Evaluating Machine Translation of the *Shan Hai Jing*: An MQM-Based Analysis of Google Translate vs. ChatGPT with Prompting Effects

(Penilaian Terjemahan Mesin bagi *Shan Hai Jing*: Analisis Berasaskan MQM terhadap Google Translate dan ChatGPT dengan Kesan Strategi Prompt)

WENQI DUAN [1], CHWEE FANG NG [1*], HAZLINA ABDUL HALIM [1] & ZHONGMING ZHANG [1]

[1] *Faculty of Modern Languages and Communication*
Universiti Putra Malaysia Serdang, Selangor, Malaysia

Received: 13 April 2026 / Accepted: 19 May 2026

ABSTRACT

Culturally dense classical texts pose persistent challenges for machine translation, particularly in reconstructing compressed semantic hierarchies and culture-specific references. Although Neural Machine Translation (NMT) and Large Language Models (LLMs) have substantially improved fluency and contextual coherence, previous studies have given limited attention to the evaluation of their performance on culturally embedded classical texts using MQM-based human evaluation alongside automatic metrics. Focusing on the English translation of the Shan Hai Jing, a culturally dense and semantically complex classical Chinese text, this study investigates whether different translation systems produce distinct error patterns in culturally compressed contexts and whether prompting strategies influence translation performance. Selected textual segments were translated using Google Translate and ChatGPT under minimal and enriched prompting strategies. Translation quality was assessed through MQM-based human evaluation alongside several automatic metrics (BLEU, chrF, BERTScore, and COMET-Kiwi). MQM analysis reveals clear differences in error patterns across systems: NMT outputs show a higher incidence of high-severity mistranslations, whereas LLM outputs tend to exhibit semantic generalisation and shifts in cultural references. By contrast, automatic metrics show limited differentiation in system rankings, with no significant main effect of system observed. Prompt enrichment does not produce consistent quality improvements and occasionally increases semantic drift. These findings suggest that translation quality in culturally compressed texts may be better interpreted through structural error patterns across MQM dimensions rather than metric-based rankings alone. Evaluation sensitivity appears to be shaped by text type, and increased prompt complexity does not necessarily enhance semantic precision in classical translation tasks.

Keywords: Neural Machine Translation (NMT); Large Language Models (LLMs); Multidimensional Quality Metrics (MQM); Shan Hai Jing; Prompt Strategies

ABSTRAK

Teks klasik yang sarat dengan unsur budaya sering menimbulkan cabaran kepada sistem terjemahan mesin, khususnya dalam membina semula hierarki semantik yang padat serta rujukan budaya yang khusus. Walaupun Terjemahan Mesin Neural (Neural Machine Translation, NMT) dan Model Bahasa Besar (Large Language Models, LLMs) telah meningkatkan kelancaran bahasa dan koherensi konteks secara ketara, kajian terdahulu masih kurang memberi perhatian terhadap penilaian prestasi sistem ini dalam teks klasik yang berunsur budaya dengan menggunakan penilaian manusia berasaskan MQM bersama metrik automatik. Dengan memfokuskan pada terjemahan bahasa Inggeris bagi Shan Hai Jing, sebuah teks klasik Cina yang terkenal dengan kepadatan semantik, rujukan mitologi, dan unsur budaya yang kompleks, kajian ini bertujuan untuk menyiasat sama ada sistem terjemahan yang berlainan menghasilkan corak kesilapan yang berbeza dalam konteks yang padat dengan unsur budaya serta

[*] *Corresponding Author: chweefang@upm.edu.my*

eISSN: 2550-2131
ISSN: 1675-8021

sama ada strategi prompt mempengaruhi prestasi terjemahan. Segmen teks terpilih diterjemahkan menggunakan Google Translate dan ChatGPT di bawah dua strategi prompt, iaitu prompt minimum dan prompt diperkaya. Kualiti terjemahan dinilai melalui penilaian manusia berasaskan Multidimensional Quality Metrics (MQM) di samping beberapa metrik automatik seperti BLEU, chrF, BERTScore dan COMET-Kiwi. Analisis MQM menunjukkan perbezaan yang jelas dalam corak kesilapan antara sistem: output NMT menunjukkan kadar kesilapan salah terjemahan berkeparahan tinggi yang lebih tinggi, manakala output LLM cenderung memperlihatkan penggeneralisasian makna serta peralihan dalam rujukan budaya. Sebaliknya, metrik automatik menunjukkan perbezaan yang terhad dalam pemeringkatan sistem tanpa kesan utama sistem yang signifikan. Pengayaan prompt tidak menghasilkan peningkatan kualiti yang konsisten dan dalam beberapa kes meningkatkan penyimpangan semantik. Dapatan ini mencadangkan bahawa kualiti terjemahan dalam teks yang padat dengan makna budaya lebih sesuai ditafsirkan melalui corak kesilapan struktur merentas dimensi MQM berbanding penilaian berasaskan metrik semata-mata. Selain itu, sensitiviti penilaian turut dipengaruhi oleh jenis teks, dan peningkatan kerumitan prompt tidak semestinya meningkatkan ketepatan semantik dalam terjemahan teks klasik.

Kata kunci: Terjemahan Mesin Neural (NMT); Model Bahasa Besar (LLMs); Multidimensional Quality Metrics (MQM); Shan Hai Jing; Strategi Prompt

INTRODUCTION

In recent years, developments in Machine Translation (MT) and Large Language Models (LLMs) have significantly advanced research in language technologies. Transformer-based Neural Machine Translation (NMT) systems have achieved substantial progress in translation fluency and semantic modeling (Wu et al., 2016; Vaswani et al., 2017), while LLMs further extend the capacity for semantic integration and context-sensitive generation (Lyu et al., 2023). At the same time, Translation Quality Assessment (TQA) remains a central concern in MT research (Jiang et al., 2024).

Despite improvements in linguistic naturalness and surface-level coherence, culturally dense semantic expressions remain a persistent challenge for current translation systems (Shen, Wang, & Birch, 2025). When translations involve historical allusions, religious symbolism, social institutions, or culturally embedded imagery, deviations often arise at the level of deeper semantic structure and cultural reference (Jiang et al., 2024; Yao & Fan, 2025). This tension is particularly evident in culture-specific items (CSIs), which reflect meanings embedded within specific cultural contexts (Newmark, 1988) and carry complex historical, social, and symbolic dimensions (Wang, Amini, & Tan, 2025). In texts marked by strong semantic and cultural condensation, translation performance is shaped not only by how meaning is generated in translation outputs but also by how those outputs are evaluated. Examining both the production of translations and the methods used to assess them therefore offers a useful perspective for analysing MT performance in culturally dense texts.

As a prototypical culturally and semantically dense classical text, the *Shan Hai Jing* contains numerous mythological entities, cultural symbols, and historically condensed expressions. It thus provides an appropriate context for examining how different translation systems represent meaning and process cultural information. Evaluating MT systems in such a textual environment enables a more refined understanding of the relationship among linguistic fluency, cultural-semantic representation, and evaluation approaches.

