Article

CADERNOS de TRADUÇÃO

How well can state-of-the-art machine translation systems render a 16th-century Chinese novel?

Mu You
University of Macau
Macau SAR, China
youmuafonso@gmail.com
https://orcid.org/0000-0001-5029-3903

Derek F. Wong
University of Macau
Macau SAR, China
derekfw@um.edu.mo
https://orcid.org/0000-0002-5307-7322

Jing Zhang
University of Macau
Macau SAR, China
jingz@um.edu.mo
https://orcid.org/0000-0003-4801-6354

Kaixin Lan
University of Macau
Macau SAR, China
nlp2ct.kaixin@gmail.com
https://orcid.org/0009-0001-4722-0743

Abstract: This study evaluates the performance of state-of-the-art machine translation systems in rendering *Journey to the West*, a culturally rich 16th-century Chinese novel, into Portuguese. Employing a mixed-methods approach, we compare translations produced by DeepSeek-V3, GPT-4o, DeepL Pro, and NovelTrans-J against a published human translation. Quantitative assessments conducted by an expert evaluator examine accuracy, fluency, stylistic elegance, cultural appropriateness, and overall translation quality at both the sentence and chunk levels. The results reveal that three MT systems (DeepSeek-V3, GPT-4o, and NovelTrans-J) produce translations of comparable or superior quality to the human translation. Among them, NovelTrans-J consistently outperforms all other participants, particularly in terms of cultural appropriateness. In contrast, DeepL Pro demonstrates significantly weaker performance across all evaluated dimensions. To complement the quantitative analysis, a qualitative investigation focuses on the rendering of culture-specific items (CSIs). NovelTrans-J exhibits outstanding performance, producing the fewest mistranslations and uniquely providing explanatory notes that facilitate reader comprehension. DeepSeek-V3 and GPT-4o also handle CSIs competently, though with less consistency, while DeepL Pro struggles considerably, showing a high rate of CSI mistranslations and generally low quality. Interestingly, the human translation also contains notable CSI-related errors, particularly in cases involving semantically opaque expressions, an area in which all participants encounter significant difficulty. These findings underscore the growing potential of MT systems to handle complex, culturally rich literary texts, although certain challenges, such as the translation of semantically

CT
Cadernos de Tradução, 45(Special Issue 3), 2025, e108394
Graduate Program in Translation Studies
Federal University of Santa Catarina, Brazil. ISSN 2175-7968
DOI https://doi.org/10.5007/2175-7968.2025.e108394

opaque expressions, remain significant obstacles. We hope this study provides an updated perspective on the current capabilities of MT and offers practical insights to guide the development of future systems that can more accurately capture and transmit the distinctive cultural nuances embedded in literary works.

Keywords: machine translation; Chinese-Portuguese literary translation; *Journey to the West*; human evaluation; culture-specific item.

1. Introduction

Literary translation, one of the most intricate and nuanced tasks even for skilled human translators, has long been considered an elusive goal for machine translation (MT)[1] (Taivalkoski-Shilov, 2019; Toral & Way, 2018). For decades, achieving high-quality automated literary translation has remained a visionary yet largely unattainable aspiration. The subpar performance of MT systems in rendering literary texts has often been cited as evidence of their limitations (Way, 2012). Critics have warned that poor-quality machine translations could harm both authors and readers by ruining the work in the target language and damaging the author's reputation (see Taivalkoski-Shilov, 2019). However, recent technological advancements have begun to transform this once-distant aspiration into a tangible reality. Promising results in translating modern web novels suggest that high-quality automated literary translation may now be within reach (Wang et al., 2023, 2024).

This paper focuses on the performance of state-of-the-art MT systems in tackling a particularly demanding task: translating *Journey to the West*, a 16th-century Chinese literary classic, into Portuguese. This novel holds profound cultural significance, both within China and globally. Its influence extends to anime, television series, literature, and video games, including iconic works such as *Dragon Ball* and the newly launched AAA game *Black Myth: Wukong*. Unlike contemporary literary texts, *Journey to the West* presents two unique challenges for Chinese-Portuguese machine translation. First, the novel is written in premodern vernacular Chinese, which can pose difficulties even for native speakers, thereby introducing additional complexity for MT systems. Second, the text is deeply embedded in a distinctive worldview that blends rich cultural references, many of which are often unfamiliar to modern Chinese readers. These features make *Journey to the West* an ideal test case for revealing the strengths and weaknesses of current MT systems in handling complex, culturally rich literary texts.

Our objective is to provide translators, educators and researchers with a clearer picture of the current capabilities of MT systems. To this end, we conduct a comprehensive human evaluation of the novel's first chapter, comparing a published human translation (Wu, 2024) with the outputs from four MT systems: DeepSeek-V3, GPT-4o, DeepL Pro and NovelTrans-J, a specialized version of NovelTrans (Liu et al., 2024). The first three are widely recognized and commonly used MT tools, while NovelTrans is among the top-performing systems in the WMT2024 Discourse-Level Literary Translation Shared Task, as determined by both automatic metrics and human assessments (Wang et al., 2024). Specifically, we employ quantitative analysis to assess accuracy, fluency, elegance, cultural appropriateness, and overall performance, and a qualitative approach to examine how each system and the human translator handle culture-specific items (CSIs).

[1] In this paper, "MT" refers broadly to any non-human system capable of performing translation.

We hope that this study offers practical insights to inform the development of more sophisticated MT systems capable of capturing and delivering the distinctive cultural nuances of Chinese classic literature. Such advancements could foster more effective intercultural dialogue and mutual understanding, especially given the substantial time and financial resources required for the manual translation of literary texts (Toral & Way, 2015).

The remainder of the paper is organized as follows: Section 2 offers an overview of Translation Quality Assessment (TQA) approaches in different contexts. Section 3 defines CSIs and presents the framework used to qualitatively analyze their translation. Section 4 outlines the methodology employed in this study, followed by the presentation and discussion of results in Section 5. Finally, Section 6 concludes with a summary of findings and final remarks.

2. Translation quality assessment

This study employs a mixed-methods design, using both quantitative and qualitative methods to evaluate translation quality. The specific evaluation procedures are described in Section 4. In the present section, we first provide a broad overview of TQA approaches within Translation Studies (TS) and the translation industry. We then examine commonly employed TQA practices for MT. The aim of this section is to contextualize and justify the evaluation methods selected for this study. As such, an in-depth discussion of specific theoretical frameworks lies beyond its scope.

2.1 Translation quality assessment in academic and industry settings

TQA is perhaps one of the most controversial and debated topics in translation scholarship and practice (Colina, 2009). Despite extensive discussion of translation quality within TS (see, for instance, Colina, 2008, 2015; Hague et al., 2011; House, 2001, 2014; Lauscher, 2000; Munday et al., 2022), a notable lack of consensus persists. As Drugan (2013, p. 36) observes, "[...] theorists disagree, even on how many *categories* of models there are" (emphasis in original). Much of this divergence can be attributed to the fact that translation constitutes a complex cognitive, linguistic, social, cultural, and technological process; any attempt to define or assess translation quality inevitably reflects such complexity (Castilho et al., 2018). As House (1997, p. 1) aptly notes, "[...] evaluating the quality of a translation presupposes a theory of translation. Thus different views of translation lead to different concepts of translational quality, and hence different ways of assessing quality". As a result, a wide array of theories has emerged, yet little agreement exists regarding what constitutes quality or how it should be measured.

Beyond this lack of consensus, many existing theories also present significant limitations when applied to practical evaluation. These limitations are especially evident in contexts where the goal is real-world assessment rather than translator training or pedagogy. First, many approaches offer only loosely defined principles for translation criticism, rather than concrete, systematic methodologies that support reliable and replicable evaluation. Second, the operationalization of these theories often proves problematic: they are either too vague to be effectively applied in practice or so complex that implementation becomes prohibitively expensive and time-consuming. For example, some models require more than 16 pages of analysis to evaluate fewer than 300 source-text words (see

Drugan, 2013), or incorporate as many as 675 parameters (see Williams, 2004). Finally, it is important to emphasize that the majority of models within TS are non-quantitative in nature, rendering them unsuitable for the kind of quantitative measurement required in the present study.

By contrast, TQA practices within the translation industry exhibit two distinctive characteristics. First, unlike research-based models, most (if not all) industry approaches lack a well-defined and explicitly articulated theoretical foundation. This is understandable, given the divergent motivations, understandings, and expectations regarding theory in academic versus professional settings (Drugan, 2013). Nevertheless, the absence of explicit theory does not mean that industry approaches operate in a theoretical vacuum. Rather, theoretical assumptions may be implicit or conveyed through non-academic terminology (Drugan, 2013). For example, the concept of fit-for-purpose translation aligns closely with Reiss and Vermeer's (2014) *Skopos* theory, while the preference for fluent, locally adapted translations resonates with Venuti's (2017) notion of domestication (Drugan, 2013). Second, industry approaches are predominantly error-based, assessing translations according to predefined error typologies and severity levels. Among these, the MQM (Multidimensional Quality Metrics) framework is currently one of the most widely used. It provides a standardized vocabulary and a hierarchical structure for TQA, applicable to both human and machine translation (Lommel, 2018). However, for the purposes of the present study, error-based frameworks like MQM may have limited utility. Focusing solely on errors can obscure the strengths of a translation, particularly when successful renderings coexist with identifiable flaws. This results in an evaluation metric that penalizes mistakes without adequately rewarding high-quality translations.

