(PERHAPS) BEYOND HUMAN TRANSLATION: HARNESSING MULTI-AGENT COLLABORATION FOR TRANSLATING ULTRA-LONG LITERARY TEXTS

Minghao Wu[1], Yulin Yuan[2], Gholamreza Haffari[1], Longyue Wang[3]*
[1]Monash University [2]University of Macau [3]Tencent AI Lab

ABSTRACT

Recent advancements in machine translation (MT) have significantly enhanced translation quality across various domains. However, the translation of literary texts remains a formidable challenge due to their complex language, figurative expressions, and cultural nuances. In this work, we introduce a novel multi-agent framework based on large language models (LLMs) for literary translation, implemented as a company called TRANSAGENTS, which mirrors traditional translation publication process by leveraging the collective capabilities of multiple agents, to address the intricate demands of translating literary works. To evaluate the effectiveness of our system, we propose two innovative evaluation strategies: *Monolingual Human Preference* (MHP) and *Bilingual LLM Preference* (BLP). MHP assesses translations from the perspective of monolingual readers of the target language, while BLP uses advanced LLMs to compare translations directly with the original texts. Empirical findings indicate that despite lower d-BLEU scores, translations from TRANSAGENTS are preferred by both human evaluators and LLMs over human-written references, particularly in genres requiring domain-specific knowledge. We also highlight the strengths and limitations of TRANSAGENTS through case studies and suggests directions for future research.

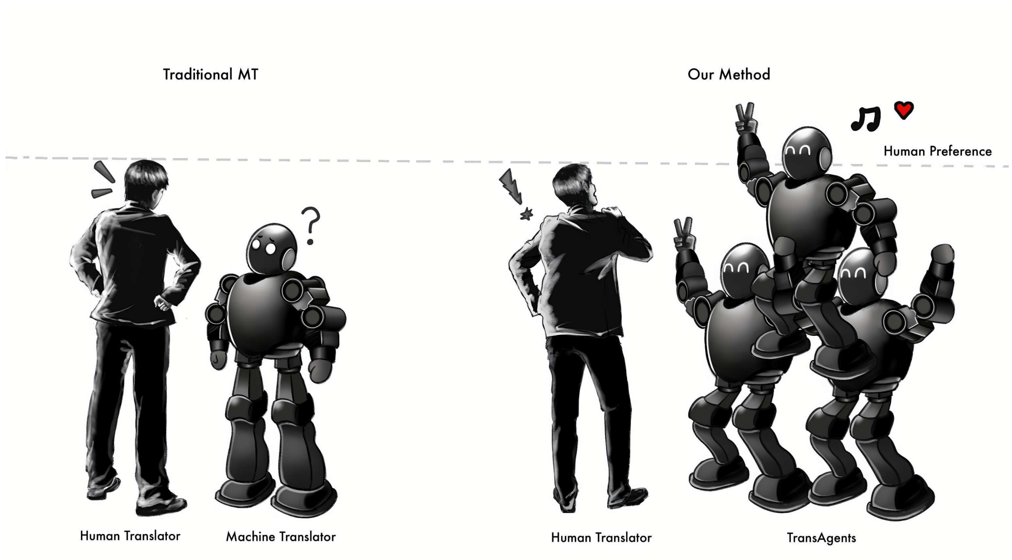

Figure 1: An illustration of our method. Traditional machine translation (MT) systems often underperform compared to human translators. In this study, we demonstrate that the translations produced by our TRANSAGENTS are more preferred by humans than those from conventional MT systems.

*Longyue Wang is the corresponding author: `vinnylywang@tencent.com`.

1 INTRODUCTION

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

In response to complex challenges across various domains, recent research in multi-agent systems, particularly those powered by large language models (LLMs), has shown significant promise (Yao et al., 2023; Wang et al., 2023e; Dong et al., 2023). These systems leverage the collective intelligence of multiple agents, enabling superior problem-solving capabilities compared to individual model approaches. Multi-agent systems excel in dynamic environments where intricate problem-solving and collaborative efforts are required.

Given the nature of literary translation, we harness the superior capabilities of multi-agent systems and establish a novel multi-agent translation company for literary translation, called TRANSAGENTS. At TRANSAGENTS, the translation process is organized into two main stages, each consisting of several sub-stages. The process begins with the selection of a Senior Editor by our pre-defined CEO agent, who chooses based on the specific requirements of each client. The selected Senior Editor then assembles a team from our roster, which includes roles such as Junior Editor, Translator, Localization Specialist, and Proofreader. Each team member collaborates through multiple sub-stages, employing strategies like *Addition-by-Subtraction Collaboration* and *Trilateral Collaboration* to refine and enhance the translation output.

Furthermore, evaluating the accuracy and quality of literary translations presents a particularly challenging task due to the subjective nature of literature and the potential imperfections in reference translations (Thai et al., 2022; Freitag et al., 2023). To effectively address these challenges, we propose two innovative evaluation strategies: *Monolingual Human Preference* (MHP) and *Bilingual LLM Preference* (BLP). Both strategies involve comparing a pair of translations from two different translation systems to determine which one is superior. The *Monolingual Human Preference* strategy simulates the realistic scenario of reading a translated work. It engages human evaluators from the target audience who assess translations without the influence of the original text. This approach focuses on how well the translation resonates with the readers in terms of fluidity, readability, and cultural appropriateness, mirroring the real-world consumption of literature. Conversely, the *Bilingual LLM Preference* leverages the capabilities of advanced LLMs, specifically GPT-4-0125-PREVIEW. In this strategy, the LLMs are provided with the original texts to facilitate a direct comparison. This method aims to harness the superior translation capabilities of advanced LLMs, mitigating the impact of imperfect reference translations.

Our empirical findings reveal that TRANSAGENTS consistently delivers the poorest performance in terms of d-BLEU scores. However, it is preferred over both human-written references and GPT-4 translations by human evaluators and an LLM evaluator. In-depth analysis shows that TRANSAGENTS excels over human-written references in genres that demand domain-specific knowledge, such as historical contexts and cultural nuances, but it falls short in contemporary genres. Additionally, we observe that TRANSAGENTS is capable of generating translations with more diverse and vivid descriptions. Our cost analysis indicates that using TRANSAGENTS for literary text translation can result in an $80\times$ reduction in costs compared to employing professional human translators. Nonetheless, we also identify significant limitations in LLM-based translation systems, including both GPT-4 and TRANSAGENTS, particularly with issues related to significant content omission.

In this work, our contributions can be summarized as follows:

- We introduces TRANSAGENTS, a novel multi-agent system for literary translation, which mirrors the traditional translation publication process. By employing a multi-agent approach, this approach addresses the complex nuances of literary works.

- We propose two novel evaluation strategies, *Monolingual Human Preference* (MHP) and *Bilingual LLM Preference* (BLP) to assess the quality of translations. MHP focuses on the translation's impact on target audience readers, emphasizing fluidity and cultural appropriateness, while BLP uses advanced LLMs to compare translations directly with the original texts.
- Despite lower d-BLEU scores, our empirical findings highlight that translations from TRANSAGENTS are preferred by both human evaluators and language models over human-written references. We also present in-depth analyses about the strengths and weaknesses of TRANSAGENTS.