Existing studies have compared MT performance across genres such as technical texts, literary prose, and poetry. These investigations have addressed various dimensions, including terminological accuracy (Peng & Yvon, 2023), cultural transmission and stylistic features (Fakih

et al., 2024; He et al., 2024), and the handling of rhetoric and imagery (Dunder et al., 2021; Gao et al., 2024).

However, compared to these genres, machine translation in Chinese classical texts has received relatively limited systematic empirical attention, particularly with regard to cultural-semantic reconstruction. Existing literature indicates that NMT continues to encounter difficulties in handling CSIs and context-dependent meanings (Jiang et al., 2024; Yao & Fan, 2025). Although LLMs demonstrate stronger generative capacity, no consistent conclusions have yet emerged regarding their performance in terms of cultural-semantic precision and semantic consistency (Shi et al., 2024; Li & Chen, 2025). This suggest that the discriminative capacity and sensitivity of automatic evaluation metrics in cultural condensed texts have not yet been adequately validated, leaving the differences in error patterns across translation systems and their impact on the transmission of cultural meaning insufficiently understood.

In this context, a key gap remains in the existing literature: there is a lack of integrated empirical investigation into translation performance in Chinese classical texts, particularly regarding how evaluation sensitivity relates to system-specific error patterns in culturally dense contexts. Building on this, prompt strategies, as a key factor influencing the generation pathways of LLMs, have yet to be systematically examined in terms of how they shape these differences in the context of classical texts.

Building on these gaps, this study aims to compare the performance of NMT and LLM-based translation systems in culturally and semantically dense texts using the *Shan Hai Jing* as an empirical case. The study combines MQM-based human assessment with automatic evaluation metrics to examine differences in overall translation quality and error patterns across systems, and further investigates how prompting strategies influence ChatGPT's semantic choices and the distribution of translation errors. Accordingly, this study addresses the following research questions:

1. To what extent do automatic evaluation metrics differentiate between NMT and LLM outputs in the English translation of the *Shan Hai Jing*?
2. How do different translation systems diverge in their error structures under the MQM framework, particularly in culturally specific and semantically dense dimensions?
3. How do different prompting strategies affect ChatGPT's translation quality in the translation of the *Shan Hai Jing*?

Through this integrated approach, the study examines the relationship between translation generation, evaluation methods, and prompting strategies in classical texts, in order to clarify how MT systems perform in culturally dense semantic environments.

LITERATURE REVIEW

COMPARATIVE STUDIES OF NMT AND LLMs IN CLASSICAL TEXT TRANSLATION

Text type is widely recognised as a key variable influencing MT performance. Classical Chinese texts are characterised by dense cultural references, culture-specific items, and strong contextual dependency. Such semantic compression and cultural embedding pose persistent challenges for machine translation systems, making these texts a useful context for examining differences across translation architectures in terms of semantic representation and cultural reconstruction.

Existing research indicates that NMT has achieved substantial improvements in fluency and surface-level coherence compared with earlier models (Rivera-Trigueros, 2022). However, these improvements are largely confined to surface-level features and do not consistently extend to culturally and semantically dense texts. This pattern is also reflected in prior research, which indicates that NMT continues to exhibit notable limitations in the representation of cultural imagery and deeper semantic meaning (Gao et al., 2024; Jiang et al., 2024). Cross-genre studies further indicate that NMT performs more effectively in technical and informational texts than in culture-loaded domains such as literary and religious discourse, where reduced semantic precision, information omission, and contextual inappropriateness are more frequently observed (Zhang et al., 2025).

LLM-based translation systems demonstrate different quality characteristics. Some studies report that ChatGPT demonstrates strengths in discourse coherence and linguistic naturalness (Wang et al., 2023; Yao & Fan, 2025). However, compared with NMT systems, LLM outputs often exhibit greater variability in semantic precision and cultural stability, with deviations arising in the rendering of culturally symbolic expressions (Shi et al., 2024; Li & Chen, 2025). The quality of LLM outputs also appears to be task-dependent, with risks of semantic drift and hallucination particularly evident in culturally embedded expressions and low-resource settings (Jiao et al., 2023). While NMT systems tend to exhibit semantic inadequacy in highly culture-specific contexts, LLMs, despite their fluency advantages, continue to generate debate regarding semantic precision and cultural stability. Overall, although previous studies have identified differences in translation quality between NMT and LLMs from a range of perspectives, such comparisons tend to remain at the level of aggregate performance or isolated dimensions and do not provide a systematic account at the cultural–semantic level. This limitation is particularly evident in culturally and semantically dense contexts, where it remains unclear whether different systems exhibit consistent error patterns and how far such differences affect the transmission of culturally embedded meanings.

CHATGPT AND PROMPTING STRATEGIES IN TRANSLATION

In LLM-based translation research, prompting strategies have emerged as an important area of investigation. ChatGPT is widely considered sensitive to prompt design, with prompting strategies potentially influencing translation quality (Jiao et al., 2023), particularly in terms of semantic selection pathways, contextual interpretation, and target-language formulation strategies. Existing research has systematically examined its impact on translation quality across multiple dimensions, including role specification, task instructions, and domain information. For instance, role-based prompts may enhance translation naturalness (He, 2024), while information about translation purpose and target audience may contribute to improved translation quality (Yamada, 2023). Task and domain information may also produce positive effects in high-resource settings, although their stability varies across linguistic conditions (Gao et al., 2023; Peng et al., 2023). These studies suggest that prompts function not merely as input but also shape how the model interprets the translation task and guides its generation process.

However, subsequent studies suggest that enriched prompting does not necessarily lead to consistent performance gains, with effects often displaying task-dependent and model-dependent characteristics (Kocmi & Federmann, 2023; Lu et al., 2024). He (2024) further suggests that even when more complex prompting elements are introduced, their influence on overall translation quality may remain limited. This finding contrasts with studies that emphasize the positive effects

of prompting, indicating that the literature remains divided regarding the stability and generalizability of prompt effects. Zhang et al. (2023) further observe that, although notable differences have been observed across prompt templates, no stable cross-task patterns have yet emerged. This suggests that conclusions regarding how prompting strategies influence translation outputs remain fragmented and lack a unified explanatory framework. This is particularly evident in culturally dense texts, where their specific role in shaping translation quality remains to be clarified.

Overall, existing research indicates that prompting strategies influence LLM-based translation, although their effects are strongly task-dependent and do not necessarily produce consistent improvements. Consequently, findings across studies remain somewhat inconsistent. This issue is particularly pronounced in culturally and semantically dense classical texts, where the ways in which prompting strategies shape semantic selection processes and the distribution of translation errors have yet to be systematically examined, and where their role in cultural-semantic representation remains insufficiently understood.

