John Benjamins Publishing Company



This is a contribution from Target 38:3
© 2026. John Benjamins Publishing Company

This electronic file may not be altered in any way. The author(s) of this material is/are permitted to use this PDF file to generate printed copies to be used by way of offprints for their personal use only.

Permission is granted by the publishers to post this file on a closed server which is accessible only to members (students and faculty) of the author's institute. It is not permitted to post this PDF on the internet, or to share it on sites such as Mendeley, ResearchGate, Academia.edu.

Please see our rights policy at **https://benjamins.com/content/customers/rights**
For any other use of this material prior written permission should be obtained from the publishers or through the Copyright Clearance Center (for USA: **www.copyright.com**).

For further information, please contact **rights@benjamins.nl** or consult our website: **www.benjamins.com**

Workflow matters

Comparing human translators and multi-agent LLMs in literary translation

Lulu Wang[1], Sanjun Sun[2], Xing Wang[3], Jinghang Gu[1] and Kanglong Liu[1]
[1] The Hong Kong Polytechnic University | [2] Beijing Foreign Studies University | [3] Tencent

Large language models (LLMs) have shown significant potential in translation tasks but often struggle with literary texts. This study compares professional human translations with translations produced by two AI-driven systems that coordinate multiple LLM-based agents. The first system mimics professional human translation practice, with distinct drafting and revision phases. The second redesigns the process specifically for LLMs' capabilities, breaking translation into granular steps with specialized AI agents handling strategic planning, stylistic refinement, and coherence checking. Expert evaluations revealed that both AI systems achieved accuracy comparable to professional human translators. The LLM-capability-driven system produced translations with superior stylistic qualities and poetic language, though it occasionally added extraneous content. Meanwhile, the practice-derived system delivered concise translations but sometimes lacked cohesive flow. Blind evaluations showed that the translations from both AI systems were frequently preferred over human translations, particularly in terms of fluency. This study demonstrates that rethinking translation workflows around LLM capabilities can yield exceptional results, sometimes surpassing human performance in certain aspects.

Keywords: large language models (LLMs), multi-agent systems, literary translation, translation quality, translation technology

1. Introduction

The translation process has long been studied as a complex cognitive activity requiring multiple competencies, from linguistic knowledge to cultural awareness

https://doi.org/10.1075/target.25081.wan | Published online: 1 June 2026
Target 38:3 (2026), pp. 422–453. ISSN 0924-1884 | E-ISSN 1569-9986
Available under the CC BY 4.0 license. © John Benjamins Publishing Company

(e.g., Muñoz Martín 2016; Carl and Schaeffer 2017). The emergence of large language models (LLMs) has introduced new possibilities for translation, with state-of-the-art models now surpassing traditional neural machine translation (NMT) systems in fluency and naturalness for many language pairs (Briva-Iglesias, Camargo, and Dogru 2024; Gao et al. 2024; Jiang et al. 2024). However, compared with expert human translation, particularly in literary contexts, LLMs still show considerable limitations (He 2024; R. Zhang, Zhao, and Eger 2025).

The integration of translation technology into professional workflows is well established in technical and commercial domains, yet its adoption in literary translation has been cautious and limited. Traditional machine translation (MT) has struggled to capture the stylistic nuances, creative metaphors, and cultural subtleties that are central to literary works, leading many literary translators to avoid computer-aided translation (CAT) tools or MT altogether (Taivalkoski-Shilov 2019; Youdale, Rothwell, and Way 2023). These tools are often perceived as incompatible with the artistic and interpretive nature of literary translation. Concerns about creativity, translator autonomy, and emerging ethical issues continue to shape skepticism toward automation in the literary field (Toral and Way 2018; Taivalkoski-Shilov 2019; Kenny and Winters 2020). However, recent quality improvements in NMT and LLMs are beginning to challenge these reservations (Youdale, Rothwell, and Way 2023). Their potential to enhance efficiency and output quality, together with their limitations in creative contexts, underscore the need to understand how such models can be effectively integrated into literary translation practice.

Research on improving LLM-based translation has largely focused on prompting, a method that provides specific instructions to guide the model's output and is more accessible than technical alternatives like fine-tuning (Elshin et al. 2024). Various prompting strategies have been tested, including providing example translations (Moslem et al. 2023), guiding models through reasoning steps (Wei et al. 2022; Peng et al. 2023), and offering detailed contextual information (He 2024; Jiang et al. 2024). Yet the results have been mixed, with some studies reporting that simpler prompts sometimes outperform complex ones (R. Zhang, Zhao, and Eger 2025). Despite these efforts, current approaches struggle to address the nuanced challenges of literary translation, where creative expression, cultural sensitivity, and stylistic appropriateness are paramount across its diverse genres and forms (Fu and L. Liu 2024; R. Zhang, Zhao, and Eger 2025).

In response to these limitations, a promising new approach involving multi-agent systems organizes multiple LLMs into collaborative structures with specialized roles (Wu, Xu, and Longyue Wang 2024). This approach divides complex tasks into manageable components and has been shown to be effective across various domains (Dorri, Kanhere, and Jurdak 2018; Guo et al. 2024). For translation

© 2026. John Benjamins Publishing Company
All rights reserved

specifically, systems like TransAgents have demonstrated strong results in fiction translation (Wu, Xu, and Longyue Wang 2024). However, many current multi-agent frameworks replicate human collaborative workflows (Qian et al. 2024; Lei Wang et al. 2024), raising questions about whether these structures optimally serve LLMs' distinct capabilities and limitations. For instance, TransAgents has been shown to omit substantial portions of source content in longer texts (Wu et al. 2025), suggesting challenges in directly transferring human workflow models to LLM systems.

The question of optimal workflow design for LLM-based translation represents a significant gap in current research. While translation process research has extensively studied human translation workflows and collaboration patterns, the equivalent knowledge for LLM-based systems remains underdeveloped. This gap becomes particularly relevant as translation technology continues to advance and the integration of LLMs into professional translation workflows becomes increasingly common.

This study examines how different multi-agent designs affect literary translation quality by comparing two systems: a practice-derived multi-agent system (PD-MAS) based on standard human translation workflows and an LLM-capability-driven multi-agent system (LCD-MAS) featuring specialized agents for planning, stylistic refinement, and coherence checking. Through expert evaluations of Chinese–English fiction translations, we assess how these different process structures influence translation accuracy, fluency, and stylistic appropriateness. Specifically, this study addresses two primary questions:

1. How does the quality of translations from the LCD-MAS compare to those from the PD-MAS in terms of accuracy, fluency, and overall rater preference?
2. Can either multi-agent system achieve translation quality comparable to that of professional human translators?

The findings contribute to our understanding of effective translation process design in the era of advanced language models and offer insights into the future of human–AI collaborative translation.

2. Related work

This section reviews three key areas underpinning our research: the challenges of literary translation, LLM capabilities in translation, and multi-agent systems that mirror collaborative translation processes. We trace how the field has shifted from viewing translation as a solitary cognitive activity to a collaborative process — an evolution now reflected in computational approaches that distribute translation tasks across specialized agents.

© 2026. John Benjamins Publishing Company
All rights reserved

2.1 Literary translation: Distinctive features and challenges

Literary translation differs fundamentally from technical translation in both its objectives and challenges. Situated at the intersection of linguistic transfer and artistic recreation, it requires not only accuracy but also the ability to reproduce literary effects, voices, and styles in the target language (Jones 2019; Kenny and Winters 2020). Translators must attend to rhythm, wordplay, metaphor, allusion, and narrative voice, all of which shape a work's artistic identity (Boase-Beier 2014; Jones 2019; Matusov 2019). These demands are further complicated by the diversity of literary genres, such as fiction, drama, and poetry, each with distinct aesthetic goals and stylistic conventions. Consequently, literary translation calls for a variety of strategies and creative skills to recreate the reading experience in another language.

Given these complexities, literary translation quality cannot be adequately assessed through error-counting or linguistic metrics alone. Instead, holistic models considering both textual and contextual factors, such as those proposed by Reiss (2000) and House (2015), emphasize the need to evaluate communicative function, effectiveness in recreating the source text's aesthetic experience, and cultural resonance. Literary translation is thus judged as much by its ability to stand as an independent work in the target culture as by its formal accuracy. Reader-response approaches, which foreground how audiences perceive translated texts, further highlight the importance of subjective and contextual factors in quality assessment (Brumme and Espunya 2012; Fonteyne, Tezcan, and Macken 2020).

The creative nature of literary translation poses significant challenges for MT systems. Despite advances in NMT, studies consistently show that automated systems struggle with the creative and culturally embedded dimensions of literary texts. They often fail to adequately handle wordplay, metaphors, register shifts, idiomatic expressions, and cultural allusions, which are central to literary effect (Toral and Way 2018; Matusov 2019; Fonteyne, Tezcan, and Macken 2020; Kenny and Winters 2020; Guerberof-Arenas and Toral 2022). These limitations stem partly from training on predominantly non-literary corpora and from an inability to recognize the cultural significance of linguistic choices (Besacier and Schwartz 2015).