2 RELATED WORK

Large Language Models Large language models (LLMs) have revolutionized the field of artificial intelligence (AI). These models are typically pretrained on a vast corpus of text data, learning to predict the next word in a sentence (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022; Anil et al., 2023b; Touvron et al., 2023a;b; Bai et al., 2023a; Anil et al., 2023a). After pretraining, the models are fine-tuned with instructions. This process, known as supervised fine-tuning (SFT) or instruction tuning (IT), allows the model to adapt its general language understanding to follow and implement instructions from humans (Sanh et al., 2022; Wei et al., 2022; Chung et al., 2022; Wang et al., 2022; Tay et al., 2023; Longpre et al., 2023; Shen et al., 2023). Thanks to the superior capabilities of large language models, recent works demonstrate that synthetic datasets generated by these models can also be used in this step (Wang et al., 2023c; Wu et al., 2023b; Li et al., 2023a; Luo et al., 2023; Lyu et al., 2023; Yue et al., 2023; Wang et al., 2023d). Furthermore, reinforcement learning from human feedback (RLHF) is used to further improve the performance of these models. In this approach, the model is fine-tuned based on feedback from humans or other large language models, who rate the quality of the model's outputs (Ouyang et al., 2022; Rafailov et al., 2023; Hejna et al., 2023; Ethayarajh et al., 2024; Hong et al., 2024). Moreover, evaluating these large language models is a complex task, often involving both automated metrics and human judgment (Hendrycks et al., 2021; Liang et al., 2022; Wu & Aji, 2023; Jiang et al., 2023; Lyu et al., 2024). Additionally, these models pose challenges in terms of efficient training (Hu et al., 2022; Dettmers et al., 2023; Liu et al., 2024), fairness (Li et al., 2023c), hallucination (Zhang et al., 2023c), and other issues, which are also active areas of research. In this work, we leverage the state-of-the-art LLM as the backbone of our multi-agent system for translating the literary texts.

Multi-Agent Systems Intelligent agents are designed to understand their environments, make informed decisions, and respond with appropriate actions (Wooldridge & Jennings, 1995). The capabilities of large language models (LLMs) align well with these expectations. The emergence of LLMs has significantly advanced research on multi-agent systems across various contexts. Multi-agent systems, compared to single-agent setups, are generally expected to either leverage collaboration among multiple agents to tackle complex problems or use diverse agents to effectively simulate complex real-world environments (Guo et al., 2024). Recent studies have shown promising outcomes in complex problem-solving areas such as software development (Qian et al., 2023; Hong et al., 2023), multi-robot collaboration (Mandi et al., 2023; Zhang et al., 2023a), evaluation (Chan et al., 2023), and fact-checking (Du et al., 2023a). Additionally, there is extensive research on using multiple agents to simulate real-world environments, including societal, economic, and gaming simulations (Park et al., 2022; 2023; Xu et al., 2023b; Li et al., 2023b; Mukobi et al., 2023). Liang et al. (2023) propose leveraging multi-agent debate for machine translation. However, their approach is limited to the sentence level. In this work, we focus on the first category, specifically on the translation of literary texts. Literary translation is considered one of the most complex and challenging translation tasks, and we aim to address this challenge using a multi-agent system powered by LLMs.

Machine Translation Machine translation (MT) has achieved significant advancements in recent years, with developments spanning general-purpose MT (Cho et al., 2014; Sutskever et al., 2014; Vaswani et al., 2017; Gehring et al., 2017; Shen et al., 2019), low-resource MT (Zoph et al., 2016; Gu et al., 2018; Haddow et al., 2022), multilingual MT (Liu et al., 2020; Fan et al., 2021; Wu et al., 2021; Li et al., 2022; Costa-jussà et al., 2022; Communication et al., 2023), and non-autoregressive MT (Gu et al., 2017; 2019a; Ghazvininejad et al., 2019), among others. However, these advancements are predominantly focused at the sentence level. Recently, efforts are made to enhance translation quality by integrating contextual information into the translation process (Wang et al., 2017; Ding et al., 2020; Sun et al., 2022; Feng et al., 2022; Wu et al., 2023a; Herold & Ney, 2023; Wu et al., 2024b), aiming to achieve more accurate and coherent translations that extend beyond individual sentences. More recently, large language models (LLMs) have demonstrated superior capabilities in various applications, including MT (Lu et al., 2023; Zhang et al., 2023b; Xu et al., 2023a; Robinson et al., 2023; Wang et al., 2023a; Wu et al., 2024a). Given the remarkable progress in machine translation (MT), the performance of MT seems to be saturating in the general domain. There is growing interest in literary translation, which is considered one of the more challenging translation tasks because it requires not only accuracy in meaning but also the conveyance of vivid expressions and cultural nuances (Thai et al., 2022; Wang et al., 2023b). Additionally, evaluating MT accurately remains a critical aspect of research in this field. While traditional metrics like BLEU are commonly used (Papineni et al., 2002), newer approaches involve utilizing pretrained language models to assess translation quality more effectively (Rei et al., 2020; Sellam et al., 2020; Juraska et al., 2023; Guerreiro et al., 2023). Kocmi & Federmann (2023) employ the state-of-the-art LLM, GPT-4, to estimate translation quality and achieve state-of-the-art quality estimation performance at WMT 2023 (Freitag et al., 2023). In this work, we establish a novel multi-agent virtual company TRANSAGENTS for translating literary texts. We also propose two evaluation strategies for assessing the quality of the translated literary texts.

3 TRANSAGENTS: A MULTI-AGENT VIRTUAL COMPANY FOR LITERARY TRANSLATION

Figure 2: TRANSAGENTS, a multi-agent virtual company for literary translation.

We establish a virtual multi-agent translation company, TRANSAGENTS, featuring a diverse range of employees including a CEO, senior editors, junior editors, translators, localization specialists, and proofreaders. When a human client assigns a book translation task, a team of selected agents from TRANSAGENTS collaborates to translate the book. This paradigm simulates the entire book translation process, where agents with different roles work together to ensure that the translation maintains high quality and consistency throughout. In this section, we describe the company overview of TRANSAGENTS in Section 3.1, the core collaboration strategies of TRANSAGENTS in Section 3.2, and the translation workflow in Section 3.3.

3.1 Company Overview

To simulate the entire book translation process, in addition to the designated CEO, we have a diverse array of roles, including senior editors, junior editors, translators, localization specialists, and proofreaders in our company TransAgents. Each of these roles carries its own set of responsibilities:

- **Senior Editors**: Senior editors are responsible for overseeing the content production process. Their primary duties encompass setting editorial standards, guiding junior editors, and ensuring that the content aligns with the company's objectives.
- **Junior Editors**: Junior editors work closely under the guidance of senior editors. Their responsibilities typically include managing the day-to-day editorial workflow, editing content, and assisting in content planning. They also handle communications with various other roles within the organization.
- **Translators**: Translators are tasked with converting written material from one language to another while preserving the tone, style, and context of the original text. Translators must possess a profound understanding of both the source and target languages, as well as a familiarity with the subject matter they are translating.
- **Localization Specialists**: Localization specialists go beyond simple translation; they adapt content for specific regions or markets. This role involves not only translating language but also adjusting cultural references, idioms, and images to resonate with local audiences.
- **Proofreaders**: Proofreaders perform final checks for grammar, spelling, punctuation, and formatting errors. Their role is crucial in ensuring that content is polished and adheres to high-quality standards before publication.

