ISSN 1799-2591 (print); ISSN 2053-0692 (online)
Theory and Practice in Language Studies, Vol. 16, No. 5, pp. 1775-1784, May 2026
DOI: https://doi.org/10.17507/tpls.1605.34

Evaluating Large Language Models and Neural Machine Translation Systems in Translating *Liaozhaizhiyi*: A Cross-Cultural Literary Study

Jing Zhao
School of Languages, Literacies and Translation, Universiti Sains Malaysia, Penang, Malaysia

Mozhgan Ghassemiazghandi*
School of Languages, Literacies and Translation, Universiti Sains Malaysia, Penang, Malaysia

Shaidatul Akma Adi Kasuma
School of Languages, Literacies and Translation, Universiti Sains Malaysia, Penang, Malaysia

***Abstract*—Neural machine translation has demonstrated strong performance in high-resource languages and commercial translation contexts. However, its effectiveness in translating classical Chinese literature remains insufficiently examined. This study conducts a comparative evaluation of three translation systems—ChatGPT, Google Translate, and Youdao Translate—using selected texts from *Liaozhaizhiyi* as the research corpus. The analysis focuses on four dimensions: semantic alignment, measured by BLEU scores; translation fluency; stylistic fidelity; and the detectability of machine-generated translation patterns. The results indicate that ChatGPT achieves superior performance in stylistic fidelity, particularly in preserving poetic tone, as well as in semantic alignment when translating idiomatic expressions and culturally embedded references. In addition, translations produced by ChatGPT exhibit fewer mechanical artefacts commonly associated with neural machine translation outputs. Further experiments demonstrate that structured prompt engineering strategies contribute to improved literary naturalness and greater cultural coherence in the translated texts. These findings suggest that large language models offer notable advantages in the translation of classical literary works and provide empirical insights into the role of artificial intelligence in facilitating cross-cultural interpretation and the international transmission of Chinese literary traditions.**

***Index Terms*—ChatGPT, Neural machine translation, Google translate, Youdao translate, *Liaozhaizhiyi*, Chinese classical literature, cross-cultural communication**

I. Introduction

Machine translation (MT), first conceptualised by Weaver (1955), has evolved from rule-based machine translation (RBMT) and statistical machine translation (SMT) to neural machine translation (NMT), driven by advances in deep learning and resulting in substantial improvements in fluency and scalability (Castilho et al., 2017; España-Bonet et al., 2016). Despite its widespread deployment in global trade, diplomacy, and education (Koehn, 2009), NMT systems continue to face persistent challenges when translating culture-specific items (CSIs), metaphorical language, poetic structures, and historically embedded expressions (Drobot, 2021; Toral & Way, 2018). These limitations are particularly pronounced in literary translation, where semantic accuracy must be accompanied by stylistic coherence and cultural resonance.

Recent advances in large language models (LLMs) have opened new possibilities for literary translation. LLMs demonstrate strong capabilities in contextual modelling, discourse-level coherence, and stylistic adaptability, all of which are essential for handling complex literary texts (Jiao et al., 2023). However, most existing studies of LLM-based translation focus on what is often termed 'universal text translation', referring to informational or procedural genres such as news reports or technical manuals that are structurally simple and culturally neutral. In contrast, literary translation involves culturally dense and stylistically intricate texts, including historical narratives and classical prose, where meaning is closely intertwined with rhetorical form and cultural symbolism. Recent scholarship indicates that the translation of such works remains underexplored, with systematic evaluations of cultural and stylistic representation still limited in scope (Akabli & Khaloufi, 2024).

The translation of classical literature requires more than the accurate transfer of lexical meaning. It demands the preservation of poetic tone, narrative rhythm, and stylistic subtlety, while minimising mechanical or template-like artefacts commonly associated with machine-generated output. Although NMT systems trained on large parallel corpora achieve strong performance in general translation tasks (Bahdanau et al., 2014; Cho et al., 2014), they frequently struggle

* Corresponding Author. Email: mozhgan.ghassemi@gmail.com

to convey figurative language, poetic structure, CSIs, and historical context, thereby limiting their ability to approximate the adaptive strategies employed by human literary translators (Toral & Way, 2018; Hadley et al., 2019). While Ahrenberg (2017) focuses on general translation, his comparison between human and machine translation suggests that effective literary translation similarly depends on preserving stylistic nuance and cultural depth.

Within this context, ChatGPT, built on a decoder-only Transformer architecture and trained using reinforcement learning from human feedback (RLHF), has attracted increasing attention for its capacity to generate coherent and stylistically adaptive text (Christiano et al., 2017; Jiao et al., 2023). Based on the Generative Pre-trained Transformer (GPT) framework, LLMs have demonstrated strong performance across a wide range of natural language processing tasks (Qin et al., 2023). Nevertheless, their effectiveness in classical literary translation, particularly for texts characterised by dense cultural embedding and stylistic complexity, remains underexplored. *Liaozhaizhiyi*, a canonical work of classical Chinese literature, exemplifies these challenges through its extensive use of allegory, culturally specific imagery, and rhythmically constrained narrative structures that resist literal translation.