Furthermore, current MT architecture faces inherent structural limitations when processing literary texts. Most systems operate at the sentence level within narrow context windows, preventing them from maintaining narrative continuity or consistent character voices across chapters (Matusov 2019; Fonteyne, Tezcan, and Macken 2020). While advanced language models can produce superficially fluent translations, they often flatten stylistic nuances and introduce semantic distortions, particularly with creative or culturally specific language (Toral and

© 2026. John Benjamins Publishing Company
All rights reserved

Way 2018; R. Zhang, Zhao, and Eger 2025). Research shows that although some machine-translated sentences may approach publishable quality, most require substantial human post-editing to address stylistic, discursive, and cultural issues (Matusov 2019). These persistent limitations underscore why human expertise remains indispensable in literary translation (Kenny and Winters 2020; Guerberof-Arenas and Toral 2022).

2.2 LLMs in translation: Capabilities and process integration

LLMs offer advantages over conventional NMT systems through large-scale transformer architectures and extensive pre-training on diverse multilingual corpora (Achiam et al. 2023). These features enable them to capture broader contextual dependencies during translation, mitigating the sentence-level fragmentation common in NMT outputs. Longyue Wang et al. (2023) and Briva-Iglesias, Camargo, and Dogru (2024) demonstrated that LLMs produce translations with improved coherence, particularly for documents requiring consistent terminology and stylistic choices. Their extensive pre-training also equips them with substantial world knowledge, enabling more nuanced translations of culturally bound expressions. Empirical studies show that LLMs outperform traditional systems across diverse genres, including legal texts (Briva-Iglesias, Camargo, and Dogru 2024), classical Chinese poetry (Gao et al. 2024), political discourse (Jiang et al. 2024), and news content (J. Yan et al. 2024).

Despite these advantages, LLMs encounter specific challenges when applied to literary texts. Their fundamental token-prediction mechanism limits creative problem-solving when confronted with novel linguistic structures or cultural references without direct equivalents (He 2024; R. Zhang, Zhao, and Eger 2025). Context windows, though expanded in recent models, still constrain narrative coherence across long texts (Karpinska and Iyyer 2023). This is particularly problematic for literary translation, where plot development and thematic motifs often span entire works. R. Zhang, Zhao, and Eger (2025) found that LLMs, while outperforming traditional NMT systems on literary texts, still produce translations characterized by overly literal renderings and stylistic deficiencies. These limitations become evident particularly in relation to the "rich points" identified in translation process research, where cultural, linguistic, and stylistic factors converge to create complex translation challenges (PACTE Group 2017).

Researchers have explored a range of prompting strategies to improve the quality of literary translation carried out by LLMs. Few-shot prompting has produced mixed results, with effectiveness depending more on the choice of examples than on their number (Moslem et al. 2023; B. Zhang, Haddow, and Birch 2023). Role-based prompting, which frames the LLM as a professional translator,

© 2026. John Benjamins Publishing Company
All rights reserved

generally produces only modest improvements (He 2024). Chain-of-thought approaches yield limited gains in translation quality (Wei et al. 2022; Peng et al. 2023). Paradoxically, studies by Puppel and Borg (2025) and R. Zhang, Zhao, and Eger (2025) report that simpler prompts can outperform more complex ones, suggesting that prompt engineering alone remains insufficient to address the creative aspects of literary translation.

2.3 Multi-agent systems: Parallels with human translation process

Recent research has proposed multi-agent systems as a solution to the persistent challenges faced by LLMs in literary translation (Wu, Xu, and Longyue Wang 2024). Building on earlier work in artificial intelligence on agent cooperation (Wooldridge 2009), these systems divide complex problems into components handled by specialized agents with distinct roles and decision-making procedures (Dorri, Kanhere, and Jurdak 2018; Guo et al. 2024). Park et al. (2023) and Chan et al. (2023) have demonstrated the effectiveness of collaborative agent frameworks in complex tasks. When applied to translation, these systems distribute translation tasks across specialized agents, often outperforming single-agent approaches (Liang et al. 2024; Wu, Xu, and Longyue Wang 2024).

Multi-agent translation systems do not simply represent a shift from single-model to collaborative computational workflows; they also parallel the evolution of translation process research from viewing translation as an individual cognitive activity to understanding it as a collaborative social process. Earlier research, such as Jakobsen (2002) and Mossop (2000), conceptualized translation as a linear progression through pre-translation analysis, drafting, and post-translation revision. As research methodologies advanced, studies by Hvelplund (2011) and Schaeffer and Carl (2013) documented the non-linear and recursive nature of the translation process, revealing how translators constantly shift between source and target texts, re-evaluating and refining their work across multiple iterations (Muñoz Martín 2016). Multi-agent systems mirror this recursiveness by assigning different agents to sequential stages, such as initial drafting, revision, and final review, allowing each agent to iteratively improve the translation.

Socio-cognitive models in Translation Studies highlight the collaborative nature of professional translation, where complex projects are distributed among teams with specialized roles through organizational workflows (Kuznik and Verd 2010; Ehrensberger-Dow and Massey 2014; Risku 2014). Multi-agent systems embody this principle by assigning distinct roles to individual agents, such as terminology management, translation, and review. This division of labor mirrors professional translation workflows, in which translators, revisers, and reviewers contribute complementary expertise to the final product (International

© 2026. John Benjamins Publishing Company
All rights reserved

Organization for Standardization 2015). The multi-agent approach thus offers a computational framework for exploring translation as a distributed cognitive activity in the context of AI.

The design of multi-agent translation systems involves a fundamental choice between human-mimicking and LLM-capability-driven workflows. Human-mimicking approaches replicate professional translation practices by assigning agents to traditional roles (Wu, Xu, and Longyue Wang 2024). While these configurations leverage established process knowledge, they often fail to fully exploit LLMs' unique capabilities. For instance, TransAgents demonstrated limitations with long texts, omitting significant portions of source content (Wu et al. 2025). In contrast, LLM-capability-driven approaches — designed around LLMs' computational strengths rather than human role divisions — remain largely unexplored but offer promising directions for developing novel translation systems (Becker 2024).

Research on LLM-based translation workflows remains nascent, with significant gaps concerning optimal agent configurations and information flow between agents. Preliminary findings suggest that different communication protocols affect both output quality and computational efficiency (Becker 2024; Q. Wang et al. 2024). This study addresses these gaps by developing and evaluating two distinct multi-agent systems: a practice-derived workflow modeled on professional translation practice (PD-MAS) and an LLM-capability-driven system (LCD-MAS) featuring more granular task decomposition. This approach enables a systematic comparison between human-mimicking and LLM-capability-driven workflows for literary translation, thereby advancing our understanding of how multi-agent translation systems can be optimized to meet the distinctive challenges of literary texts.

3. Methods

3.1 Multi-agent translation systems

3.1.1 *Practice-derived multi-agent translation system (PD-MAS)*

The PD-MAS implements a workflow aligned with the ISO 17100:2015 (International Organization for Standardization 2015) translation service requirements. We designed agent profiles to reflect industry standards, enabling direct comparison with professional human workflows.

The system operates through two sequential stages: pre-production and production (Figure 1). In the pre-production stage, two specialized agents prepare essential resources: the text analyst analyzes source text characteristics (genre,

© 2026. John Benjamins Publishing Company
All rights reserved

domain, purpose, and stylistic features), while the term expert creates bilingual terminology lists for consistency. In this literary context, we use 'terminology' operationally to include any recurring element requiring consistent translation. This is particularly important for handling character names, place names, and recurring motifs, which have been identified as a key challenge for literary translation quality (Matusov 2019).

The production stage encompasses translation and quality assurance. The translator generates target text guided by the analysis and terminology resources, applying criteria such as terminological consistency, genre appropriateness, and cultural adaptation. After self-checking, the translator passes the text to the reviser, who conducts comparative analysis between source and target texts, focusing on accuracy and completeness. The reviewer then ensures linguistic and stylistic coherence before the proofreader performs the final quality check.