2.2 Machine translation quality assessment

As discussed in Section 2.1, TQA in academic and professional contexts differs substantially in both focus and evaluative methods, owing to the complex nature of translation. This complexity is similarly evident in the domain of MT, where a diverse array of evaluation methods is employed. These methods are typically classified into two categories: human evaluation and automatic evaluation.

Recent editions of the Conference on Machine Translation (also known as WMT, a major event for MT and MT research) have employed two primary types of human evaluation approaches (see Akhbardeh et al., 2021; Barrault et al., 2019, 2020; Bojar et al., 2017, 2018; Kocmi et al., 2022, 2023, 2024): direct assessment (Graham et al., 2013), i.e., scoring translations directly on a 100-point scale, and error-based evaluation following the MQM framework. While specific implementation details may vary, such as evaluator selection, contextual information availability, and the use of reference translations, the core evaluation mechanisms of these two approaches have remained consistent.

In addition to human evaluation, automatic tools capable of performing TQA with or without reference translations play a crucial role in MT evaluation. Compared to human evaluation, automatic evaluation is not only significantly more efficient, but also more consistent in performance and less susceptible to extralinguistic biases. This is particularly relevant in the MT context, where "professional evaluators tend to be the exception, rather than the rule", and the involvement of amateur evaluators with varying levels of language proficiency and translation expertise is common (Castilho et al., 2018, p. 23). However, a major limitation of automatic evaluation lies in the reliability of its results. Since these tools must still be validated against human judgments, their applicability to the evaluation of complex, culturally rich literary texts remains highly questionable.

Given these considerations, this study adopts direct assessment by an expert evaluator as the primary evaluation method, with automatic tools used solely to provide supplementary insights. The direct assessment focuses not only on overall translation quality but also on four specific dimensions: accuracy, fluency, elegance, and cultural appropriateness. As the study aims to examine the capacity of MT systems to bridge cultural gaps, a qualitative analysis is also conducted, with particular attention to how CSIs are handled in translation.

3. Translation of culture-specific items

The great abundance of CSIs in *Journey to the West* presents a significant challenge for both MT systems and human translators. These CSIs are deeply embedded in the novel's mythological world, which is itself closely tied to the cultural context of 16th-century China. Furthermore, the novel's use of premodern Chinese adds an additional layer of complexity to the interpretation and translation of these elements. By analyzing how these CSIs are translated, we can gain deeper insights into the intercultural competence of current MT systems, i.e., their ability to "perceive and handle difference" as culture mediators (Katan, 2009). In this section, we first briefly define the notion of culture-specific item, followed by a discussion of various translation techniques used to address these cultural references.

In the literature, a range of terms has been used to refer to the concept of CSI and related notions, including culture-specific or culture-bound references, elements, terms, items, or expressions; *realia*; allusions; culture-loaded terms, expressions, or elements; cultural references or referents; cultural elements; and culturemes (see Amenador & Wang, 2023; Marco, 2019; Ranzato, 2013). Despite this terminological variety, as Marco (2019) argues, the concept of difference seems to be fundamental: for an element to be considered culturally specific, it must either be nonexistent in the target culture or possess a significantly different intertextual status within it (Franco Aixelá, 1996; Marco, 2019; Olk, 2013). In light of this, the present study adopts Franco Aixelá's (1996, p. 57) classical definition of CSIs as "[...] any linguistically represented reference in a source text which, when transferred to a target language, poses a translation problem due to the nonexistence or to the different value [...] of the given item in the target language culture".

Regarding translation techniques (i.e., micro-level textual procedures; see Marco, 2007, for a discussion of the relevant terminology) employed to address the complexities associated with CSIs, a substantial body of classification has been developed by various scholars. Examples include Davies (2003), Díaz Cintas and Remael (2007), Franco Aixelá (1996), Gottlieb (2009), Leppihalme (2011), Marco (2019), Olk (2013), Pedersen (2005), and Ranzato (2013). When general-purpose translation techniques (e.g., Newmark, 1988; Vinay & Darbelnet, 2000) are also considered, the list expands even further. The large number of taxonomies suggests that no single classification can be universally applied. As Gottlieb (2009) aptly points out, classification in the arts and humanities is inherently somewhat arbitrary; nevertheless, it must be adapted to meet the specific needs of a given study. In this regard, we select and adjust existing classifications to suit the analysis of *Journey to the West*'s translation (see Table 1).

Table 1: Analytical categories for qualitative analysis

Category	Definition
Borrowing	The transfer of a source text CSI directly into the target language (TL), which can be pure or naturalized (i.e. adapted to the spelling and morphology of the target language) (Marco, 2019)
Direct translation	A word-for-word translation where the semantic load of the CSI remains unchanged (Pedersen, 2005)
Extratextual gloss	The inclusion of additional explanation outside the main text, such as footnotes, endnotes, glossary entries, or commentary/translation in brackets or italics (Franco Aixelá, 1996)
Intratextual gloss	The integration of a gloss as an indistinct part of the text (Franco Aixelá, 1996)
Substitution	Replacing a CSI with another element, either a different CSI or some sort of paraphrase, which may or may not involve a CSI (Pedersen, 2005)
Omission	The removal of a CSI from the target text
Mistranslation	An incorrect translation of a CSI

Source: Authors (2025)

Two key points need to be highlighted. First, in addition to borrowing, direct translation, extratextual gloss, intratextual gloss, substitution, and omission, we have also included "mistranslation" as a separate analytical category in our analysis. While mistranslation is not, strictly speaking, a translation technique, its inclusion allows us to account for instances where CSIs are rendered inaccurately or misleadingly. Second, we do not intend to provide an exhaustive taxonomy. While it is suitable for our qualitative analysis of a relatively small number of CSIs, it may require further elaboration if applied to other contexts or larger datasets.

4. Methodology

This study employed both quantitative and qualitative methods to evaluate translation quality. For the quantitative analysis, the evaluation was based on five criteria: accuracy, fluency, elegance, cultural appropriateness, and overall performance. Among these, accuracy and fluency are the most widely used in MT evaluation. Accuracy (also referred to as adequacy or fidelity) measures the extent to which the translation faithfully conveys the meaning, content, and intent of the original source material. Fluency, on the other hand, assesses the naturalness, coherence, and readability of the TL text, focusing on how well the translation conforms to the linguistic norms and conventions of the TL, independent of the source text (Castilho et al., 2018). Elegance pertains to the stylistic sophistication of the translation. Given that *Journey to the West* is a literary work, aesthetic quality is considered an important evaluation dimension, though it poses significant challenges for MT systems.

Cultural appropriateness evaluates how adequately and sensitively cultural elements are represented to the audience. Finally, overall performance refers to a holistic assessment of translation quality, which is not necessarily equivalent to the arithmetic average of the other four criteria.

All evaluation criteria were clearly explained to the expert evaluator, who is a native speaker of Portuguese, sinologist, and highly experienced Chinese-Portuguese translator. With decades of professional experience and a record of widely acclaimed published translations, the evaluator brought a high level of linguistic and cultural expertise to the task. The evaluation was conducted in a self-paced manner, with the evaluator taking breaks as needed to minimize fatigue and maintain concentration. To ensure fairness, the evaluation was conducted blindly: the identities of all participants were anonymized, and all translations were presented simultaneously without a predetermined order. This approach enabled the evaluator to make direct comparisons and promoted more consistent judgment across translations from different sources.

In line with standard practices in MT evaluation, this study examined a sample drawn from the first chapter of the novel. Specifically, 20 consecutive sentences were selected from the beginning, middle, and end of the chapter, accounting for roughly one-fifth of the chapter in total. The evaluator rated each sentence across five dimensions, using a 100-point scale. In addition to the sentence-level evaluation, a chunk-level evaluation was performed. For the chunk-level evaluation, the evaluator assessed six chunks, each consisting of 10 consecutive sentences that had already been rated individually at the sentence level. Each chunk was also rated on the same five dimensions, using a 100-point scale. This chunk-level evaluation aimed to provide a more comprehensive understanding of translation quality beyond the sentence level. As previously mentioned, the evaluation included five participants: four MT systems (DeepSeek-V3, GPT-4o, DeepL Pro and NovelTrans-J) and a published human translation (Wu, 2024). For each participant, the evaluation resulted in a total of 300 sentence-level ratings (60 sentences × 5 dimensions) and 30 chunk-level ratings (6 chunks × 5 dimensions).

Following the human evaluation, relevant statistics were computed and visualized separately for the sentence-level and chunk-level data. For the sentence-level data, the distribution and mean of the evaluation scores were analyzed, and the rankings of the five participants were determined using bootstrap methods (Efron, 1992). For the chunk-level data, the distribution and mean of the evaluation scores were likewise examined; however, relative rankings are discussed descriptively without significance testing due to the small sample size.