Although ChatGPT has demonstrated promising results in general-purpose translation, its performance in translating classical Chinese literature such as *Liaozhaizhiyi* has not been systematically examined. The tales in *Liaozhaizhiyi* present substantial difficulties for MT systems due to their condensed syntax, rich symbolic content, and pervasive metaphorical expression, reflecting broader challenges in translating idiomatic, culturally embedded, and context-dependent language identified in recent MT research (Naveen & Trojovský, 2024). Traditional NMT systems, including Google Translate and Youdao Translate, rely primarily on large-scale parallel corpora and often fail to capture cultural nuance, interpret figurative language, or preserve literary aesthetics, particularly when confronted with non-compositional expressions and idioms that challenge current Transformer-based approaches (Dankers et al., 2022).

In contrast, ChatGPT offers potential advantages in contextual sensitivity and stylistic adaptability through prompt-based interaction and structured guidance, consistent with recent findings on chain-of-thought prompting in LLMs (Wei et al., 2022). However, whether ChatGPT can consistently outperform conventional NMT systems in classical literary translation, and how prompt engineering influences its output, remain open empirical questions.

Recent studies have begun to address these issues. Gao et al. (2024) report that ChatGPT outperforms Google Translate and DeepL in preserving poetic rhythm and imagery in Chinese classical poetry when guided by customised prompts. Zhou and Cheng (2025), comparing ChatGPT with human translation by Ken Liu in Chinese science fiction, find that although ChatGPT produces coherent and structurally sound translations, it often lacks deeper cultural insight and stylistic refinement. Extending this line of inquiry to ancient prose, Si et al. (2024) highlight both the contextual strengths of ChatGPT under prompt support and its limitations in rendering classical style and resolving named entities. Collectively, these findings suggest that while ChatGPT demonstrates advantages in fluency and structural coherence, its capacity to capture the cultural and aesthetic complexity of classical Chinese literature warrants further systematic investigation.

To address this gap, the present study compares ChatGPT with traditional NMT systems in the translation of classical Chinese literary texts into English. Focusing on selected tales from *Liaozhaizhiyi*, the evaluation is structured around four core dimensions: semantic alignment, fluency, stylistic fidelity, and MT style detectability. By combining automatic metrics with expert-based human evaluation, this study aims to provide empirical evidence on the strengths and limitations of LLM-based translation in culturally embedded literary contexts. In addition, the study examines how prompt engineering, particularly context-enriched prompting strategies, shapes the translation behaviour of ChatGPT, thereby contributing to the development of AI-assisted translation approaches that are more sensitive to literary style and cultural nuance.

II. Methodology

A. Materials and Data

This study is based on the bilingual Chinese–English edition of *Liaozhaizhiyi* (2007), published in the Chinese Classics Library series and annotated by literary scholars to ensure cultural authenticity and linguistic fidelity. This edition provides complete original texts, authoritative English translations, and scholarly annotations and therefore serves as a reliable reference for translation analysis. As shown in Table 1, the selected texts include *Niexiaoqian* (《聂小倩》), *The Cricket* (《促织》), *Fox Girl Qingfeng* (《狐女》), *The Snake Eater* (《蛇癖》), and *Candidate for the Post of City God* (《考城隍》). The selected tales were chosen to represent a range of narrative forms and stylistic complexities relevant to the evaluation of literary translation. Each tale foregrounds distinct thematic and stylistic concerns. *The Cricket* satirises imperial bureaucracy and social injustice; *Fox Girl Qingfeng* explores emotional intimacy and moral tension through romantic fantasy; *The Snake Eater* presents a grotesque portrayal of obsession and compulsive desire; and *Candidate for the Post of City God* offers a moral reflection on justice through a supernatural examination system. *Niexiaoqian* centres on the intersection of human–spirit intimacy, moral integrity, and social norms, combining romantic fantasy with ethical restraint and narrative realism. As the longest text in the corpus (2,325 Chinese characters), it provides a comparatively extended narrative structure that enables discourse-level analysis of characterisation, dialogue, and stylistic continuity.

TABLE 1
OVERVIEW OF FIVE SELECTED TEXTS USED IN THE TRANSLATION EXPERIMENTS

Title (English)	Title (Chinese)	Length (Chinese characters)	Notable Features
Niexiaoqian	《聂小倩》	2,325	Extended romantic supernatural narrative focusing on human–spirit relationships; strong ethical orientation and moral restraint; dialogue-driven storytelling; sustained discourse-level stylistic coherence
The Cricket	《促织》	1,816	Satirical portrayal of imperial bureaucracy; culture-specific references, such as tribute systems; integration of supernatural motifs; allegorical narration in a classical prose style
Fox Girl Qingfeng	《狐女》	1,648	Romantic supernatural narrative structured around transformation motifs; culturally embedded symbolism; allegorical expression combined with lyrical and poetic rhythm
The Snake Eater	《蛇癖》	78	Highly condensed grotesque narrative; vivid corporeal imagery; metaphorical depiction of obsession and taboo; intense sensory emphasis
Candidate for the Post of City God	《考城隍》	461	Satirical treatment of religious belief and bureaucracy; evaluative and judgement-oriented language; dreamlike narrative framing moral reasoning; dense use of poetic idioms

