Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-026-06868-y OPEN

Evaluating literary translation by large language models: a multidimensional quality assessment of Shen Congwen's *Border Town*

Wei Yang[1] & Mingxing Yang[1]✉

Large language models (LLMs) have exhibited remarkable abilities in understanding and generating human language, which is applied in transferring languages. However, the translation of literary works presents unique challenges. The translation quality of literary works generated by LLMs is yet to be explored and tested. Therefore, this study aims to evaluate the quality of translations produced by various LLMs in comparison to a well-established human-translated work. The famous Chinese literary work *Border Town* by Shen Congwen was selected as the source text. ChatGPT 4, ChatGPT 4o, WXYY 4.0 Turbo, and Gemini were adopted as the models to process the translation. Jeffrey Kinkley's translation was chosen as the human translation for comparison. This research employs Multi-dimensional Quality Metrics to evaluate translation quality by providing detailed error typologies. We focused on error analysis from three key dimensions of translation quality: accuracy, fidelity, and cultural appropriateness. The results showed that five types of errors were identified: mistranslation, omission, over-translation, cultural mistranslation, and discourse-level errors. Mistranslation has the top frequency in all models, omission occurs the most in Gemini, over translation and cultural mistranslation appear the most in GPT4. Discourse-level error occurred in WXYY 4.0 turbo the most. GPT-4o appears to yield comparatively higher translation quality under the MQM framework. The research reveals that literary translation by LLMs requires more specific training prompt strategies and human post-editing to improve its accuracy, fidelity, and cultural appropriateness.

[1] Moutai Institute, Zunyi, China. ✉email: yangmingxing@mtxy.edu.cn

Introduction

The rapid development of large language models (LLMs) such as OpenAI's GPT series, Google's Gemini, and the leading Chinese LLM Baidu's 文心一言 (WXYY, hereinafter referred to as WXYY) has revolutionized various natural language processing tasks, including machine translation (MT). These models have demonstrated remarkable abilities in understanding and generating human language, allowing them to excel in tasks like text generation, summarization, and especially translation (Jiao et al., 2023). The translation of literary works, however, presents unique challenges. Unlike technical or socialized translations, literary texts contain complex layers of meaning, style, tone, and cultural differences that are difficult to convey fully in another language (Ponzio, 2007; Mikoyan, 2019). For this reason, analyzing the translation quality of LLMs when applied to literary works offers a fascinating insight into their potential and limitations and fills the existing gap.

This study aims to evaluate the quality of translations produced by various LLMs in comparison to a well-established human translation. Specifically, we compare the translation of the Chinese work *Border Town* (Chinese title: 边城Bian Cheng) by renowned translator Kinkley with translations generated by LLMs such as ChatGPT4, GPT4o, Gemini, and WXYY. *Border Town*, written by Shen Congwen, is a classic Chinese novel that is known for its poetic language and deep cultural context. Unlike other modern Chinese novels, this one is widely studied for its depiction of rural Chinese customs, folk values, and regional identity. Translating this text poses a significant challenge, as it requires not only linguistic accuracy but also the preservation of cultural and stylistic elements unique to Chinese literature (Xu, 2012, 2019). Therefore, it is highly suitable for testing the limits of LLM translation performance in capturing literary subtleties. The contrast between human and machine translations will provide valuable insights into the current capabilities and limitations of LLMs in handling such a complex task.

Border Town's English translation has undergone multiple attempts. The most prominent and widely studied translation is Jeffrey Kinkley's version, which represents the most recent major translation effort. Kinkley is a well-known American scholar and translator. This version was originally published by Harper Collins Publishers in America in 2009. Kinkley is a specialist studying Shen Congwen and his works. Kinkley's translation is noted for its fidelity to Shen Congwen's poetic style and cultural subtleties. He approached the translation with the intent to maintain the emotional depth and cultural resonance of the original, while also making the text accessible to English-speaking readers. Scholars have praised Kinkley's ability to convey the novel's philosophical undertones and the pastoral beauty that is central to Shen's work (Xu, 2012; Xu and Yu, 2019). They point out that Kinkley's translation carefully balances the need for linguistic accuracy with an appreciation of the novel's aesthetic qualities, making it one of the most respected translations of *Border Town* and offering a strong benchmark for comparison.

In conducting this study, we focus on error analysis from several key dimensions of translation quality: accuracy, fidelity, and cultural appropriateness (Lommel, 2018). Accuracy refers to the model's ability to render the meaning of the source text (ST) correctly, without significant errors or omissions (Sun, 2015). Fidelity concerns the preservation of the author's intent, tone, and style, which is particularly important in literary translation (Thompson and Dooley, 2019). Finally, cultural appropriateness involves how well the translation conveys the cultural elements of the ST, an essential aspect when translating between languages as distant as Chinese and English.

This research employs Multidimensional Quality Metrics (MQM), a robust framework designed to evaluate translation quality by providing detailed error typologies that account for various aspects of translation performance. MQM has emerged as a standard in translation quality assessment, particularly for machine translation (MT) systems, due to its flexibility and fine-grained evaluation approach. The MQM framework gained prominence as it addressed the limitations of previous evaluation metrics, which focused primarily on lexical overlap but failed to capture more errors related to accuracy, style, and cultural adaptation. Unlike traditional metrics, MQM allows for the evaluation of translation output by categorizing errors across multiple dimensions (accuracy, style, etc.) and assigning different weights depending on the severity of the error, namely minor, major, and critical (Lommel, 2018). This comprehensive approach enables a more detailed analysis of machine translation outputs and improves alignment with human judgment.

Literature review

English translation of *Border Town*. Literary translation is a subfield of translation. It "represents a distinctive kind of translating because it is concerned with a distinctive kind of text" (Hermans, 2007, p. 78). Literary translation can be considered as a literary creation and re-creation. When translating a piece of literary work, the translator develops a unique artistic image using his deep knowledge of life and literature before expressing the image with the right words and sentences (Wang, 2014). Wright (2016) holds that the quality of literary translation relies on the fully-played function of translating a work into another form of literary work mediated for the target audience.

The translation studies on *Border Town* emphasize that in a narrative text, the construction of the characters' image must take precedence over the settings, stories, and plots. Hu (2019), in her thesis, examined the continuous rewriting of China's rural image in foreign countries. Ma (2022), in her submission, also discussed China's image from the perspective of imagology by emphasizing the translation of the folklore in the English translation of *Bian Cheng*. It was concluded that the translator constructed the Chinese images of high responsibility, enjoying harmony between humans and nature, and pursuing a happy life. Chen (2019) put an effort into the understanding of the translators' role in the translation by exemplifying the English translation of *Bian Cheng*. The researcher drew the concepts from psychoanalytic theories and put forward a framework of creative translation, positing that translation is a creative act and translators are creative agents. The author also agreed with Xu's (2012) classifying Kinkley as a scholarly translator.