EVALUATING CLASSICAL CHINESE TRANSLATION: AUTOMATIC METRICS AND MQM-BASED HUMAN EVALUATION

MT evaluation is commonly divided into automatic and human evaluation paradigms (Chatzikoumi, 2020). Automated metrics broadly distinguish surface-form measures (e.g., BLEU, chrF) and semantic metrics based on contextual representations (e.g., BERTScore, COMET) (Zhang et al., 2025; Chen et al., 2022). Although these metrics are widely used in system comparisons, their evaluation outcomes often exhibit inconsistency. On the one hand, studies based on BLEU and chrF report relatively high scores for mainstream NMT systems (Chow et al., 2025); on the other hand, some analyses indicate that under conditions of high textual variation BLEU may overestimate system performance, while chrF tends to be more sensitive to surface-level differences (Rei et al., 2020; Zhao et al., 2025). These findings suggest that system rankings may remain unstable even within surface-based metrics. These discrepancies suggest that existing automatic evaluation metrics remain unstable in both their evaluative criteria and the interpretation of their results.

From the combined perspectives of methodological mechanisms and translation theory, most existing automatic evaluation metrics rely on surface matching or semantic similarity calculations, and their evaluative logic is not well suited to capturing fine-grained variations at the cultural-semantic level. Drawing on translation semantics and cultural semantics (Baker, 1992; Sharifian, 2017), culture-specific concepts often undergo semantic generalisation or hierarchical shifts during translation. As these changes do not necessarily affect surface-level similarity, they are difficult for automatic evaluation metrics to detect effectively. Although semantic-oriented metrics are generally considered more sensitive in capturing meaning correspondence (Rei et al., 2020), and neural semantic metrics such as COMET and MetricX have been shown to outperform BLEU and chrF in distinguishing translation quality in classical text contexts (Bennett et al.,2025), system-level comparisons reveal a different pattern. Metrics such as BLEURT, BERTScore, and COMET-QE tend to assign higher scores to Google Translate, while showing relatively low agreement with human judgments of ChatGPT-4 outputs (Zhang et al., 2025). Such discrepancies suggest that even semantically oriented evaluation methods may diverge from human judgments of cultural meaning, thereby highlighting the limited explanatory power of current evaluation methods at the cultural-semantic level.

By contrast, in translation quality research, human evaluation is widely regarded as a necessary complement to automatic evaluation methods in translation quality research, and the Multidimensional Quality Metrics (MQM) framework provides a hierarchical scheme for classifying translation errors while maintaining analytical flexibility (Lommel et al., 2014). Existing MQM-based studies commonly analyse system outputs across dimensions such as Accuracy, Fluency, and Style. In applied contexts, related research has found that translation problems in literary texts tend to concentrate in deeper semantic dimensions. For example, Zhang et al. (2025) report a higher error density in literary translations, with deviations mainly occurring in the dimensions of Accuracy and Style; Fakih et al. (2024), using the MQM framework, identify key issues such as semantic deviations and imbalances in expression in literary translation evaluation. Building on this, some studies have further extended the MQM framework. For instance, Wang (2024) introduces a Fidelity dimension to better capture emotional tone and cultural nuance. These studies indicate that MQM has certain advantages in uncovering deeper semantic issues in translation, particularly demonstrating strong explanatory potential in the analysis of literary and culturally dense texts. However, although the MQM framework includes dimensions such as Audience Appropriateness, existing research has seldom undertaken detailed analysis or operationalisation of constructs related to cultural reference and cultural appropriateness. As a result, culture-related deviations are often subsumed within existing categories rather than being systematically examined as independent analytical objects.

Overall, automatic evaluation methods show limited capacity to identify cultural-semantic phenomena, while the MQM framework has not been fully exploited in modelling cultural factors. This, to some extent, weakens the explanatory power of existing evaluation systems in cross-system comparisons, particularly in revealing how different translation systems diverge in cultural-semantic expression and the associated error types.

METHODOLOGY

RESEARCH DESIGN

This study combines automatic metrics and expert human evaluation to analyse the performance of NMT and LLMs in the translation of classical texts. Human evaluation was conducted using the MQM framework with fine-grained annotation of error types. Automatic evaluation employed a complementary set of metrics, including reference-based metrics (BLEU, chrF, and BERTScore) and a reference-free semantic metric (COMET-Kiwi). This multi-metric design allows system performance to be examined from different evaluation perspectives, while the integration of automatic and human evaluation facilitates the comparison of translation systems and the assessment of their applicability in classical text translation.

As illustrated in Figure 1, this study introduces a conceptual framework in which cultural semantics is realised through CSIs and systematically mapped onto MQM evaluation dimensions. In the evaluation process, cultural meaning is primarily examined in terms of Accuracy, Linguistic Conventions, Style, and Terminology, supplemented by Audience Appropriateness, to capture the transfer of cultural meaning, appropriateness of expression, consistency of lexical choices, and the accessibility and naturalness of the target text. The assignment of culturally embedded meanings to specific MQM dimensions was guided by predefined annotation criteria. To minimise potential overlap between dimensions and ensure analytical consistency, explicit annotation guidelines were established prior to evaluation, based on MQM definitions and refined through pilot annotation.