B. Translation Systems and Prompt Design

Three translation systems were examined in this study: ChatGPT (GPT-5.2) as a representative large language model, and Google Translate and Youdao Translate as NMT systems based on encoder–decoder architectures. The two NMT systems rely on large-scale parallel corpora and demonstrate strong performance in general-purpose translation tasks but remain limited in preserving stylistic nuance, discourse coherence, and cultural references in literary texts. By contrast, ChatGPT is based on a decoder-only Transformer architecture and incorporates reinforcement learning from human feedback, enabling stronger contextual modelling and adaptive language generation (Christiano et al., 2017; Wei et al., 2022). Its output is highly sensitive to prompt design, making prompting strategies a critical factor in determining translation quality. To examine the role of prompting, two prompt conditions were employed. Prompt 1 was a standard prompt consisting of a direct instruction to translate the source text into English without additional contextual or stylistic guidance. Prompt 2 was an enhanced prompt that guided the model to first interpret the semantic and symbolic content of the text and then produce a style-adaptive translation that preserved literary rhythm, metaphor, and cultural imagery. This two-stage prompt design integrates advanced prompting strategies, such as chain-of-thought and related multi-step decomposition methods, which have been shown to enhance output coherence and other qualitative aspects in LLMs (Gozzi & Di Maio, 2024).

To support transparency and reproducibility, representative prompt formulations are provided. Prompt 1 used a minimal instruction, asking the model to translate a classical Chinese literary text into English. Prompt 2 added historical context, thematic framing, and stylistic constraints to guide the translation process more explicitly. This controlled variation establishes well-defined experimental conditions, allowing system performance to be compared while isolating the effects of structured prompt guidance on translation quality.

The exact prompt formulations used in the experiments are reported below in order to document the input conditions.

Prompt 1 consisted solely of the instruction:

"Translate the following classical Chinese literary text into English."

Prompt 2 required the model to carry out a sequence of internal analytical operations before producing a translation:

Step 1: Semantic interpretation.

Identify the core events, characters, and causal relations in the text. Summarise the central narrative meaning in neutral terms, without introducing interpretation beyond what is implied by the text itself.

Step 2: Cultural and symbolic reasoning.

Identify culturally embedded concepts, symbolic imagery, idiomatic expressions, and historically grounded references present in the text. Explain how these elements function within the narrative context, drawing only on information available in the source text.

Step 3: Stylistic and rhetorical analysis.

Analyse the dominant narrative tone, rhetorical structure, and literary style of the original text, including features such as narrative rhythm, figurative language, evaluative stance, and moral implication.

Step 4: Translation.

On the basis of the preceding analysis, produce a fluent and coherent English literary translation that preserves the original meaning, stylistic tone, and cultural resonance.

The model was instructed to output only the final English translation.

The overall design is consistent with recent developments in interactive MT using LLMs (Wang, 2025) and supports systematic analysis of how prompt engineering shapes semantic reasoning and stylistic simulation in literary translation (Kulkarni et al., 2023).

C. Evaluation Metrics and Dimensions

In this study, translation quality was evaluated across four dimensions: semantic alignment, fluency, stylistic fidelity, and MT style detectability. Semantic alignment was measured using BLEU and BERTScore, which capture surface-level overlap and contextual semantic similarity with reference translations, respectively (Papineni et al., 2002; Zhang et al., 2019). Fluency was assessed through human judgement, with evaluators focusing on grammatical well-formedness and overall readability. BLEU and BERTScore were reported alongside these judgements as complementary automatic indicators of similarity (Papineni et al., 2002; Zhang et al., 2019).

Stylistic fidelity and MT style detectability were evaluated manually. Stylistic fidelity concerned the extent to which translations preserved literary tone, rhetorical structure, and cultural context, with particular attention to narrative and poetic expression (Jing et al., 2019). MT style detectability focused on the presence of mechanical, formulaic, or templated linguistic patterns, with lower scores indicating output perceived as more human-like (Aharoni et al., 2014).

To improve interpretability and balance across evaluation dimensions, a hybrid framework was adopted. Automatic metrics offered efficient and reproducible assessments of surface-level quality, while expert human evaluation addressed rhetorical, contextual, and aesthetic features that remain difficult for automated measures to capture reliably (Toral & Way, 2018; Lau et al., 2024). Manual scoring criteria drew on principles from MT-assisted language learning research, including rhetorical coherence, intonational plausibility, cultural reproduction, and linguistic naturalness (Deng & Yu, 2022).

BLEU scores were calculated using the SacréBLEU Python library with default settings and four-gram overlap, and were reported on a 0–100 scale (Post, 2018). BERTScore was computed using the official implementation with the RoBERTa-large encoder and inverse document frequency rescaling. Manual evaluations employed a five-point Likert scale for both stylistic fidelity and MT style detectability. Automatic and human evaluation results were reported separately but interpreted jointly in order to provide complementary perspectives on translation quality.

The selection of these four dimensions reflects both the practical demands of literary translation and recent calls for more comprehensive, behaviour-oriented evaluation paradigms in natural language processing (Ribeiro et al., 2020). Taken together, they capture semantic accuracy, linguistic naturalness, stylistic depth, and perceived human likeness in translated literary texts.

