–2:22. ACM. https://doi.org/10.1145/3586183.3606763

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *The 35th Annual ACM Symposium on User Interface Software and Technology, UIST 2022, Bend, OR, USA, 29 October 2022–2 November 2022*, pages 74:1–74:18. ACM. https://doi.org/10.1145/3526113.3545616

Celia Rico Pérez. 2002. Translation and project management. *Translation Journal*, 6(4):38–52.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and

Maosong Sun. 2023. Communicative agents for software development. *CoRR*, abs/2307.07924.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.40

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Mary F. Terhaar, Rachael Crickman, and Deborah S. Finnell. 2019. Project management for translation. *Translation of Evidence Into Nursing and Healthcare*, page 199. https://doi.org/10.1891/9780826147370.0009

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.672

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick S. H. Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *CoRR*, abs/2404.18796.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rob Voigt and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.

Callum Walker. 2022. *Translation Project Management*. Routledge. https://doi.org/10.4324/9781003132813-1, https://doi.org/10.4324/9781003132813

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024a. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.428

Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024b. Findings of the WMT 2024 shared task on discourse-level literary

translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.58

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.1036

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.3

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1301

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2024c. Delta: An online document-level translation agent based on multi-level memory. *CoRR*, abs/2410.08143.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023c. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *CoRR*, abs/2302.01560.

Michael J. Wooldridge and Nicholas R. Jennings. 1995. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10(2):115–152. https://doi.org/10.1017/S0269888900008122

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.33

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024b. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.44

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024c. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-demo.14

Yuhao Xie, Zongyao Li, Zhanglin Wu, Daimeng Wei, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, Hengchao Shang, Jiaxin Guo, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. HW-TSC's submissions to the WMT23 discourse-level literary translation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 302–306, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.32

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou,

Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. Effective long-context scaling of foundation models. *CoRR*, abs/2309.16039.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023a. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *CoRR*, abs/2401.08417.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023b. Exploring large language models for communication games: An empirical study on werewolf. *CoRR*, abs/2309.04658.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. GPT-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *CoRR*, abs/2407.03658.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. 2023. Building cooperative embodied agents modularly with large language models. *CoRR*, abs/2307.02485.

Ran Zhang, Wei Zhao, and Steffen Eger. 2024. How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. *CoRR*, abs/2410.18697.

Anqi Zhao, Kaiyu Huang, Hao Yu, and Degen Huang. 2023. DUTNLP system for the WMT2023 discourse-level literary translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 296–301, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.31

## A Interface for Monolingual Human Preference

We present the user interface for MHP in Figure 8.

```
Q: Which of the following writing style do you prefer?

[x] Chapter 455: Turnaround 3 "Allow me to demonstrate the
↪  sensing of Formless Fluctuation; it's remarkably
↪  straightforward," interjected another sorcerer, a
↪  smile evident in his voice. "Your assistance is
↪  appreciated," Lin Sheng responded, offering a nod of
↪  gratitude. Time was of the essence in finding the
↪  remaining Fragments. He had initially planned to
↪  conquer an array of Great Evil Spirits to amass
↪  substantial reserves of pure soul power. Yet, the
↪  present opportunity necessitated an immediate and
↪  decisive acquisition. Promptly, the sorcerer leader
↪  brought Lin Sheng to a daunting Evil Spirit Gate. Both
↪  extended their hands, gently touching the gate's
↪  enigmatic frame, eyes closed as one. The leader
↪  rapidly employed his Special Ability to establish a
↪  Spatial Foundation, thus setting a Coordinate Code.

[ ] Chapter 455 Reversion 3 "This is to let you feel the
↪  fluctuation of aura. It's really simple." Another
↪  Warlock couldn't help but interrupt with a smile.
↪  "Then I'll have to trouble you." Lin Sheng nodded. He
↪  needed to find the other fragments as soon as
↪  possible. Originally, he had planned to conquer more
↪  evil spirits and obtain more pure soul power. But now
↪  that he encountered such an opportunity, the most
↪  important thing for him was to get it as soon as
↪  possible. Soon, the Warlock Commander led Lin Sheng to
↪  an Evil Spirit Gate. The two reached out, touched the
↪  frame of the Evil Spirit Gate at the same time, and
↪  closed their eyes. The Warlock Commander quickly used
↪  his ability to build the space base as a coordinate.

[ ] No Preference
```

Figure 8: The user interface for Monolingual Human Preference (MHP). `[x]` indicates the selection of human evaluator.