eISSN: 2550-2131
ISSN: 1675-8021

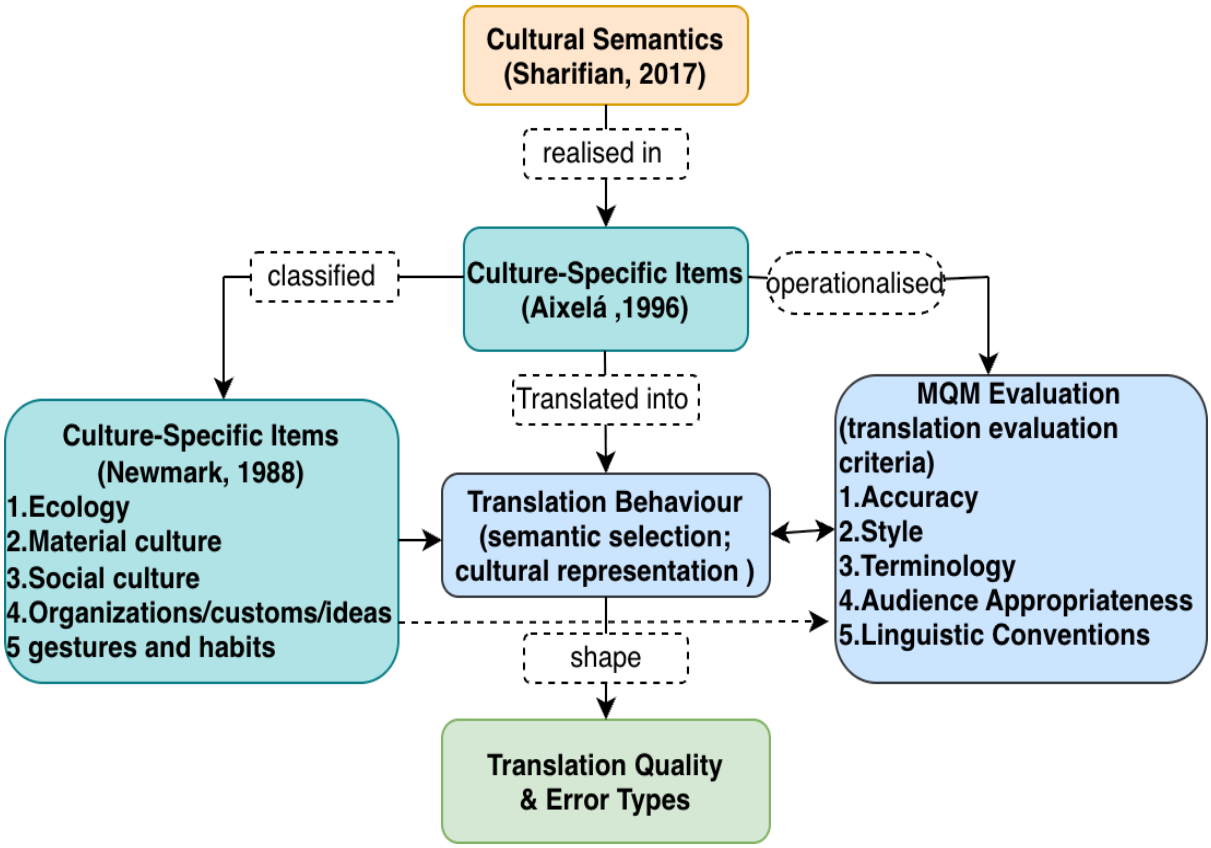

FIGURE 1. Conceptual framework

As illustrated in Figure 2, the study follows a four-stage procedure. First, fifteen sentences containing CSIs, proper nouns, and metaphorical imagery were selected from the Shan Hai Jing as the source corpus. Translations were then generated using Google Translate and ChatGPT-5.2 under two prompting conditions (minimal prompting and enriched prompting). The outputs were subsequently evaluated using automatic metrics and assessed by two professional translators following the MQM framework across the dimensions of Accuracy, Linguistic Conventions, Terminology, Style, and Audience Appropriateness. Inter-annotator agreement between the two evaluators was subsequently calculated. Finally, the results of automatic and human evaluation were compared to identify performance patterns across systems and prompting strategies, as well as recurring challenges in translating culturally embedded expressions.

eISSN: 2550-2131
ISSN: 1675-8021

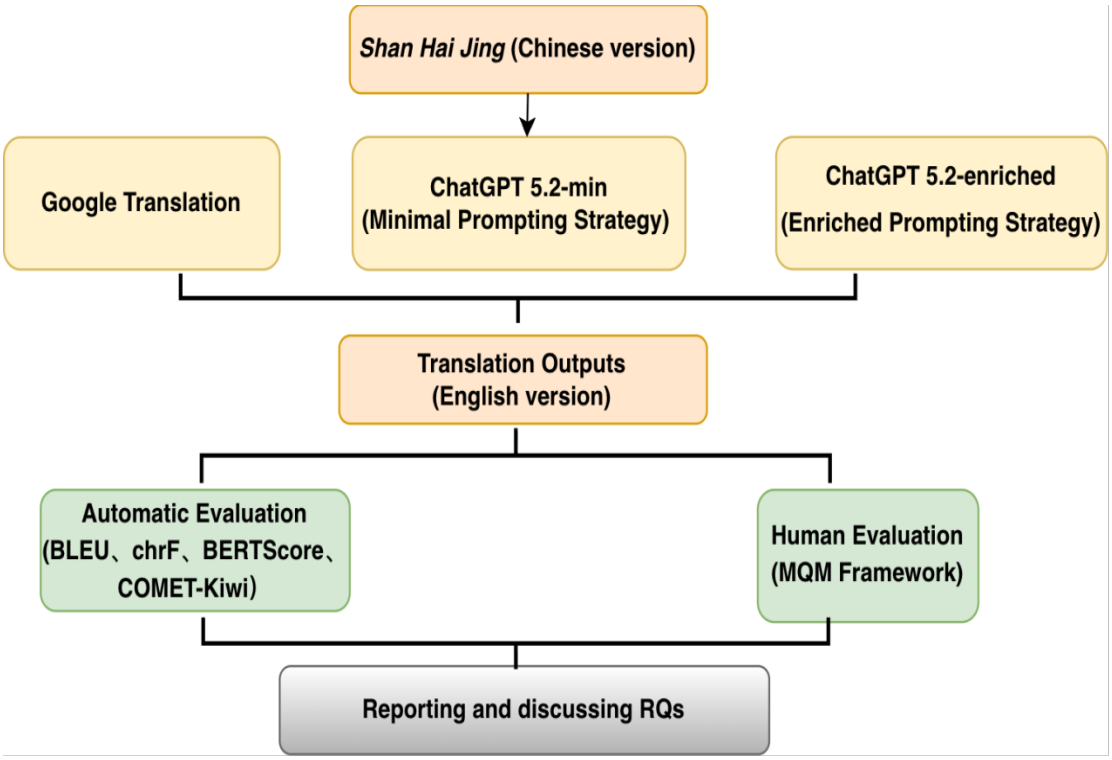

FIGURE 2. Translation evaluation flowchart

DATA SELECTION

The corpus was drawn from the classical Chinese text *Shan Hai Jing*, characterised by culturally dense semantics, frequent proper names, and rich metaphorical imagery—features widely recognised as challenging for MT systems (Strassberg, 2002). The source text follows Yuan Ke's annotated edition Shan Hai Jing Jiaozhu, whose textual collation and scholarly annotations provide a reliable basis for interpretation and sentence-level analysis (Strassberg, 2002).

Purposive sampling was adopted to select information-rich cases for detailed analysis (Creswell & Poth, 2018). The selected sentences cover the five major categories of CSIs, following Newmark's (1988) classification (ecological, material, social, organisational/customs/ideas, and gestures and habits). They are characterised by a high density of culturally embedded expressions, proper names, and metaphorical structures, thereby forming analytical units with high informational density and diagnostic value. The final dataset comprises 15 sentences (502 Chinese characters), with sentence lengths ranging from 17 to 63 characters (mean ≈ 33). The dataset is kept relatively small to enable fine-grained MQM-based error analysis, which requires detailed manual annotation and systematic comparison across translation outputs. Diagnostic analyses of this type commonly rely on smaller, carefully selected datasets to maintain annotation reliability and allow close examination of translation deviations. The study is therefore positioned as a small-scale diagnostic comparative analysis with a methodological focus on detailed error patterns rather than statistical generalisation at the system level (Creswell & Poth, 2018). Sentences serve as the basic unit of analysis, with each source sentence and its three system outputs forming a comparable evaluation unit.