To enhance the realism and efficacy of our simulation in the translation process, we strategically utilize GPT-4-TURBO to generate a diverse set of 30 virtual agent profiles for each distinct role. As illustrated in Figure 3, these profiles are comprehensively designed to include a wide array of attributes that extend well beyond language skills. Key characteristics such as gender, nationality, rate per word, educational background, years of experience, and areas of specialization are thoughtfully incorporated. This detailed and personalized approach not only enriches the authenticity of the translation process simulation but also mirrors the complexity and diversity found in real-world translation settings. The inclusion of such rich, detailed metadata about the agents not only enhances current simulation strategies but is also designed to support and inspire future research.

```
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 3: An example profile of **Senior Editor**.

3.2 Agent Collaboration Strategies

In this section, we introduce two collaboration strategies used in this work, including *Addition-by-Subtraction Collaboration* (Algorithm 1) and *Trilateral Collaboration* (Algorithm 2).

Addition-by-Subtraction Collaboration In our framework, we propose the *Addition-by-Subtraction Collaboration* between two agents. Unlike the debate-style strategy (Liang et al., 2023; Du et al., 2023a; Chan et al., 2023), where multiple agents propose their own answers and a third-party agent concludes the discussion, our strategy involves only two agents. One acts as an *Addition agent*, responsible for extracting as much relevant information as possible, while the other agent serves as a *Subtraction agent*, tasked with reviewing the extracted information, eliminating redundant details, and providing feedback to the Addition agent. We present the details of our collaboration strategy in Algorithm 1. The Addition agent $\mathbf{A}$ first generates the initial response, aiming to include as much informative content as possible. Subsequently, the Subtraction agent $\mathbf{S}$ reviews the response and removes any redundant information. The conversation iterates until no further revisions are needed for the response.

Algorithm 1: Addition-by-Subtraction Collaboration

Input : Context $\mathbf{C}$; Instruction $\mathbf{I}$; Maximum number of iterations $\mathbf{M}$; Addition agent $\mathbf{A}$; Subtraction agent $\mathbf{S}$;
Output: The final response $\mathbf{R}$ that both agents agree upon.

```
H ← [C; I]    ▷ Initialize the conversation history;
R ← ∅    ▷ Initialize the response;
m ← 0    ▷ Current round;
while m ≤ M do
    m ← m + 1;
    R′ ← A(H)    ▷ Generate detailed response;
    F ← S(H, R′)    ▷ Review and remove redundant information;
    H ← H + [R′; F]    ▷ Append R′ and F to the conversation history H;
    if R = R′ then
        Break    ▷ Stop iterating as no further revisions are needed;
    R ← R′;
Return the final response R;
```

Algorithm 2: Trilateral Collaboration

Input : Context $\mathbf{C}$; Instruction $\mathbf{I}$; Maximum number of iterations $\mathbf{M}$; Action agent $\mathbf{P}$; Critique agent $\mathbf{Q}$; Judgment agent $\mathbf{J}$;
Output: The final response $\mathbf{R}$ that is approved by the Judgment agent $\mathbf{J}$;