Other studies are from different perspectives. Guo et al. (2020) studied the translator's voice through the translation of characters' names in *Bian Cheng*. A contrastive study of three English translations of *Bian Cheng* from the perspective of corpus-based critical translation studies was conducted by Liu (2023), examining the ideological factors embedded in these translations in terms of character designation, cultural expression, nominalization, passive structure, and modal verbs. The same perspective was taken by Liu (2019) on four English translations, and the focus was on the translators' style, through the length of sentences and richness of words, exploring the factors influencing their translation style and features.

It can be seen that the English translation of this work involves complicated factors and studies of it contribute to the knowledge body of literary translation across cultures. However, due to the complexity of literary translation, with the spreading application of LLMs, it is enlightening to evaluate the translation quality of it in the literary field.

LLMs and translation studies. Recent studies on LLMs have explored their potential in machine translation (MT), with

promising results but also notable challenges. Zhang et al. (2023) systematically analyzed the role of prompting strategies in improving translation quality, highlighting that factors like prompt design and the use of optimal examples significantly enhance performance. Similarly, Lu et al. (2023) demonstrated that error analysis prompting improves translation evaluation by generating human-like assessments in models like ChatGPT. These findings suggest that the translation abilities of LLMs can be fine-tuned through specific strategies (e.g., Xu et al. 2024, Li et al. 2024), yet challenges remain in achieving consistency, especially at the segment level. Kocmi and Federmann (2023) extended this analysis by showing that LLMs, particularly GPT4 and GPT4o, outperform traditional MT systems in evaluation tasks, especially at the system level, suggesting that these models can serve as robust evaluators of translation quality. However, the performance of LLMs in low-resource languages continues to lag behind, a limitation also observed by Zhu et al. (2023), who noted that while LLMs perform well in high-resource languages, they struggle with rare and complex language pairs.

Recent studies have started to address the application of LLMs in the domain of literary translation, although this area remains underexplored. Jiao et al. (2023) introduced ParroT, a framework that enhances translation by incorporating human feedback and instruction fine-tuning, particularly effective in mitigating common translation errors such as mistranslation and omission. This aligns with He et al. (2024), who explored human-like translation strategies using the Multi-Aspect Prompting and Selection (MAPS) framework, demonstrating that mimicking human translation processes can reduce errors such as hallucination and awkward phrasing. Their study found that some LLM outputs while demonstrate surface-level textual adequacy, the translations often lack cohesion and contextual appropriateness when examined in detail. Wang et al. (2023) in their study stated that LLMs like ChatGPT show potential as a new paradigm for document-level machine translation, outperforming commercial systems and demonstrating a stronger ability for discourse modeling. Zuo et al. (2024) examined the performance of ChatGPT in translating Thai literary texts into Chinese and observed that, despite producing syntactically and lexically acceptable outputs, the model struggled to accurately convey culturally embedded elements and rhetorical devices. Their analysis suggests that the lack of interpretive depth in machine-generated translations limits their applicability to texts characterized by rich stylistic and cultural features. Collectively, these studies provide a foundation for further empirical inquiry, particularly in evaluating model outputs using comprehensive quality assessment frameworks that extend beyond sentence-level fidelity. Despite these advances, the translation of literary texts remains particularly challenging for LLMs, and the number of studies in this field is limited. Literary works, with their intricate cultural and stylistic elements, require translations that go beyond linguistic accuracy. Therefore, while LLMs offer exciting possibilities for MT, especially in technical domains, their application in literary translation still requires further investigation.

MQM in translation evaluation. This study utilizes MQM, a robust framework designed to evaluate translation quality. Studies have applied the MQM framework to evaluate the performance of MT systems in various contexts. A notable application of MQM comes from Klubicka et al. (2018), who used it to compare different MT systems in English-to-Croatian translation. By adopting a fine-grained MQM-compliant error taxonomy tailored to Slavic languages, the study demonstrated that neural MT systems performed significantly better in reducing long-distance agreement errors, which were more common in phrase-based systems. This highlights the utility of MQM for identifying specific linguistic challenges in translations, providing insights that are often missed by automatic metrics (Klubicka et al., 2018).

Freitag et al. (2021) conducted a large-scale study using MQM to evaluate the outputs of top MT systems from the Workshop on Machine Translation (WMT) 2020 shared task. By annotating translation errors using professional translators with full document context, they found that MQM offered a more detailed understanding of translation quality than crowd-sourced evaluations. Their analysis revealed that neural machine translation systems outperformed traditional models, particularly in terms of fluency and long-distance error correction, although MQM also identified areas where even the best systems struggled, such as contextual consistency and idiomatic translation (Freitag et al., 2021).

Moreover, MQM has been instrumental in evaluating translations for linguistically diverse regions, such as India. Sai et al. (2023) introduced an MQM-based dataset to evaluate machine translation performance for Indian languages, showing that neural metrics like COMET, which integrate MQM annotations, significantly outperformed traditional n-gram-based metrics like Bilingual Evaluation Understudy (BLEU) in capturing translation nuances for morphologically rich languages. This study emphasizes MQM's adaptability to multilingual contexts and its ability to address the complexity of language-specific translation challenges (Sai et al., 2023).

In addition to its applications in MT evaluation, MQM has also been used to bridge the gap between human and machine translation evaluations. For example, a study by Fernandes et al. (2023) proposed AutoMQM, a system that leverages LLMs like PaLM-2 to perform automatic error detection based on MQM's error typologies. AutoMQM demonstrated that combining LLMs with MQM could provide fine-grained, human-aligned error spans for translation evaluation without relying on reference translations, significantly advancing the use of MQM in automated contexts (Fernandes et al., 2023). This work highlights MQM's capacity to be integrated with cutting-edge AI technologies to enhance translation quality assessment.

The MQM framework has established itself as a vital tool for assessing translation quality, offering a flexible approach to error categorization. Its fine-grained evaluation methodology has proven effective across diverse language pairs, translation systems, and even LLM-based translation models. With its capacity to integrate human and machine evaluations and adapt to specific linguistic challenges, MQM remains one of the most reliable frameworks for evaluating both human and machine translation outputs. However, studies on applying the MQM framework to LLM translation of literary works remain underexplored.

Based on the above discussion, this research posed two research questions:

(1) What are the errors in accuracy, fidelity, and cultural appropriateness identified in the LLM translations of *Border Town?*
(2) In what aspects do LLM translations deviate from the original text compared to the human translation?

Methodology

This research combines a qualitative approach with content analysis. This section outlines the methodology used to compare the translation quality of Shen Congwen's *Border Town* by LLMs and human translator Kinkley. It aims to evaluate which approach performs better in capturing the essence of literary

translation, with a focus on accuracy, fidelity, and cultural appropriateness.