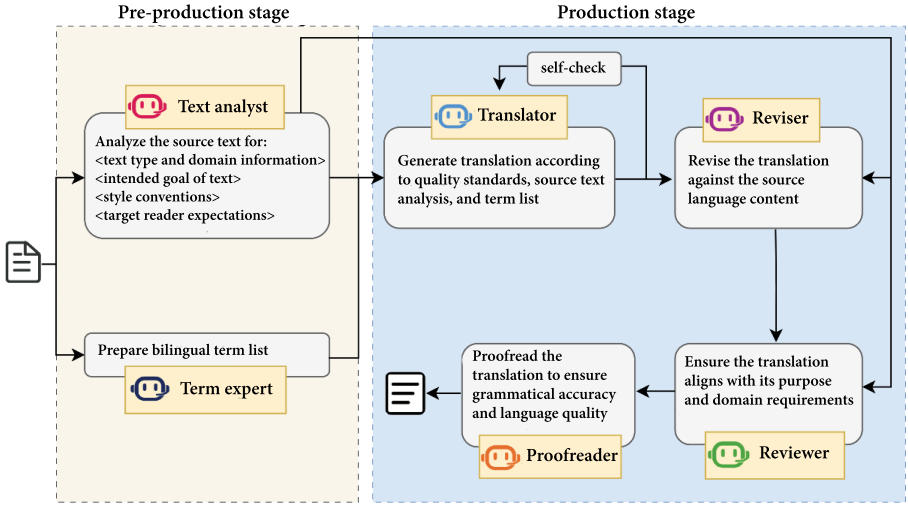

Figure 1. Practice-derived multi-agent workflow

We structured the agent instructions as itemized lists rather than prose descriptions to optimize LLM performance, following recommended prompt engineering practices (Phoenix and Taylor 2024). The workflow progresses through defined stages of preparation, translation, revision, and review, enabling assessment of whether practice-derived translation processes remain effective when implemented through LLM-based agents.

© 2026. John Benjamins Publishing Company
All rights reserved

3.1.2 *LLM-capability-driven multi-agent translation system (LCD-MAS)*

The LCD-MAS was designed around the computational characteristics of large language models, featuring granular task decomposition and dedicated stylistic processing (Figure 2).

A key challenge in the translation of long texts is LLMs' context limitations. Despite impressive technical context windows (e.g., GPT-4o's 128K tokens), models show degraded performance in coherence, instruction-following, and accuracy at much shorter context lengths (Hsieh et al. 2024; Levy, Jacoby, and Goldberg 2024). Levy, Jacoby, and Goldberg (2024) report that the reasoning accuracy of LLMs, including GPT-4, declines gradually as input length increases, with measurable degradation even at around 3000 tokens. Such performance deterioration poses consistency challenges for LLM-based translation systems (Liang et al. 2024).

This system addresses context window limits by dividing source texts into semantically coherent units of approximately 300 Chinese characters each before the translation stage. To counterbalance potential loss of global context caused by source text chunking, we implemented specialized mechanisms at pre-translation and finalization stages.

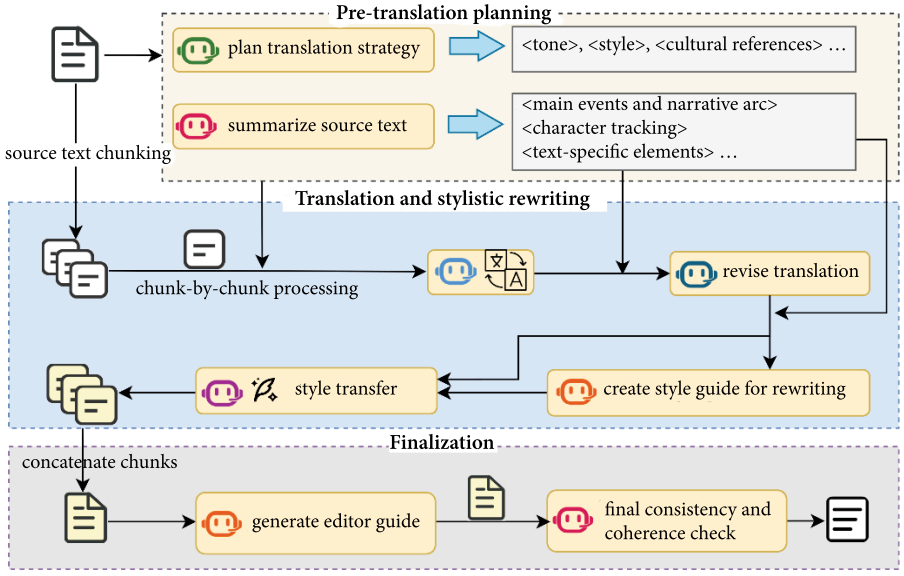

Figure 2. LLM-capability-driven multi-agent workflow

© 2026. John Benjamins Publishing Company
All rights reserved

The system operates through three interconnected stages. Pre-translation planning establishes global context through two agents: the summarizer generates a concise narrative summary capturing main events, characters, and themes, and the strategy planner develops a comprehensive translation plan addressing audience expectations, text type, and cultural references.

Translation and stylistic rewriting separate semantic transfer from stylistic refinement. For each source chunk, the translator produces an initial translation using the summary and strategy plan from the pre-translation stage. The reviser checks for accuracy, and then the style guide generator identifies appropriate stylistic enhancements. Finally, the stylistic rewriter applies these recommendations, incorporating literary devices while preserving semantic content. This pipeline addresses known limitations of fluency and style in LLM translations (He 2024; Jiang et al. 2024; R. Zhang, Zhao, and Eger 2025). It breaks down the translation process into two phases: an initial interlingual translation that conveys the semantic content, followed by an intralingual translation (i.e., stylistic rewriting within the target language) (Jakobson 1959; Whyatt 2017), where the stylistic rewriter enhances literary expression. By leveraging LLMs' strengths in text style unbundling (Phoenix and Taylor 2024) and text style transfer (Reif et al. 2022; Tao et al. 2024), this design aims to achieve stylistic improvement without compromising semantic fidelity.

Finalization ensures coherence across independently translated segments, which are concatenated at this stage. After text concatenation, the style guide generator detects inconsistencies and awkward transitions and produces guidelines for the editor to implement, maintaining global coherence while preserving the established stylistic qualities.

This architecture reconfigures the translation process around LLMs' computational characteristics rather than human cognitive patterns, combining global context-setting, granular task division, dedicated stylistic processing, and systematic finalization.

3.2 Materials

This study evaluated translation quality using a corpus of contemporary Chinese fiction with existing professional English translations. We selected twenty-eight chapters from fifteen works by fourteen prominent Chinese authors, including Nobel laureate Mo Yan (e.g., 天堂蒜薹之歌 *Tiantang suantai zhi ge*, translated by Howard Goldblatt as *The Garlic Ballads*; see Appendix A for the complete list). These established translations served as benchmarks for evaluating whether machine-generated translations could match or exceed professional human translation quality in literary contexts.

© 2026. John Benjamins Publishing Company
All rights reserved

The corpus included diverse subgenres to ensure broad representativeness: general fiction, mystery and detective fiction, science fiction, romance, and 仙侠 *xianxia* (a genre featuring cultivation and martial arts elements). To capture potential variations across narrative progression, we selected chapters from the beginning, middle, and end of each work.

To control for text length effects, we standardized chapters to approximately 3000 Chinese characters, with longer chapters truncated at narrative breaks and shorter ones supplemented with adjacent content. This standardization ensured comparable processing conditions across all texts.

3.3 Technical implementation

Both multi-agent systems were implemented using OpenAI's GPT-4o (version gpt-4o-2024-11-20) via Microsoft Azure OpenAI API, with Python 3.12.0 as the development environment. All twenty-eight source chapters were processed on 1 January 2025, ensuring consistent model performance across the evaluation corpus. This setup provided a controlled experimental environment where workflow architecture, rather than model capability, served as the independent variable.

Temperature settings were strategically configured based on agent function across both systems. Temperature is a parameter that controls the randomness of model outputs, where lower values produce more deterministic results and higher values allow for more variation. Analytical agents (text analyst, term expert, summarizer, and editor) operated at a temperature of 0 to produce deterministic outputs with high consistency. Translation and stylistic agents operated at a temperature of 0.5, balancing creative language generation with semantic fidelity. The top_p parameter, which controls the range of vocabulary the model considers when generating text, remained at its default value. The max_tokens parameter was left unrestricted to avoid artificial truncation of outputs.

Agent interactions were coordinated through an orchestration layer that maintained contextual continuity across processing stages, allowing outputs from earlier phases to be seamlessly incorporated into subsequent ones. This implementation ensured that any observed differences in translation quality could be attributed to workflow design rather than technical variables.

3.4 Text selection and quality assessment framework

In the evaluation phase, we extracted thirty text samples from the translated chapters, carefully selecting passages that represented both narrative and dialogue elements from the beginning, middle, and end sections of the source texts to ensure comprehensive coverage of the stylistic and register variations typical in literary

© 2026. John Benjamins Publishing Company
All rights reserved

fiction (Egbert and Mahlberg 2020; Chou and K. Liu 2024). Table 1 presents summary statistics showing length distributions across the source texts and all three translation versions. As Figure 3 shows, LCD-MAS produced generally longer translations than both human translators and PD-MAS — a pattern examined in our discussion of stylistic tendencies.