```
H ← [C; I]    ▷ Initialize the conversation history;
m ← 0    ▷ Current round;
while m ≤ M do
    m ← m + 1;
    R ← P(H)    ▷ Generate response;
    F ← Q(H, R)    ▷ Generate critiques;
    H ← H + [R; F]    ▷ Append R′ and F to the conversation history H;
    if m > 1 then
        D ← J(C, I, R)    ▷ The Judgment agent J evaluate the response quality;
        if D = TRUE then
            Break    ▷ Stop iterating if the Judgment agent J thinks the response is of high
              quality;
Return the final response R;
```

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

- **Action**: The power to follow the instruction and implement the required actions.
- **Critique**: The power to review the generated response and provide constructive feedback to the Action branch.
- **Judgment**: The power to make the final decision on whether the response is satisfactory or requires further revision.

We assign one agent for each branch and present the details of the collaboration among these agents in Algorithm 2. The *Action agent* $\mathbf{P}$ generates a response $\mathbf{R}$ given the context $\mathbf{C}$ and instruction $\mathbf{I}$. The *Critique agent* $\mathbf{Q}$ then writes critiques $\mathbf{F}$ against the response $\mathbf{R}$. The Action agent $\mathbf{P}$ has the option to either accept the critiques and update the response or maintain the original response. At the end of the iteration, the *Judgment agent* $\mathbf{J}$ evaluates the response $\mathbf{R}$ to determine if the discussion can be concluded or if further deliberation is required.

3.3 TRANSLATION WORKFLOW

In this section, we introduce the book translation workflow in our company TRANSAGENTS, including two main stages: preparation (Section 3.3.1) and execution (Section 3.3.2).

3.3.1 Preparation

Project Members Selection System prompts or messages are used to assign roles to individual agents during the role-playing process. In our company's setup, we create 30 agent profiles, each accompanied by a unique role assignment prompt, as illustrated in Figure 3. These prompts are essential for assigning specific roles to the agents before the dialogues begin. Within our framework, the initial step involves the CEO selecting a Senior Editor for the book translation project. This selection process takes into account both the client's requirements and the qualifications of potential Senior Editors. Once the Senior Editor is chosen, they work closely with the CEO to assemble the rest of the project team, carefully considering the skill sets and backgrounds of the candidates. Furthermore, we introduce a *self-reflection* strategy (Yao et al., 2023; Shinn et al., 2023; Qian et al., 2023). This strategy involves incorporating a "ghost agent" whose task is to prompt the CEO to reconsider their decision, as we observe that they sometimes struggle to select a Senior Editor with the desired language skills.

Translation Guideline Documentation To maintain consistency throughout the entire translation workflow, which involves multiple agents, we need to have a translation guideline. In TransAgents, there are five components: the glossary, the book summary, the tone, the style, and the target audience. We have designed different strategies to process them:

- **Glossary**: The primary purpose of a glossary in book translation is to compile essential terms from the source language and provide their corresponding translations in the target language. This ensures consistency and accuracy in the usage of these terms throughout the book, especially since some terms may have multiple acceptable translations. In our process, we leverage the *Addition-by-Subtraction Collaboration*, as described in Algorithm 1, for collecting the key terms. For each chapter, the Junior Editor, serving as the Addition agent $\mathbf{A}$, makes an exhaustive attempt to identify all potential key terms initially. Subsequently, the Senior Editor, serving as the Subtraction agent $\mathbf{S}$, reviews the identified key terms and removes any that are generic. The conversation continues until the list of collected key terms does not need further revision. Next, the collected key terms are translated by the Senior Editor, with consideration of their context.
- **Book Summary**: Generating a book summary is crucial to provide a comprehensive overview of the narrative. This task is 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. In this process, the Junior Editor aims to retain as much detail as possible in the chapter summaries, while the Senior Editor focuses on removing superfluous information. Following the compilation of chapter summaries, the Senior Editor then crafts the book summary, mirroring the process of gathering a glossary.
- **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 translated book based on a randomly selected chapter.

Overall, the glossary, book summary, tone, style, and target audience collectively constitute the comprehensive translation guidelines. **These guidelines serve as an essential part of the prompts for all roles involved in the book translation process, ensuring consistency and coherence throughout the entire work.**

3.3.2 Execution

In the execution phase, the process is divided into four distinct sub-stages: translation, cultural adaptation, proofreading, and final review. During the first three sub-stages, our approach utilizes the collaborative strategy as illustrated in Algorithm 2. Within this framework, the roles of Action agents $\mathbf{P}$ are assigned to the Translator, the Localization Specialist, and the Proofreader, in that order. Meanwhile, the responsibilities of the Critique agent $\mathbf{Q}$ and the Judgment agent $\mathbf{J}$ are fulfilled by the Junior Editor and the Senior Editor, respectively. Finally, the Senior Editor performs the final checks before publication.

Translation, Localization, and Proofreading The translation stage involves three key roles: the Translator, the Junior Editor, and the Senior Editor. These roles collaborate to translate the book from the source language to the target language on a chapter-by-chapter basis. The translation process begins with the Translator (the Action agent $\mathbf{P}$) initially translating the chapter content from the source language to the target language. Next, the Junior Editor (the Critique agent $\mathbf{Q}$) undertakes a thorough review of the translation, ensuring it adheres to the guidelines while also identifying any potential errors or areas for improvement. Lastly, the Senior Editor (the Judgment agent $\mathbf{J}$) evaluates the translation and determines if further revision is needed. Following the translation, the cultural adaptation process begins. The Localization Specialist tailors the translated content to fit the cultural context of the target audience, ensuring that it resonates well and maintains the intended meaning. Next, the Proofreader perform the checks for language errors. Throughout the cultural adaptation and proofreading stages, both the Junior Editor and the Senior Editor continue to offer critiques and evaluations to refine the content further.

Final Review The final review is the concluding step in the editorial process. At this point, the Senior Editor evaluates the translation quality of each chapter and also examines how pairs of adjacent chapters flow into each other. The Senior Editor not only verifies that each chapter is internally coherent and meets quality standards on its own but also ensures that the transitions between chapters are smooth, thereby maintaining narrative consistency.

On the Importance of the Judgment Agent We introduce the Judgment Agent in Algorithm 2, which is responsible for evaluating the quality of the response and determining whether further revision is needed, without requiring the conversation history. Owing to the nature of web novels, each turn of dialogue is likely to contain a few thousand words. Although recent advances in large language models (LLMs) claim that LLMs are capable of processing extremely lengthy sequences of up to millions of tokens, we still observe that our agents are not able to effectively leverage the information in the context as the conversation expands. Additionally, we observe that the meaning of translations tends to deviate from the original text after several iterations of revision. Therefore, it is critical to have the Judgment agent within the *Trilateral Collaboration* to ensure the overall quality of the response.

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). The following sections introduce the baselines (Section 4.1), datasets (Section 4.2), and evaluation approaches (Section 4.3) used in our study.

4.1 BASELINES

We leverage the state-of-the-art LLM GPT-4-TURBO as the backbone of our agents,[1] and compare our approach with the unconstrained systems in WMT2023 shared task on DLLT:

- **LLAMA-MT**: Du et al. (2023b) fine-tune LLAMA-7B for literary translation. The fine-tuned LLAMA-MT model translates 2,048 consecutive tokens at a time.
- **GPT-4**: While recent versions of GPT-4 models claim to support a context size of up to 128K tokens, they are restricted to generating a maximum of 4,096 tokens per response (OpenAI, 2023). Therefore, we employ the GPT-4-0613 and GPT-4-1106-PREVIEW models to translate the documents on a chapter-by-chapter basis.
- **GOOGLE**: We employ the GOOGLE TRANSLATE system to translate the documents on a sentence-by-sentence basis.
- **DUT**: Zhao et al. (2023) explore several techniques to enhance the performance of large language models (LLMs) in discourse-level translation tasks.
- **HW-TSC**: Xie et al. (2023) initially train a sentence-level Transformer to establish a baseline, subsequently enhancing its discourse-level capabilities through domain adaptation and discourse modeling, employing a variety of techniques.

[1]Model signature: `gpt-4-1106-preview`

4.2 DATASETS

In this work, we do not need to train new models and all the agents is GPT-4-TURBO with various roles. Hence, we only leverage the official test set of WMT2023 shared task on DLLT. The official test set is collected from 20 web novels, each of which consists 20 consecutive chapters, totaling 240 chapters. 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 page.

4.3 EVALUATION

Translating literary works differs significantly from translating standard machine translation (MT) corpora, such as news articles or parliamentary proceedings. Thai et al. (2022) present a comprehensive list of techniques employed by literary translators, which largely differ from those used in common MT domains. Furthermore, literary translators have the freedom and the burden of both semantic and critical interpretation, resulting in the absence of a single, unique best translation for literary texts. In this work, we employ two evaluation approaches:

- **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,[2], 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. We present the results in Section 5.
- **Preference Evaluation**: Acknowledging the concern that there is no single, universally preferred translation for literary texts, we ask human raters or LLMs to select their preferred translation without giving them a reference translation. Further details regarding this novel evaluation approach are discussed in Section 6.

5 STANDARD EVALUATION

We present the automatic evaluation results in Table 1. Interestingly, our approach performs poorly in terms of the d-BLEU metric, achieving the lowest scores among the compared methods. However, it is important to consider that d-BLEU has limitations and may not fully capture the quality and coherence of the generated text. As pointed out by Freitag et al. (2020), typical references used for calculating d-BLEU scores often exhibit poor diversity and tend to concentrate around translationese language. This suggests that a low d-BLEU score does not necessarily imply poor performance of our approach.

	d-BLEU ↑
LLAMA-MT (Du et al., 2023b)	43.1
GPT-4-0613 (OpenAI, 2023)	43.7
GPT-4-1106-PREVIEW (OpenAI, 2023)	47.8
GOOGLE	47.3
DUT (Zhao et al., 2023)	50.2
HW-TSC (Xie et al., 2023)	52.2
TRANSAGENTS (Ours)	**25.0**

Table 1: Automatic evaluation (d-BLEU) results on WMT2023 DLLT test set. ↑ indicates higher is better. **The worst result is highlighted in bold**.

Our results align with the findings from Thai et al. (2022), who argue that automatic metrics cannot accurately reflect human preference in the context of literary translation. Furthermore, while automatic metrics are typically highly correlated with human judgments based on the Multidimensional Quality Metrics (MQM) framework (Burchardt, 2013), this framework may not be suitable for assessing translation quality in the context of literary translation.[3] The unique characteristics and creative aspects of literary texts require a more nuanced evaluation approach that goes beyond the scope of standard automatic metrics and MQM-based human assessments.

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

[3] In our preliminary study, we conduct small-scale MQM-based human evaluation and also observe that our approach, TRANSAGENTS, receives a low MQM score.

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

6 PREFERENCE EVALUATION

It is crucial to acknowledge that a literary text does not possess a single, universal translation. Conventional translation evaluation methodologies, which typically rely on direct comparisons to a standard reference translation, fail to accommodate the multifaceted and subjective nature of literary texts. Following Thai et al. (2022), we engage both human evaluators and large language models (LLMs) to assess translations based on their preferences. In this section, we describe our methods for preference evaluation in Section 6.1 and present our results in Section 6.2.

6.1 EVALUATION METHODS

In this section, we propose two preference evaluation methods, *monolingual human preference* (MHP, Section 6.1.1) and *bilingual LLM preference* (BLP, Section 6.1.2). For both methods, we use the winning rate (%), which is the percentage of instances where a model's generated chapter is preferred by either the human evaluators (in MHP) or the LLM (in BLP), to measure the model performance.

6.1.1 MONOLINGUAL HUMAN PREFERENCE

When reading a translated book, it is not necessary for the reader to understand the original language. Therefore, a better translation should naturally be preferred by readers without needing to refer to the text in its original language.

Preprocessing In this work, the translations of each chapter are first manually split into several segments containing approximately 150 words each, based on the story's plot. This translation segmentation step is necessary because the full translations contain thousands of words, and human evaluators may struggle to stay focused when evaluating such long passages at once.

Evaluation The human evaluators are tasked with comparing pairs of translation segments describing the same part of the story and selecting their preferred translation for each segment pair, with the user interface shown in Figure 4. To ensure evaluations consider the full context, each evaluator is required to evaluate all the segments within a chapter in their original order, as segments may depend on information from previous segments.

Implementation In this study, we collect human preferences on translations through SurveyMonkey.[4] To ensure the evaluators are from the target audience, we ask if they are interested in Chinese

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

web novels before starting the evaluation.[5] 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. 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;
- Giving "No Preference" for all selections;
- Taking less than 20 seconds for the evaluation.

After filtering, we collect at least 5 responses per segment pair.

Mitigating Positional Bias Human evaluators may exhibit a positional bias when evaluating response quality. To mitigate this bias in our translation evaluations, the positions of the translation segments being compared are randomly swapped for each selection, as shown in Figure 4. Furthermore, the "No Preference" (Tie) option, indicating that the evaluator does not prefer one translation over the other, is always presented as the third option.

Response Aggregation We aggregate the human evaluations using majority voting, where the most selected option is considered the final preference. If two translation systems receive the same number of votes, we record the final preference as "No Preference" (Tie).

6.1.2 BILINGUAL LLM PREFERENCE

The nature of literary texts, with their inherent complexities, artistic expression, and cultural nuances, makes it virtually impossible to produce a single, universally correct translation. As a result, multiple translations of the same literary text can coexist, each offering a unique perspective and interpretation. Recent works demonstrate that the reference translations are likely to be of low quality (Freitag et al., 2023; Xu et al., 2024). Kocmi & Federmann (2023) demonstrate that GPT-4 is capable of accurately estimating translation quality without the need for human reference translations. Their proposed GEMBA-MQM metric achieves state-of-the-art performance in WMT 2023 Metric Shared task (Freitag et al., 2023).