Data selection. The text used for this study is Shen Congwen's *Border Town*, a widely regarded modern Chinese literature known for its rich cultural and emotional depth. This study adopts a case-based exploratory approach and does not intend to generalize to all forms of literary translation. Therefore, we selected the first two chapters of the text that are representative of the novel's linguistic and stylistic challenges. The two chapters amount to 6785 Chinese characters and 204 sentences. In the two chapters, the descriptive passages are rich in descriptions of nature and rural life, highlighting the use of poetic language and imagery, which can represent the overall linguistic style and cultural richness of the whole work. Furthermore, the rich narrative, stylistic complexity, and cultural content are representative of Shen Congwen's work.

The translations are divided into two categories: Large language model translation (LLMT) produced by advanced LLMs, specifically GPT4, GPT4o, Gemini, and WXYY Turbo 4.0, were used for comparison. WXYY Turbo 4.0 is a leading Chinese-developed model and serves as our representative of domestically trained LLMs. They were selected for their availability and prominence (at the time of experimentation, February–June 2024). All four LLMs used the same prompt: "You are a professional literary translator. Please translate the text into English." This prompt was deliberately chosen to serve as a baseline condition, allowing us to evaluate each model's translation competence without engineered guidance. The corresponding translations exhibit variation in length and sentence count across systems. Specifically, the WXYY Turbo 4.0 translation contains 5889 words and 306 sentences; the GPT4 translation contains 6292 words and 328 sentences; the GPT4o translation contains 6188 words and 312 sentences; and the Gemini translation contains 4850 words and 312 sentences. The existing published translation of *Border Town* by professional literary translator Jeffrey Kinkley was used as the benchmark of human translation (HT) and to compare the HT and LLMTs. His translation was chosen for its recognition and critical acclaim within the field of literary translation.

Coding and criteria. To conduct the error annotation, a rigorous and systematic process was followed to ensure reliability and consistency. First, the original Chinese text and the translated versions produced by the four LLMs (ChatGPT 4, ChatGPT 4o, WXYY 4.0 Turbo, and Gemini) were aligned at the sentence level with the original text using ChatGPT 4o. But ChatGPT 4o was only used to facilitate initial alignment. All alignments were manually reviewed and corrected by the annotation team to eliminate any model-induced bias in segmenting or matching sentences. This alignment enabled coders to evaluate each translation unit in direct comparison to the source. A shared annotation manual based on the MQM guidelines was developed and used to guide the annotation process.

Three major dimensions are assessed to determine the quality of each translation, focusing on the error analysis. Accuracy evaluates translating ST correctly, without significant errors or omissions. Fidelity concerns the preservation of the author's intent, tone, and style. Cultural appropriateness involves translating the cultural elements of the ST. Following the MQM framework, we categorized and quantified errors in each translation based on the aforementioned criteria (accuracy, fidelity, cultural appropriateness). Errors were marked as mistranslation, omission, cultural mistranslation, etc., based on their impact on the overall translation quality. Each error can be assigned a severity level (e.g., M1-minor, M2-major, C-critical). The severity level was defined following these rules: Minor (M1): Errors that have a limited impact on meaning or style. The sentence remains understandable and fluent, though slightly awkward or imprecise. Major (M2): Errors that significantly alter meaning, mislead readers, or disrupt the flow or tone of the translation. Critical (C): Errors that cause complete misinterpretation of the ST or result in highly misleading, incoherent, or culturally inappropriate output.

Data analysis. The analysis process is primarily descriptive and comparative. Based on the codes and to answer the first research question, we analyzed the errors presented in the LLMT from three dimensions: accuracy, fidelity, and cultural appropriateness. Examples were selected to further analyze and compare the LLMTs and HT. In the example tables, underlined segments indicate the focal error(s) discussed in the analysis. They do not necessarily represent all possible deviations within the sentence. Error identification was conducted with reference to the broader narrative context. Then, we discussed how the LLMT deviates from the original intent of the ST, and special attention was given to whether the LLMs can preserve these culture-specific elements compared to the HT.

Trustworthiness. Ensuring trustworthiness is crucial in this research to validate the findings. To enhance credibility and reduce subjective bias in evaluating literary translation quality, a structured and iterative annotation procedure was adopted. Two trained annotators with backgrounds in translation studies participated in the annotation process. Both annotators jointly covered all sentence-aligned units across the selected chapters of *Border Town*. Prior to full-scale annotation, calibration rounds were conducted on a subset of the data using a shared MQM-based annotation manual. These rounds aimed to align interpretations of error categories and severity levels, particularly for context-sensitive categories such as omission and discourse-level errors.

During the main annotation phase, the annotators first formed initial judgments with reference to the source text and the aligned translations. Discrepancies were then identified and addressed through iterative rounds of discussion and reconciliation. In cases where disagreements could not be readily resolved, a senior expert with over 30 years of experience in literary translation was consulted to guide the final decision-making process.

The annotation was not fully blind with respect to the source of the translations, which is acknowledged as a methodological limitation, as knowledge of translation authorship may influence evaluative perception. To mitigate this risk, severity judgments were not based on isolated sentences alone. Annotators consistently considered broader narrative, stylistic, and cultural context before assigning error types and severity levels, particularly for omissions and discourse-level phenomena with downstream narrative effects.

Because the annotation relied on iterative discussion and consensus-based reconciliation rather than retaining independent pre-consensus coding records, a formal inter-annotator agreement coefficient could not be validly calculated. This is recognized as a limitation of the study. However, given the interpretive and context-dependent nature of literary translation quality assessment, the consensus-based approach was considered more appropriate for standardizing category application and reducing individual cognitive bias. Finally, thick descriptions of the research context, including the literary features and cultural significance of Border Town, the analytical framework, and the annotation procedures, are provided to support transparency and

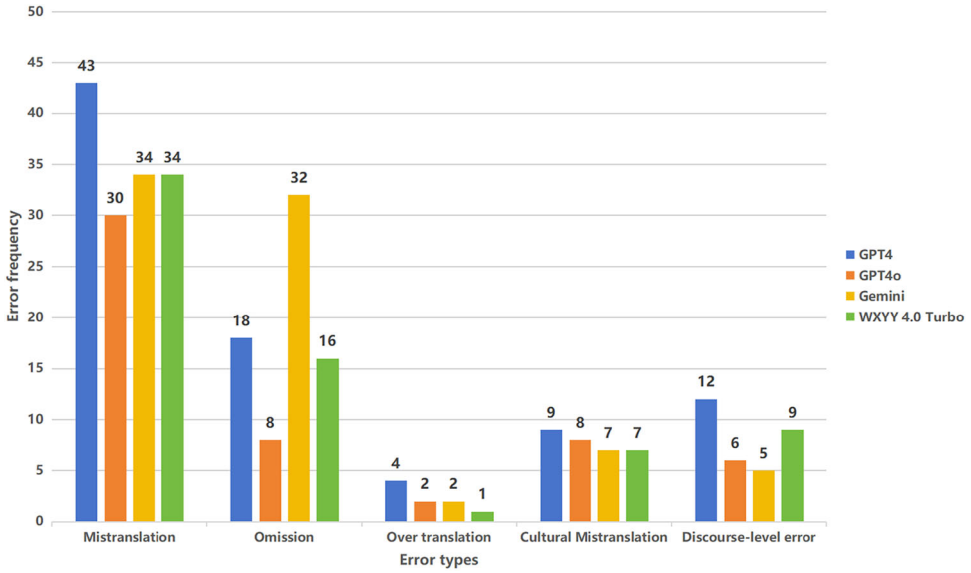

Fig. 1 Identified overall errors of LLMTs.