## B Rephrasing Experiment

We acknowledge that a drop from 47.8 to 25.0 in $d$-BLEU may intuitively suggest poorer translation performance. However, we argue that this significant decline in BLEU does not necessarily indicate poorer translation quality but rather reflects the inherent limitations of BLEU as a metric. BLEU primarily measures surface-level n-gram overlap, which often fails to capture deeper semantic and stylistic qualities, especially in nuanced domains like literary translation. To illustrate this limitation, we conduct an additional experiment on the FLORES Chinese-English test set. First, we used GPT-4O to generate translations and then rephrase these translations to use entirely different vocabulary while preserving the intended meaning. We evaluate the original references, the translations generated by GPT-4O, and their rephrased versions using BLEU and COMETKIWI.[8] As shown in Table 10, the BLEU scores decline significantly after rephrasing, while the COMETKIWI scores remain almost unchanged. These results highlight BLEU's inability to account for semantic preservation when lexical variation occurs. Therefore, **the large drop in BLEU does not suggest poor translation performance by our approach but rather underscores the BLEU's limitations in evaluating nuanced translations and poor diversity of references, as discussed by Freitag et al. (2020).**

[8] BLEU signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1` and COMETKIWI signature: `Unbabel/wmt22-cometkiwi-da`.

| | BLEU | COMETKIWI |
|---|---|---|
| Reference 1 | 100.0 | 82.7 |
| Rephrased Reference 1 | 12.5 | 82.7 |
| GPT-4O | 29.4 | 84.9 |
| Rephrased GPT-4O | 11.5 | 82.8 |

Table 10: FLORES Chinese-English translation results given by the reference, rephrased reference, GPT-4O, and rephrased GPT-4O. The COMETKIWI is scaled up by $\times$ 100.

## C More Cases

In this section, we present additional case study on undertranslation and comments from professional translators.

**Undertranslation** Our TRANSAGENTS is generally preferred over both REFERENCE 1 and GPT-4-1106-PREVIEW according to evaluations by human judges and LLMs (Figure 5 and Figure 6). However, despite its higher preference, the translations produced by TRANSAGENTS are not without flaws. A detailed analysis of the translated chapters, when divided into smaller segments, reveals that both GPT-4-1106-PREVIEW and TRANSAGENTS exhibit significant issues with undertranslation, as illustrated in Table 11. While these undertranslations do not seem to impact the overall development of the story plot, they could potentially influence other critical aspects of the narrative. For example, missing content could diminish the depth of character development or alter the intended emotional impact of the text. Such undertranslations, therefore, raise concerns about the completeness and fidelity of the translation in preserving the nuanced expressions and thematic elements of the original texts.