A multi-reference evaluation dataset was constructed using three English translations by Wang Hong, Birrell, and Strassberg. After sentence-level alignment, the dataset yielded 43 reference sentences (1,576 words), as two source sentences lacked corresponding reference translations in the selected editions. This minor omission does not affect the computation of automatic evaluation metrics. Multi-reference configurations can improve evaluation robustness when semantic variation is limited, while the degree of semantic similarity among references affects evaluation stability (Wu et al., 2025). The translations were selected based on textual completeness, sentence-level alignability, and semantic comparability to ensure consistency in metric calculation.

TRANSLATION SYSTEMS AND PROMPT DESIGN

This study compares the performance of an NMT system and an LLM in translating culturally dense classical texts. Google Translate was selected as the representative NMT baseline because of its widespread deployment and its frequent use as a benchmark in MT evaluation studies (Son & Kim, 2023). GPT-5.2 was chosen as the LLM under investigation given its sensitivity to prompt variation and responsiveness to explicit instructions (He, 2024).

To examine the influence of prompt structure on translation behaviour, prompting was implemented under two experimental conditions. The minimal prompt condition (ChatGPT-min) followed the basic translation instruction used by Jiao et al. (2023): "Please provide the English translation for the following sentence." This prompt served as the low-constraint baseline condition.

The enriched prompt condition (ChatGPT-enriched) was designed following task-specific and domain-oriented prompting principles (Peng et al., 2023) and incorporated contextual information to strengthen the model's understanding of the Shan Hai Jing context. The prompt instructed ChatGPT to interpret the meaning prior to translation: "You are a machine translation system that translates sentences in the Chinese classical mythology domain. The following is a passage from Shan Hai Jing. Please understand its meaning first and then translate it into English."

The two prompts differ in structural complexity and task constraints, enabling the examination of how prompt strategies may influence translation output. Under default deterministic model settings, each source sentence was translated once under three system configurations (Google Translate, ChatGPT-min, and ChatGPT-enriched). No repeated sampling was performed. With prompt type treated as the primary independent variable, a total of 45 translation outputs were obtained (1,649 English tokens), ensuring comparability across conditions.

EVALUATION METHODS

The evaluation framework consists of two components: MQM-based human evaluation and automatic metric-based assessment. MQM was adopted to enable fine-grained analysis beyond surface-level metrics. Accordingly, five evaluation dimensions from the MQM core typology were adopted (see Figure 3): Accuracy, Linguistic Conventions, Terminology, Style, and Audience Appropriateness. Accuracy and Linguistic Conventions reflect semantic correspondence and linguistic acceptability in the target language (Lommel et al., 2014; Koby et al., 2014). Given the frequent occurrence of proper names and CSIs in classical texts, the Terminology dimension was included to examine conceptual consistency (Fakih et al., 2024). As quality deviations in literary

eISSN: 2550-2131
ISSN: 1675-8021

and culturally dense texts often manifest at the stylistic level, the Style dimension was introduced to assess stylistic reproduction (Zhang et al., 2025). Audience Appropriateness was used to evaluate cultural intelligibility and reader acceptability (Fakih et al., 2024).

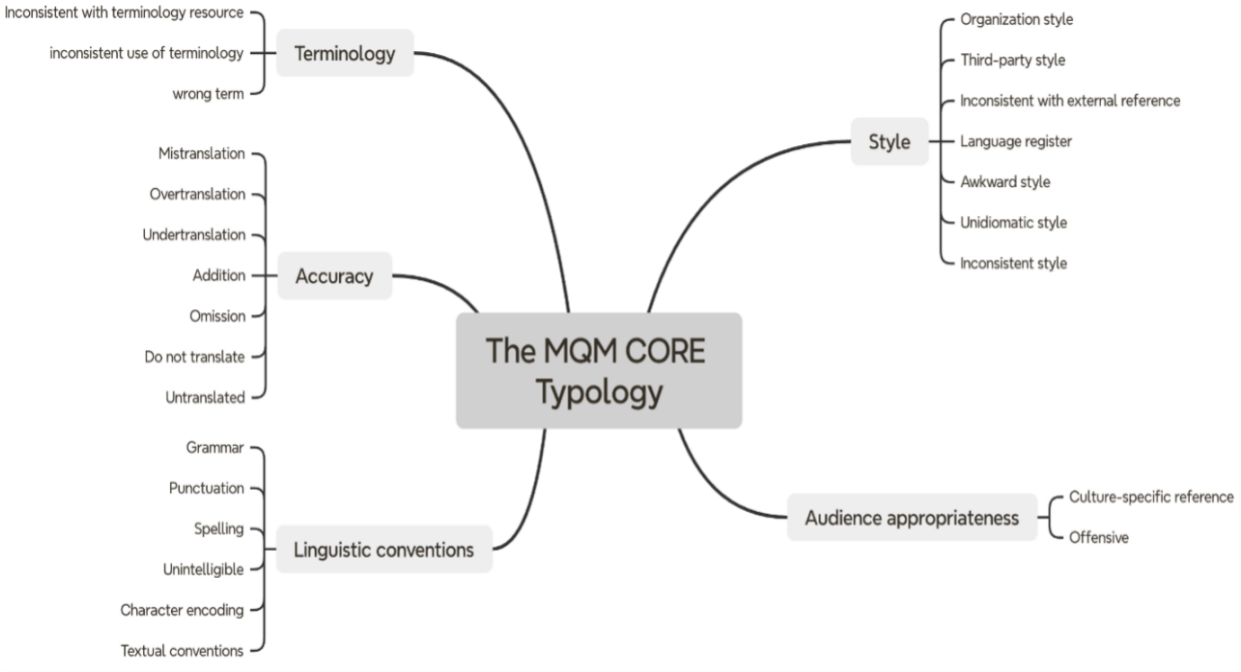

FIGURE 3. MQM Error Dimensions Applied in This Study (https://themqm.org/the-mqm-typology/)

The MQM dimensions, definitions, and severity levels followed the official MQM typology, with errors classified as minor, major, or critical (weights: 1, 5, and 25) according to their impact on comprehension and semantic integrity. The overall MQM score was calculated by computing a weighted penalty total— (Minor × 1) + (Major × 5) + (Critical × 25)—which was then normalised by the number of evaluated words and converted into the final MQM quality score following standard settings. Error annotation was conducted independently by two evaluators with translation backgrounds. Errors were annotated at the segment level, allowing multiple errors within the same sentence. Cohen's kappa was calculated to assess inter-annotator agreement, and disagreements were resolved through discussion to produce the final dataset.

As a complement to human evaluation, automatic evaluation provided quantitative evidence. Four complementary metrics were employed: three reference-based metrics (BLEU, chrF, and BERTScore) and one reference-free quality estimation metric (COMET-Kiwi), capturing translation performance at both surface-form and semantic levels. BLEU and chrF were computed using the SacreBLEU framework (Post, 2018), while BERTScore was implemented using the bert-score package (Zhang et al., 2020). All reference-based metrics were calculated under a multi-reference setting. Reference-free evaluation was conducted using COMET-Kiwi within the COMET-22 framework (Rei et al., 2022). Automatic evaluation was performed at the sentence level, with preprocessing limited to whitespace normalisation and the removal of empty lines.