Table 1 Mistranslation frequency of LLMTs.

LLMT	Mistranslation
GPT4	43
GPT4o	30
Gemini	34
WXYY 4.0 Turbo	34

to enable readers to evaluate the applicability of the findings to other literary translation contexts involving LLMs.

Findings

In analyzing the performance of the four LLMTs - GPT4, GPT4o, Gemini, and WXYY 4.0 Turbo - in translating the first two chapters of *Border Town* by Shen Congwen, the data reveals distinct patterns across five types of errors: mistranslation, omission, over-translation, cultural mistranslation, and discourse-level errors, as shown in Fig. 1. By comparing the frequency of these errors, this analysis seeks to uncover each model's strengths and weaknesses in literary translation and to understand their implications for translation quality.

Mistranslation. As presented in Table 1, GPT4 exhibited the highest frequency of mistranslations (43) among all the TT of LLMTs, primarily at lexical levels. It suggests limitations in its ability to handle vocabulary specific to *Border Town*. This may indicate a lack of domain-specific training in this model. Mistranslations in literary works often change the original meaning and distort the intended accuracy, a significant drawback for high-quality translation.

In comparison, GPT4o, an optimized variant of GPT4, has the fewest mistranslation errors (30), showing the comparatively stronger performance among the four models. This suggests that specific optimizations in GPT4o, possibly through better parameter tuning, targeted training data, or task-specific fine-tuning, have enhanced its ability to handle the specifics in translation. Both Gemini and WXYY 4.0 Turbo exhibit the same number of mistranslation errors (34), suggesting that they may share similarities in training approach. However, their mistranslation rates are still higher than GPT4o, indicating room for improvement in accuracy. Previous research has shown that smaller or more optimized versions of LLMs, like GPT4o, can sometimes outperform larger models in specific tasks due to targeted optimization and tuning.

Example 1

ST: 我有了口粮, 三斗米, 七百钱, 够了。谁要这个!

In this example, exhibited in Table 2, "七百钱" specifically refers to "seven hundred copper coins," a traditional Chinese currency unit that underscores the historical and cultural setting of the novel. Maintaining this detail in translation is essential to preserve the historical background. In the GPT4 and GPT4o translations, "seven hundred coins" conveys the idea of money but fails to specify that these are copper coins, losing the material and traditional aspect of the currency, which is significant to the period. This lack of specificity results in minor inaccuracy.

The Gemini and WXYY 4.0 Turbo translations, however, use the term "cash," which in modern English often implies either paper currency or a general monetary value, potentially distorting the original meaning by erasing the material-specific reference. This choice risks misleading readers into perceiving a more contemporary form of wealth and shifts the setting, thus introducing critical inaccuracy. By contrast, the HT, "seven hundred coppers," conveys the essence of "七百钱" with accuracy, using "coppers" to reflect both the material and the historical context of the currency, preserving the social background of the original text.

Omission. Omission errors in literary translation can result in the loss of essential lexical expressions, plot details, or stylistic elements, thereby diminishing the richness of the source material. Gemini shows the highest omission frequency (32 occurrences), indicating a tendency to simplify or compress content, which may result in the loss of detail, atmosphere, and plot, thereby reducing fidelity to the ST. Particularly in some crucial paragraphs, Gemini omitted a relatively large portion of sentences.

GPT4 and WXYY 4.0 Turbo fall in between (with 18 and 16 occurrences, respectively), reflecting a balance but still room for improvement in preserving ST detail. As shown in Table 3, in contrast, GPT4o had the lowest omission rate (8), suggesting that this model preserved the content more effectively. This data suggests that GPT4o's optimizations may make it more reliable for literary translations in terms of minimizing omissions, while

Table 2 Example of mistranslation of LLMTs.

LLM	TT	Severity level	HT
GPT4	I have enough: three dou of rice, seven hundred coins. Who needs this!	M1	I'm paid for my work: three pecks of rice and seven hundred coppers. That's enough for me. Who needs this charity?
GPT4o	I already have my share, three dou of rice and seven hundred coins, that's enough. Who wants this!	M1	
Gemini	I have a fixed price, three dou of rice, seven hundred cash, that's enough. Who needs this!	C	
WXYY 4.0 Turbo	I have enough to eat, three dou of rice, seven hundred cash, that's enough. Who needs this!	C	

Table 3 Omission frequency of LLMTs.

LLMT	Omission
GPT4	18
GPT4o	8
Gemini	32
WXYY 4.0 Turbo	16

Gemini's high omission rate indicates challenges in maintaining textual completeness.

Example 2

ST: 气运好, 半年之内船不坏事, 于是他从所赚的钱上, 又讨了一个略有产业的白脸黑发小寡妇。

As illustrated in Table 4, this example illustrates an omission in the Gemini translation, where the descriptors "白脸" ("fair-faced" or "pale-faced") and "黑发" ("black-haired") are left out. The descriptors "白脸黑发" are essential as they foreshadow an upcoming detail about the widow's son, the main character Nuosong, who shares similar features. Omitting these descriptors in the Gemini translation disrupted the continuity and impacted readers' understanding of character relations. Thus, this omission is categorized as a critical error. In contrast, the HT retains "pretty and black-haired," aligning with the original text by including essential descriptors that are significant for character consistency later in the narrative.

Example 3

ST: 傩送美丽得很, 茶峒船家人拙于赞扬这种美丽, 只知道为他取出一个诨名为"岳云"。虽无什么人亲眼看到过岳云, 一般的印象, 却从戏台上小生岳云, 得来一个相近的神气。

In Table 5, the translations by Gemini and WXYY 4.0 Turbo differ in accuracy and cultural depth compared to HT's because of omission. The ST emphasizes Nuosong's beauty and the local boatmen's struggle to express it, leading them to nickname him "Yue Yun," a figure they associate with handsome opera characters as a cultural reference. Gemini's translation simplifies this depiction, omitting "虽无什么人亲眼看到过岳云"(no one has seen Yue Yun in person) and "得来一个相近的神气"(saw a resemblance to), which are significant because they add a layer of mystique and conveys the local residents' imaginative impression of the historical figure. WXYY's translation also omitted the same two sentences, reducing the richness of cultural elements embedded in the original and losing the subtle description of the ST. Consequently, both Gemini and WXYY 4.0 Turbo produce translations that are inaccurate and lack the cultural richness of the original, resulting in critical errors in accuracy, fidelity, and cultural appropriateness.