```
[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 5: The prompt used for bilingual LLM preference evaluation.

Motivated by Kocmi & Federmann (2023), we evaluate the translation segment pairs using GPT-4-0125-PREVIEW without providing the reference translations. Recent research demonstrates that even state-of-the-art LLMs may struggle to process extremely long sequences (Bai et al., 2023b; Song et al., 2024; Li et al., 2024). Therefore, we require GPT-4-0125-PREVIEW to determine which translation segment is better as described in Section 6.1.1, using the prompt shown in Figure 5, instead of directly comparing the quality of two entire chapters. We employ a different variant of GPT-4 for evaluation to avoid the potential bias. Given concerns about positional bias in LLM evaluation raised by recent studies (Wu & Aji, 2023; Zheng et al., 2023a; Dubois et al., 2024), we evaluate each translation segment pair in both forward and reversed directions.

6.2 EXPERIMENTS

Setup As described in Section 4.2, there are 12 web novels consisting of 240 chapters in our test set. Due to the high cost of human evaluation, we only compare our TRANSAGENTS with the REFERENCE 1 and GPT-4-1106-PREVIEW models. We evaluate the first two chapters of each of the 12 web novels in our test set using both of our preference evaluation methods.

[5]We initially attempt to collect responses directly from web novel forums, such as the `r/WebNovels` subreddit on Reddit. However, this approach proves to be too slow and sometimes violates the community rules of these platforms.

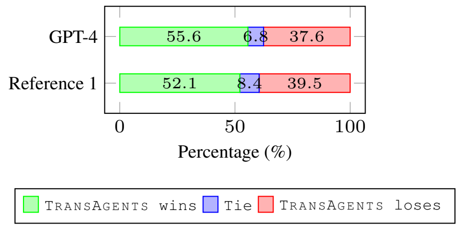

Figure 6: Monolingual Human Preference evaluation results. GPT-4 indicates GPT-4-1106-PREVIEW.

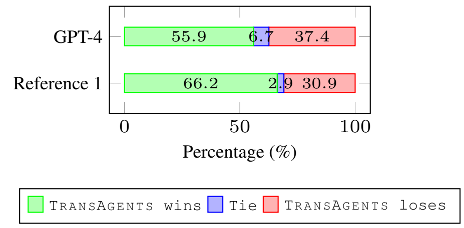

Figure 7: Bilingual LLM Preference evaluation results. GPT-4 indicates GPT-4-1106-PREVIEW.

	Overall	VG	EF	SR	CR	F	SF	HT	FR
Monolingual Human Preference									
GPT-4-1106-PREVIEW	55.6	64.5	68.2	63.3	44.6	68.2	<u>39.1</u>	48.0	**77.8**
REFERENCE 1	52.1	**67.7**	63.6	56.7	42.9	63.6	<u>37.0</u>	40.0	66.7
Bilingual LLM Preference									
GPT-4-1106-PREVIEW	55.9	**74.1**	56.8	58.3	47.3	70.5	47.8	<u>34.0</u>	66.7
REFERENCE 1	66.2	**88.7**	59.1	70.0	54.5	83.0	<u>53.3</u>	62.0	61.1

Table 3: The breakdown winning rate of TRANSAGENTS against GPT-4-1106-PREVIEW and REFERENCE 1. The best results are highlighted in **bold**. The worst results are highlighted in <u>underline</u>.

Results We compare the performance of our TRANSAGENTS with REFERENCE 1 and GPT-4-1106-PREVIEW using monolingual human preference evaluations. The results, presented as winning rates, are shown in Figure 6. The translations produced by TRANSAGENTS are marginally preferred by human evaluators compared to both REFERENCE 1 and GPT-4-1106-PREVIEW. Additionally, we evaluate the models using bilingual LLM preference, with the results presented in Figure 7. The translations generated by TRANSAGENTS are also more preferred by GPT-4-0125-PREVIEW compared to the other models. Referring to the results in Table 4, we observe that GPT-4-0125-PREVIEW appears to have a strong preference for diverse and vivid descriptions when evaluating literary translations. We leave the further investigation to the future work.

7 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. The results, presented in Table 2, reveal that, although the backbone of the agents in TRANSAGENTS is GPT-4-1106-PREVIEW, the initial translation produced by TRANSAGENTS achieves a significantly lower d-BLEU score. This suggests that the translation guideline is the main contributor to the final translation quality. Moreover, the localization step further reduces the d-BLEU score, while the proofreading step only minimally modifies the translation.

	d-BLEU
GPT-4-1106-PREVIEW	47.8
TRANSAGENTS	
– translation	28.8
– localization	25.5
– proofreading	25.0

Table 2: d-BLEU results given by each stage in TRANSAGENTS on WMT2023 DLLT test set. Note that the "proofreading" translation is the final translation of TRANSAGENTS.