In addition, we conducted an automatic evaluation based on CometKiwi 2022 (Rei et al., 2022), a tool capable of providing evaluation without relying on reference translations. As discussed in Section 2.2, automatic evaluation is generally less reliable than human evaluation, particularly in the context of culturally rich literary translation. Nevertheless, it can still offer a supplementary perspective that helps to substantiate certain findings from our human evaluation. Furthermore, while developing automatic evaluation tools for literary translation is beyond the scope of this paper, we hope that the comparison between human and automatic evaluations provides useful data for future research in this area.

The qualitative analysis complemented the quantitative evaluation, focusing on the translation of CSIs. Given that *Journey to the West* is rich in CSIs, analyzing the first fifth of the chapter was sufficient for the purposes of this qualitative investigation. We first examined the original text to identify CSIs, defined as any element in the source text that presents a translation challenge due to its absence or differing significance in the target culture (Franco Aixelá, 1996). The corresponding translations for each identified CSI were then collected. Each CSI and its translation were analyzed according to the classification outlined in Table 1. Repeated translations of the same CSI were excluded from the analysis, preventing any single translation choice from being repeatedly penalized or rewarded. Finally, the results were examined to compare the performance of the five participants in dealing with CSIs.

5. Results and discussion

5.1 Sentence-level quantitative evaluation

For the sentence-level evaluation, we collected 300 evaluation scores across five dimensions for each participant. The mean scores for each dimension are presented in Table 2 and illustrated in Figure 1. To determine whether differences in mean scores between participants were statistically significant ($p < .05$), we employed bootstrap methods (Efron, 1992). According to the bootstrapping results, we ranked the five participants across the five dimensions, as shown in Figure 2.

Taking the ranking for accuracy as an example: NovelTrans-J ranks first, significantly outperforming all other participants. DeepL Pro ranks fifth, with performance significantly lower than that of the other four participants. DeepSeek-V3 ranks between second and third, GPT-4o between second and fourth, and the human translator between third and fourth. These rankings are derived from the following pairwise comparisons:

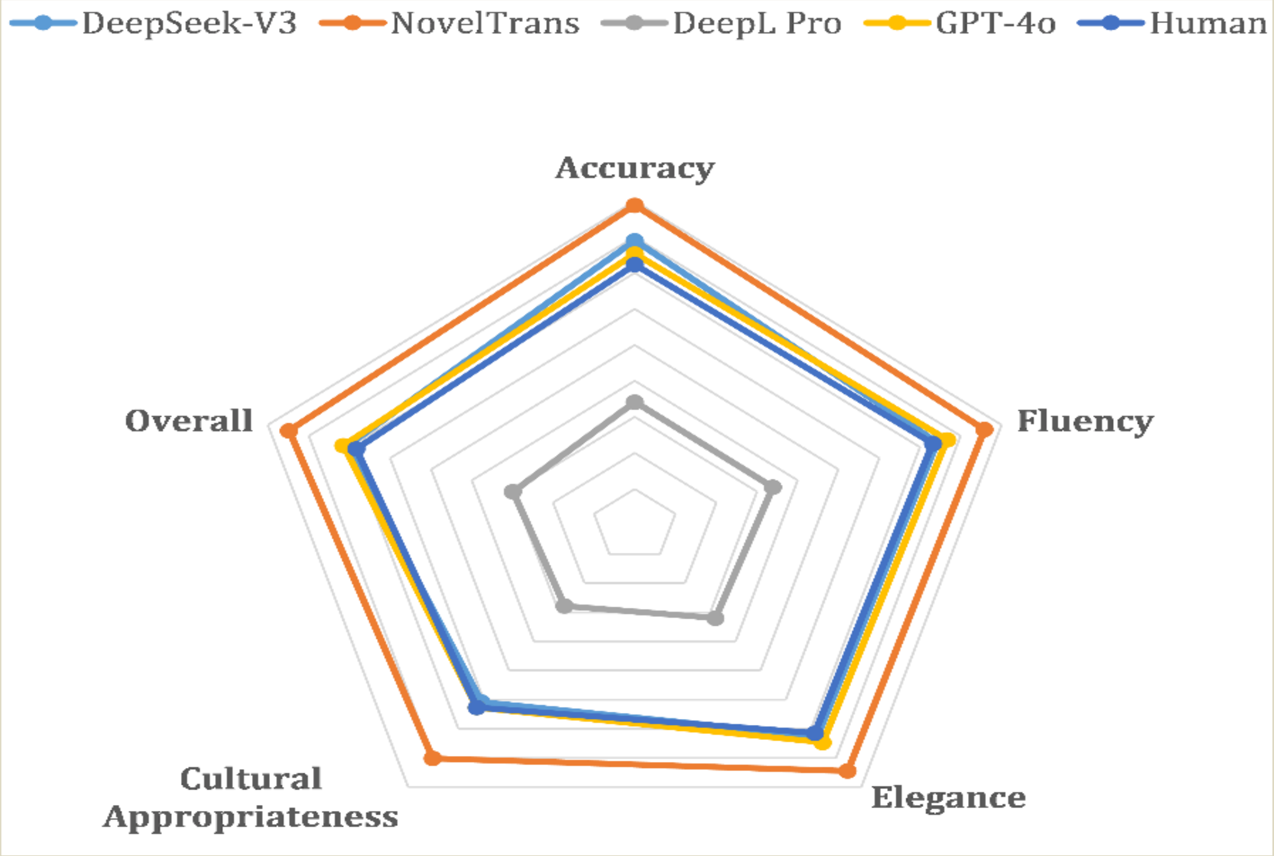
- DeepSeek-V3 does not significantly outperform GPT-4o (79 vs. 75.33), but it significantly outperforms the human translator (79 vs. 72.33).
- GPT-4o's score is neither significantly higher than the human translator's (75.33 vs. 72.33) nor significantly lower than DeepSeek-V3's (75.33 vs. 79).
- The human translator performs significantly worse than DeepSeek-V3 (72.33 vs. 79), but the difference is not significant when compared to GPT-4o (72.33 vs. 75.33).

Table 2: Mean scores across five dimensions (sentence-level evaluation)

	Accuracy	Fluency	Elegance	Cultural Appropriateness	Overall
DeepSeek-V3	79	74.17	72.5	60.83	70.67
NovelTrans-J	88.83	85.83	84.5	80.17	84.83
DeepL Pro	34.17	33.92	32	27.83	29.92
GPT-4o	75.33	76.67	74.67	62.67	71.5
Human	72.33	73.17	71.5	62.67	68.33

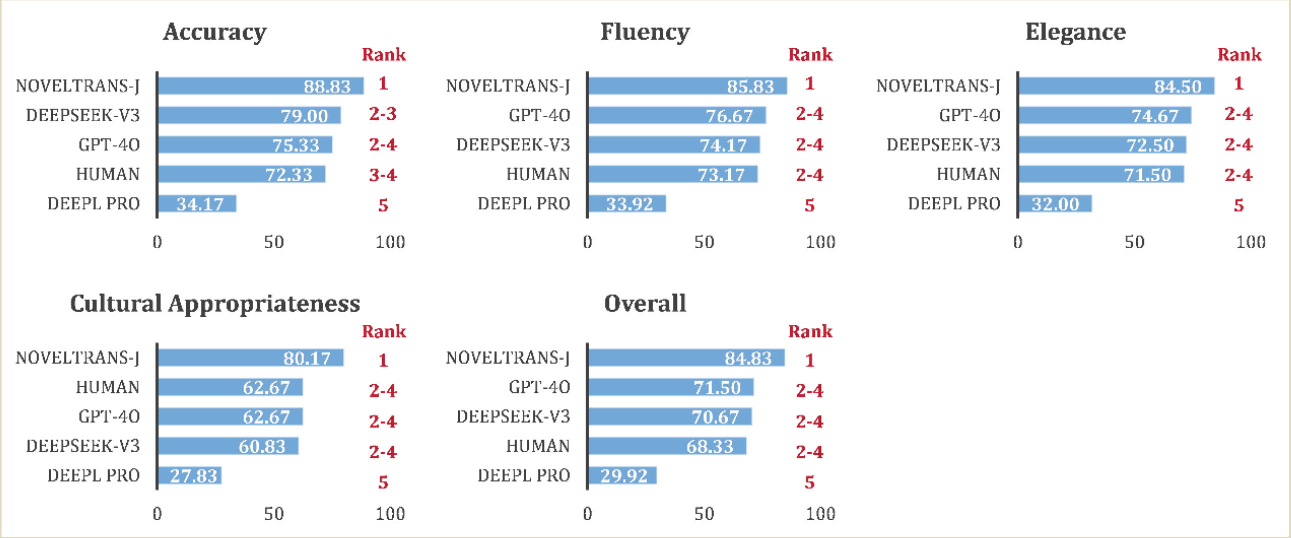
Source: Authors (2025)

Figure 1: Mean scores across five dimensions (sentence-level evaluation)



Source: Authors (2025)

Figure 2: Ranking with statistical testing (sentence-level evaluation)