D. Expert Evaluation Protocol

Expert evaluation was conducted to address the limitations of automatic metrics in capturing literary tone and stylistic nuance. Two dimensions were assessed: stylistic fidelity and MT style detectability. Evaluations were carried out using a five-point Likert scale ranging from very poor to excellent. All source texts were translated in full, and each translation was assessed in its entirety.

Initial scoring was undertaken by the research team. To ensure reliability and consistency, three bilingual scholars with professional expertise in translation studies or comparative literature independently reviewed and validated the scores. These reviewers did not assign new ratings but assessed the coherence and consistency of the existing evaluations against standardised guidelines. Any discrepancies exceeding two points were resolved through discussion prior to finalising the scores.

To reduce potential evaluation bias, assessments were conducted under double-blind conditions. All system identifiers and prompt information were removed, and evaluators were unaware of the translation source. Final scores for each system output were calculated as the mean of the validated ratings.

The evaluation process did not involve human participants, personal data collection, or physiological measurement. All assessments were based solely on anonymised textual materials.

E. Data Collection and Statistical Procedures

ChatGPT under two prompt conditions, Google Translate, and Youdao Translate were used to generate English translations of the selected tales, yielding a total of twenty complete translation samples. Each translation was evaluated using the four predefined dimensions described above. BLEU and BERTScore were employed as automatic indicators of semantic alignment and fluency, while stylistic fidelity and MT style detectability were assessed through expert manual scoring.

To examine overall differences between systems under the standard prompt condition, a one-way analysis of variance was performed. Within-system comparisons for ChatGPT under Prompt 1 and Prompt 2 were analysed using paired t tests. In addition, a linear mixed-effects model was constructed to account for random variation associated with source text and rater identity, with translation system and prompt type specified as fixed effects. Model specification followed the recommendations of Barr et al. (2013), with the alpha threshold set at 0.2 to balance statistical power and control of Type I error (Matuschek et al., 2017).

All analyses were conducted on anonymised textual data. No personal data were collected.

III. Results and Analysis

A. Overall System Performance

Across all evaluation dimensions, ChatGPT outperformed the two NMT systems, Google Translate and Youdao Translate. For cross-system comparisons, the ChatGPT results reported in this section are based on outputs generated

under Prompt 2 (the context-enriched prompt), which was identified as the model's most effective configuration in the automatic evaluation (see Table 2).

As illustrated in Figure 1, performance was compared across four dimensions—semantic alignment, fluency, stylistic fidelity, and MT style detectability—for ChatGPT under both prompt conditions (Prompt 1 and Prompt 2), as well as for Google Translate and Youdao Translate. Figure 1 presents aggregated expert ratings across all systems and conditions, providing an overview of relative performance patterns.

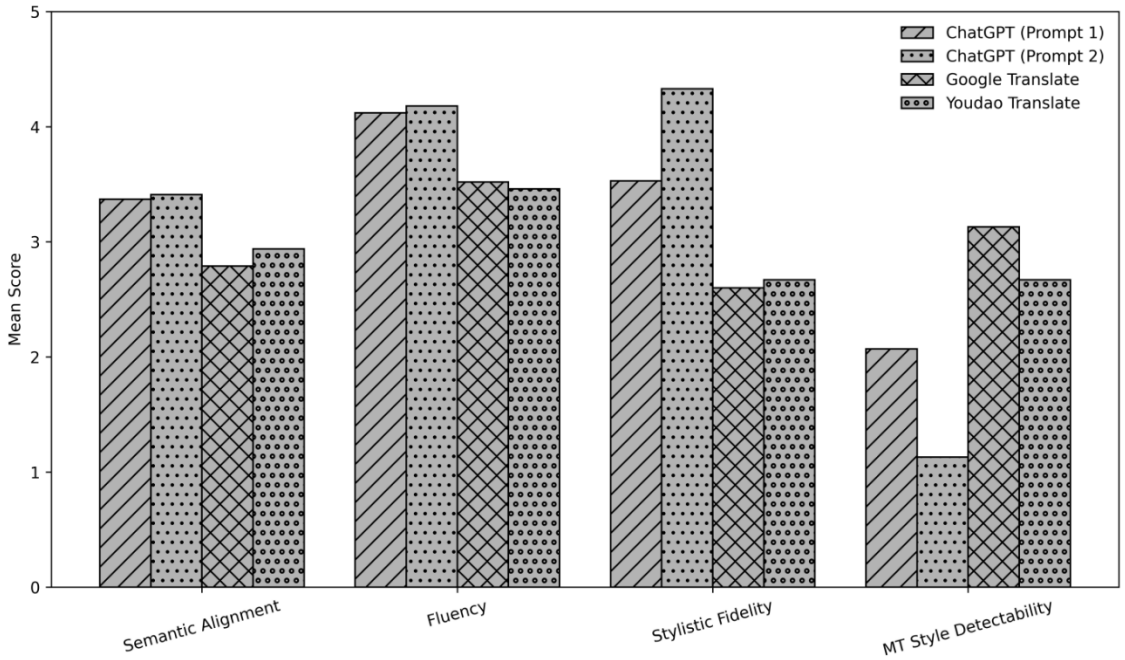

Figure 1. Expert Evaluation Results of Translation Quality Across Systems and Evaluation Dimensions

With respect to semantic alignment and fluency, expert evaluators relied on qualitative judgement informed by close reading and comparative interpretation rather than numerical scoring. Across the texts examined, ChatGPT outputs were consistently judged to display stronger discourse coherence, smoother information flow, and more contextually appropriate reconstruction of source meaning than those produced by Google Translate and Youdao Translate. By contrast, the NMT systems more frequently exhibited fragmented semantic relations and weaker sensitivity to implicit narrative context, particularly in passages containing culturally specific references.