In contrast, HT excels in accuracy, cultural fidelity, and expressive detail. It not only mentions the Song Dynasty origin of "Yue Yun," adding historical context, but also conveys the villagers' imaginative perception derived from local opera, painting a vivid cultural landscape. The omitted sentences by Gemini and WXYY were transferred. HT translates the nickname as more than a mere label, giving readers insight into the cultural heritage and local storytelling practices. Overall, HT's translation surpasses the others by balancing literary quality with cultural nuance, aligning more closely with ST, and containing the emotional and cultural depth of the original text.

Over-translation. Over-translation occurs where the model adds extraneous information or repetitive descriptions of the original text that have already been translated loyally. However, in literary contexts, this often disrupts the fluency and surpasses the intended tone.

GPT4 has the highest frequency of over-translation, with four occurrences. WXYY 4.0 Turbo has the least over-translation error, as seen in Table 6. Gemini and GPT4o both showed a low frequency of over-translation (2). Compared to other errors, the over-translation error is the least among all the four LLMs, which may suggest that their optimization parameters might include mechanisms to prevent elaboration beyond the ST.

Example 4

ST: 女孩子的母亲, 老船夫的独生女, 十五年前同一个茶峒军人, 很秘密的背着那忠厚爸爸发生了暧昧关系。

In this example, as shown in Table 7, the ST does not necessarily make moral condemnation or emotional judgment. In Chinese, "暧昧关系" is an euphemism referring to an ambiguous or emotionally vague relationship, often suggestive but not explicitly romantic or sexual. Therefore, in the HT, Kinkley transferred it into "sung by each other," which conveys the implicitness and echos with the local culture, where the locals sing to each other to show their affection. However, in the TT, "a secret affair" implies a clear sexual relationship or extramarital affair, which is stronger and more direct than the original. Moreover, "背着" means "behind someone's back" or "without someone's knowledge." It implies secrecy but does not necessarily depict moral condemnation. In contrast, "betraying" carries a heavy emotional and moral weight, implying intentional harm or moral failure. They contribute to an over-translated tone, with overly moralized portrayal, distorting the narrative voice and character portrayal. Thus, this over-translation can be marked as "C", meaning critical error in translation.

Cultural mistranslation. The literary work *Border Town* by Shen Congwen is deeply rooted in Chinese cultural contexts. The rich cultural elements in describing characters and folk customs made this work unique and stylistic. This research differentiates mistranslation and cultural mistranslation because of their different focus. The former concentrates on mistranslation at lexical or syntactical levels without less culture-bound elements, while the latter focuses more on culturally charged elements.

In the four LLMs, as shown in Table 8, GPT4 showed the highest rate of cultural mistranslation errors (9), indicating its potential

Table 4 Example of omission in Gemini.

LLM	TT	Severity level	HT
Gemini	With good luck, his boat remained undamaged for half a year, and he used the earnings to marry a young widow with some property.	C	Luck was with him; the boat sailed safely, and in six months he'd saved money enough to marry a pretty, black-haired young widow.

Table 5 Example of omission in Gemini and WXYY 4.0 Turbo.

LLM	TT	Severity level	HT
Gemini	Nuansong was very handsome, and the people of Chadong, not skilled in describing beauty, simply called him "Yue Yun," after the handsome character in Chinese operas.	C	Nuosong was exquisitely handsome. The boat people of Chadong were hard put to find words for his good looks. The best they could come up with was the nickname Yue Yun. None of them had ever seen Yue Yun, that most handsome warrior of the Song dynasty a thousand years earlier, but they thought they saw a resemblance to the dashing Yue Yun figure who appeared onstage in local opera.
WXYY 4.0 Turbo	Nuosong was very handsome. The boatmen of Chadong, not skilled in praising beauty, nicknamed him "Yue Yun," inspired by the young opera character's demeanor.	C	

Table 6 Over translation frequency of LLMTs.

LLMT	Over translation
GPT4	4
GPT4o	2
Gemini	2
WXYY 4.0 Turbo	1

difficulty in handling culturally specific content. WXYY 4.0 Turbo and Gemini demonstrated relatively fewer errors (7 each), possibly due to more specialized training or fine-tuning on East Asian languages and cultures. However, the overall frequency difference is relatively small. Moreover, WXYY, as a Chinese LLM, performed similarly to Gemini. This raises the doubt whether the LLM trained on predominantly English-centric data sets struggles to capture non-Western cultural contexts.

Example 5

ST: 杂货铺卖美孚油及点美孚油的洋灯, 与香烛纸张。

In GPT 4, translating 美孚油 as "Mobil oil" is misleading, as in this context, it refers to kerosene rather than modern Mobil-brand products like gasoline or lubricants. The phrase点美孚油的洋灯as "lamps that use Mobil oil" shares the same error as the former phrase. 香烛纸张was translated as "scented candles and paper," as seen in Table 9, which is inaccurate. In this context, 香烛refers to incense and ritual candles used for religious or ceremonial purposes, which lost the ritual and cultural connotations of this object. In GPT4o translation. 香烛was translated as "incense," which omitted candles and can not capture the ritual nature of these goods. In Gemini translation, 美孚油was translated as "Meifu oil" through transliteration, which is confusing for English readers, as "Meifu" doesn't convey the intended meaning of kerosene. 洋灯 translated as "Meifu oil lamps" loses the cultural context and does not clearly indicate traditional kerosene lamps. 香烛纸张was transferred as "incense and paper" which is incorrect, since it omitted the "candles", which are significant in a religious or ceremonial context. In WXYY translation, 美孚油 was also translated as "Mobil oil," same as GPT 4. 香烛纸张 was translated as "incense, paper, and candles," which is closer to the intended meaning. In contrast, the HT version kept the most connotations of the culture-bound elements and aligned with the original text.

Discourse-level errors. Discourse-level errors affect the coherence and flow of translation, resulting in inaccuracy in the TT. In the LLMTs of *Border Town*, errors in the discursive level primarily present as inconsistency of the grammatical usages.

As shown in Table 10, GPT4 has the highest rate of errors (12 occurrences), likely impacting the readability and overall quality of translation. WXYY 4.0 Turbo ranks second in terms of errors, with 9 errors. While it performs better than GPT4, it still displays noticeable challenges in managing discourse-level coherence.GPT4o exhibits a much lower frequency of errors, with only 6 errors. Gemini exhibited the fewest discourse-level errors (5), suggesting stronger coherence and accuracy in its translations compared to the other three LLMTs. However, the difference between GPT 4o and Gemini is slight. This may be due to better alignment strategies or contextual embedding improvements, as observed in models optimized for text comprehension.