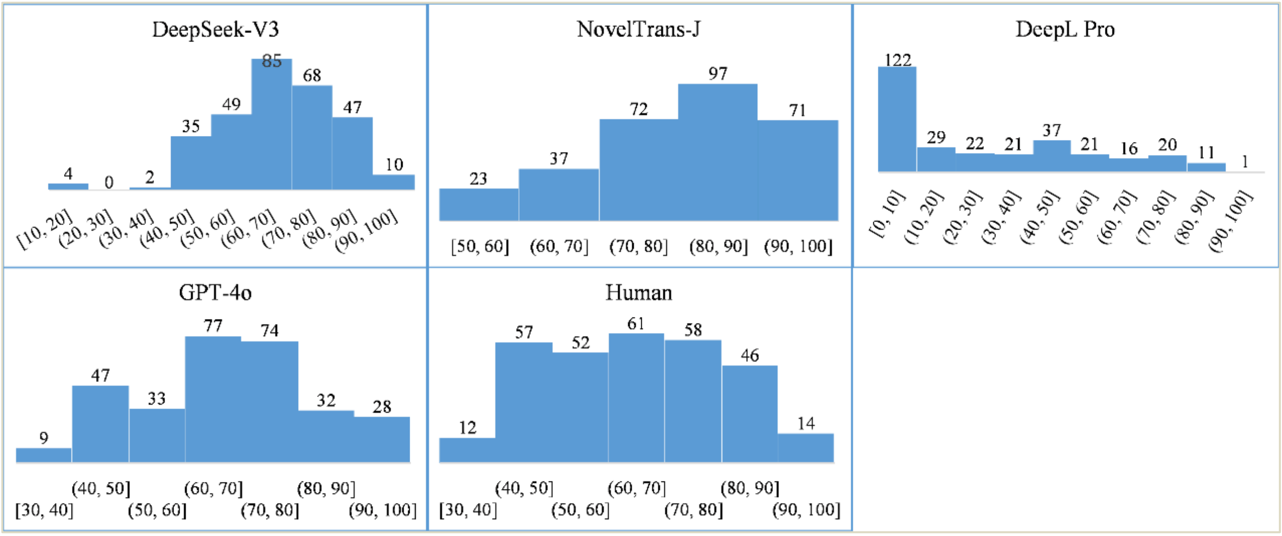


Source: Authors (2025)

Figures 1 and 2 reveal a clear performance hierarchy among the participants: NovelTrans-J consistently outperforms all others across the five evaluated dimensions, while DeepL Pro falls significantly behind, showing a considerable margin of underperformance compared to the other participants. DeepSeek-V3, GPT-4o, and the human translator exhibit no substantial differences across four of the five dimensions, except for accuracy, where, as previously noted, DeepSeek-V3 surpasses the human translator. These results indicate that, at least at the sentence level, the published human translation is not superior to the outputs of three MT systems and is, in fact, outperformed by one. Another noteworthy finding is that, in terms of cultural appropriateness, which is typically considered a weakness for MT, NovelTrans-J achieves a higher mean score than the human translator. Indeed, the cultural dimension represents NovelTrans-J's strongest area of advantage, with the largest margin observed across the five dimensions.

Figure 3 displays the distribution of all evaluation scores. It can be observed that the scores for DeepL Pro are heavily concentrated at the lower end of the 100-point scale, in contrast to the more normally distributed scores of the other four participants. Among them, NovelTrans-J exhibits a slightly skewed distribution toward the higher end of the scale.

Figure 3: Score distribution (sentence-level evaluation)



Source: Authors (2025)

Focusing on high scores (> 80), NovelTrans-J demonstrates a clear advantage over DeepSeek-V3, GPT-4o, and the human translator, with 168 (97+71) instances compared to 57 (47+10), 60 (32+28), and 60 (46+14), respectively. This provides strong evidence of NovelTrans-J's ability to produce high-quality translations.

5.2 Chunk-level quantitative evaluation

For the chunk-level evaluation, we collected 30 evaluation scores across the five dimensions for each participant. Following the approach used in the previous sentence-level evaluation, we calculated the mean scores for each dimension, which are presented in Table 3 and illustrated in Figure 4. To depict the participants' relative rankings across the five dimensions, we present Figure 5. Due to the limited sample size, no significance testing was conducted.

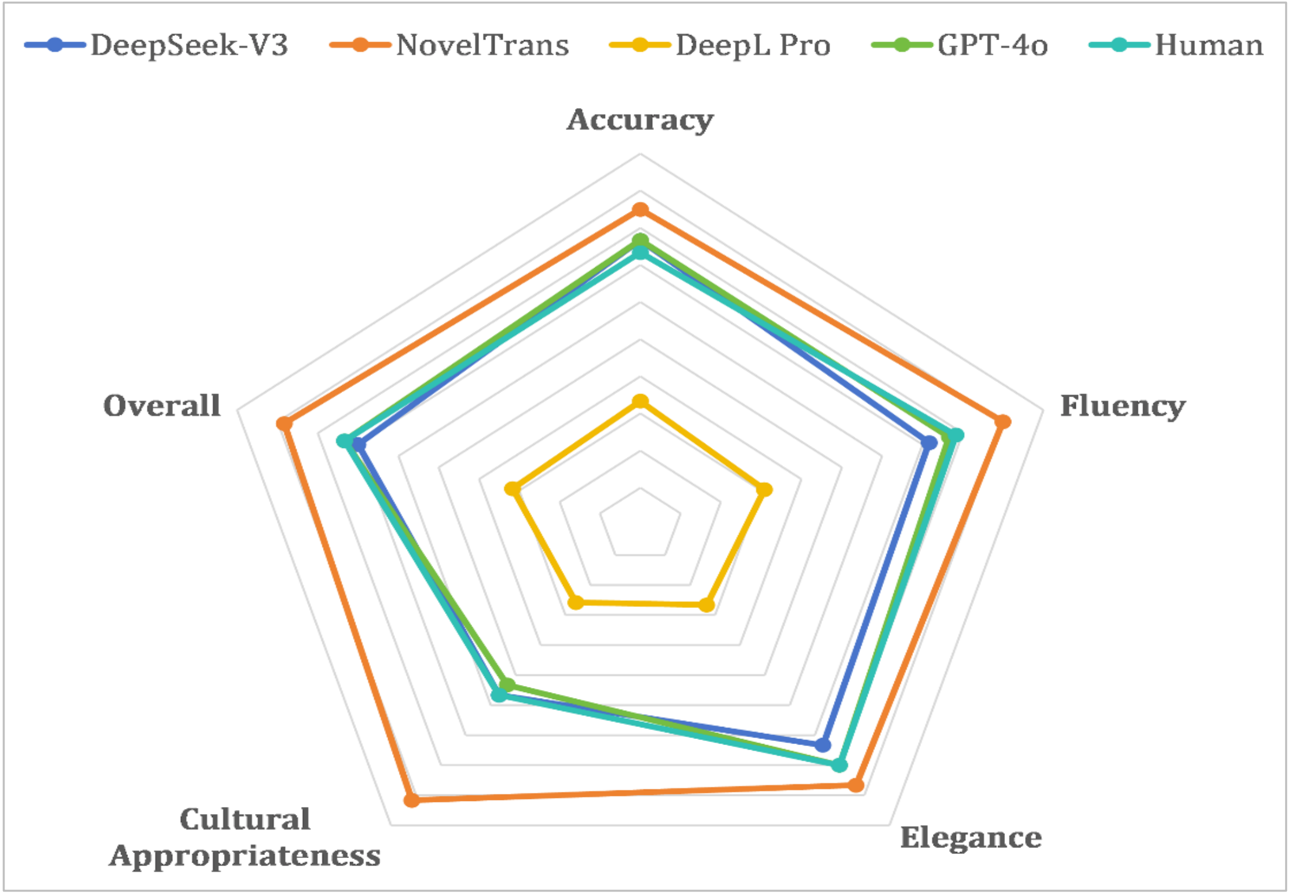
Figures 4 and 5 indicate that the participants' performances at the chunk level largely mirror those observed at the sentence level. The superior performance of NovelTrans-J and the underperformance of DeepL Pro remain clearly evident, while DeepSeek-V3, GPT-4o, and the human translator continue to receive very similar evaluation scores across the five dimensions. Combined with the findings from sentence-level evaluation, we can conclude that, at both the sentence and chunk levels, these three MT systems demonstrate a comparable or even higher level of quality to the published human translation. In terms of cultural appropriateness, NovelTrans-J once again achieves a higher mean score than the human translator, with the cultural dimension continuing to represent its strongest area of advantage. This phenomenon will be further investigated in the subsequent quantitative analysis.