The same set of sentences was translated by all three systems; therefore, the data follow a repeated-measures design. To control for sentence-level variation and compare system mean scores, a one-way repeated-measures ANOVA was conducted to test differences in automatic metric scores across systems. Normality and sphericity assumptions were examined to assess model suitability. The null hypothesis was formulated as follows: H_0: There is no significant difference in automatic metric scores among the translation systems (RQ1).

RESULTS

AUTOMATIC EVALUATION RESULTS

The following section reports the main findings of the automatic and human evaluations based on the research design and procedures outlined above. The numerical results of the automatic evaluation are detailed in Table 1, with a visual comparison illustrated in Figure 4. Overall, score distributions across the three translation systems appear to show substantial overlap within each metric, with only minor differences in median values. BLEU scores are generally low and exhibit greater variability across systems. In contrast, BERTScore and COMET-Kiwi tend to yield consistently higher scores with relatively concentrated distributions. chrF shows slight variation among systems but does not produce a clear performance ranking. Overall, the automatic metrics may provide limited differentiation across the three systems.

TABLE 1. Scores of automatic evaluation metrics

Metric	System	Mean	SD
BLEU	Google Translate	25.31	17.86
	ChatGPT-min	23.75	11.98
	ChatGPT-enriched	23.98	12.82
chrF	Google Translate	40.49	9.35
	ChatGPT-min	45.88	9.21
	ChatGPT-enriched	44.36	9.10
BERTScore	Google Translate	53.38	9.19
	ChatGPT-min	53	11.65
	ChatGPT-enriched	53.13	10.07
COMET-Kiwi	Google Translate	66.51	8.71
	ChatGPT-min	64.39	8.36
	ChatGPT-enriched	66.14	9.38

eISSN: 2550-2131
ISSN: 1675-8021

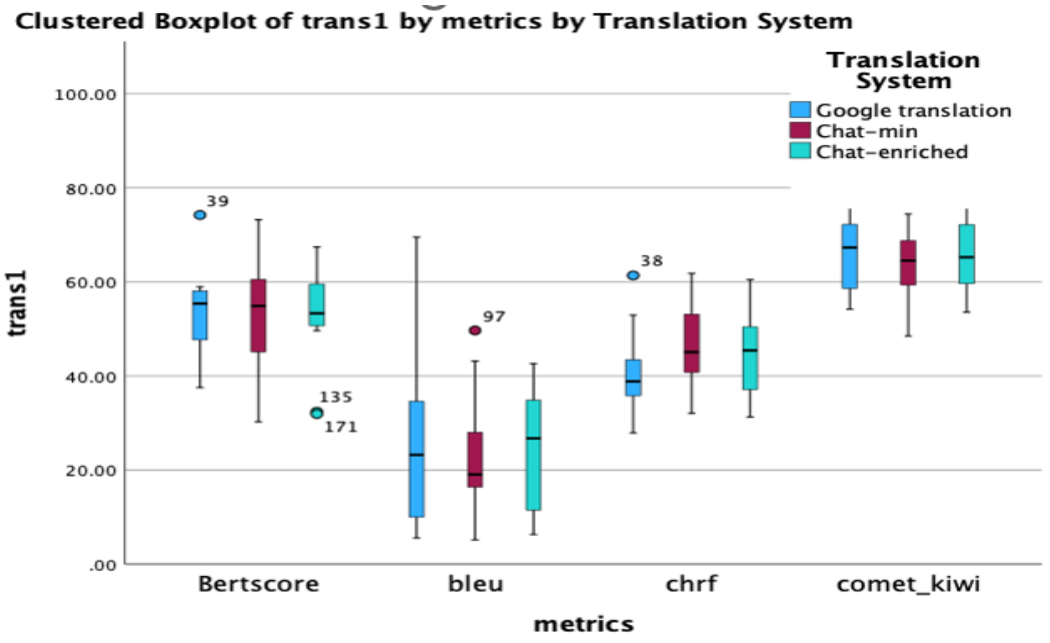

FIGURE 4. Clustered boxplot of evaluation scores by metric and translation system

The one-way repeated-measures ANOVA (see Table 2) indicated a significant main effect of evaluation metric type, F (1.912, 80.314) = 161.512, p < .001, ηp^2 = .794. Neither the main effect of translation system nor the interaction between metric and system was found to be statistically significant. These results indicate that the observed differences are primarily associated with variation across evaluation metrics rather than differences among translation systems. Across the four automatic metrics, no significant performance differences were observed among the three systems. In short, current automatic metrics may be insufficient to fully capture the translation quality of culturally dense classical texts.

TABLE 2. Results of Repeated-Measures ANOVA

Source	df	F	p-value	ηp^2
metrics	1.912	161.512	<.001	0.794
metrics * system	3.824	0.561	0.684	0.026
system	2	0.016	0.984	0.001

Note. Shapiro–Wilk tests indicated normality (p > .05). Mauchly's test showed that sphericity was violated (W = 0.42, χ^2 (5) = 35.343, p < .001); therefore, Greenhouse–Geisser correction was applied (ε = .637).

INTER-ANNOTATOR AGREEMENT

Inter-annotator agreement for MQM error annotation was assessed using Cohen's κ (Table 3). Across the three system conditions, κ values ranged from 0.434 to 0.478. According to Landis and Koch (1977), this range corresponds to moderate agreement, indicating an acceptable level of consistency in error category identification between the evaluators. Similar κ ranges have been reported in previous MQM studies (Zhang et al., 2025), reflecting the inherent variability of manual translation error annotation. Following the agreement test, disagreements were resolved through discussion to produce the final dataset. Overall, the annotation results provide a reliable basis for subsequent analysis.

eISSN: 2550-2131
ISSN: 1675-8021

TABLE 3. Inter-annotator agreement (Cohen's κ)

Translation System	Cohen's k	p-value	Error number
Google Translate	0.434	<.001	32
ChatGPT-min	0.472	<.001	30
ChatGPT-enriched	0.478	<.001	26

MQM ERROR ANALYSIS

Under the MQM framework (see Table 4), the systems exhibit clear differences in error structure. Google Translate produced the highest overall number of errors, with high-severity errors largely concentrated in the Accuracy dimension, particularly in cases of mistranslation. ChatGPT-enriched showed the lowest error count, with errors mainly occurring in culture-specific reference and Accuracy categories. Notably, no substantial errors were observed in the Linguistic Conventions dimension, indicating that all systems maintained relatively stable performance in basic grammatical and formal conventions. However, errors related to culture-specific references persisted across all systems, suggesting that cultural–semantic mapping remains a consistent challenge regardless of system type. The following sections examine the distribution of errors across different dimensions in relation to RQ2.