Quantitative expert ratings reinforce these qualitative observations (Matuschek et al., 2017). For stylistic fidelity, ChatGPT under Prompt 2 achieved the highest mean score ($M = 4.33$, $SD = 0.49$), substantially exceeding the scores for Google Translate ($M = 2.60$, $SD = 0.51$) and Youdao Translate ($M = 2.67$, $SD = 0.49$). A one-way analysis of variance confirmed a statistically significant effect of system type on stylistic fidelity, $F(2, 42) = 59.18$, $p < 0.001$. This result indicates a markedly stronger capacity on the part of ChatGPT to preserve literary tone, rhythmic patterning, and stylistic features characteristic of classical Chinese prose.

A comparable pattern emerged for MT style detectability, where lower scores correspond to translations that are less readily perceived as machine-generated. ChatGPT under Prompt 2 recorded the lowest mean detectability score ($M = 1.13$, $SD = 0.35$), compared with Google Translate ($M = 3.13$, $SD = 0.64$) and Youdao Translate ($M = 2.67$, $SD = 0.49$). The effect of system type was again statistically significant, $F(2, 42) = 63.86$, $p < 0.001$, suggesting that ChatGPT outputs contain fewer mechanical or formulaic features typically associated with NMT.

Results from automatic evaluation further support these findings. As reported in Table 2, ChatGPT, particularly under the context-enriched prompt condition (Prompt 2), achieved higher average BLEU and BERTScore values across the classical texts. These scores indicate stronger surface-level semantic overlap and embedding-based similarity with the reference translations. Consistent with previous research, however, automatic metrics remain limited in their ability to capture stylistic nuance and cultural resonance, underscoring the importance of complementary expert evaluation.

TABLE 2
AUTOMATIC EVALUATION RESULTS

System	BLEU	BERTScore
ChatGPT (Prompt 1: standard prompt)	15.2	0.789
ChatGPT (Prompt 2: context-enriched prompt)	14.8	0.795
Google Translate	15.3	0.664
Youdao Translate	13.9	0.703

B. Effect of Prompt Design on ChatGPT Performance

As Google Translate and Youdao Translate do not support prompt-controlled input, the impact of prompt design was examined solely within ChatGPT by comparing outputs generated under Prompt 1 (standard prompt) and Prompt 2 (context-enriched prompt).

Expert evaluations indicate that prompt design has a pronounced effect on dimensions requiring higher-level interpretive judgement. For stylistic fidelity, translations produced under Prompt 2 were consistently rated more highly than those generated under Prompt 1, reflecting stronger preservation of literary tone, stylistic features, and rhetorical coherence. A similar pattern was observed for MT style detectability, with Prompt 2 yielding lower detectability scores, suggesting a reduced presence of mechanical or formulaic linguistic patterns commonly associated with machine-generated translation.

By contrast, automatic evaluation metrics reveal only modest differences between the two prompt configurations. As shown in Table 2, ChatGPT achieved a BLEU score of 15.2 under Prompt 1 and 14.8 under Prompt 2, while BERTScore increased slightly from 0.789 to 0.795. These small variations suggest that context-enriched prompting does not substantially affect surface-level semantic overlap or embedding-based similarity with reference translations, and may involve minor trade-offs in n-gram overlap.

Taken together, these findings indicate that the principal contribution of context-enriched prompting lies not in improving surface-level accuracy as captured by automatic metrics, but in supporting higher-order stylistic reconstruction and cultural adaptation, as reflected in expert judgement. Explicit cultural and contextual guidance appears to enable ChatGPT to move beyond baseline linguistic competence and more closely approximate human literary translation practices.

C. Dimension-Based Case Analysis

To illustrate system differences while maintaining analytical conciseness, one representative case is presented for each evaluation dimension.

(a). Semantic Alignment

A representative example was drawn from *Fox Girl Qingfeng*, focusing on the culturally embedded expression '秋波流慧'. Both Google Translate and Youdao Translate rendered the phrase as 'bright and intelligent eyes', which conveys the basic propositional meaning but fails to capture its metaphorical richness. ChatGPT under Prompt 1 produced 'her intelligent eyes shone with wisdom', partially restoring figurative nuance. Under Prompt 2, the phrase was translated as 'her gaze filled with an ethereal intelligence', offering a more effective reconstruction of metaphor and cultural resonance. This example illustrates ChatGPT's stronger capacity for context-sensitive semantic alignment in literary discourse.