Table 3: Mean scores across five dimensions (chunk-level evaluation)

	Accuracy	Fluency	Elegance	Cultural Appropriateness	Overall
DeepSeek-V3	76.67	71.67	73.33	56.67	70
NovelTrans-J	85	90	86.67	91.67	88.33
DeepL Pro	33.33	30.83	26.67	25.83	31.67
GPT-4o	76.67	76.67	80	53.33	73.33
Human	73.33	78.33	80	56.67	73.33

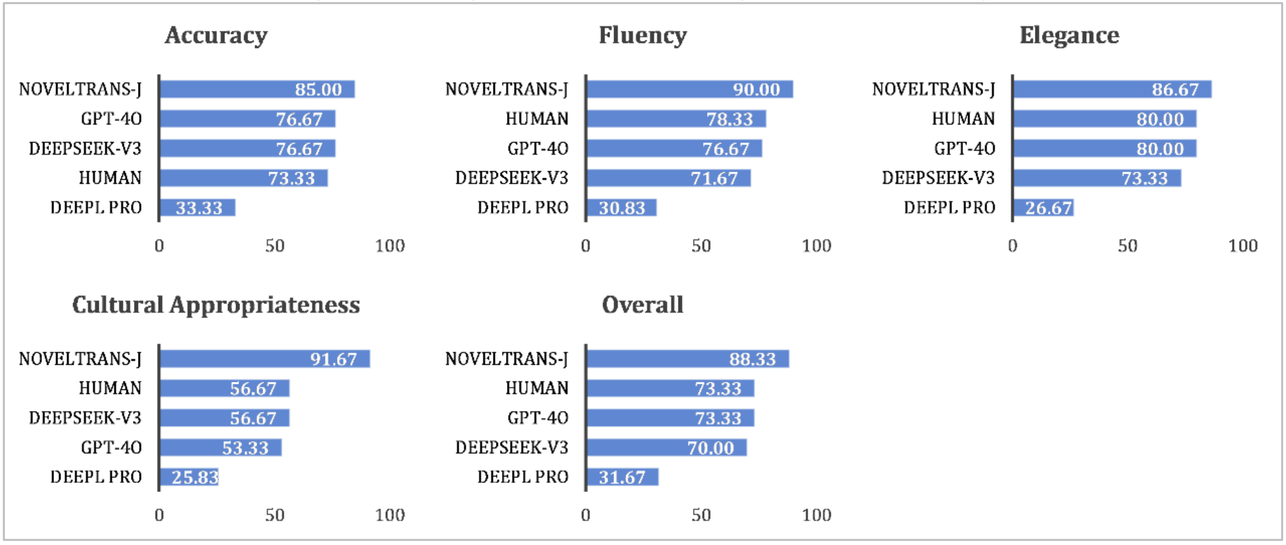
Source: Authors (2025)

Figure 4: Mean scores across five dimensions (chunk-level evaluation)



Source: Authors (2025)

Figure 5: Ranking across five dimensions (chunk-level evaluation)

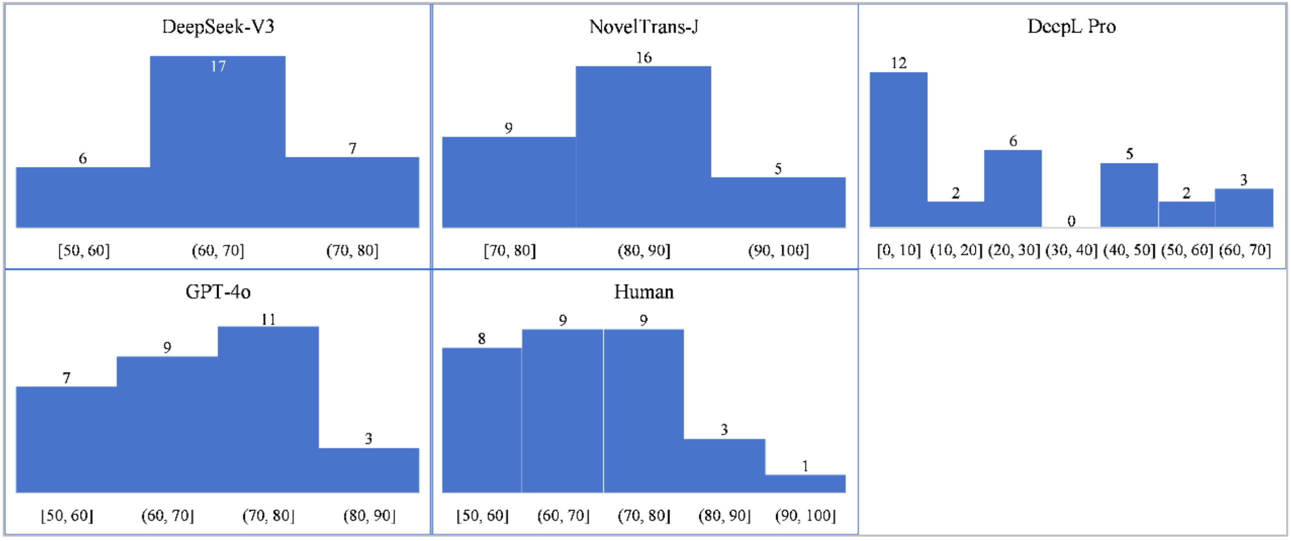


Source: Authors (2025)

Figure 6 displays the score distribution. DeepL Pro continues to perform poorly, whereas NovelTrans-J again demonstrates a marked advantage over DeepSeek-V3, GPT-4o, and the human translator in the high-score range (> 80), achieving 21 instances compared to 0, 3, and 4, respectively. This outcome suggests that NovelTrans-J consistently maintains high translation quality at the chunk level.

A comparison of the evaluation results across the two levels further reveals that the use of a larger evaluation unit appears to enhance the perceived quality of the human translation. In other words, sentence-level evaluation may obscure certain strengths of human output that become more salient when larger textual units are assessed.

Figure 6: Score distribution (chunk-level evaluation)



Source: Authors (2025)

Finally, we present the results of the automatic evaluation using CometKiwi (Rei et al., 2022) to offer a complementary perspective. As shown in Table 4, the resulting ranking (NovelTrans-J >

DeepSeek-V3 > GPT-4o > Human > DeepL Pro) largely aligns with the findings from the human evaluation, thereby lending additional support to its conclusions. However, a notable discrepancy warrants attention. The four better-performing participants (NovelTrans-J, DeepSeek-V3, GPT-4o, and the human translator) are significantly underestimated by the automatic metric, while DeepL Pro is slightly overestimated.

Table 4: Automatic evaluation via CometKiwi

Participant	Score (0-1) ↑
NovelTrans-J	0.4711
DeepSeek-V3	0.4637
GPT-4o	0.4617
Human	0.4579
DeepL Pro	0.3876

Source: Authors (2025)

Consequently, the scores fail to accurately reflect the true performance differences: the distinctions among the top four participants appear minimal, and the performance gap between them and DeepL Pro is considerably narrower than in the human evaluation. These findings highlight the need for caution when interpreting automatic evaluation results, particularly in the context of complex, culturally nuanced literary translation. At the same time, they offer valuable insights to inform the future development of more reliable evaluation methods.

5.3 Qualitative analysis

Due to the substantial cultural and linguistic differences, culture-specific items in *Journey to the West* present significant challenges for translators. This section analyzes and categorizes the translations of these CSIs as outlined in Section 4, with a summary of the results presented in Table 5. We begin with a detailed analysis of mistranslations and omissions, supported by concrete examples, followed by an examination of the main characteristics of each participant's translation.

The statistics on mistranslations align with the findings of previous quantitative analysis. NovelTrans-J shows the fewest mistranslations (9/77), followed by DeepSeek-V3 (12/77), GPT-4o (14/79), the human translator (15/77), and DeepL Pro (40/79). Notably, DeepL Pro produces a significantly higher number of errors than the other participants, with more than half of its CSI translations being incorrect. Surprisingly, the human translation also exhibits a relatively high rate of mistranslations. To better understand these results, we analyze representative examples of mistranslations for each participant. Given that omissions can affect translation quality in ways similar to mistranslations, we examine both phenomena together.

The mistranslations produced by DeepL Pro indicate that this translation task may exceed the system's current capabilities. DeepL Pro not only makes a considerable number of errors, but the nature of these errors also suggests that the system frequently fails to understand the CSIs. For instance, DeepL Pro renders "天气" (celestial *qi*, or celestial energy) as "*tempo*" (weather) and "地气" (terrestrial *qi*, or terrestrial energy) as "*gás da terra*" (gas of the earth). It also incorrectly translates "一阳" (the first Yang energy) as "*um sol*" (a sun). Additionally, it mistranslates "子", "丑"

and “亥”, three of the Twelve Earthly Branches, as “*filho*” (son), “*feio*” (ugly), and, amusingly, “Ohio”, respectively. In terms of omissions, the only instance results in an incomplete sentence, suggesting that the omission might also reflect a lack of comprehension of the original text. Collectively, these mistranslations and the omission contribute to the overall poor quality of DeepL Pro’s output.