Table 1. Summary statistics for sample lengths ($N=30$) (Chinese source texts in characters; English translations in words)

Source and target texts	Min	Max	Median	Mean	SD
Chinese source texts	106	360	163	166	55.2
PD-MAS translations	36	186	94	99	42.8
LCD-MAS translations	81	313	143	149.5	58.1
Human translations	65	204	112	113.4	30.5

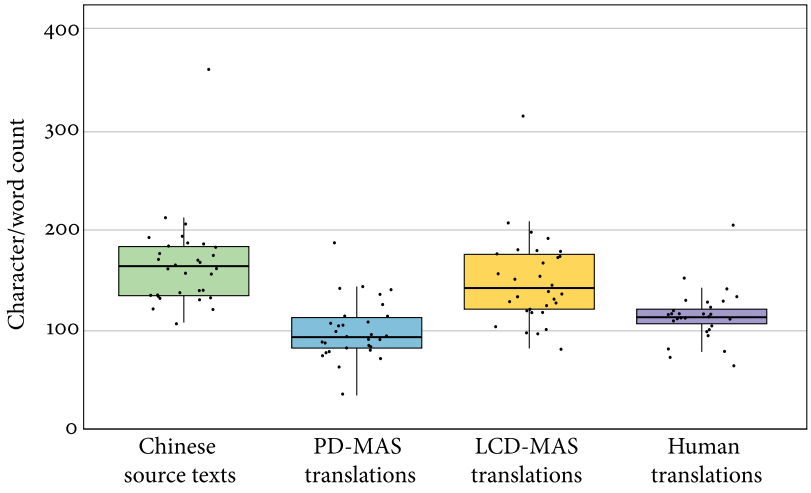

Figure 3. Distribution of sample lengths ($N=30$) (Chinese source texts in characters; English translations in words)

The thirty text samples were evaluated along two key dimensions: accuracy (faithful conveyance of source text meaning) and fluency (naturalness and adherence to target language norms) (Castilho et al. 2018; Salmi 2020).

© 2026. John Benjamins Publishing Company
All rights reserved

Four expert raters with complementary expertise conducted the evaluations. Two native English speakers with extensive experience teaching English writing assessed fluency (Raters 1 and 2), while two professors of translation from Chinese universities, each with over ten years of experience teaching literary translation, evaluated accuracy (Raters 3 and 4). We adapted Waddington's (2001) five-level scoring rubric to better suit professional translation evaluation, imposing stricter requirements for higher scores (see Appendix B). Raters used a 1–10 scale with whole-number increments to enhance scoring reliability.

To ensure consistent application of assessment criteria, all raters participated in training and calibration sessions. The evaluation employed a blind design. Each rater received a document containing the source texts alongside three anonymized and randomized translations (produced by human translators and the two multi-agent systems). Raters were not informed that any translations were machine-generated. In addition to assigning numerical scores, raters selected their preferred translation for each sample and provided written comments explaining their evaluations, considering factors including accuracy, fluency, stylistic appropriateness, creativity, and any other aspects they deemed relevant to translation quality.

This framework allowed assessment of technical quality through numerical ratings and of subjective reception through preference votes and qualitative feedback, providing a comprehensive view of how different translation approaches performed on literary texts.

4. Results

Our comparison of PD-MAS and LCD-MAS translations with professional human translations revealed distinct patterns in quality and reception. The evaluation examined three dimensions: accuracy (semantic fidelity to source texts), fluency (naturalness and readability in the target language), and overall preference as determined by expert raters. Statistical analyses for each dimension, complemented by qualitative assessments of translation characteristics, revealed both expected and unexpected patterns across the three translation approaches.

4.1 Accuracy analysis

We first assessed the consistency of accuracy evaluations using the intra-class correlation coefficient (ICC). A two-way random-effects model for average ratings, ICC(2,k), showed good interrater reliability between the two translation experts (ICC = .74, 95% CI [.52, .85], $F(89, 89) = 4.50$, $p < .001$), indicating consistent application of the evaluation criteria.

© 2026. John Benjamins Publishing Company
All rights reserved

Statistical comparisons of average accuracy scores across the three translation approaches were conducted using non-parametric tests, since the Shapiro-Wilk test indicated non-normal distributions for all three groups. The Friedman test showed no statistically significant differences in accuracy among professional human translations, PD-MAS translations, and LCD-MAS translations ($\chi^2(2) = 0.37$, $p = .832$).

Follow-up pairwise comparisons using Bonferroni-corrected Wilcoxon signed-rank tests confirmed this result. Median accuracy scores were identical across all three approaches ($Mdn = 8$), with no significant differences detected between any pair: LCD-MAS versus PD-MAS ($p = .739$, $r = .06$), LCD-MAS versus human translators ($p = .920$, $r = .02$), and PD-MAS versus human translators ($p = .837$, $r = .04$).

4.2 Fluency analysis

Fluency evaluations showed high consistency between raters, with interrater reliability analysis yielding an ICC of 0.83 (95% CI [.74, .89], $F(89, 89) = 6.30$, $p < .001$). This indicates strong agreement between the two native English-speaking evaluators in their assessment of how naturally the translations read in English.

In contrast to accuracy scores, fluency ratings revealed significant differences across translation approaches. A Friedman test indicated statistically significant variation among the three translation types ($\chi^2(2) = 14.92$, $p < .001$). To identify specific differences, we conducted pairwise comparisons using Wilcoxon signed-rank tests. The p-values were Bonferroni-corrected for multiple comparisons.

These pairwise tests showed that LCD-MAS received significantly higher fluency scores ($Mdn = 8$) than both human translators ($Mdn = 7$, $p < .001$, $r = .71$) and PD-MAS ($Mdn = 7.5$, $p = .024$, $r = .49$). The effect size for LCD-MAS versus human translations ($r = .71$) indicates a large effect. No significant difference was observed between human and PD-MAS translations ($p = .164$, $r = .35$).

Figure 4 presents a comparison of both accuracy and fluency scores across all three translation approaches. While the accuracy distribution shows identical median scores ($Mdn = 8$), the box plots reveal that LCD-MAS achieved significantly higher fluency ratings than both human translators and PD-MAS.

© 2026. John Benjamins Publishing Company
All rights reserved

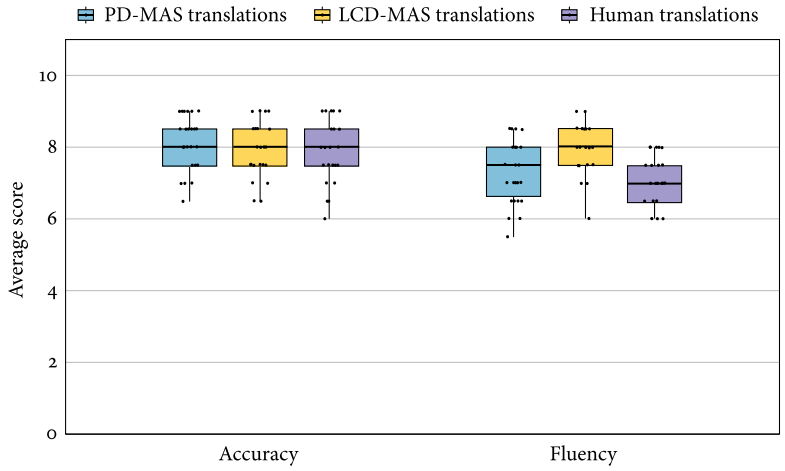

Figure 4. Average accuracy and fluency scores of LLM-based multi-agent and human translations

4.3 Translation preference analysis

Beyond numerical ratings, we examined overall translation preferences through direct comparison. Raters selected their preferred translation for each sample, yielding clear preference patterns across the 120 total evaluations (30 samples × 4 raters). LCD-MAS emerged as the most frequently preferred translation approach, receiving fifty-two votes (43.33%), followed by PD-MAS with thirty-nine votes (32.50%). Professional human translations were least preferred, with only twenty-nine votes (24.17%) (see Table 2).

Table 2. Rater preferences for translation approaches ($N = 120$)

Translator	Rater 1	Rater 2	Rater 3	Rater 4	Total votes	Percentage
PD-MAS	9	11	7	12	39	32.50
LCD-MAS	14	10	15	13	52	43.33
Humans	7	9	8	5	29	24.17

Although LCD-MAS received the highest number of preference votes overall, individual rater preferences showed some variation. Three of the four raters consistently preferred LCD-MAS, while one rater (Rater 2) slightly favored PD-MAS. This variation suggests that despite the overall preference trend, translation quality assessment remains somewhat subjective, with different evaluators prioritizing different aspects of translation.

© 2026. John Benjamins Publishing Company
All rights reserved

The preference data align with the fluency results reported in Section 4.2, indicating that fluency may have exerted a stronger influence on overall preference than accuracy. This relationship is particularly noteworthy given that no significant differences were found in accuracy scores across the three translation approaches, while LCD-MAS demonstrated significantly higher fluency.

Raters' written justifications for their preferences revealed distinctive characteristics associated with each translation approach. These qualitative insights are examined in detail in Section 5, where we analyze how specific translation qualities influenced overall preference patterns and what this suggests about effective translation process design.

5. Discussion

Our comparison of practice-derived and LLM-capability-driven translation systems reveals insights that challenge conventional assumptions about literary translation. This section interprets the observed performance patterns, analyzes distinctive stylistic characteristics, discusses persistent challenges in cultural translation, and considers broader implications for translation practice and technology.