(b). Fluency

Fluency was examined using a long, verb-dense sentence from *The Cricket*. Google Translate and Youdao Translate preserved core propositional content but generated syntactically rigid and rhythmically uneven output. ChatGPT under Prompt 1 improved clause variation and logical cohesion, while Prompt 2 further refined narrative rhythm by reshaping sentence flow into more natural literary prose. This case highlights ChatGPT's advantage in managing sentence rhythm and discourse coherence in the translation of classical narrative texts.

(c). Stylistic Fidelity

Stylistic fidelity was analysed through a symbolic passage from *The Cricket* involving ritualised action and narrative suspense. The NMT outputs were semantically adequate but stylistically flat, lacking allegorical tone and atmospheric tension. ChatGPT under Prompt 1 preserved narrative pacing but remained stylistically restrained. By contrast, Prompt 2 achieved a closer approximation of the source text's symbolic cadence and tonal restraint, demonstrating improved stylistic fidelity through controlled elaboration.

(d). MT Style Detectability

MT style detectability was illustrated using a concise emotional statement from *Candidate for the Post of City God*. The NMT systems produced literal translations with clear traces of structural alignment to the source text. ChatGPT under Prompt 1 reduced some mechanical features but remained closely tied to source structure. Prompt 2 generated a more idiomatic and rhetorically elevated rendering, substantially reducing the detectability of machine-generated style. At the same time, this example highlights the need to balance stylistic enhancement with semantic discipline, as excessive elaboration may introduce elements not fully grounded in the source text.

D. Interim Summary

Taken together, the qualitative and quantitative analyses indicate that ChatGPT consistently outperforms traditional NMT systems in semantic alignment, fluency, stylistic fidelity, and MT style detectability when translating classical Chinese literature. While enhanced prompting does not substantially affect surface-level accuracy, it markedly improves stylistic reconstruction and cultural sensitivity. These findings underscore the importance of prompt design as a key mechanism for aligning large language model translation output with the norms and expectations of literary translation.

IV. Discussion

This study examined the performance of ChatGPT and two widely used NMT systems, Google Translate and Youdao Translate, in translating four short tales from *Liaozhaizhiyi* across four dimensions: semantic alignment, fluency, stylistic fidelity, and MT style detectability. The findings show that ChatGPT consistently outperforms the two NMT systems across all dimensions, with particularly marked advantages in stylistic fidelity and MT style detectability. To account for these differences, this section interprets the results from several perspectives, including underlying model mechanisms, prompt design strategies, structural limitations of conventional NMT systems, and insights derived from the case-based analysis.

A. Sources of ChatGPT's Translation Advantages

ChatGPT's stronger performance in literary translation can be attributed primarily to its capacity for contextual modelling and flexible language generation as a large language model (Brown et al., 2020; Karpinska & Iyyer, 2023). Training techniques such as reinforcement learning with human feedback and the use of chain-of-thought-style reasoning contribute to its ability to handle ambiguous metaphors, culturally embedded expressions, and stylistic nuance (Wei et al., 2022; Gao et al., 2024). Under Prompt 2, the structured, context-enriched design explicitly engages processes of semantic interpretation and stylistic reconstruction, resulting in marked improvements in stylistic fidelity and cultural adaptation. These mechanisms allow ChatGPT to represent narrative shifts and symbolic imagery more effectively, elements that are central to the literary texture of *Liaozhaizhiyi*.

B. Structural Limitations of NMT Systems

By contrast, Google Translate and Youdao Translate rely primarily on encoder–decoder architectures trained on large-scale parallel corpora. Although effective for general-purpose translation, such systems face structural limitations when applied to literary genres characterised by symbolic language, classical syntax, and culturally specific rhetoric. The recurrent production of stylistically flat output, template-like constructions, and semantically simplified renderings observed in the translations reflects these constraints (Guerberof-Arenas & Toral, 2022; Karpinska & Iyyer, 2023). For example, the expression '母年七十无所依倚' in *Candidate for the Post of City God* was translated as a neutral statement of information, failing to convey the Confucian ethical framework and emotional restraint embedded in the original narrative.

C. Mechanism and Risks of Prompt Strategy

Prompt design plays a central role in shaping the performance of LLM-based translation. In this study, Prompt 2, which combines chain-of-thought-style reasoning with a least-to-most structure, proved effective in enhancing textual coherence and cultural embedding (Zhang et al., 2023). At the same time, the analysis highlights potential risks associated with high-flexibility prompting strategies. In some instances, ChatGPT produced stylistically elaborate renderings that introduced details not fully supported by the source text, resulting in overly literary output and reduced semantic fidelity. This tendency reflects the broader risk of hallucination associated with generative models and underscores the importance of maintaining balanced prompt constraints when translating culturally sensitive literary material (Zhang et al., 2019).

D. Implications From Case-Based Analysis

Analysis of representative passages characterised by dense cultural and rhetorical features indicates that ChatGPT under Prompt 2 performs most strongly in stylistic fidelity and MT style detectability. Under this configuration, the model is better able to reconstruct emotionally charged and culturally nuanced expressions, effectively conveying themes such as moral conflict, compassion, and reversals of fate that are central to classical Chinese narrative. At the same time, the comparatively stable semantic performance of the NMT systems contrasts with their recurrent stylistic flattening and distortion, reinforcing limitations identified in earlier work (Toral & Way, 2018). Taken together, these findings extend recent discussions on the role of prompt mechanisms in translation tasks involving high levels of cultural embedding (Lau et al., 2024).