**Comments from Professional Translators** We anonymize the translations from TRANSAGENTS,

| | |
|---|---|
| Original Text | 她招来女仆带叶琛和程安雅下去洗漱，小奶包虽然很想跟着去，不过他还是留在这里，白夜作势就要揍人了，小奶包赶紧拉着他的袖子。“白夜，你能有办法救我爹地妈咪吗？”小孩子的眼睛很亮，如两颗黑葡萄镶嵌在白嫩的脸上，充满了期盼，仿佛白夜一摇头，他眸中的亮光就会黯淡了。杰森一把揪起小奶包抱在怀里，豪气万千，“宝贝儿，你放心，小白死人都能救，别说活生生的人了，你担心个屁，有空过来给我轰了黑手党的防护。”“刚是谁质疑白夜的医术的？”黑杰克对此表示疑惑，杰森一掌过去，他敏捷闪开。小奶包被大高个子抱着，异常的纠结，踢了踢杰森，“放我下来。”“老子也想要这么个儿子，宁宁，你来当我儿子吧？老子垂涎你很久了。”杰森湛蓝色的眸迸发出澎湃的光芒，活似小奶包就是一块肥肉。众人，“.....”白夜微笑说道，“杰森，你中文再让你妈教教，别老说长官不会用词语，你也好不到哪儿去。”“我和长官不是一个级别的好吧？”杰森很不满意有人把他和长官联系在一起，所谓官寇不一家，这是原则问题。小奶包挣扎一下，杰森放他下来，小奶包问道：“白夜.....”“宁宁，等我给他们做过检查才能确定，你先别着急。”白夜说道，揉揉小奶包的头，“我保证会尽全力。” |
| REFERENCE 1 | She called a maid to take Ye Chen and Cheng Anya downstairs to freshen up. Little Cutie really wanted to go with them, but he stayed behind as Bai Ye seemed ready to start fighting someone, so Little Cutie quickly grabbed his sleeve. “Bai Ye, can you save my Mommy and Daddy?” The child’s eyes were bright, like two black grapes embedded in his fair face, full of expectation. It seemed that if Bai Ye shook his head, the light in his eyes would dim. Jason grabbed Little Cutie and hugged him fiercely, feeling quite confident. ”Darling, don’t worry. Little Bai can save even the dead, not to mention living people. Stop worrying about them and come help me take down the Mafia’s line of defense when you have time.” “Who was questioning Bai Ye’s medical skills just now?” Black Jack expressed doubt about this. Jason threw a punch at him, but he agilely dodged it. Being carried by the tall man, Little Cutie felt extremely conflicted and kicked Jason. ”Put me down.” “I also want a son like you. Ningning, will you be my son? I’ve been coveting you for a long time.” Jason’s azure eyes burst with enthusiasm as if Little Cutie was a rare piece of jewelry. Everyone was speechless.... Bai Ye smiled and said, ”Jason, let your mother help you improve your Chinese. Don’t always complain that the commander doesn’t know how to use words. You’re not much better.” “I’m not on the same level as the commander, okay?” Jason was very dissatisfied that someone had compared him to the commander. It was a matter of principle that officials and bandits were not on the same side. Little Cutie struggled a little, and Jason put him down. Little Cutie asked, “Bai Ye....” “Ningning, don’t worry. I’ll check them up before making a conclusion.” Bai Ye said, rubbing the head of Little Cutie, ”I promise I will do my best.” |
| GPT-4-1106-PREVIEW | She summoned the maids to take Ye Chen and Cheng Anya to wash up. Xiao Nai Bao wanted to follow, but he stayed here, as Bai Ye looked like he was about to beat someone up, so Xiao Nai Bao quickly grabbed his sleeve. Bai Ye, can you save my daddy and mommy? Xiao Nai Bao’s eyes were bright, like two black grapes embedded in his tender face, full of hope, as if the light in his eyes would dim if Bai Ye shook his head. Jason grabbed Xiao Nai Bao and hugged him, full of pride, “Baby, don’t worry, Xiao Bai can even save the dead, let alone the living. Stop worrying about nothing, come help me bombard the mafia’s defenses when you’re free.” Who was questioning Bai Ye’s medical skills just now? Black Jack expressed doubt, and with a palm strike, Jason swiftly dodged. Xiao Nai Bao struggled a bit, and Jason put him down. Xiao Nai Bao asked, “Bai Ye...” Ningning, wait until I’ve examined them before I can be sure, don’t be anxious, Bai Ye said, ruffling Xiao Nai Bao’s head, “I promise to do my best.” |
| TRANSAGENTS | At her behest, maids ushered Ye Chen and Cheng An Ya away for a bath. Xiao Naibao, torn between following and staying, ultimately remained as Bai Ye appeared poised for a scuffle. Tugging at Bai Ye’s sleeve, Xiao Naibao looked up at him with hopeful eyes. “Bai Ye, is there a way to cure my daddy and mommy?” “Ning Ning, let me conduct a thorough examination first. Stay calm,” Bai Ye soothed, patting the boy’s head. “I’ll do everything in my power.” |