5.1 Performance patterns

The equivalence in accuracy scores across all three translation approaches challenges long-standing assumptions about literary translation requirements. The statistical parity between both AI systems and professional human translators suggests that well-designed multi-agent systems can effectively transfer semantic content from source to target language.

The superior fluency performance of LCD-MAS demonstrates how architectural design can influence translation quality beyond semantic fidelity. Its dedicated stylistic processing stage produced translations that raters consistently preferred over both PD-MAS outputs and professional human translations. This finding diverges from previous research suggesting that human translators retain significant advantages over LLMs in stylistic aspects (R. Zhang, Zhao, and Eger 2025). Raters' preference patterns align more closely with fluency than with accuracy, suggesting that they prioritized natural, engaging language over strict semantic equivalence when evaluating literary translation quality.

These findings indicate that translation workflow design significantly impacts output quality. The LCD-MAS architecture, which separates semantic transfer from stylistic refinement, appears particularly effective for literary translation,

© 2026. John Benjamins Publishing Company
All rights reserved

where both meaning and artistic expression matter. This challenges the conventional integrated approach where translators often address meaning and style simultaneously.

5.2 Stylistic profiles and translation quality

The quantitative performance metrics reveal only part of the story. Analysis of raters' qualitative feedback uncovered distinct stylistic profiles associated with each translation approach, which significantly influenced quality assessments and raters' preferences. These stylistic differences emerged consistently across the corpus, revealing characteristic approaches to literary translation that extended beyond individual texts.

LCD-MAS's translations displayed a distinctive literary quality characterized by rich and vivid imagery, figurative language, and dynamic expression. Example (1) illustrates this stylistic profile.

(1) ST: 来日的晨时，秋天里的晨时。晨时里的日光，在豫东的平原上，因着晨时，它就血成一团，漫天漫地红着。铺红着，就有了这一天的晨时。
lai ri de chen shi qiu tian li de chen shi chen shi li de ri guang zai yu dong de ping yuan shang yin zhe chen shi ta jiu xue cheng yi tuan man tian man di hong zhe pu hong zhe jiu you le zhe yi tian de chen shi
'The morning of the coming day, a morning in autumn. The sunlight in the morning, on the plains of eastern Henan, because of the morning, it congealed into a ball of blood, reddening the entire sky and earth. Spreading red, thus came this day's morning.' (L. Yan 2006, Chapter 3)

LCD-MAS: The next morning arrived, an autumn morning steeped in quiet beauty. The sunlight stretched across the eastern plains of Henan, painting the horizon in hues of glowing red, like molten fire. The sky and earth seemed to merge in a vivid crimson, as if the day itself was being born in flames. This fiery red heralded the start of a day that carried with it a fragile sense of hope.

PD-MAS: The coming morning marked the onset of autumn. The early sunlight spread across the plains of Eastern Henan, painting the sky and earth in vivid crimson hues. With this brilliant red, the day began.

Human translator: It is late autumn, the dawn of a new day. The sun rises above the East Henan plain. A blood-red ball turning the earth and sky a deep shade of crimson. As red unfurls, so follows morning. Another day begins. (L. Yan 2011, Chapter 3)

© 2026. John Benjamins Publishing Company
All rights reserved

The source text is highly poetic and atmospheric, characterized by rhythmic repetition (晨时 *chenshi* 'morning time'), which creates an almost incantatory effect. Its language is deeply descriptive and evocative, employing vivid imagery such as 血成一团 *xue cheng yi tuan* 'congealed like blood' and 漫天漫地红着 *man tian man di hong zhe* 'reddening the whole sky and earth'. This repetition and imagery work together to evoke a cyclical and ritualistic sense of dawn on the eastern plain.

The LCD-MAS translation amplifies this effect through vivid figurative language ("like molten fire," "born in flames") that heightens the text's evocative quality. Phrases such as "steeped in quiet beauty" and "a fragile sense of hope" enhance the atmosphere while remaining thematically consistent with the source text. In contrast, PD-MAS delivers a more direct and concise translation that conveys the core information efficiently but lacks the literary flair of the LCD-MAS version. The human translation employs a distinctive staccato and fragmented style, creating a different rhythm that diverges from the more poetic and incantatory mood of the original.

However, this literary embellishment occasionally led LCD-MAS to introduce elements absent from the source text, as Example (2) demonstrates.

(2) ST: 半个小时后，当他们轻易译出密电后，方才还是莫名无实的慌惶，顿时像剥掉了皮肉，露出血淋淋、狰狞的本质，把他们都吓瘫了。
ban ge xiao shi hou dang ta men qing yi yi chu mi dian hou fang cai hai shi mo ming wu shi de huang huang dun shi xiang bo diao le pi rou lu chu xue lin lin zheng ning de ben zhi ba ta men dou xia tan le
'Half an hour later, when they easily deciphered the encrypted message, the vague and groundless panic from just moments before suddenly seemed to have its skin and flesh stripped away, revealing its bloody, hideous true nature, leaving them all paralyzed with fear.' (Mai 2018, Chapter 1)

LCD-MAS: Half an hour later, as they effortlessly deciphered the encrypted message, the vague disquiet that had settled over them earlier took on a far more sinister shape. It was as if a mask had been torn away, exposing something raw, bloody, and grotesque beneath. The revelation struck with such force that it froze them in place, their minds paralyzed by a terror so overwhelming it seemed to drain the air from the room.

PD-MAS: Half an hour later, when they easily cracked the code, the vague unease they had felt earlier transformed into raw, visceral terror — an unrelenting nightmare that left them paralyzed with fear.

© 2026. John Benjamins Publishing Company
All rights reserved

Human translator: Half an hour later, when they had deciphered the dictionary message with perfect ease, that earlier sense of bemusement was replaced by an all-consuming, paralysing terror. It was as if they'd been flayed, as if they'd been stripped of their surface equilibrium and reduced to raw emotion. (Mai 2020, Chapter 1)

Here, the LCD-MAS translation develops the original metaphor ('skin and flesh being peeled off') into an extended series of images. While this creates dramatic tension, it adds elements not present in the source text, such as "drain the air from the room." This tendency toward embellishment sometimes crossed into over-translation, with raters describing such passages as "florid" or "superfluous."

PD-MAS consistently produced more concise translations that effectively conveyed the core meaning. Its approach prioritized accuracy and directness, often condensing source text information into efficient target language expressions. However, this conciseness occasionally led to reduced cohesion and stylistic nuance, with raters noting "fragmented syntax" and a "lack of cohesion."

Human translations exhibited yet another stylistic profile, characterized by accurate rendering of meaning with varying levels of fluency. While human translators generally captured cultural nuances effectively, their stylistic choices sometimes resulted in what raters described as "overly literal" renderings or "fragmented syntax." The human translation in Example (1) shows this tendency toward fragmentation, with short, choppy sentences that accurately convey content but can appear abrupt.

These stylistic profiles help explain why LCD-MAS received higher fluency scores and preference ratings despite all three approaches achieving comparable accuracy. Its emphasis on literary quality and engaging language appears to have resonated with evaluators, even when it occasionally expanded beyond the source text's literal meaning. This finding confirms that literary translation quality depends not only on semantic accuracy but also on the stylistic and affective impact of the target text.

However, the dominant criticism raised by raters against LCD-MAS warrants careful consideration. They observed that it systematically added content, from small descriptive details to entirely new information. This was perceived as its primary flaw, often sacrificing fidelity for a "dramatic" style criticized as "superfluous" and "unwarranted." Rater 4's comment that some passages read more like "transcreation" than translation highlights a key tension in our findings: while raters frequently preferred the more engaging prose, they simultaneously questioned its deviation from translation norms.

© 2026. John Benjamins Publishing Company
All rights reserved

This tendency toward embellishment raises questions about the boundaries of translation and the ethics of AI-mediated creativity. LCD-MAS's output, though successful by certain metrics, blurs the lines between translation, adaptation, and creative rewriting. Optimizing AI systems for stylistic effect may inadvertently privilege fluency over the preservation of authorial voice and cultural specificity, which is an especially delicate issue in literary contexts (Taivalkoski-Shilov 2019; Kenny and Winters 2020). When an LLM introduces its own metaphors or dramatic flourishes, it risks misrepresenting the original author's voice, style, and intended meaning, even if the translation output achieves stylistic appeal. It may also homogenize diverse authorial styles into a recognizable "AI voice," inadvertently erasing the very cultural and stylistic nuances that make literary works unique. This aligns with concerns that technology could flatten diverse voices "to sound like one and the same person" (Taivalkoski-Shilov 2019, 697). Such embellishment also raises ethical concerns regarding readers who expect a faithful rendering of the original work (Taivalkoski-Shilov 2019).

Ultimately, the system's success in fluency and preference ratings highlights a promising direction for translation technology, but its content addition signals a departure from established translational ethics. The challenge lies in striking an appropriate balance between aesthetic effect and semantic fidelity in AI-assisted literary translation, while preserving authorial integrity and cultural authenticity.