TABLE 4. Error types and severity in *Shan Hai Jing*

	Error Type	Sub-Error Type	Minor	Major	Critical
Google Translate	Accuracy	Mistranslation	2	4	4
		Under-translation	2	1	1
		Over-translation	0	2	1
	Style	Unidiomatic style	2		
		Awkward style	1		
	Audience appropriateness	Culture-specific reference	4	3	3
	Terminology	Wrong term	1		
ChatGPT-min	Accuracy	Under-translation	2	1	
		Addition	2	1	
		Over-translation	1		
	Style	Mistranslation	1		
		Language register	3		
		Unidiomatic style	1		
	Terminology	Wrong term	3		
	Audience appropriateness	Culture-specific reference	9	1	
ChatGPT-enriched	Accuracy	Mistranslation	1	3	1
		Under-translation	2	1	
		Addition	1		
		Over-translation	1		
	Style	Unidiomatic style	1		
		Language register	1		
		Inconsistent style	1		
	Audience appropriateness	Culture-specific reference	6	2	

ACCURACY

The examples presented in Tables 5–8 are intended to illustrate recurring patterns observed in the dataset. In the Accuracy dimension, Google translate system generated a higher number of errors and greater severity overall, particularly in cases of mistranslation and under-translation. By comparison, ChatGPT under both prompting conditions generated fewer Accuracy errors, with a lower proportion of high-severity cases, primarily involving under-translation and addition. In the ChatGPT-enriched condition, the number of Accuracy errors decreased further (10 cases), with under-translation remaining the dominant subtype. Cross-system comparison indicates that under-translation occurred consistently across all three systems, whereas mistranslation was more concentrated in the Google outputs. In contrast, Accuracy deviations in the ChatGPT outputs more often appeared as compensatory additions or instances of semantic generalisation.

As shown in Table 5, the case analysis further illustrates differences in error structure. For example, the expression "不眯 (bù mī)" conveys a semantic reference associated with mental or spiritual states. The Google Translate output reflects semantic misselection, whereas the ChatGPT outputs tend toward semantic generalisation. Although the underlying mechanisms differ, both types of deviation result in semantic shifts within the Accuracy dimension.

TABLE 5. Examples of evaluation results for Accuracy errors

ST：是多冉遗之鱼，鱼身蛇首六足，其目如马耳，食之使人不眯，可以御凶。
Google Translate: It is a fish with the body of a fish, the head of a snake, and six legs. Its eyes are like horse ears. Eating it will prevent people from squinting and can ward off evil. ChatGPT-min: There are many fish called Ranyi here. They have the body of a fish, the head of a snake, and six feet; their eyes resemble a horse's ears. Eating them prevents people from dozing and enables them to ward off misfortune. ChatGPT-enriched: There are many Ran-yi fish here: they have the body of a fish, the head of a snake, and six legs; their eyes resemble horse's ears. Eating them prevents drowsiness and enables one to ward off misfortune.

STYLE

In the Style dimension, all identified errors were minor, primarily involving unidiomatic style and language register. No major or critical errors were observed. ChatGPT-min showed slightly more stylistic deviations than the other two systems. Overall, differences across systems were mainly reflected in register choice and narrative reference, but these variations did not affect the core meaning.

The example of "应龙 (Yìnglóng)" (Table 6) illustrates these stylistic differences. Google Translate adopts a relatively modern tone, whereas ChatGPT-min uses an impersonal construction ("it was unable"). In contrast, ChatGPT-enriched employs an explicit subject ("he"), resulting in greater narrative coherence and stylistic consistency. This example suggests that although systems vary in register and referential strategy, the underlying semantic content remains stable. Compared with stylistic deviations, culture–semantic issues appear more pronounced in the Audience Appropriateness dimension.

eISSN: 2550-2131
ISSN: 1675-8021

TABLE 6. Examples of evaluation results for Style errors

ST：☒☒☒南极，☒蚩尤与夸父，不得复上，故下数旱。
Google Translate: Yinglong resided in the South Pole, where he killed Chiyou and Kuafu, and could not ascend again, hence the frequent droughts that followed. ChatGPT-min: Yinglong dwelt at the southern extremity. After slaying Chiyou and Kuafu, it was unable to return to the heavens; therefore, droughts repeatedly occurred below. ChatGPT-enriched: Yinglong dwelt at the southern extremity; after killing Chiyou and Kuafu, he was unable to ascend again, and therefore droughts repeatedly occurred below.

AUDIENCE APPROPRIATENESS

In the Audience Appropriateness dimension, all systems exhibited deviations related to culture-specific references. Google translate system produced both a higher number and greater severity of such errors. By contrast, ChatGPT-min and ChatGPT-enriched showed fewer instances. No critical cases were observed for ChatGPT-enriched, although deviations involving culture-specific references persisted.

Table 7 illustrates the phrase "食者不蛊" (shí zhě bù gǔ). The term "蛊" (gǔ) refers to a culturally specific concept denoting a malignant illness associated with *gu* poison or sorcery. Google translate renders the expression as "will not be harmed," reflecting semantic generalisation, whereas ChatGPT-enriched translates it as "will not be afflicted by poisoning," representing semantic approximation. In both cases, the cultural reference is not entirely removed, but a shift in semantic level occurs. Overall, deviations involving culture-specific references constitute a persistent challenge across systems, typically manifested as semantic generalisation or substitution with a proximate concept.

TABLE 7. Examples of evaluation results for Audience Appropriateness errors

ST：有☒焉其状如狐而九尾，其音如☒儿，能食人，食者不☒。
Google Translate: There is a beast that looks like a fox but has nine tails. Its cry is like that of an infant. It can eat people, and those who eat it will not be harmed. ChatGPT-min: There is a beast whose form resembles that of a fox but with nine tails. Its cry is like that of an infant. It is capable of devouring humans; those who eat it are said not to suffer from gu (poisonous sorcery). ChatGPT-enriched: There is a beast whose form resembles a fox but with nine tails; its cry sounds like that of an infant. It can devour humans, yet those who eat its flesh will not be afflicted by poisoning.

TERMINOLOGY

In the Terminology dimension, the ChatGPT-min condition shows a relatively higher number of Wrong term deviations. These issues primarily concern the selection of translation equivalents for culture-specific terms and their conventionalised rendering.

For example, in the case of "太岁" (Tàisuì) (Table 8), ChatGPT-min produces the translation "Great Year (Taisui)". Although the source reference *Taisui* is retained, the English expression does not form a conventional cultural equivalent and may weaken its cultural–chronological connotations.

eISSN: 2550-2131
ISSN: 1675-8021

TABLE 8. Examples of evaluation results for Terminology errors

ST: 地之所⊠, 六合之⊠, 四海之內, 照之以日月, ⊠之以星辰, ⊠之以四⊠, 要之以太⊠
Google Translate: The earth contains everything within the six directions and the four seas, illuminated by the sun and moon, guided by the stars, marked by the four seasons, and governed by the Grand Duke Jupiter. ChatGPT-min: The earth bears all that exists; within the space of the six directions and inside the four seas, it is illuminated by the sun and moon, traversed by the stars, regulated by the four seasons, and governed by the Great Year (Taisui). ChatGPT-enriched: That which the earth bears—within the space of the Six Directions, inside the Four Seas—is illuminated by the sun and moon, traversed by the stars, ordered by the four seasons, and governed by the Great Year.