Table 5: Classification of CSI translation based on proposed analytical categories

Type	DeepSeek-V3	NovelTrans-J	DeepL Pro	GPT-4o	Human
Mistranslation	12	9	40	14	15
Omission	0	1	1	3	1
Total Mistranslation/Omission	12	10	41	17	16
Retention	7	0	9	8	10
Retention + Extratextual Gloss	0	8	0	0	0
Retention + Intratextual Gloss	0	0	0	0	1
Total Retention	7	8	9	8	11
Direct Translation	33	24	15	24	23
Direct Translation + Extratextual Gloss	0	7	0	0	0
Direct Translation + Retention	2	0	0	0	4
Direct Translation + Omission	2	1	6	1	5
Direct Translation + Intratextual Gloss	0	1	0	0	0
Total Direct Translation	37	33	21	25	32
Substitution	19	18	8	29	18
Substitution + Extratextual Gloss	0	1	0	0	0
Substitution + Intratextual Gloss	0	1	0	0	0
Substitution + Retention	2	6	0	0	0
Total Substitution	21	26	8	29	18
Total Number	77	77	79	79	77

Source: Authors (2025)

The mistranslations in the human translation primarily result from the misinterpretation of CSIs. Notably, the human translator inaccurately renders several relatively simple CSIs that are correctly processed by the three higher-performing MT systems (DeepSeek-V3, NovelTrans-J, and GPT-4o). This may help explain why the human translation exhibits a relatively high rate of mistranslation, surpassed only by the lowest-performing system, DeepL Pro. Three illustrative examples highlight this issue: (i) “天气” (celestial *qi* or celestial energy) and “地气” (terrestrial *qi* or terrestrial energy) are incorrectly translated as “*ar do céu*” (air of the sky) and “*ar da terra*” (air of the earth), respectively; (ii) “来龙” (geomantic origin) is interpreted literally and translated as “*dragão que veio*” (dragon that came), which deviates completely from the meaning of the CSI; (iii) “三才” (three elements, a cosmological concept comprising Heaven, Earth, and Humanity) is mistranslated literally as “*três talentos*” (three talents), reflecting a literal but culturally uninformed interpretation.

In addition to these relatively straightforward cases, the human translation also contains errors involving more complex and semantically opaque CSIs that none of the participants manage to translate accurately. One such example is “木火方隅” (literally “wood-fire corner”, where wood and fire symbolize the East and South, respectively). In the human translation, this CSI is misinterpreted: it is split into two parts and rendered as "*a água e o fogo*" (the water and the fire) and "*extremidades*" (extremities). This not only fails to capture the intended directional and cosmological meaning, but also introduces "water", an element absent from the source text. The inclusion of such an unwarranted element resembles an AI hallucination and may help explain the relatively low ranking of the human translation.

In terms of omissions, the human translation contains a single instance. Although this omission does not stem from misinterpretation, it affects the logical coherence of the text. The phrase "三丈六尺五寸高，按周天三百六十五度" is translated as "*a altura de 3 zhang e 6 chi corresponde aos 365 graus do céu*" (the height of 3 zhang and 6 chi corresponds to the 365 degrees of the sky), omitting "五寸" (five *cun*), thereby disrupting the numerical correspondence between height and degrees. Overall, these mistranslations and the omission significantly diminish the perceived quality of the human translation and are key factors contributing to its being outperformed by NovelTrans-J and positioned on par with DeepSeek-V3 and GPT-4o.

Regarding GPT-4o's outputs, aside from those semantically opaque CSIs that none of the participants were able to translate correctly, the system generally demonstrates a strong capacity to comprehend CSIs. However, this comprehension is occasionally only partial, resulting in translations that are not entirely accurate. For example, "乾元" (the *qian* principle, the creative force of Heaven) and "坤元" (the *kun* principle, the receptive force of Earth) are translated as "*o ciclo do céu*" (the cycle of the sky) and "*o ciclo da terra*" (the cycle of the earth), respectively. These renderings reflect an inadequate grasp of the philosophical context embedded in these terms within classical Chinese cosmology. Moreover, GPT-4o exhibits three instances of omission, which can disrupt the coherence and philosophical completeness of the text. One notable example is the sentence "又五千四百岁, 亥会将终, 贞下起元, 近子之会, 而复逐渐开明", which is translated as: "*após mais 5.400 anos, no final do período Hai, a clareza começa a surgir novamente*" (after another 5,400 years, at the end of the Hai period, the clarity begins to emerge once again). This translation omits two significant expressions: "贞下起元" (after the decline, a new cycle begins, an astrological notion signifying cyclical renewal), and "近子之会" (around the Zi period, referencing a specific cosmological timeframe). These omissions obscure key temporal and philosophical markers that are essential for a full understanding of the passage.

DeepSeek-V3 and NovelTrans-J produce the fewest mistranslations, with most errors arising from particularly challenging CSIs that prove difficult for all participants. One such example, as previously discussed, is "木火方隅". DeepSeek-V3 renders this CSI as "*no canto da madeira e do fogo*" (in the corner of the wood and the fire), a translation that fails to convey the symbolic association of "wood" and "fire" with the eastern and southern directions. NovelTrans-J, similar to the human translator, splits the term into two components—"*madeira e fogo*" (wood and fire) and "*extremidades*" (extremities)—thereby also missing the intended cosmological and directional meaning.

Another illustrative example involves the names of the Four Great Continents, which originate from Buddhist cosmology and Sanskrit. Introduced into Chinese through a combination of interpretation and transliteration, these terms are only partially semantically transparent. For instance, in "北俱芦洲" (Pinyin: *Běi jù lú zhōu*; Sanskrit: *Uttarakuru*) and "东胜神洲" (Pinyin: *Dōng* *shèng shén zhōu*; Sanskrit: *Pūrvavideha*), components such as “俱芦” and “胜神” are historical transliterations of “*kuru*” and “*Videha*”, respectively, whereas elements like “北洲” (Northern Continent) and “东洲” (Eastern Continent) retain clear, interpretable meanings. This blending of transliterated and semantically transparent components generates ambiguity, making it difficult to discern which parts should be interpreted literally. Without sufficient background knowledge, this lack of clarity can obscure the intended meaning and hinder accurate translation. DeepSeek-V3 renders “东胜神洲” as “*o Continente de Deus Vitorioso do Leste*” (the Eastern Victorious God Continent), interpreting “胜神” according to the literal meanings of the individual characters rather than reflecting the term’s intended significance.

Notably, DeepSeek-V3 mistranslates “西牛贺洲” as “*o Continente de Boas Novas do Oeste*” (the Western Continent of Good News), “南赡部洲” as “*o Continente de Abundância do Sul*” (the Southern Continent of Abundance), and “北俱芦洲” as “*o Continente de Lótus do Norte*” (the Northern Lotus Continent), fabricating meanings (Good News, Abundance, and Lotus) that are not present in the original CSIs. These errors constitute clear instances of AI hallucination. Similarly, NovelTrans-J fails to accurately translate “东胜神洲”, rendering it as “*o Continente Divino do Leste*” (the Eastern Divine Continent), again interpreting “胜神” literally. NovelTrans-J, however, correctly translates the other three continent names.

A noteworthy phenomenon emerges in the human translation of the four continental names. The translator renders them as follows: “*o continente de Niuhe no oeste*” (the continent of Niuhe in the west), “*o continente de Jambu no sul*” (the continent of Jambu in the south), “*o continente de Aojou no leste*” (the continent of Aojou in the east), and “*o continente de Loujue no norte*” (the continent of Loujue in the north). While the first two translations are acceptable, the latter two contain unusual transliterations that deviate notably from the original names, presenting an inconsistency difficult to explain.

We now proceed to the analysis of other translation techniques employed by each participant. An examination of DeepL Pro’s output reveals that a considerable proportion of its translations of CSIs are of suboptimal quality. While these instances do not strictly qualify as “mistranslations”, they nevertheless hinder comprehension and detract from reading experience. Illustrative examples include the translations of “乾元”, “坤元”, and “仙桃”. DeepL Pro translates “乾元” (the *qian* principle, the creative force of Heaven) as “*Qian yuan*,” and “坤元” (the *kun* principle, the receptive force of Earth) as “*Kun Yuan*”. In the absence of further explanation, such transliterations are likely to cause confusion among readers. Moreover, DeepL Pro translates “仙桃” (immortal peach) merely as “*pêssego*”, thereby omitting the essential adjective that distinguishes a common fruit from one endowed with magical properties.

Regarding the human translation, the techniques employed are generally appropriate. Nevertheless, two issues warrant attention. First, the technique of retention is used more frequently than by the other participants. Since no explanatory notes are provided, some of the CSIs are likely to cause confusion and fail to fulfill their intended textual function. For example, “千里眼” (thousand-mile eyes) and “顺风耳” (wind-listener), two immortals with extraordinary abilities, are simply transliterated as “Qianliyan” and “Shunfenger”, which are likely unintelligible to Portuguese readers. Similarly, “麒麟”, a mythical creature, is rendered as “qilin” without further explanation, which may also hinder comprehension. Second, the technique Direct Translation + Omission is employed five times, more frequently than in the three higher-performing MT systems, which use this technique only once or twice. Although these omissions generally do not result in significant negative effects, they do lead to a loss of meaning. For instance, “修竹” (slender bamboos) is translated as “*bambus*”, thereby losing the connotation of elegance conveyed by the adjective.