E. Practical Prospects and Future Research Directions

Although this study concentrates on classical narrative prose, the applicability of LLM-based translation to other genres remains to be examined systematically. This includes poetry, low-resource languages, and texts with complex syntactic or rhetorical structures. Future work could explore adaptive prompt designs, cross-linguistic transferability of prompt strategies, and hybrid translation workflows that incorporate human post-editing. Further research is also needed to assess the role of LLM-based translation systems in educational settings, cultural dissemination, and broader forms of cross-cultural communication. Taken together, the prompt enhancement approach examined in this study offers a practical avenue for improving LLM performance in culturally embedded literary translation and provides a methodological reference point for subsequent empirical research and system development.

V. Conclusion

This study evaluated the performance of ChatGPT and two widely used NMT systems, Google Translate and Youdao Translate, in translating four classical narrative tales from *Liaozhaizhiyi*. Translation quality was assessed across four dimensions: semantic alignment, fluency, stylistic fidelity, and MT style detectability. The findings indicate that ChatGPT consistently outperforms the traditional NMT systems across all dimensions, with particularly pronounced advantages in stylistic fidelity and MT style detectability.

Further analysis demonstrates that the structured, context-enriched prompt (Prompt 2) plays a decisive role in enhancing ChatGPT's capacity to reconstruct literary style and cultural nuance. By explicitly guiding semantic interpretation and stylistic reconstruction, this prompt strategy reduces mechanical language patterns while improving aesthetic quality and cultural resonance in translated literary texts. These results support the practical effectiveness of chain-of-thought reasoning and least-to-most prompting in complex literary translation tasks. At the same time, the findings also indicate that context-enriched prompting may introduce risks of excessive rhetorical elaboration or source-unsupported additions in certain cases. This underscores the need to balance stylistic enrichment with semantic fidelity and interpretive discipline.

The primary limitation of this study lies in its exclusive focus on classical narrative prose, without extending the analysis to other challenging text types such as poetry, low-resource languages, or abstract philosophical writing. Future research could examine adaptive prompt strategies across genres, cross-linguistic transferability, and hybrid translation workflows that incorporate human post-editing. Overall, the stylistic sensitivity and cross-cultural expressive capacity demonstrated by ChatGPT point to the broader potential of large language models as tools for the international dissemination of Chinese literary heritage and for fostering cross-cultural understanding. With carefully calibrated prompt design and output control, LLM-based translation systems may become an important technical medium for cross-linguistic literary exchange and global cultural communication.

References

[1] Aharoni, R., Koppel, M., & Goldberg, Y. (2014). Automatic Detection of Machine-translated Text and Translation Quality Estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 289–295). https://doi.org/10.3115/v1/P14-2048

[2] Ahrenberg, L. (2017). Comparing Machine Translation and Human Translation: A Case Study. In *Proceedings of the Workshop on Human-Informed Translation and Interpreting Technology* (pp. 21–28). Association for Computational Linguistics: Copenhagen, Denmark. Retrieved December 29, 2025, from https://aclanthology.org/W17-7903/

[3] Akabli, J., & Khaloufi, R. (2024). Translating identity in Leila Abouzeid's Return to Childhood. *AWEJ for Translation & Literary Studies*, *8*(2), 2–17. https://doi.org/10.24093/awejtls/vol8no2.1

[4] Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate.* arXiv preprint, arXiv:1409.0473. https://arxiv.org/abs/1409.0473

[5] Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://arxiv.org/abs/2005.14165

[7] Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art? *Prague Bulletin of Mathematical Linguistics*, *108*, 109–120. https://doi.org/10.1515/pralin-2017-0013

[8] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). https://doi.org/10.3115/v1/D14-1179

[9] Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*, 4299–4307. https://arxiv.org/abs/1706.03741

[10] Drobot, I.-A. (2021). Translating literature using machine translation: Is it really possible? *Scientific Bulletin of the Politehnica University of Timişoara: Transactions on Modern Languages*, *20*(1), 57–64. https://doi.org/10.59168/FUAP6124

[11] Dankers, V., Lucas, C. G., & Titov, I. (2022). *Can transformers be too compositional? Analysing idiom processing in neural machine translation.* arXiv preprint, arXiv:2205.15301. https://doi.org/10.48550/arXiv. 2205.15301

[12] Deng, X., & Yu, Z. (2022). A systematic review of machine-translation-assisted language learning for sustainable education. *Sustainability*, *14*(13), 7598. https://doi.org/10.3390/su14137598

[13] España-Bonet, C., Costa-Jussà, M. R., Rapp, R., Lambert, P., Eberle, K., Banchs, R. E., & Babych, B. (2016). Hybrid machine translation overview. In *Hybrid Approaches to Machine Translation* (pp. 1–24). Springer Cham. https://doi.org/10.1007/978-3-319-21311-8

[14] Gao, R., Lin, Y., Zhao, N., & Cai, Z. G. (2024). Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, *11*(1), Article 835. https://doi.org/10.1057/s41599-024-03363-0