5.3 Challenges in translating cultural references

Despite the impressive performance of both multi-agent systems in overall accuracy and fluency, our analysis revealed persistent difficulties in translating culturally specific references. These challenges represent a significant limitation of current LLM-based approaches to literary translation.

Both multi-agent systems struggled with culturally bound expressions that require deep contextual understanding rather than linguistic knowledge alone. For example, when translating the temporal reference "用了两炷香的时间" *yong le liang zhu xiang de shijian* 'in the time it took to burn two incense sticks', both AI systems opted for literal renderings — "took the time of two incense sticks" and "took him two incense sticks' worth of time." While comprehensible, these translations fail to convey the idiomatic meaning readily understood by readers familiar with Chinese culture. The human translator appropriately rendered this as "took him two hours," demonstrating cultural competence beyond literal transfer.

Similar patterns emerged with titles and proper names. When translating "上神" *shangshen* 'high god/supreme deity' in "青丘的那位九尾狐的上神" *Qingqiu de na wei jiuwei hu de shangshen* 'that nine-tailed fox high god from Qingqiu', LCD-MAS produced "that Nine-Tailed Fox Shangshen from Qingqiu," while PD-

© 2026. John Benjamins Publishing Company
All rights reserved

MAS rendered it as "that Nine-Tailed Fox High God from Qingqiu." The human translation — "this Qingqiu goddess" — better conveys the meaning to English readers. Likewise, both systems translated "土司太太" *tusi taitai* 'chieftain's wife' literally ("Tusi Madam" and "Tusi's wife"), whereas the human translator used the more culturally appropriate "the chieftain's wife."

Notably, these culturally inappropriate translations did not stem from a lack of relevant knowledge. Our examination of agents' outputs revealed that the systems often recognized the cultural references but suggested suboptimal translation strategies. For instance, the strategy planner in LCD-MAS correctly identified "上神" *shangshen* 'high god/supreme deity' as referring to "hierarchical relationships in the celestial realm," yet explicitly recommended transliteration with explanatory notes, but these notes did not subsequently appear in the final translation.

This disconnect between cultural knowledge and translation execution points to a critical limitation in current multi-agent translation systems: while cultural information is available, it is not effectively incorporated into the final translation. Even when individual agents proposed appropriate strategies for handling cultural references, these were not consistently implemented in the translation pipeline. This challenge highlights the continuing importance of human expertise and suggests that fully automated literary translation still faces significant obstacles where cultural competence is required.

5.4 Implications for translation technology and practice

Our findings have important implications for translation technology development and professional practice. LCD-MAS's superior performance suggests that appropriately designed multi-agent architectures can produce high-quality literary translations that raters may prefer to human translations in certain aspects.

The effectiveness of separating semantic transfer from stylistic refinement demonstrates the value of workflow designs tailored to computational strengths rather than modeled on human cognitive processes. This architectural insight could guide future translation technology development toward specialized processing stages rather than end-to-end approaches.

The persistent difficulties observed in handling cultural references indicate that fully automated literary translation still faces challenges. The findings indicate that optimal approaches may involve human–AI collaboration rather than full automation, with human translators focusing on cultural adaptation while AI systems handle drafting and stylistic enhancement.

© 2026. John Benjamins Publishing Company
All rights reserved

For translation theory, our findings invite reconsideration of translation processes. The effectiveness of non-human workflow design, which breaks translation into specialized sub-tasks, challenges traditional models and opens new theoretical directions for translation process design. This perspective shifts translation from being viewed primarily as an individual cognitive activity to a collaborative, functionally distributed process, whether performed by humans or AI agents.

6. Conclusion

This study compared the performance of two multi-agent translation systems against professional human translations for literary texts. The findings demonstrate that LLM-based multi-agent systems can achieve accuracy comparable to human translators while potentially surpassing them in fluency and rater preference. The LLM-capability-driven system, designed around LLMs' computational capabilities rather than standard human practice, produced translations with enhanced literary quality and stylistic richness, though sometimes at the cost of introducing content absent from the source text. The human-practice-derived system generated more concise translations but often lacked cohesion and natural flow. Notably, both AI approaches struggled with cultural references despite demonstrating understanding of these elements, suggesting a gap between cultural knowledge and effective translation strategy implementation. These results challenge fundamental assumptions about literary translation requirements and indicate that rethinking translation workflows specifically for LLM capabilities can yield exceptional results in certain aspects of translation quality.

Our study has several limitations that should be acknowledged. First, we focused exclusively on Chinese-to-English translation with a single LLM (GPT-4o), limiting the generalizability of our findings to other language pairs and model architectures. Second, the evaluation was based on relatively short text samples rather than full-length novels, leaving questions about how these systems can maintain consistency across longer narratives. Third, our study evaluated each multi-agent system as a holistic unit and did not isolate the performance of individual agents within the pipeline. Finally, our evaluation, while incorporating both quantitative ratings and qualitative assessments from expert raters, still captures only certain dimensions of translation quality and may not fully represent how different audiences would perceive the translations.

Future research should explore a broader range of language pairs, text types, and LLM architectures to assess the generalizability of our findings. Developing methods to address the cultural reference challenges we identified represents

© 2026. John Benjamins Publishing Company
All rights reserved

a particularly important direction, perhaps through enhanced coordination between agents responsible for strategic planning and those implementing the translation. Studies examining longer texts or complete literary works would also help determine whether multi-agent systems can maintain consistency across book-length translations. Research into human–AI collaborative translation workflows that combine the stylistic strengths of LLM systems with human cultural expertise could lead to particularly productive approaches. Moreover, our study has a potential confounding variable in the design of the LLM-capability-driven system, as it simultaneously introduced text chunking and a more sophisticated agentic architecture. Consequently, our results cannot fully disentangle whether the observed improvements in translation quality stem from the granular processing of smaller text units, the specialized multi-agent architecture, or their synergistic effect. Future research should aim to isolate these variables to determine their independent contributions. These directions can further expand our understanding of how LLM-based systems can contribute to literary translation while addressing their current limitations. By reimagining translation processes around the capabilities of advanced language models rather than simply replicating human workflows, researchers and developers can continue to push the boundaries of what MT can achieve in even the most challenging domains.

Funding

This work was supported by funding from The Hong Kong Polytechnic University (P0046370; P0051009).
Open Access publication of this article was funded through a Transformative Agreement with Hong Kong Polytechnic University.

Acknowledgements

The authors thank the reviewers and editors for their constructive comments, which greatly improved the quality of this paper. The first author would also like to thank Professors Ricardo Muñoz Martín, Bogusława M. Whyatt, Joss Moorkens, and Christopher D. Mellinger for their valuable input during the individual tutorial sessions at the MC2 Lab's 3rd International Summer School on Cognitive Translation & Interpreting Studies in July 2025.

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, et al. 2023. "GPT-4 Technical Report." *arXiv* preprint, arXiv:2303.08774v6. Accessed 1 August 2024.

© 2026. John Benjamins Publishing Company
All rights reserved

Becker, Jonas. 2024. "Multi-Agent Large Language Models for Conversational Task-Solving." *arXiv* preprint, arXiv:2410.22932v2. Accessed 27 November 2024.

Besacier, Laurent, and Lane Schwartz. 2015. "Automated Translation of a Literary Work: A Pilot Study." In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, edited by Anna Feldman, Anna Kazantseva, Stan Szpakowicz, and Corina Koolen, 114–122. Denver, CO: Association for Computational Linguistics.

Boase-Beier, Jean. 2014. *Stylistic Approaches to Translation*. Abingdon: Routledge.

Briva-Iglesias, Vicent, Joao Lucas Cavalheiro Camargo, and Gokhan Dogru. 2024. "Large Language Models 'ad referendum': How Good are They at Machine Translation in the Legal Domain?" *MonTI* 16: 75–107.

Brumme, Jenny, and Anna Espunya. 2012. "Background and Justification: Research into Fictional Orality and Its Translation." In *The Translation of Fictive Dialogue*, edited by Jenny Brumme and Anna Espunya, 7–31. Leiden: Brill.

Carl, Michael, and Moritz J. Schaeffer. 2017. "Models of the Translation Process." In *The Handbook of Translation and Cognition*, edited by John W. Schwieter and Li Wei, 50–70. Hoboken, NJ: John Wiley & Sons.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. "Approaches to Human and Machine Translation Quality Assessment." In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 9–38. Cham: Springer.

Chan, Chi-Min, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. "ChatEval: Towards Better LLM-Based Evaluators Through Multi-Agent Debate." *arXiv* preprint, arXiv:2308.07201v1. Accessed 1 June 2024.

Chou, Isabelle, and Kanglong Liu. 2024. "Style in Speech and Narration of Two English Translations of Hongloumeng: A Corpus-Based Multidimensional Study." *Target* 36 (1): 76–111.