Example 6

ST: 代替了天, 使他在日头升起时, 感到生活的力量, 当日头落下时, 又不至于思量与日头同时死去的, 是那个伴在他身旁的女孩子。

In Example 6, the source text conveys a deep emotional bond between the granddaughter Cuicui and her grandfather, emphasizing Cuicui's symbolic role in "replacing heaven" by providing emotional support and existential reassurance to the old boatman. The metaphor in the source text assigns this functional role to the girl, rather than suggesting that the grandfather himself assumes such a role. In the GPT-4 translation (Table 11), this metaphorical relation is rendered as "he replaces the heaven," which shifts the experiential and functional focus to the grandfather. While this formulation may be interpreted as a lexical misselection of replaces, it also restructures participant roles by foregrounding the male protagonist as the agent of the metaphor. A similar pattern is observed in the WXYY 4.0 Turbo translation, where "In place of heaven, he feels the strength of life" likewise centers the grandfather as the grammatical subject. In contrast, the human translation preserves a cleft structure ("It is the girl… that…") which explicitly maintains Cuicui as the functional agent. In the literary context of Border Town, this shift in subject assignment weakens the symbolic and relational structure between characters. For this reason, the error is interpreted as having discourse-level implications, rather than being treated as a purely lexical or stylistic variation.

Table 7- Example of over-translation in Gemini.

LLM	TT	Severity level	HT
Gemini	The girl's mother, the old ferryman's only daughter, had a secret affair with a Chadong soldier fifteen years ago, betraying her kind-hearted father.	C	The girl's mother, the old ferryman's only child, had some fifteen years earlier come to know a soldier from Chadong through the customary exchange of amorous verses, sung by each in turn across the mountain valley. And that had led to trysts carried on behind the honest ferryman's back.

Table 8 Cultural mistranslation frequency of LLMTs.

LLMT	Cultural Mistranslation
GPT4	9
GPT4o	8
Gemini	7
WXYY 4.0 Turbo	7

Discussion

This study set out to evaluate the translation performance of four LLMs, GPT4, GPT4o, Gemini, and WXYY 4.0 Turbo, using an MQM framework and error analysis framework (Lommel, 2018). Our findings are analyzed through three key dimensions of translation quality: accuracy, fidelity, and cultural appropriateness. By integrating a multidimensional analysis into translation quality analysis, this study contributes a more targeted framework for evaluating LLMs in literary contexts, which is still underexplored in academics.

Literary translation is inherently more demanding than technical or specialized translation because it involves not only conveying the literal meaning of words but also preserving style, tone, cultural context, and emotional resonance. The results demonstrate that positive outputs are generated by these LLMs, despite the recorded errors, suggesting a robust performance in rendering specific literary features, such as imagery and narrative clarity. According to Venuti (1995), literary translation requires "domestication" and "foreignization" strategies to balance cultural integrity, considering the target audiences. However, LLMs, including state-of-the-art models, often lack training in these strategies, resulting in issues such as mistranslation, cultural mistranslation, and omission.

Border Town is a literary work deeply embedded with vivid and poetic descriptions of the characters, scenery in Chadong, and local folk customs of Chinese cultural elements. The analysis indicates that accuracy is a persistent challenge for all LLMs, although with different severity levels. Mistranslation and omission were prevalent in the LLMTs, especially in GPT4, which reveals the model's limitations in contextualizing complex sentences and phrases within a literary framework, showing lexical inaccuracy and misinterpretation. GPT4 and its optimized variant, GPT4o, exhibited different strengths in terms of accuracy. However, accuracy remains a significant challenge for all LLMs. Even though the newer and optimized models like GPT4o can reduce low-level errors, accuracy is still affected by sentence complexity, idiomatic expressions, and ambiguous phrasing in the ST. The critical severity level caused by mistranslation and omission demonstrates that the LLMTs require more training in Chinese-English translation in the literary field.

One of the insights of this study is the imbalance between model size and translation specificity. GPT4o, a more optimized and compact version of GPT4, generally exhibited fewer errors across categories compared to GPT4, which suggests a smaller, more specialized model may handle difficult tasks better. This finding proves that LLMs with targeted optimizations can outperform larger models in domain-specific applications. There is still potential for LLMs to perform near-human translation when equipped with strategic guidance, though challenges remain in highly cultural contexts (He et al., 2024). Gemini and WXYY 4.0 Turbo still suffer from high omission and mistranslation, likely due to their limited training in extensive and diverse literary contexts. Interestingly, WXYY 4.0 Turbo, a Chinese-developed model, was expected to perform better in this dimension, but it still demonstrated comparable cultural inaccuracies, which performed similarly to Gemini in terms of cultural mistranslation errors. Even though WXYY is a primary Chinese LLM, its improvements in certain types of accuracy remain limited, as it does not fully address the unique challenges posed by literary translation in the Chinese context. The large portion of omission is not beneficial to construct the images in the original text as the HT by Kinkley (Hu, 2019; Ma, 2022).

Another finding concerning fidelity is the prevalence of discourse-level errors in LLM translations, particularly in GPT4 and WXYY 4.0 Turbo. Traditional MT models, and even LLMs like GPT4, face significant challenges in maintaining coherence over longer passages, particularly in literary texts where narrative continuity and thematic unity are paramount. GPT4 had the highest frequency of discourse-level errors. Over-translation errors were also observed in GPT4, where unnecessary additions or moralized interpretations altered the original tone. These exhibit that GPT4 has the least satisfactory translation results. Literature often employs complex temporal shifts, metaphorical language, and layered meanings, all of which demand a deep understanding of discourse structure. Despite achieving surface-level fluency, LLMs still face limitations in grasping authorial intention, which is a critical part of literary fidelity.

Cultural appropriateness is another significant limitation for LLMs, particularly when dealing with culturally embedded symbols and regional references. Cultural transformation is paramount in literary translation, as it allows readers to access not just the language but also the cultural essence of the ST. Shen Congwen's *Border Town* is rooted in rural Chinese culture, with themes, values, and societal structures that may be unfamiliar to non-Chinese readers. Cultural misinterpretations in the LLMTs indicate a deficiency in processing culturally charged elements. For instance, the mistranslation of "美孚油" as "Mobil oil" in several LLM outputs reflects a failure to contextualize historical and cultural elements, leading to distortions in meaning. The data reflects the failure to deal with implied cultural references of LLMTs, with cultural misinterpretation errors observed across all models. The findings suggest that LLMs require not only larger, diverse datasets but also enhanced capabilities for handling cultural elements across languages and cultures. This study, therefore, contributes to the ongoing discussion on the need for cultural training in LLMs, especially when applied to literary translation.