Strengths and Weaknesses of TRANSAGENTS The original texts of the test examples are publicly accessible online and span a variety of genres, including Video Games (VG), 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. On the other hand, TRANSAGENTS tends to underperform in contemporary domains, which may not require as much specialized knowledge. This performance trend underscores the model'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 & McFall, 2010) and the Measure of Textual Lexical Diversity (MTLD) (McCarthy & Jarvis, 2010). As shown in Table 4, assisted by our translation guidelines, our initial translation significantly improves linguistic diversity compared to the source 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.

	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.

Cost Analysis The cost of human translation services can be influenced by several factors, including the genre of the text, the translator's location, and their level of experience. The American Translators Association recommends a minimum rate of US$0.12 per word for professional translation services.[6] The REFERENCE 1 from the WMT2023 DLLT test set contains an average of 1,404 English words per chapter, resulting in a translation cost of $168.48 USD per chapter. In comparison, translating using TRANSAGENTS costs approximately $500 USD for the entire test set, which is equivalent to $2.08 USD per chapter. **Translating literary text using TRANSAGENTS can lead to an $80\times$ reduction in translation costs.**

8 CASE STUDY

In this section, we explore two case studies with regard to cultural adaptation and content omission, shedding light on both the strengths and weaknesses of our approach. Additionally, we enrich our analysis by incorporating insights from interviews with two experienced professional translators.

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 5, TRANSAGENTS is the only system that accurately reflects this cultural context in its translations. In contrast, both REFERENCE 1 and GPT-4-1106-PREVIEW fail to correctly adjust the order of names and job titles, thus not adhering to the cultural norms expected in the target language. The ability to produce translations that are not only linguistically accurate but also culturally appropriate is crucial. This emphasizes the capability of TRANSAGENTS to provide translations that are culturally appropriate, ensuring an immersive reading experience for readers in the target language.

Global Consistency It is important to maintain consistency throughout the book translation from the start to the end. As shown in Table 6, the chapter titles are consistent, with the exception of the index. Both REFERENCE 1 and TRANSAGENTS successfully produce consistent translations.

[6] We could not find the direct source of this information from the American Translators Association. Our source of information is available at `https://tinyurl.com/bdze92xr`. We assume that the recommended rate of $0.12 USD per word is based on the number of words in the English language text.

Original 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-4-1106-PREVIEW	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	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 5: Case study for cultural adaptation. The text highlighted in red indicates that the translation is accurate in meaning but not in cultural context. The text highlighted in blue indicate that the translation is accurate both in meaning and in cultural context.

Original Text	第1906章不思量，自难忘（十二）[OMITTED] 第1907章不思量，自难忘（十三）[OMITTED]
REFERENCE 1	Chapter 1906: Unforgettable Memories (12) [OMITTED] Chapter 1907: Unforgettable Memories (13)
GPT-4-1106-PREVIEW	Chapter 1906: It's Hard to Forget Without Thinking (Twelve) [OMITTED] Chapter 1907: Without Intention, Unforgettable (Thirteen)
TRANSAGENTS	Chapter 1906: Without Intention, Unforgettable (Twelve) [OMITTED] Chapter 1907: Without Intention, Unforgettable (Thirteen)

Table 6: Case study for global consistency. The text highlighted in red indicates that GPT-4-1106-PREVIEW generates inconsistent translations across different chapters.

However, GPT-4-1106-PREVIEW struggles with maintaining consistency across different chapters. This demonstrates that TRANSAGENTS is capable of maintaining consistency throughout the entire translation process, similar to human translators.

Content Omission Our TRANSAGENTS is generally preferred over both REFERENCE 1 and GPT-4-1106-PREVIEW according to evaluations by human judges and large language models (LLMs) (Figure 6 and Figure 7). 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 content omission, as illustrated in Table 7. While these omissions 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 omissions, 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, 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 8, 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.

9 LIMITATIONS

The primary limitation of our study centers on the evaluation methods used. Extensive literature has highlighted the issues in conventional machine translation (MT) evaluation techniques, such as poor evaluation metrics and the reliability of reference translations (Papineni et al., 2002; Post, 2018; Rei et al., 2020; Freitag et al., 2020; 2021; 2022; Kocmi et al., 2023; Freitag et al., 2023). Beyond traditional MT evaluation metrics such as d-BLEU, we propose additional methods, namely

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 7: Case study for content omission. 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 8: 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.

Monolingual Human Preference and *Bilingual LLM Preference*, to assess translation quality. However, the implementation of these novel evaluation strategies introduces several challenges that may undermine the validity of our findings:

- **Document Segmentation**: Evaluating ultra-long texts introduces distinct challenges in human evaluation. In our preliminary study, we observe that human evaluators often struggle to maintain focus when reading documents containing thousands of words, which could potentially compromise the accuracy of their evaluations. Moreover, while segmenting these lengthy texts into smaller, content-based portions may simplify the task, this method risks disrupting the narrative flow and connections between different sections, potentially resulting in a loss of overall coherence. We strategically segmented the documents for this study. However, developing more effective methods for human evaluation of ultra-long texts remains an area for future research.

- **Target Audience**: Literary texts are crafted with specific target audiences in mind. In our study, we initially aim to distribute our questionnaires through an online forum dedicated to web novels, intending to gather feedback directly from the target audience. However, this approach faced challenges, either due to community regulations or the slow pace of feedback collection. Additionally, although we confirm the interest of human evaluators in Chinese web novels before they participate in the evaluation, there is a possibility that evaluators might claim interest simply to qualify for the job, regardless of their true preferences. Consequently, this could mean that our evaluation results might not accurately reflect the true preferences of the target audience.
- **Evaluation Scale**: Due to constrained resources, the scope of our evaluation scale may be inadequate. We segment only the first two chapters of each book in the test set and gather a minimum of five valid responses per segment. Recent studies highlight the significant diversity in human preferences (Zheng et al., 2023b; Wu & Aji, 2023; Hosking et al., 2023). Consequently, the limited scale of our evaluation could affect the outcomes.
- **Human-Written References**: Although the reference translations are said to be authored by professional human translators, there is a likelihood that these translators may use commercial machine translation systems, such as GOOGLE TRANSLATE, to reduce their workload. Unfortunately, we cannot verify whether the reference translations are genuinely created by humans.

We acknowledge these limitations and leave them to the future studies.

10 CONCLUSION

In this paper, we introduce TRANSAGENTS, a novel multi-agent virtual company designed for literary translation that reflects the traditional translation publication process. Utilizing a multi-agent approach, this system effectively tackles the intricate nuances inherent in literary texts. We propose two innovative evaluation strategies: *Monolingual Human Preference* (MHP) and *Bilingual LLM Preference* (BLP), to assess the quality of the translations. MHP evaluates how the translation resonates with the target audience, focusing on fluidity and cultural appropriateness, whereas BLP employs advanced language models to directly compare the translations with the original texts. Although the d-BLEU scores are lower, our empirical results demonstrate that translations produced by TRANSAGENTS are favored by both human evaluators and language models over human-written references. We also provide detailed analyses of the strengths and weaknesses of TRANSAGENTS, highlighting possible directions for future research.