Dorri, Ali, Salil S. Kanhere, and Raja Jurdak. 2018. "Multi-Agent Systems: A Survey." *IEEE Access* 6: 28573–28593.

Egbert, Jesse, and Michaela Mahlberg. 2020. "Fiction: One Register or Two? Speech and Narration in Novels." *Register Studies* 2 (1): 72–101.

Ehrensberger-Dow, Maureen, and Gary Massey. 2014. "Translators and Machines: Working Together." In *Proceedings of the XXth World Congress of the International Federation of Translators (Volume I)*, edited by Wolfram Baur, Brigitte Eichner, Sylvia Kalina, Norma Keßler, Felix Mayer, and Jeanette Ørsted, 199–207. Berlin: BDÜ.

Elshin, Denis, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, and Ekaterina Enikeeva. 2024. "From General LLM to Translation: How We Dramatically Improve Translation Quality Using Human Evaluation Data for LLM Finetuning." In *Proceedings of the Ninth Conference on Machine Translation*, edited by Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, 247–252. Miami, FL: Association for Computational Linguistics.

Fu, Linling, and Lei Liu. 2024. "What Are the Differences? A Comparative Study of Generative Artificial Intelligence Translation and Human Translation of Scientific Texts." *Humanities and Social Sciences Communications* 11 (1): 1–12.

© 2026. John Benjamins Publishing Company
All rights reserved

Fonteyne, Margot, Arda Tezcan, and Lieve Macken. 2020. "Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, et al., 3790–3798. Marseille: European Language Resources Association.

Gao, Ruiyao, Yumeng Lin, Nan Zhao, and Zhenguang G. Cai. 2024. "Machine Translation of Chinese Classical Poetry: A Comparison Among ChatGPT, Google Translate, and DeepL Translator." *Humanities and Social Sciences Communications* 11 (1): 1–10.

Guerberof-Arenas, Ana, and Antonio Toral. 2022. "Creativity in Translation: Machine Translation as a Constraint for Literary Texts." *Translation Spaces* 11 (2): 184–212.

Guo, Taicheng, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. "Large Language Model Based Multi-Agents: A Survey of Progress and Challenges." In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, edited by Kate Larson, 8048–8057. Darmstadt: International Joint Conferences on Artificial Intelligence.

He, Sui. 2024. "Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts." In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, edited by Carolina Scarton, Charlotte Prescott, Chris Bayliss, et al., 316–326. Sheffield: European Association for Machine Translation.

House, Juliane. 2015. *Translation Quality Assessment: Past and Present*. Abingdon: Routledge.

Hsieh, Cheng-Ping, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. "RULER: What's the Real Context Size of Your Long-Context Language Models?" *arXiv* preprint, arXiv:2404.06654v3. Accessed 1 September 2024.

Hvelplund, Kristian Tangsgaard. 2011. *Allocation of Cognitive Resources in Translation: An Eye-Tracking and Key-Logging Study*. PhD diss. Copenhagen Business School.

International Organization for Standardization. 2015. *ISO 17100:2015: Translation Services — Requirements for Translation Services*. Geneva: ISO.

Jakobsen, Arnt Lykke. 2002. "Translation Drafting by Professional Translators and by Translation Students." *Copenhagen Studies in Language* 27: 191–204.

Jakobson, Roman. 1959. "On Linguistic Aspects of Translation." In *On Translation*, edited by Reuben Arthur Brower, 232–239. Cambridge: Harvard University Press.

Jiang, Zhaokun, Qianxi Lv, Ziyin Zhang, and Lei Lei. 2024. "Convergences and Divergences Between Automatic Assessment and Human Evaluation: Insights from Comparing ChatGPT-Generated Translation and Neural Machine Translation." *arXiv* preprint, arXiv:2401.05176v3. Accessed 1 November 2024.

Jones, Francis R. 2019. "Literary Translation." In *Routledge Encyclopedia of Translation Studies*, 3rd ed., edited by Mona Baker and Gabriela Saldanha, 294–299. Abingdon: Routledge.

Karpinska, Marzena, and Mohit Iyyer. 2023. "Large Language Models Effectively Leverage Document-Level Context for Literary Translation, but Critical Errors Persist." In *Proceedings of the Eighth Conference on Machine Translation*, edited by Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, 419–451. Singapore: Association for Computational Linguistics.

© 2026. John Benjamins Publishing Company
All rights reserved

Kenny, Dorothy, and Marion Winters. 2020. "Machine Translation, Ethics and the Literary Translator's Voice." *Translation Spaces* 9 (1): 123–149.

Kuznik, Anna, and Joan Miquel Verd. 2010. "Investigating Real Work Situations in Translation Agencies: Work Content and Its Components." *HERMES — Journal of Language and Communication in Business* 44: 25–43.

Levy, Mosh, Alon Jacoby, and Yoav Goldberg. 2024. "Same Task, More Tokens: The Impact of Input Length on the Reasoning Performance of Large Language Models." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Lun-Wei Ku, Andre Martins, and Vivek Srikumar, 15339–15353. Bangkok: Association for Computational Linguistics.

Liang, Tian, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 17889–17904. Miami, FL: Association for Computational Linguistics.

Mai, Jia. 2018. 风声 [The message]. Beijing: October Arts and Literature Publishing House.

Mai, Jia. 2020. *The Message* [orig. 风声]. Translated by Olivia Milburn. London: Head of Zeus.

Matusov, Evgeny. 2019. "The Challenges of Using Neural Machine Translation for Literature." In *Proceedings of the Qualities of Literary Machine Translation*, edited by James Hadley, Maja Popović, Haithem Afli, and Andy Way, 10–19. Dublin: European Association for Machine Translation.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. "Adaptive Machine Translation with Large Language Models." In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, edited by Mary Nurminen, Judith Brenner, Maarit Koponen, et al., 227–237. Tampere: European Association for Machine Translation.

Mossop, Brian. 2000. "The Workplace Procedures of Professional Translators." In *Translation in Context: Selected Papers from the EST Congress, Granada 1998*, edited by Andrew Chesterman, Natividad Gallardo San Salvador, and Yves Gambier, 39–48. Amsterdam: John Benjamins.

Muñoz Martín, Ricardo. 2016. "Reembedding Translation Process Research: An Introduction." In *Reembedding Translation Process Research*, edited by Ricardo Muñoz Martín, 1–20. Amsterdam: John Benjamins.

PACTE Group. 2017. "PACTE Translation Competence Model: A Holistic, Dynamic Model of Translation Competence." In *Researching Translation Competence by PACTE Group*, edited by A. Hurtado Albir, 35–42. Amsterdam: John Benjamins.

Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." In *UIST '23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, edited by Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche, 1–22. New York: Association for Computing Machinery.

© 2026. John Benjamins Publishing Company
All rights reserved

Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. "Towards Making the Most of ChatGPT for Machine Translation." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, edited by Houda Bouamor, Juan Pino, and Kalika Bali, 5622–5633. Singapore: Association for Computational Linguistics.

Phoenix, James, and Mike Taylor. 2024. *Prompt Engineering for Generative AI: Future-Proof Inputs for Reliable AI Outputs at Scale*. Sebastopol: O'Reilly.

Puppel, Melissa, and Claudine Borg. 2025. "Evaluating ChatGPT's Performance in Creative Text Translation for Communication: A Case Study from English into German." *Media and Intercultural Communication* 3 (1): 1–27.

Qian, Chen, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, et al. 2024. "ChatDev: Communicative Agents for Software Development." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Lun-Wei Ku, Andre Martins, and Vivek Srikumar, 15174–15186. Bangkok: Association for Computational Linguistics.

Reif, Emily, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. "A Recipe for Arbitrary Text Style Transfer with Large Language Models." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, edited by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, 837–848. Dublin: Association for Computational Linguistics.

Reiss, Katharina. 2000. *Translation Criticism — The Potentials and Limitations: Categories and Criteria for Translation Quality Assessment* [orig. *Möglichkeiten und Grenzen der Übersetzungskritik: Kategorien und Kriterien für eine sachgerechte Beurteilung von Übersetzungen*]. Translated by Erroll F. Rhodes. Manchester: St. Jerome.

Risku, Hanna. 2014. "Translation Process Research as Interaction Research: From Mental to Socio-Cognitive Processes." *MonTI* 7 (2): 331–353.

Salmi, Leena. 2020. "Fluency in Evaluating and Assessing Translations." In *Fluency in L2 Learning and Use*, edited by Pekka Lintunen, Maarit Mutta, and Pauliina Peltonen, 146–165. Bristol: Multilingual Matters.

Schaeffer, Moritz, and Michael Carl. 2013. "Shared Representations and the Translation Process: A Recursive Model." *Translation and Interpreting Studies* 8 (2): 169–190.

Taivalkoski-Shilov, Kristiina. 2019. "Ethical Issues Regarding Machine(-Assisted) Translation of Literary Texts." *Perspectives* 27 (5): 689–703.