[15] Gozzi, M., & Di Maio, F. (2024). Comparative analysis of prompt strategies for large language models: Single-task vs. multitask prompts. *Electronics*, *13*(23), 4712. https://doi.org/10.3390/electronics13234712

[16] Guerberof-Arenas, A., & Toral, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, *11*, 184–212. https://doi.org/10.1075/ts.21025.gue

[17] Hadley, J., Popović, M., Afli, H., & Way, A. (2019). *Proceedings of the Qualities of Literary Machine Translation.* European Association for Machine Translation, Dublin, Ireland. Retrieved December 7, 2025, from https://aclanthology.org/W19-7300/
[18] Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023). *Is ChatGPT a good translator? Yes with GPT-4 as the engine: A preliminary study.* arXiv preprint, arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745
[19] Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2019). Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics*, *26*(11), 3365–3385. https://doi.org/10.1109/TVCG.2019.2921336
[20] Koehn, P. (2009). *Statistical machine translation.* MIT Press.
[21] Kulkarni, A., Shivananda, A., Kulkarni, A., & Gudivada, D. (2023). The ChatGPT architecture: An in-depth exploration of OpenAI's conversational language model. In *Applied generative AI for beginners: Practical knowledge on diffusion models, ChatGPT, and other LLMs* (pp. 55–77). Apress. https://doi.org/10.1007/978-1-4842-9994-4_4
[22] Karpinska, K., & Iyyer, M. (2023). Large language models effectively leverage document-level context for machine translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023), Volume 1: Research Papers* (pp. 478–489). Retrieved December 7, 2025, from https://aclanthology.org/2023.wmt-1.41/
[23] Lau, J., Wang, Y., & Tang, G. (2024). Improving BERTScore for machine translation evaluation through contrastive learning. *IEEE Access*, *12*, 77739–77749. https://doi.org/10.1109/ACCESS.2024.3406993
[24] Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001
[25] Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience*, *27*(10), 110878. https://doi.org/10.1016/j.isci.2024.110878
[26] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). https://doi.org/10.3115/1073083.1073135
[27] Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT) Volume 1: Research Papers* (pp. 186–191). https://doi.org/10.48550/arXiv.1804.08771
[28] Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). *Is ChatGPT a general-purpose natural language processing task solver?* arXiv preprint, arXiv:2302.06476. https://doi.org/10.48550/arXiv.2302.06476
[29] Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4902–4912). Retrieved December 7, 2025, from https://aclanthology.org/2020.acl-main.442/
[30] Si, S., Zhou, S., & Zhang, Y. (2024). Exploring the capabilities of ChatGPT in ancient Chinese translation and person name recognition. *Corpus-Based Studies across Humanities*, *2*, 221–234. https://doi.org/10.1515/csh-2024-0017
[31] 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. https://doi.org/10.1007/978-3-319-91241-7_12
[32] Weaver, W. (1955). Translation. In W. N. Locke & A. D. Booth (Eds.), *Machine translation of languages: Fourteen essays* (pp. 15–23). MIT Press.
[33] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903
[34] Wang, Q. (2025). Evaluating Uighur literary translation: A comparative study of ChatGPT, Google Translate, and Bing Translator. *PLoS ONE*, *20*, e0335261. https://doi.org/10.1371/journal.pone.0335261
[35] Zhou, P., & Cheng, J. (2025). Stylistic variation across English translations of Chinese science fiction: Ken Liu versus ChatGPT. *Frontiers in Artificial Intelligence*, *8*, Article 1576750. https://doi.org/10.3389/frai.2025.1576750
[36] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). *BERTScore: Evaluating text generation with BERT.* arXiv preprint, arXiv:1904.09675. https://doi.org/10.48550/arXiv.1904.09675
[37] Zhang, B., Haddow, B., & Birch, A. (2023). *Prompting large language model for machine translation: A case study.* arXiv preprint, arXiv:2301.07069. https://doi.org/10.48550/arXiv.2301.07069

Zhao Jing is a Ph.D. candidate in Translation Studies at the School of Languages, Literacies and Translation, Universiti Sains Malaysia. She holds a Master's degree in Translation and Interpreting and a Bachelor's degree in English Language and Literature. She also completed postgraduate studies in Civil and Commercial Law. She has professional experience as an English lecturer and translator in higher education and industry settings. Her research interests include literary translation, machine translation evaluation, large language models, and Chinese classical literature.

Mozhgan Ghassemiazghandi, Ph.D., is a Senior Lecturer at the School of Languages, Literacies and Translation, Universiti Sains Malaysia. Her research interests focus on Translation Technology, Machine Translation, and Audiovisual Translation. In addition to her academic work, she is a professional translator and subtitler with more than a decade of industry experience, combining theoretical insight with practical expertise in translation practice and technology.

Shaidatul Akma Adi Kasuma is an Associate Professor at the School of Languages, Literacies and Translation, Universiti Sains Malaysia (USM). She obtained her Bachelor in Education with Computer Science (TESL) from Universiti Teknologi Malaysia (UTM), Master in Linguistics from Universiti Malaya (UM), and PhD in Arts Education from the University of Warwick, United Kingdom. She is interested and has published extensively in the fields of TESL, Applied Linguistics, Technology in English Language Learning, Applied Linguistics, and the language of sustainability communication to fit in with the global agenda.