It is useful to relate these findings to recent large-scale evaluation efforts. The 2025 Findings of the WMT report that, in human evaluation, the literary domain appears comparatively easy to translate (Kocmi et al., 2025). This outcome, however,

Table 9 Example of cultural mistranslation of LLMTs.

LLM	TT	Severity level	HT
GPT4	The grocery stores sell Mobil oil and lamps that use Mobil oil, along with scented candles and paper.	C	The general store sold American kerosene, the Standard Oil lamps that burned it, incense, candles, and paper goods.
GPT4o	The general store sells kerosene and kerosene lamps, as well as incense and paper goods.	C	
Gemini	Grocery stores sold Meifu oil and Meifu oil lamps, as well as incense and paper.	C	
WXYY 4.0 Turbo	The grocery store sold Mobil oil and oil lamps that used Mobil oil, as well as incense, paper, and candles.	C	

Table 10 Discourse-level error frequency of LLMTs.

LLMT	Discourse-level error
GPT4	12
GPT4o	6
Gemini	5
WXYY 4.0 Turbo	9

Table 11 Example of discourse-level error in GPT4 and WXYY 4.0 Turbo.

LLM	TT	Severity level	HT
GPT4	It's as though he replaces the heavens, feeling the vigor of life at sunrise, and at sunset, he does not consider dying with the day, thanks to the girl who stays by his side.	C	It was the girl keeping him company who was Heaven's agent, letting him feel the power of life as the sun rose, and stopping him from thinking of expiring along with the sunlight when it faded at night.
WXYY 4.0 Turbo	In place of heaven, he feels the strength of life when the sun rises. And when the sun sets, he does not have to contemplate dying with it. It is the girl who accompanies him.	C	

likely reflects limitations in evaluation sensitivity, as large-scale assessments tend to prioritize overall fluency and readability, potentially overlooking finer literary qualities such as narrative perspective, character relations, and culturally embedded meaning. Our results show that seemingly fluent LLM outputs may still contain critical omissions, discourse-level distortions, and cultural misinterpretations that substantially affect literary meaning. In this respect, the study complements rather than contradicts WMT findings by foregrounding dimensions of literary translation quality that are less visible in large-scale human evaluation.

These findings extend recent studies (e.g., He et al. 2024, Wang et al. 2023, Zuo et al. 2024) that highlight the surface adequacy but deeper contextual and cultural limitations of LLM-generated translations, suggesting that while LLMs have made significant strides in translation, they are not yet ready to replace human translators in literary contexts. Distinct from LLMs, human translators often make deliberate adjustments to ensure coherence, especially when dealing with texts that have complex narrative structures. Comparing Kinkley's translation with the LLMTs manifests that LLMs lack human flexibility and capacity for holistic textual interpretation, which further limits their effectiveness in translating literature without significant post-editing by human translators to improve translation quality. However, LLMs could still play a valuable tool in translation. Error analysis is not an end in itself but serves as a foundation for future work. As was suggested, integrating human feedback into the translation process can enhance the overall quality of LLM outputs (He et al., 2024), and the prompt strategies can also be improved (Lu et al., 2023; Zhang et al., 2023).

Conclusion

This study has provided a comprehensive evaluation of the performance of LLMs in translating the first two chapters of Shen Congwen's *Border Town*, focusing on accuracy, fidelity, and cultural appropriateness. GPT4, GPT4o, Geimini, and WXYY 4.0 Turbo were selected as the models. This study contributes valuable insights into the field of MT by providing empirical evidence on the specific challenges faced by LLMs in literary translation. It applies a framework for assessing translation quality in machine-translated literary texts by categorizing errors across mistranslation, omission, over-translation, cultural mistranslation, and discourse-level errors and comparing their frequency across different LLMs. The findings reflect insights specific to the two selected chapters of *Border Tow*n, which were chosen due to their high cultural density and representative style. It contributes to the emerging body of research on LLM-based literary translation by providing a multidimensional, comparative assessment rooted in qualitative and functional perspectives.

The findings show the performance patterns across the four LLMs. GPT4o had the least errors overall, and it was relatively effective at preserving cultural and stylistic elements. While GPT4 showed reasonable lexical fidelity, it produced inconsistencies at the discourse level. Gemini exhibited the highest omission rate, especially in metaphorical and cultural elements, indicating challenges in handling abstract or less literal content. WXYY 4.0 Turbo demonstrated frequent cultural misinterpretation and literal translations of idioms, which may suggest a lack of adaptation to cross-cultural communication.

The findings also reveal the limitations of LLMs in handling culturally charged texts, highlighting the need for ongoing need for hybrid approaches and domain-specific training of LLMs. However, this study is limited because of the relatively small sample size and single literary work comparison. Moreover, the LLMs are evolving rapidly, so the generalization to newer models or beyond the selected text may not be appropriate. Therefore, in the future, further investigation can be conducted concerning other literary works, such as prose, plays, and poems, and may employ larger-scale annotations for statistical validation and allow for broader MT benchmarking. Moreover, a collaborative approach is required that combines the strengths of LLMs and human translators, such as human post-translation editing. We also suggest longitudinal evaluations to track model improvement over time and comparative studies involving professional, novice, and machine-generated translations. Additionally, this study focused primarily on localized negative error categorization, and MQM is used for diagnostic purposes rather than holistic literary evaluation, which may overlook some document-level qualities. Future studies could include an affirmative assessment of literary qualities such as stylistic beauty, metaphorical fidelity, and emotional resonance. It should be noted that this study adopted a zero-shot prompt to assess baseline model competence. Future studies should investigate whether context-enriched or task-specific prompts could enhance cultural fidelity and stylistic alignment in literary translation.

Data availability

The materials analyzed in this study include excerpts from Shen Congwen's Border Town, the large language model–generated translations, and the corresponding MQM-based error annotation data. Due to copyright restrictions on the original literary text, the source text excerpts cannot be publicly shared. The LLM-generated outputs, annotated evaluation data, coding manual, and other related files are available from the corresponding author upon reasonable request.