REFERENCES

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023a. doi: 10.48550/ARXIV.2312.11805. URL `https://doi.org/10.48550/arXiv.2312.11805`.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James

Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. Palm 2 technical report. *CoRR*, abs/2305.10403, 2023b. doi: 10.48550/ARXIV.2305.10403. URL https://doi.org/10.48550/arXiv.2305.10403.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *CoRR*, abs/2309.16609, 2023a. doi: 10.48550/ARXIV.2309.16609. URL https://doi.org/10.48550/arXiv.2309.16609.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. *CoRR*, abs/2308.14508, 2023b. doi: 10.48550/ARXIV.2308.14508. URL https://doi.org/10.48550/arXiv.2308.14508.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Aljoscha Burchardt. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib. URL https://aclanthology.org/2013.tc-1.6.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *CoRR*, abs/2308.07201, 2023. doi: 10.48550/ARXIV.2308.07201. URL https://doi.org/10.48550/arXiv.2308.07201.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022. doi: 10.48550/arXiv.2204.02311. URL https://doi.org/10.48550/arXiv.2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/ARXIV.2210.11416. URL https://doi.org/10.48550/arXiv.2210.11416.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Y. Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. Seamlessm4t-massively multilingual & multimodal machine translation. *CoRR*, abs/2308.11596, 2023. doi: 10.48550/ARXIV.2308.11596. URL https://doi.org/10.48550/arXiv.2308.11596.

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. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672, 2022. doi: 10.48550/ARXIV.2207.04672. URL https://doi.org/10.48550/arXiv.2207.04672.

Michael A Covington and Joe D McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100, 2010.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Liang Ding, Longyue Wang, Di Wu, Dacheng Tao, and Zhaopeng Tu. Context-aware cross-attention for non-autoregressive translation. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4396–4402, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.389. URL https://aclanthology.org/2020.coling-main.389.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. Self-collaboration code generation via chatgpt. *CoRR*, abs/2304.07590, 2023. doi: 10.48550/ARXIV.2304.07590. URL https://doi.org/10.48550/arXiv.2304.07590.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *CoRR*, abs/2305.14325, 2023a. doi: 10.48550/ARXIV.2305.14325. URL https://doi.org/10.48550/arXiv.2305.14325.

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

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: model alignment as prospect theoretic optimization. *CoRR*, abs/2402.01306, 2024. doi: 10.48550/ARXIV.2402.01306. URL https://doi.org/10.48550/arXiv.2402.01306.

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. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48, 2021. URL http://jmlr.org/papers/v22/20-1307.html.

Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. Learn to remember: Transformer with recurrent memory for document-level machine translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1409–1420, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.105. URL https://aclanthology.org/2022.findings-naacl.105.

Markus Freitag, David Grangier, and Isaac Caswell. BLEU might be guilty but references are not innocent. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 61–71, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.5. URL https://aclanthology.org/2020.emnlp-main.5.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.2.

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. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 578–628, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.51. URL https://aclanthology.org/2023.wmt-1.51.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1243–1252. PMLR, 2017. URL http://proceedings.mlr.press/v70/gehring17a.html.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6112–6121, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1633. URL https://aclanthology.org/D19-1633.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation. *CoRR*, abs/1711.02281, 2017. URL http://arxiv.org/abs/1711.02281.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. Meta-learning for low-resource neural machine translation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3622–3631, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1398. URL https://aclanthology.org/D18-1398.

Jiatao Gu, Changhan Wang, and Junbo Zhao. Levenshtein transformer. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11179–11189, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/675f9820626f5bc0afb47b57890b466e-Abstract.html.

Jiatao Gu, Changhan Wang, and Junbo Zhao. Levenshtein transformer. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11179–11189, 2019b. URL https://proceedings.neurips.cc/paper/2019/hash/675f9820626f5bc0afb47b57890b466e-Abstract.html.

Nuno Miguel Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *CoRR*, abs/2310.10482, 2023. doi: 10.48550/ARXIV.2310.10482. URL https://doi.org/10.48550/arXiv.2310.10482.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *CoRR*, abs/2402.01680, 2024. doi: 10.48550/ARXIV.2402.01680. URL https://doi.org/10.48550/arXiv.2402.01680.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindrich Helcl, and Alexandra Birch. Survey of low-resource machine translation. *Comput. Linguistics*, 48(3):673–732, 2022. doi: 10.1162/COLI_A_00446. URL https://doi.org/10.1162/coli_a_00446.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. Contrastive preference learning: Learning from human feedback without RL. *CoRR*, abs/2310.13639, 2023. doi: 10.48550/ARXIV.2310.13639. URL https://doi.org/10.48550/arXiv.2310.13639.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Christian Herold and Hermann Ney. Improving long context document-level machine translation. In Michael Strube, Chloe Braud, Christian Hardmeier, Junyi Jessy Li, Sharid Loaiciga, and Amir Zeldes (eds.), *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pp. 112–125, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.codi-1.15. URL https://aclanthology.org/2023.codi-1.15.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: monolithic preference optimization without reference model. *CoRR*, abs/2403.07691, 2024. doi: 10.48550/ARXIV.2403.07691. URL https://doi.org/10.48550/arXiv.2403.07691.

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. Metagpt: Meta programming for multi-agent collaborative framework. *CoRR*, abs/2308.00352, 2023. doi: 10.48550/ARXIV.2308.00352. URL https://doi.org/10.48550/arXiv.2308.00352.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. *CoRR*, abs/2309.16349, 2023. doi: 10.48550/ARXIV.2309.16349. URL https://doi.org/10.48550/arXiv.2309.16349.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for large language models. *CoRR*, abs/2310.20410, 2023. doi: 10.48550/ARXIV.2310.20410. URL https://doi.org/10.48550/arXiv.2310.20410.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 756–767, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.63. URL https://aclanthology.org/2023.wmt-1.63.

Jeremy Klemin. The last frontier of machine translation. *The Atlantic*, 2024. URL https://www.theatlantic.com/technology/archive/2024/01/literary-translation-artificial-intelligence/677038/.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 768–775, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.64. URL https://aclanthology.org/2023.wmt-1.64.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *CoRR*, abs/2305.15011, 2023a. doi: 10.48550/ARXIV.2305.15011. URL https://doi.org/10.48550/arXiv.2305.15011.

Nian Li, Chen Gao, Yong Li, and Qingmin Liao. Large language model-empowered agents for simulating macroeconomic activities. *CoRR*, abs/2310.10436, 2023b. doi: 10.48550/ARXIV.2310.10436. URL https://doi.org/10.48550/arXiv.2310.10436.