DeepSeek-V3 and GPT-4o exhibit similar issues concerning retention and partial omission. For example, both systems transliterate “盘古” (a mythological figure regarded as the creator of the universe) and “邵康节” (a philosopher from the Song dynasty) without explanatory notes. Such retentions may impede reader comprehension and generate confusion for readers unfamiliar with these figures. In terms of Direct Translation + Omission, both systems translate “青松翠柏” (green pines and verdant cypresses) as “*(os) pinheiros verdes e (os) ciprestes*” (green pines and cypresses), omitting the adjective “翠” (verdant). This omission diminishes the vivid and refined visual imagery that emphasizes the vibrancy and resilience of the trees, and also disrupts the parallel, symmetrical structure of the phrase, which holds aesthetic value. However, the frequency of problematic translations is lower in DeepSeek-V3 and GPT-4o compared to the human translation. For instance, GPT-4o uses retention to translate “阴” and “阳” as “yin” and “yang”, concepts that widely known in the West and therefore unlikely to cause confusion. Similarly, DeepSeek-V3 renders “阳气” (yang *qi*, or yang energy) as “yang”; although this involves partial omission, the meaning remains sufficiently clear and does not negatively impact overall understanding.

By contrast, NovelTrans-J stands out by providing explanatory notes, a feature absent from the outputs of other participants. By providing explanatory notes, NovelTrans-J offers an enriched reading experience, making alien cultural references more accessible to Portuguese readers and bridging the gap between the two languages and cultures. To illustrate this characteristic, we present two examples of the footnotes generated by NovelTrans-J:

- *Pangu*: Pangu é uma figura mitológica chinesa considerada o criador do universo, responsável por separar o céu e a terra.
  (*Pangu*: Pangu is a Chinese mythological figure regarded as the creator of the universe, responsible for separating the sky and the earth).
- *Hongmeng*: Termo filosófico chinês que representa o estado primordial e caótico do universo antes de sua separação em céu e terra.
  (*Hongmeng*: A Chinese philosophical term representing the primordial and chaotic state of the universe before its separation into sky and earth).
- *Doze ramos terrestres* (Zi, Chou, Yin, Mao, Chen, Si, Wu, Wei, Shen, You, Xu, Hai): Sistema tradicional chinês usado para dividir o tempo em ciclos de doze períodos, associados a diferentes características naturais e simbólicas.
  (*Twelve Earthly Branches* (Zi, Chou, Yin, Mao, Chen, Si, Wu, Wei, Shen, You, Xu, Hai): Traditional Chinese system used to divide time into cycles of twelve periods, each associated with different natural and symbolic characteristics).

This section concludes with a summary of the key findings. First, the quantitative evaluation reveals a high degree of stability in the rankings of the five participants, both at the sentence and chunk levels. NovelTrans-J consistently outperforms all other participants across the five evaluated dimensions. By contrast, DeepL Pro exhibits a substantial performance gap relative to the other participants. DeepSeek-V3, GPT-4o, and the human translator show no significant differences in performance, although the human translation receives slightly higher scores at the chunk level.

Second, in terms of cultural appropriateness, an area commonly identified as a limitation in MT systems, NovelTrans-J attains a higher mean score than the human translator. The cultural dimension constitutes the area of greatest advantage for NovelTrans-J, with the most pronounced margin observed across all five dimensions at both the sentence and chunk levels.

Third, the qualitative analysis corroborates the findings of the quantitative evaluation. Specifically: (i) NovelTrans-J and DeepL Pro exhibit the highest and lowest incidence of errors for CSI translation, respectively; (ii) NovelTrans-J further distinguishes itself by providing explanatory notes, a feature absent from the outputs of the other participants. This inclusion of additional information enhances both the quality of the translation and the reading experience for Portuguese readers.

Finally, the analysis suggests that semantic transparency plays a crucial role in the accurate translation of CSIs. A lack of semantic transparency challenges both the MT systems and the human translator, potentially leading to AI hallucinations.

6. Conclusion

This study evaluates the performance of state-of-the-art MT systems in rendering *Journey to the West*, a culturally and linguistically complex 16th-century Chinese novel, into Portuguese. Through a combination of quantitative and qualitative evaluations, several key findings emerged.

First, at both the sentence and chunk levels, three MT systems (DeepSeek-V3, GPT-4o, and NovelTrans-J) demonstrate translation quality comparable to or exceeding that of a published human translation, with NovelTrans-J consistently outperforming all participants across all evaluation dimensions. DeepL Pro, by contrast, significantly lags behind. Second, regarding cultural appropriateness, a dimension traditionally viewed as particularly challenging for MT, NovelTrans-J not only surpasses its machine counterparts but also outperforms the human translator of a published translation. This suggests that, with appropriate system design, MT can effectively mediate complex cultural content. Qualitative analysis of CSIs reinforces these findings: NovelTrans-J made the fewest errors and uniquely provided explanatory notes alongside translations, thereby enhancing both clarity and cultural comprehension for readers. Third, the qualitative analysis highlights that semantic opacity in source texts poses challenges for both the human translator and MT systems, sometimes even leading, in the latter case, to AI hallucination. Overall, these results suggest that high-quality, culturally rich literary translation via MT is becoming increasingly attainable, although certain challenges, such as the handling of semantically opaque expressions, remain significant obstacles.

Traditionally, it has been assumed that human translators possess an inherent advantage over MT systems in literary translation, particularly when dealing with CSIs embedded in complex social and historical contexts. Successfully translating such cultural references requires not only advanced linguistic proficiency but also deep understanding of both the source and target cultures. However, real-world human translation is often constrained by factors such as time pressure and limited compensation, as well as individual differences in language proficiency and cultural awareness. As a result, human translations can fall short of ideal standards, frequently exhibiting notable inaccuracies, as demonstrated in our earlier analysis. By contrast, MT systems powered by large language models have become increasingly competitive. As shown in the case of *Journey to the West*, current MT output can achieve a level of quality comparable to that of published human translations. Given these developments, MT is well positioned to play an increasingly prominent role as a cultural mediator, bridging cultural divides across different literary traditions.

While these findings are promising, several limitations must be acknowledged. First, the scope of the evaluation was relatively limited. Focusing on selected excerpts rather than the full novel may constrain the generalizability of the results. Second, although the evaluation was conducted by an expert, reliance on a single evaluator may introduce bias and affect the reliability of the findings. Future research should address these limitations by expanding the dataset to include larger and more varied literary corpora, incorporating multiple evaluators to enhance reliability, and investigating how different audiences perceive machine-translated literary texts.

Continued innovation in MT holds great potential for widening access to world literatures and fostering intercultural dialogue. We hope this study provides an updated perspective on the current capabilities of MT and offers practical insights to guide the development of future systems that can more accurately capture and transmit the distinctive cultural nuances embedded in literary works.

Acknowledgements

The authors would like to sincerely thank Ana Cristina Alves, for her valuable contribution to the translation evaluation, which represents an essential part of this work. We are also grateful to the reviewers for their constructive comments and insightful suggestions, which improved the quality of this manuscript.

References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., & Zampieri, M. (2021). Findings of the 2021 Conference on Machine Translation (WMT21). In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita & C. Monz (Eds.), *Proceedings of the Sixth*

Conference on Machine Translation (pp. 1–88). Association for Computational Linguistics. https://aclanthology.org/2021.wmt-1.1/

Amenador, K. B., & Wang, Z. (2023). The Image of China as a Destination for Tourist: Translation Strategies of Culture-Specific Items in the Chinese-English Food Menus. *SAGE Open*, *13*(3), 1–14. https://doi.org/10.1177/21582440231196656

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., & Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi & K. Verspoor (Eds.), *Proceedings of the Fourth Conference on Machine Translation – Volume 2: Shared Task Papers, Day 1* (pp. 1–61). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5301

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., & Zampieri, M. (2020). Findings of the 2020 Conference on Machine Translation (WMT20). In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa & M. Negri (Eds.), *Proceedings of the Fifth Conference on Machine Translation* (pp. 1–55). Association for Computational Linguistics. https://aclanthology.org/2020.wmt-1.1/

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., & Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn & J. Kreutzer (Eds.), *Proceedings of the Second Conference on Machine Translation* (pp. 169–214). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4717

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., & Monz, C. (2018). Findings of the 2018 Conference on Machine Translation (WMT18). In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi & K. Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (pp. 272–303). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6401

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari & S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 9-38). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_2

Colina, S. (2008). Translation Quality Evaluation: Empirical Evidence for a Functionalist Approach. *The Translator*, *14*(1), 97–134. https://doi.org/10.1080/13556509.2008.10799251

Colina, S. (2009). Further Evidence for a Functionalist Approach to Translation Quality Evaluation. *Target*, *21*(2), 235–264. https://doi.org/10.1075/target.21.2.02col

Colina, S. (2015). *Fundamentals of Translation*. Cambridge University Press.

Davies, E. E. (2003). A Goblin or a Dirty Nose? The Treatment of Culture-Specific References in Translations of the Harry Potter Books. *The Translator*, *9*(1), 65–100. https://doi.org/10.1080/13556509.2003.10799146

Díaz Cintas, J., & Remael, A. (2007). *Audiovisual Translation: Subtitling*. Routledge. https://doi.org/10.4324/9781315759678

Drugan, J. (2013). *Quality in Professional Translation: Assessment and Improvement*. Bloomsbury.