Table 11: Case study for undertranslation. The text highlighted in red and blue indicates the content that is omitted by TRANSAGENTS. The text highlighted in red represents the content that is omitted by GPT-4-1106-PREVIEW.

| | |
|---|---|
| Translator A | TRANSAGENTS’s translation style is similar to that of a novel, with sophisticated wording and personal flair. Despite some omissions, it makes the text more concise and effectively conveys the original text’s mood and meaning. REFERENCE 1 and GPT-4-1106-PREVIEW’s translations are more conventional, adhering strictly to the original text word for word. However, GPT-4-1106-PREVIEW’s translation is more grammatically precise than REFERENCE 1’s, and its wording is slightly better, making its translation aesthetically superior to REFERENCE 1’s but still not reaching the literary expressiveness of TRANSAGENTS. From their translation habits, TRANSAGENTS appears to have a solid foundation in English, REFERENCE 1 seems to rely on machine translation, and GPT-4-1106-PREVIEW behaves like a standard, rule-abiding translator. |
| Translator B | TRANSAGENTS’s translation breaks away from the constraints of the original language, using the language freely with ample additions and expansions, and the choice of vocabulary also demonstrates a deeper understanding of the language. REFERENCE 1 remains faithful to the original text, translating directly and succinctly without adding personal interpretations. GPT-4-1106-PREVIEW’s translation style is similar to REFERENCE 1’s, both strictly adhering to the original without much personal interpretation or embellishment. Overall, TRANSAGENTS’s translation shows the greatest depth and sophistication, followed by REFERENCE 1, while GPT-4-1106-PREVIEW performs most ordinarily among the three. |

Table 12: Comments from two experienced professional translators on the translations from TRANSAGENTS, REFERENCE 1, and GPT-4-1106-PREVIEW. We present both the original text and the anonymized translations to two experienced professional translators. The original comments are written in Chinese, and we make adaptations while preserving their original meaning. We replace the anonymized system names with the actual system names to improve readability. The translation systems are highlighted in red.

REFERENCE 1, and GPT-4-1106-PREVIEW for a randomly selected chapter and present both the original text and the translations to two experienced professional translators. We request that they assess and rank the quality of each translation and provide their comments on the translations. As shown in Table 12, both Translator A’s and Translator B’s comments highlight the novel-like, expressive translation style of TRANSAGENTS, which uses sophisticated language, though it sometimes omits parts of the original text. REFERENCE 1, and GPT-4-1106-PREVIEW stick closer to the original text. Overall, TRANSAGENTS’s translations are viewed as the most expressive and engaging, REFERENCE 1’s as straightforward, and GPT-4-1106-PREVIEW’s as the most traditional. These comments confirm that TRANSAGENTS is capable of producing more expressive and engaging translations, compared to REFERENCE 1 and GPT-4-1106-PREVIEW.