Received: 25 February 2025; Accepted: 24 February 2026;
Published online: 14 March 2026

References

Chen X (2019) When translation meets psychoanalysis: a study in contemporary Chinese literary translation [Doctoral thesis]. State University of New York

Fernandes P, Deutsch D, Finkelstein M, Riley P, Martins AFT, Neubig G, Garg A, Clark J, Freitag M, Firat O (2023) The devil is in the errors: leveraging large language models for fine-grained machine translation evaluation. In: Conference on machine translation - proceedings. https://doi.org/10.48550/arXiv.2308.07286

Freitag M, Foster G, Grangier D, Ratnakar V, Uszkoreit J (2021) Experts, errors, and context: A large-scale study of human evaluation for machine translation. Trans Assoc Comput Linguist. https://doi.org/10.1162/tacl_a_00437

Guo X, Ang LH, Rashid MS, Ser WH (2020) The translator's voice through the translation of characters' names in Bian Cheng. Southeast Asian J Engl Lang Stud 26:81–95

He Z, Liang T, Jiao W, Zhang Z, Yang Y (2024) Exploring human-like translation strategy with large language models. Trans Assoc Comput Linguist. https://doi.org/10.1162/tacl_a_00642

Hermans T (2007) Literary translation. In: Kuhiwczak P, Littau K (eds), A companion to translation studies. Multilingual Matters, p 77–91 https://translationjournal.net/journal/45review.htm

Hu F (2019) A study on English translations of Bian Cheng from the perspective of imagology [Master thesis]. Shanghai International Studies University

Jiao W, Huang J-T, Wang W, He Z, Wang X, Tu Z (2023) ParroT: Translating during chat using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. pp.15009−15020. https://doi.org/10.18653/v1/2023.findings-emnlp.1001

Klubicka F, Toral A, Sánchez-Cartagena V (2018) Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. Mach Transl 32:195–215. https://doi.org/10.1007/s10590-018-9214-x

Kocmi T, Bojar O, Federmann C, Graham Y, Grundkiewicz R, Haddow B, Zampieri M (2025) Findings of the 2025 conference on machine translation (WMT25). In: *Proceedings of the eighth conference on machine translation* (WMT)

Kocmi T, Federmann C (2023) Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. pp. 193-203. https://aclanthology.org/2023.eamt-1.19/

Li J, Zhou H, Huang S, Cheng S (2024) Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. Trans Assoc Comput Linguist. https://doi.org/10.1162/tacl_a_00655

Liu L (2019) A corpus-based study of the translator's style in four English translations of Biancheng [Master Thesis]. East China University of Science and Technology

Liu N (2023) A contrastive study of three English versions of Biancheng from the perspective of corpus-based critical translation studies. *Modern Linguistics*. pp. 145–157. https://doi.org/10.12677/ML.2023.111021

Lommel A (2018) Metrics for translation quality assessment: a case for standardising error typologies. In: Moorkens J, Castilho S, Gaspari F, Doherty S (eds) Translation quality assessment. machine translation: technologies and applications, Vol 1. Springer, Cham. https://doi.org/10.1007/978-3-319-91241-7_6

Lu Q, Qiu B, Ding L, Xie L, Tao D (2023) Error analysis prompting enables human-like translation evaluation in large language models: a case study on ChatGPT. In *Findings of the Association for Computational Linguistics: ACL 2024*. pp. 8801−8816. https://doi.org/10.18653/v1/2024.findings-acl.520

Ma R (2022) China's image in Jefferey Kinkley's translation of Border Town: An imagological approach [Master thesis]. Beijing Foreign Studies University

Mikoyan A (2019) Understanding in literary translation. Armenian Folia Anglistika. 64–85. https://doi.org/10.46991/afa/2019.15.2.064

Ponzio A (2007) Translation and the literary text. TTR 20:89–119. https://doi.org/10.7202/018823AR

Sai A, Nagarajan V, Dixit T, Dabre R, Kunchukuttan A, Kumar P, Khapra M (2023) IndicMT Eval: a dataset to meta-evaluate machine translation metrics for indian languages. In: *Proceedings of the 61st annual meeting of the association for computational linguistics* (Volume 1: Long Papers), https://doi.org/10.18653/v1/2023.acl-long.795

Sun S (2015) Measuring translation difficulty: theoretical and methodological considerations. Across Lang Cult 16:29–54. https://doi.org/10.1556/084.2015.16.1.2

Thompson G, Dooley K (2019) Ensuring translation fidelity in multilingual research. In: The Routledge handbook of research methods in applied linguistics. Routledge, p 63–75. https://doi.org/10.4324/9780367824471-6

Venuti L (1995) The translator's invisibility: a history of translation. In: The translator's invisibility. Routledge. https://doi.org/10.4324/9780203360064

Wang L, Lyu, C, Ji, T, Zhang, Z, Yu D, Shi S, Tu Z (2023) Document-level machine translation with large language models. In: *Proceedings of the 2023 conference on empirical methods in natural language processing*. https://doi.org/10.18653/v1/2023.emnlp-main.1036

Wang Z (2014) The translator's subjectivity in literary translation. Comp Lit 19:96–111. https://doi.org/10.1080/25723618.2014.12015489

Wright C (2016) Literary translation, 1st edn. Routledge. https://doi.org/10.4324/9781315643694

Xu H, Kim YJ, Sharaf A, Awadalla HH (2024) A paradigm shift in machine translation: boosting translation performance of large language models. Preprint at https://arxiv.org/abs/2309.11674

Xu M (2012) On scholar translators in literary translation–a case study of Kinkley's translation of 'Biancheng. Perspect Stud Translatol 20:151–163. https://doi.org/10.1080/0907676X.2011.554610

Xu M (2019) Translation of modern Chinese literature in America: an interview with Jeffrey C. Kinkley. ARIEL 50:127–138. https://doi.org/10.1353/ari.2019.0036

Xu M, Yu J (2019) Sociological formation and reception of translation: the case of Kinkley's translation of Biancheng. Transl Interpret Stud 14:333–350. https://doi.org/10.1075/tis.19039.xu

Zhang B, Haddow B, Birch A (2023) Prompting large language models for machine translation: a case study. In: Proceedings of the 40th international conference on machine learning (ICML'23). https://doi.org/10.5555/3618408.3620130

Zhu W, Liu H, Dong Q, Xu J, Kong L, Chen J, Huang S (2023) Multilingual machine translation with large language models: Empirical results and analysis. https://doi.org/10.48550/arXiv.2304.04675

Zuo Y, Ching GS, Khotsing R (2024) The application of ChatGPT in literary translation: a case study from Thai to Chinese. In: Uden L, Liberona D (eds) Learning technology for education challenges. LTEC 2024. Communications in computer and information science, vol 2082. Springer, Cham. https://doi.org/10.1007/978-3-031-61678-5_24

Acknowledgements
This work was supported by the project "Brand Image Construction of Guizhou Liquor Culture Translation in the Global Beverage Cultural Context", Grant No. 25GZQN93.

Author contributions
In terms of author contributions, WY wrote the main manuscript text, MXY prepared tables, and both conducted the analysis of the translation errors. All authors reviewed and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Ethical approval
This study analyzes literary texts and machine-generated translations and does not involve human participants, personal data, or animal subjects. Therefore, ethical approval was not required.

Informed consent
This article does not contain any studies with human participants performed by any of the authors therefore, informed consent was not required.

Additional information
Correspondence and requests for materials should be addressed to Mingxing Yang.