Pengfei Li, Liangyou Li, Meng Zhang, Minghao Wu, and Qun Liu. Universal conditional masked language pre-training for neural machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6379–6391, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.442. URL https://aclanthology.org/2022.acl-long.442.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *CoRR*, abs/2308.10149, 2023c. doi: 10.48550/ARXIV.2308.10149. URL https://doi.org/10.48550/arXiv.2308.10149.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *CoRR*, abs/2211.09110, 2022. doi: 10.48550/ARXIV.2211.09110. URL https://doi.org/10.48550/arXiv.2211.09110.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *CoRR*, abs/2305.19118, 2023. doi: 10.48550/ARXIV.2305.19118. URL https://doi.org/10.48550/arXiv.2305.19118.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *CoRR*, abs/2402.09353, 2024. doi: 10.48550/ARXIV.2402.09353. URL https://doi.org/10.48550/arXiv.2402.09353.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22631–22648. PMLR, 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. Chain-of-dictionary prompting elicits translation in large language models. *CoRR*, abs/2305.06575, 2023. doi: 10.48550/ARXIV.2305.06575. URL https://doi.org/10.48550/arXiv.2305.06575.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *CoRR*, abs/2306.08568, 2023. doi: 10.48550/ARXIV.2306.08568. URL https://doi.org/10.48550/arXiv.2306.08568.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *CoRR*, abs/2306.09093, 2023. doi: 10.48550/ARXIV.2306.09093. URL https://doi.org/10.48550/arXiv.2306.09093.

Chenyang Lyu, Minghao Wu, and Alham Fikri Aji. Beyond probabilities: Unveiling the misalignment in evaluating large language models. *CoRR*, abs/2402.13887, 2024. doi: 10.48550/ARXIV.2402.13887. URL `https://doi.org/10.48550/arXiv.2402.13887`.

Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. *CoRR*, abs/2307.04738, 2023. doi: 10.48550/ARXIV.2307.04738. URL `https://doi.org/10.48550/arXiv.2307.04738`.

Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010.

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation. *CoRR*, abs/2310.08901, 2023. doi: 10.48550/ARXIV.2310.08901. URL `https://doi.org/10.48550/arXiv.2310.08901`.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774. URL `https://doi.org/10.48550/arXiv.2303.08774`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040`.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In Maneesh Agrawala, Jacob O. Wobbrock, Eytan Adar, and Vidya Setlur (eds.), *The 35th Annual ACM Symposium on User Interface Software and Technology, UIST 2022, Bend, OR, USA, 29 October 2022 - 2 November 2022*, pp. 74:1–74:18. ACM, 2022. doi: 10.1145/3526113.3545616. URL `https://doi.org/10.1145/3526113.3545616`.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (eds.), *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*, pp. 2:1–2:22. ACM, 2023. doi: 10.1145/3586183.3606763. URL `https://doi.org/10.1145/3586183.3606763`.

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL `https://aclanthology.org/W18-6319`.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *CoRR*, abs/2307.07924, 2023. doi: 10.48550/ARXIV.2307.07924. URL `https://doi.org/10.48550/arXiv.2307.07924`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. ChatGPT MT: Competitive for high- (but not low-) resource languages. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 392–418, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.40. URL https://aclanthology.org/2023.wmt-1.40.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022. doi: 10.48550/ARXIV.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *CoRR*, abs/2305.14705, 2023. doi: 10.48550/ARXIV.2305.14705. URL https://doi.org/10.48550/arXiv.2305.14705.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. Mixture models for diverse machine translation: Tricks of the trade. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5719–5728. PMLR, 2019. URL http://proceedings.mlr.press/v97/shen19c.html.

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

Neural Information Processing Systems, 2023. URL https://openreview.net/forum?id=vAElhFcKW6.

Mingyang Song, Mao Zheng, and Xuan Luo. Counting-stars: A simple, efficient, and reasonable strategy for evaluating long-context large language models. *CoRR*, abs/2403.11802, 2024. doi: 10.48550/ARXIV.2403.11802. URL https://doi.org/10.48550/arXiv.2403.11802.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. Rethinking document-level neural machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3537–3548, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.279. URL https://aclanthology.org/2022.findings-acl.279.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3104–3112, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=6ruVLB727MC.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023a. doi: 10.48550/ARXIV.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b. doi: 10.48550/ARXIV.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on*

Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Rob Voigt and Dan Jurafsky. Towards a literary machine translation: The role of referential cohesion. In David Elson, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz (eds.), *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pp. 18–25, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-2503.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. Exploiting cross-sentence context for neural machine translation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2826–2831, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1301. URL https://aclanthology.org/D17-1301.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16646–16661, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1036. URL https://aclanthology.org/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. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 55–67, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.3. URL https://aclanthology.org/2023.wmt-1.3.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL https://aclanthology.org/2022.emnlp-main.340.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754.

Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation. *CoRR*, abs/2311.16511, 2023d. doi: 10.48550/ARXIV.2311.16511. URL https://doi.org/10.48550/arXiv.2311.16511.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *CoRR*, abs/2302.01560, 2023e. doi: 10.48550/ARXIV.2302.01560. URL https://doi.org/10.48550/arXiv.2302.01560.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Michael J. Wooldridge and Nicholas R. Jennings. Intelligent agents: theory and practice. *Knowl. Eng. Rev.*, 10(2):115–152, 1995. doi: 10.1017/S0269888900008122. URL https://doi.org/10.1017/S0269888900008122.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. *CoRR*, abs/2307.03025, 2023. doi: 10.48550/ARXIV.2307.03025. URL https://doi.org/10.48550/arXiv.2307.03025.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7291–7305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.580. URL https://aclanthology.org/2021.emnlp-main.580.

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

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *CoRR*, abs/2304.14402, 2023b. doi: 10.48550/ARXIV.2304.14402. URL https://doi.org/10.48550/arXiv.2304.14402.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468, 2024a. doi: 10.48550/ARXIV.2401.06468. URL https://doi.org/10.48550/arXiv.2401.06468.

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

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. HW-TSC's submissions to the WMT23 discourse-level literary translation shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 302–306, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.32. URL https://aclanthology.org/2023.wmt-1.32.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674, 2023a. doi: 10.48550/ARXIV.2309.11674. URL https://doi.org/10.48550/arXiv.2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *CoRR*, abs/2401.08417, 2024. doi: 10.48550/ARXIV.2401.08417. URL https://doi.org/10.48550/arXiv.2401.08417.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf. *CoRR*, abs/2309.04658, 2023b. doi: 10.48550/ARXIV.2309.04658. URL https://doi.org/10.48550/arXiv.2309.04658.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 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, 2023. URL https://openreview.net/pdf?id=WE_vluYUL-X.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *CoRR*, abs/2309.05653, 2023. doi: 10.48550/ARXIV.2309.05653. URL https://doi.org/10.48550/arXiv.2309.05653.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. *CoRR*, abs/2307.02485, 2023a. doi: 10.48550/ARXIV.2307.02485. URL https://doi.org/10.48550/arXiv.2307.02485.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *CoRR*, abs/2306.10968, 2023b. doi: 10.48550/ARXIV.2306.10968. URL https://doi.org/10.48550/arXiv.2306.10968.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023c. doi: 10.48550/ARXIV.2309.01219. URL https://doi.org/10.48550/arXiv.2309.01219.

Anqi Zhao, Kaiyu Huang, Hao Yu, and Degen Huang. DUTNLP system for the WMT2023 discourse-level literary translation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 296–301, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.31. URL https://aclanthology.org/2023.wmt-1.31.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685, 2023b. doi: 10.48550/arXiv.2306.05685. URL https://doi.org/10.48550/arXiv.2306.05685.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1575, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1163. URL https://aclanthology.org/D16-1163.