## D Prompts

In this section, we present the prompts used for translation, localization, and proofreading in Figure 9, Figure 10, and Figure 11, respectively.

```
# Translation Guidelines

## Glossary
罗德: Rhode
虚空之龙: Void Dragon
星月佣兵团: Star Moon Mercenary Corps
[TRUNCATED]

## Book Summary
The book centers on Rhode Alante, initially a Summoner
↪  Swordsman in the game 'Dragon Soul Continent,'
↪  [TRUNCATED]

## Tone
The tone of the book is adventurous and immersive with
↪  elements of fantasy and suspense. [TRUNCATED]

## Style
The book is a gripping blend of fantasy and litRPG,
↪  characterized by its immersive world-building, dynamic
↪  combat scenes, and a clear progression system.
↪  [TRUNCATED]

## Target Audience
The target audience for this book includes young adults
↪  and adults who enjoy fantasy and adventure genres,
↪  particularly those who are fans of MMORPG [TRUNCATED]

# Chapter Text
序章 传奇落幕
乌云笼罩着天空，昏暗无光的地面上四周都是一片狼籍。
[TRUNCATED]

# Instruction
Translate the chapter text from Chinese into English
↪  Ensure that your translation closely adheres to the
↪  provided translation guidelines, including the
↪  glossary, book summary, tone, style, and target
↪  audience, for consistency and accuracy. Remember to
↪  maintain the original meaning and tone as much as
↪  possible while making the translation understandable
↪  in English.
```

Figure 9: An example prompt for the Translator, including the translation guidelines, the chapter text in the source language, and the instruction.

```
# Translation Guidelines

## Glossary
罗德: Rhode
虚空之龙: Void Dragon
星月佣兵团: Star Moon Mercenary Corps
[TRUNCATED]

## Book Summary
The book centers on Rhode Alante, initially a Summoner
↪  Swordsman in the game 'Dragon Soul Continent,'
↪  [TRUNCATED]

## Tone
The tone of the book is adventurous and immersive with
↪  elements of fantasy and suspense. [TRUNCATED]

## Style
The book is a gripping blend of fantasy and litRPG,
↪  characterized by its immersive world-building, dynamic
↪  combat scenes, and a clear progression system.
↪  [TRUNCATED]

## Target Audience
The target audience for this book includes young adults
↪  and adults who enjoy fantasy and adventure genres,
↪  particularly those who are fans of MMORPG [TRUNCATED]

# Chapter Text
序章 传奇落幕
乌云笼罩着天空，昏暗无光的地面上四周都是一片狼籍。
[TRUNCATED]

# Chapter Translation
Prologue, the end of the legend
Dark clouds hung over the sky, and there was a mess all
↪  around on the dim ground.
[TRUNCATED]

# Instruction
Guided by our translation guidelines, including glossary,
↪  book summary, tone, style, and target audience,
↪  localize the chapter translation for English context.
↪  You MUST maintain all the details and the orginal
↪  writing style of the chapter text.
```

Figure 10: An example prompt for the Localization Specialist, including the translation guidelines, the chapter text in the source language, the chapter translation in the target language, and the instruction.

```
# Translation Guidelines

## Glossary
罗德: Rhode
虚空之龙: Void Dragon
星月佣兵团: Star Moon Mercenary Corps
[TRUNCATED]

## Book Summary
The book centers on Rhode Alante, initially a Summoner
↪  Swordsman in the game 'Dragon Soul Continent,'
↪  [TRUNCATED]

## Tone
The tone of the book is adventurous and immersive with
↪  elements of fantasy and suspense. [TRUNCATED]

## Style
The book is a gripping blend of fantasy and litRPG,
↪  characterized by its immersive world-building, dynamic
↪  combat scenes, and a clear progression system.
↪  [TRUNCATED]

## Target Audience
The target audience for this book includes young adults
↪  and adults who enjoy fantasy and adventure genres,
↪  particularly those who are fans of MMORPG [TRUNCATED]

# Chapter Text
序章 传奇落幕
乌云笼罩着天空，昏暗无光的地面上四周都是一片狼籍。
[TRUNCATED]

# Chapter Translation
Prologue End of the Legend
Black clouds covered the sky. The ground was in darkness,
↪  with everything around in chaos.
[TRUNCATED]

# Instruction
Guided by our translation guidelines, including the
↪  glossary, book summary, tone, style, and target
↪  audience, proofread the chapter translation.
```

Figure 11: An example prompt for the Proofreader, including the translation guidelines, the chapter text in the source language, the chapter translation in the target language, and the instruction.