Efron, B. (1992). Bootstrap Methods: Another Look at the Jackknife. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Methodology and Distribution* (pp. 569-593). Springer. https://doi.org/10.1007/978-1-4612-4380-9_41

Franco Aixelá, J. F. (1996). Culture-Specific Items in Translation. In Á. Román & M. C.-Á. Vidal (Eds.), *Translation, power, subversion* (pp. 52-78). Multilingual Matters.

Gottlieb, H. (2009). Subtitling Against the Current: Danish Concepts, English Minds. In J. Díaz Cintas (Ed.), *New Trends in Audiovisual Translation* (pp. 21-43). Multilingual Matters. https://doi.org/10.21832/9781847691552-004

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Continuous Measurement Scales in Human Evaluation of Machine Translation. In A. Pareja-Lora, M. Liakata & S. Dipper (Eds.), *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 33–41). Association for Computational Linguistics. https://aclanthology.org/W13-2305/

Hague, D., Alan, M., & Zheng, W. (2011). Surveying Translation Quality Assessment: a Specification Approach. *The Interpreter and Translator Trainer*, *5*(2), 243–267. https://doi.org/gr5qdq

House, J. (1997). *Translation Quality Assessment: a Model Revisited*. Narr.

House, J. (2001). Translation Quality Assessment: Linguistic Description versus Social Evaluation. *Meta*, *46*(2), 243–257. https://doi.org/10.7202/003141ar

House, J. (2014). *Translation Quality Assessment: Past and Present*. Routledge. https://doi.org/10.4324/9781315752839

Katan, D. (2009). Translator Training and Intercultural Competence. In S. Cavagnoli, E. Di Giovanni & R. Merlini (Eds.), *La Ricerca nella Comunicazione Interlinguistica. Modelli Teorici e Metodologici* (pp. 282–301). Franco Angeli.

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., & Popović, M. (2022). Findings of the 2022 Conference on Machine Translation (WMT22). In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi & M. Zampieri (Eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 1–45). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.1/

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Koehn, P., Marie, B., Monz, C., Morishita, M., Murray, K., Nagata, M., Nakazawa, T., Popel, M., Popović, M., Shmatova, M., & Suzuki, J. (2023). Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In P. Koehn, B. Haddow, T. Kocmi & C. Monz (Eds.), *Proceedings of the Eighth Conference on Machine Translation* (pp. 1–42). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.1

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., Shmatova, M., Steingrímsson, S., & Zouhar, V. (2024). Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In B. Haddow, T. Kocmi, P. Koehn & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 1–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.1

Lauscher, S. (2000). Translation Quality Assessment: Where Can Theory and Practice Meet? *The Translator*, *6*(2), 149–168. https://doi.org/10.1080/13556509.2000.10799063

Leppihalme, R. (2011). Realia. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of Translation Studies* (pp. 126–130). John Benjamins Publishing Company. https://doi.org/10.1075/hts.2.rea1

Liu, Y., Yao, Y., Zhan, R., Lin, Y., & Wong, D. F. (2024). NovelTrans: System for WMT24 Discourse-Level Literary Translation. In B. Haddow, T. Kocmi, P. Koehn & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 980–986). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.98

Lommel, A. (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In J. Moorkens, S. Castilho, F. Gaspari & S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 109–127). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_6

Marco, J. (2007). The Terminology of Translation: Epistemological, Conceptual and Intercultural Problems and Their Social Consequences. *Target*, *19*(2), 255-269. https://doi.org/10.1075/target.19.2.06mar

Marco, J. (2019). The Translation of Food-Related Culture-Specific Items in the Valencian Corpus of Translated Literature (COVALT) Corpus: A Study of Techniques and Factors. *Perspectives*, *27*(1), 20–41. https://doi.org/10.1080/0907676X.2018.1449228

Munday, J., Pinto, S. R., & Blakesley, J. (2022). *Introducing Translation Studies: Theories and Applications*. Routledge. https://doi.org/10.4324/9780429352461

Newmark, P. (1988). *A Textbook of Translation*. Prentice Hall.

Olk, H. M. (2013). Cultural References in Translation: A Framework for Quantitative Translation Analysis. *Perspectives*, *21*(3), 344–357. https://doi.org/10.1080/0907676X.2011.646279

Pedersen, J. (2005). *How is Culture Rendered in Subtitles?* In H. Gerzymisch-Arbogast & S. Nauert (Ed.), *Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation* (pp. 2–6). MuTra.

Ranzato, I. (2013). *The Translation of Cultural References in the Italian Dubbing of Television Series* [Doctoral thesis]. Imperial College London.

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., C. de Souza, J. G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., & Martins, A. F. T. (2022). CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes,

T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi & M. Zampieri (Eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 634–645). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.60/

Reiss, K., & Vermeer, H. J. (2014). *Towards a General Theory of Translational Action: Skopos Theory Explained* (C. Nord, Trans.). Routledge. https://doi.org/10.4324/9781315759715

Taivalkoski-Shilov, K. (2019). Ethical Issues Regarding Machine(-assisted) Translation of Literary Texts. *Perspectives*, *27*(5), 689–703. https://doi.org/10.1080/0907676X.2018.1520907

Toral, A., & Way, A. (2015). Machine-assisted Translation of Literary Text: a Case Study. *Translation Spaces*, *4*(2), 240–267. https://doi.org/10.1075/ts.4.2.04tor

Toral, A., & Way, A. (2018). What Level of Quality Can Neural Machine Translation Attain on Literary Text? In J. Moorkens, S. Castilho, F. Gaspari & S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 263–287). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_12

Venuti, L. (2017). *The Translator's Invisibility: A History of Translation*. Routledge. https://doi.org/10.4324/9781315098746

Vinay, J.-P., & Darbelnet, J. (2000). A Methodology for Translation (J. C. Sager & M.-J. Hamel, Trans.). In L. Venuti (Ed.), *The Translation Studies Reader* (pp. 84–93). Routledge.

Wang, L., Tu, Z., Gu, Y., Liu, S., Yu, D., Ma, Q., Lyu, C., Zhou, L., Liu, C.-H., Ma, Y., Chen, W., Graham, Y., Webber, B., Koehn, P., Way, A., Yuan, Y., & Shi, S. (2023). Findings of the WMT 2023 Shared Task on Discourse-Level Literary Translation: A Fresh Orb in the Cosmos of LLMs. In P. Koehn, B. Haddow, T. Kocmi & C. Monz (Eds.), *Proceedings of the Eighth Conference on Machine Translation* (pp. 55–67). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.3

Wang, L., Liu, S., Lyu, C., Jiao, W., Wang, X., Xu, J., Tu, Z., Gu, Y., Chen, W., & Wu, M. (2024). Findings of the WMT 2024 Shared Task on Discourse-Level Literary Translation. In B. Haddow, T. Kocmi, P. Koehn & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 699–700). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.58

Way, A. (2012). David Bellos (ed): Is that a fish in your ear: translation and the meaning of everything. *Machine Translation*, *26*(3), 255–269. https://doi.org/10.1007/s10590-012-9129-x

Williams, M. (2004). *Translation Quality Assessment: An Argumentation-Centred Approach*. University of Ottawa Press.

Wu, C-E. (2024). *Jornada ao Oeste* (I. Aldebrand, Trans.). [Kindle edition]. Amazon Kindle.

Editorial notes

Authorship contribution

Conceptualization: D. F. Wong, J. Zhang, M. You
Supervision: J. Zhang, D. F. Wong
Data curation: M. You, K. Lan
Original draft: M. You
Writing – review and editing: J. Zhang, D. F. Wong

Research dataset

Not applicable.

Funding

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0070/2022/AMJ and FDCT/060/2022/AFJ), the National Natural Science Foundation of China (Grant Nos. 62261160648 and 62266013), the China Strategic Scientific and Technological Innovation Cooperation Project (Grant No. 2022YFE0204900), and the University of Macau and the University of Macau Development Foundation (Grant Nos. MYRG-GRG2023-00006-FST-UMDF, MYRG-GRG2024-00165-FST-UMDF, and MYRG-GRG2024-00148-FAH).

Image copyright

Not applicable.

Approval by ethics committee

This research is part of the project *Translation and Research of Journey to the West into Portuguese* (Grant No. MYRG-GRG2024-00148-FAH), approved by the Ethics Committee of the University of Macau.

Conflicts of interest

Not applicable.

Data availability statement

The data from this research, which are not included in this work, may be made available by the author upon request.

Publisher

Cadernos de Tradução is a publication of the Graduate Program in Translation Studies at the Federal University of Santa Catarina. The journal *Cadernos de Tradução* is hosted by the Portal de Periódicos UFSC. The ideas expressed in this paper are the responsibility of its authors and do not necessarily represent the views of the editors or the university.

Guest editors

Xiang Zhang – Li Ye

Section editors

Andréia Guerini – Willian Moura

Style editors

Alice S. Rezende – Ingrid Bignardi – João G. P. Silveira – Kamila Oliveira

Article history

Received: 29-04-2025
Approved: 21-06-2025
Revised: 27-07-2025
Published: 09-2025