Tao, Zhen, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. 2024. "CAT-LLM: Prompting Large Language Models with Text Style Definition for Chinese Article-Style Transfer." *arXiv* preprint, arXiv:2401.05707v1. Accessed 1 August 2024.

Toral, Antonio, and Andy Way. 2018. "What Level of Quality Can Neural Machine Translation Attain on Literary Text?" In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 263–287. Cham: Springer.

Waddington, Christopher. 2001. "Different Methods of Evaluating Student Translations: The Question of Validity." *Meta* 46 (2): 311–325.

Wang, Lei, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, and Yankai Lin. 2024. "A Survey on Large Language Model-Based Autonomous Agents." *Frontiers of Computer Science* 18 (6): 186345.

© 2026. John Benjamins Publishing Company
All rights reserved

Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. "Document-Level Machine Translation with Large Language Models." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, edited by Houda Bouamor, Juan Pino, and Kalika Bali, 16646–16661. Singapore: Association for Computational Linguistics.

Wang, Qineng, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. "Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key?" In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Lun-Wei Ku, Andre Martins, and Vivek Srikumar, 6106–6131. Bangkok: Association for Computational Linguistics.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35: 24824–24837.

Whyatt, Boguslawa. 2017. "Intralingual Translation." In *The Handbook of Translation and Cognition*, edited by John W. Schwieter and Li Wei, 176–192. Hoboken, NJ: John Wiley & Sons.

Wooldridge, Michael. 2009. *An Introduction to Multiagent Systems*. Hoboken, NJ: John Wiley & Sons.

Wu, Minghao, Jiahao Xu, and Longyue Wang. 2024. "TransAgents: Build Your Translation Company with Language Agents." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, edited by Delia Irazu Hernandez Farias, Tom Hope, and Manling Li, 131–141. Miami, FL: Association for Computational Linguistics.

Wu, Minghao, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. "(Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts." *Transactions of the Association for Computational Linguistics* 13: 901–922.

Yan, Jianhao, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. "GPT-4 vs. Human Translators: A Comprehensive Evaluation of Translation Quality Across Languages, Domains, and Expertise Levels." *arXiv* preprint, arXiv:2407.03658v1. Accessed 1 August 2024.

Yan, Lianke. 2006. 丁庄梦 [Dream of Ding Village]. Shanghai: Shanghai Literature and Art Publishing House.

Yan, Lianke. 2011. *Dream of Ding Village* [orig. 丁庄梦]. Translated by Cindy Carter. Melbourne: Text Publishing.

Youdale, Roy, Andrew Rothwell, and Andy Way. 2023. "Why More Literary Translators Should Embrace Translation Technology." *Revista Tradumàtica* 21: 87–102.

Zhang, Biao, Barry Haddow, and Alexandra Birch. 2023. "Prompting Large Language Model for Machine Translation: A Case Study." In *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, edited by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, 41092–41110. Honolulu: JMLR.org.

© 2026. John Benjamins Publishing Company
All rights reserved

Zhang, Ran, Wei Zhao, and Steffen Eger. 2025. "How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, edited by Luis Chiruzzo, Alan Ritter, and Lu Wang, 10961–10988. Albuquerque: Association for Computational Linguistics.

Appendix A. Sources of the text samples used for the experiment

Book title	Author	Publisher	Publication year	Chapter(s)	Translation title	Translator(s)	Publisher	Publication year
尘埃落定 *Chen'ai luoding* 'Dust settles'	Alai	People's Literature Publishing House	2012	3, 12	*Red Poppies*	Howard Goldblatt, Sylvia Li-Chun Lin	Houghton Mifflin Harcourt Publishing Company	2002
第七天 *Di qi tian* 'The seventh day'	Yu Hua	New Star Press	2013	1, 4	*The Seventh Day: A Novel*	Allan H. Barr	Pantheon Books	2015
丁庄梦 *Ding zhuang meng* 'Dream of Ding Village'	Yan Lianke	Shanghai Literature and Art Publishing House	2006	2, 3	*Dream of Ding Village*	Cindy Carter	Text Publishing	2011
我们家 *Women jia* 'Our family'	Yan Ge	Zhejiang Literature and Art Publishing House	2013	6	*The Chilli Bean Paste Clan: A Novel*	Nicky Harman	Balestier Press	2018
高兴 *Gaoxing* 'Happy'	Jia Pingwa	People's Literature Publishing House	2008	9, 10	*Happy Dreams*	Nicky Harman	AmazonCrossing	2017
天堂蒜薹之歌 *Tiantang suantai zhi ge* 'Song of garlic scapes in paradise'	Mo Yan	China Writers Publishing House	2012	9, 10	*The Garlic Ballads*	Howard Goldblatt	Arcade Publishing	2011
风声 *Feng sheng* 'The sound of wind'	Mai Jia	Beijing October Arts and Literature Publishing House	2018	1, 4	*The Message*	Olivia Milburn	Head of Zeus	2020
无证之罪 *Wu zheng zhi zui* 'Crime without evidence'	Zijin Chen	Hunan People's Publishing House	2014	1	*The Untouched Crime*	Michelle Deeter	AmazonCrossing	2016
北京折叠 *Beijing zhedie*	Hao Jingfang	Zhejiang Education	2023	2, 4	*Folding Beijing*	Ken Liu	Uncanny Magazine	2015

© 2026. John Benjamins Publishing Company
All rights reserved

Appendix A. *(continued)*

Book title	Author	Publisher	Publication year	Chapter(s)	Translation title	Translator(s)	Publisher	Publication year
'Folding Beijing'		Publishing House						
流浪地球 *Liulang diqiu* 'The wandering earth'	Liu Cixin	Changjiang Literature and Art Publishing House	2008	1	*The Wandering Earth*	Ken Liu, Elizabeth Hanlon, Zac Haluza, Adam Lanphier, and Holger Nahm	Head of Zeus	2017
三体 *San ti* 'The three-body [problem]'	Liu Cixin	Chongqing Publishing House	2016	21	*The Three-Body Problem*	Ken Liu	Head of Zeus	2015
荒潮 *Huang chao* 'Waste tide'	Chen Qiufan	Shanghai Literature and Art Publishing House	2019	3, 4	*Waste Tide*	Ken Liu	Tom Doherty Associates	2019
盗墓笔记1：七星鲁王宫 *Daomu biji 1: Qixing Lu wang gong* 'Tomb-robbing notes 1: Seven-star palace of King Lu'	Nanpai Sanshu	Shanghai Culture Publishing House	2011	2, 3, 8	*The Grave Robbers' Chronicles: Cavern of the Blood Zombies*	Kathy Mok	ThingsAsian Press	2011
我欲封天 *Wo yu feng tian* 'I shall seal the heavens'	Er Gen	21st Century Publishing Group	2015	1, 5	*I Shall Seal the Heavens*	Jeremy Bai	Wuxiaworld Publishing	2021
三生三世十里桃花 *Sansheng sanshi shili taohua* 'Three lifetimes, three worlds, ten miles of peach blossoms'	Tang Qi	Changjiang Publishing House	2016	2, 15, 16	*To the Sky Kingdom*	Poppy Toland	AmazonCrossing	2016

© 2026. John Benjamins Publishing Company
All rights reserved

Appendix B. Scoring rubric for translation quality evaluation

Level	Accuracy	Fluency	Score
Level 5	Complete transfer of source text information.	Translation reads like a piece originally written in English.	9–10
Level 4	Almost complete transfer; there may be one or two insignificant inaccuracies; some revision needed to reach professional standard.	Large sections read like a piece originally written in English, but minor lexical, grammatical, or spelling errors are present.	7–8
Level 3	General ideas of the source text are conveyed, but with a number of lapses in accuracy; considerable revision required to reach professional standard.	Certain parts read like a piece originally written in English, but others clearly read like a translation. A considerable number of errors are present.	5–6
Level 2	Transfer of content is undermined by serious inaccuracies; thorough revision required to reach professional standard.	Almost the entire text reads like a translation, with continual lexical, grammatical, or spelling errors.	3–4
Level 1	Transfer of content is totally inadequate; the translation is not worth revising.	Text reveals a total lack of ability to express ideas adequately in English.	1–2

Address for correspondence

Kanglong Liu
Department of Language Science and Technology
The Hong Kong Polytechnic University
HHB812c, Hung Hom Bay Campus
Hung Hom, Kowloon
Hong Kong

kl.liu@polyu.edu.hk
https://orcid.org/0000-0003-3962-2563

© 2026. John Benjamins Publishing Company
All rights reserved

Co-author information

Lulu Wang
The Hong Kong Polytechnic University
lulu99.wang@connect.polyu.hk

Xing Wang
Tencent
xingwsuda@gmail.com

Sanjun Sun
Beijing Foreign Studies University
sunsanjun@bfsu.edu.cn

Jinghang Gu
The Hong Kong Polytechnic University
jinghang.gu@polyu.edu.hk

Publication history

Date received: 28 March 2025
Date accepted: 20 April 2026
Published online: 1 June 2026

© 2026. John Benjamins Publishing Company
All rights reserved