COMPARISON OF MQM SCORECARDS

The MQM scorecard results are presented in Table 9. Overall, the ChatGPT-min condition achieves the highest overall quality score, corresponding to the lowest penalty score. ChatGPT-enriched ranks second, whereas Google Translate shows a relatively higher penalty level and the lowest overall quality score.

TABLE 9. Comparative metrics for translation system quality

Metric	Google Translate	ChatGPT-min	ChatGPT-enriched
Evaluation Word Count	495	576	545
Reference Word Count	475	475	475
ET Weight	1	1	1
Absolute Penalty Total	287.00	37	69
Per-Word Penalty Total	0.5798	0.0642	0.1266
Overall Quality Score	42.02	93.58	87.34

This pattern is consistent with the preceding error distribution analysis. The concentration of Accuracy-related deviations and high-severity errors in the NMT output has a substantial impact on the overall quality score. By contrast, although LLM outputs exhibit instances of culture–semantic generalisation, the overall number of structural errors remains comparatively lower. This result also addresses RQ3, suggesting that under the present conditions, simpler prompts appear to contribute to more stable translation quality than enriched prompts.

DISCUSSION

Building on the results reported above, the following discussion interprets the findings in relation to the research questions and relevant literature. RQ1 focused on the scoring behaviour of automatic evaluation metrics across translation systems. Overall, the results indicate that scores across systems tend to converge across multiple automatic metrics, and the differences between systems do not reach statistical significance. The variation observed arises primarily from differences among the evaluation metrics themselves rather than from performance differences between translation systems. This finding therefore supports the null hypothesis (H_0) and suggests that, in culturally and semantically dense texts, the evaluation framework itself may limit the discriminative capacity of system comparisons.

However, this finding does not fully align with the results reported by Chow et al. (2025) and Zhang et al. (2025). Previous studies based on BLEU and chrF have reported relatively higher scores for systems such as Google Translate and DeepL, suggesting that certain semantic metrics

may exhibit a degree of bias towards NMT outputs. By contrast, the present study does not observe a consistent scoring advantage for any particular system type across the automatic metrics examined.

This pattern is consistent with prior discussions indicating that automatic metrics may exhibit relatively limited sensitivity, especially when differences in system quality are small (Mathur et al., 2020). This limitation may be more pronounced under small-sample conditions, where the number of systems is limited and performance differences are modest, potentially affecting the stability of metric-based rankings. Moreover, in the context of classical texts, reference-based metrics may struggle to capture subtle shifts in cultural reference. Taken together, the observed score convergence is more likely to reflect the measurement properties of the evaluation metrics, rather than necessarily indicating substantive equivalence in system performance.

Therefore, from a methodological perspective, these results indicate that in evaluating translation quality in culturally dense texts, it is necessary to integrate human evaluation to address the limitations of automatic metrics. These findings further highlight the need to develop automatic evaluation approaches better suited to such texts. Given the limited sample size and potential uncontrolled variables, these findings are best interpreted as task-specific observations.

RQ2 examined the quality characteristics of different translation systems under the MQM framework. Overall, translations produced under the ChatGPT conditions achieved higher scores than those generated by Google Translate. This trend broadly aligns with Wang (2025), who reported that ChatGPT demonstrates relative advantages in fidelity and fluency. However, the principal contribution of the present study lies less in overall score differences than in the distinct error structures observed across systems. Specifically, the two systems display different error distribution patterns. Errors produced by Google Translate are concentrated primarily in the Accuracy dimension, whereas deviations in ChatGPT outputs occur mainly in the culture-specific reference and Accuracy categories.

This observation is consistent with previous studies highlighting persistent limitations of NMT systems in handling CSIs and deep semantic representation (Jiang et al., 2024). Similar limitations have also been reported in the translation of cultural imagery (Wang, 2024), as well as in relation to insufficient semantic depth and hallucination risks in the translation of classical texts (Zhao et al., 2025). These patterns may be associated with, at least in part, the substantial demands for background knowledge and contextual reasoning posed by classical texts (Jiao et al., 2023), which affect both cultural reference and semantic precision.

The observed differences in error patterns may be partly associated with differences in how the systems generate translation. NMT systems such as Google Translate are generally based on learned correspondences between source and target segments, which may contribute to a higher likelihood of mistranslation when culturally dense expressions lack direct lexical equivalents (Wu et al., 2016; Vaswani et al., 2017). By contrast, LLMs generate translations through probabilistic language modelling combined with contextual inference (Wang et al., 2023). Such generative processes may be linked to semantic approximation or generalisation rather than explicit mistranslation, a tendency also observed in recent analyses of LLM-based translation behaviour (Karabayeva & Kalizhanova, 2024).

At the level of semantic deviation types, the two systems display different tendencies. Google Translate appears to show a tendency towards semantic misselection, whereas ChatGPT may be more inclined towards semantic hypernymisation, where culturally specific expressions are replaced by semantically broader or adjacent terms. For example, in the expression "蚩 (gǔ)",

which refers to a culturally specific concept associated with ritual poisoning in classical Chinese culture, Google Translate renders the term as "harm", whereas ChatGPT produces "poisoning". While both translations retain part of the semantic orientation, the culturally specific concept embedded in the original term is partially generalised in the target expression. In MQM, this deviation falls under the culture-specific reference category.

Unlike Naveen and Trojovský (2024), who emphasised cultural meaning erosion, the present study does not observe systematic loss of cultural meaning. In most cases, culture-specific references retain their basic referential orientation but tend to shift towards more generalised expressions. This tendency is also consistent with Chen and Lin (2025), who observed that ChatGPT demonstrates greater cultural sensitivity than traditional NMT systems, although some degree of semantic-level displacement. This phenomenon may be interpreted in light of a probability-driven generation strategy, whereby broader and higher-frequency expressions may be more likely to be selected (Karabayeva & Kalizhanova, 2024), potentially weakening the hierarchical precision of culturally embedded semantics.

From the perspective of cultural meaning transmission, semantic hypernymization may contribute to a reduction in the cultural specificity embedded in the source text. In classical works such as the *Shan Hai Jing*, many expressions function not only as lexical units but also as culturally embedded references that point to specific mythological entities and cultural knowledge. When such expressions are replaced with semantically broader terms, the overall semantic orientation may be preserved, yet the precision of cultural reference tends to be reduced. Consequently, this may weaken the transmission of culturally embedded knowledge.

These findings may suggest that translation quality evaluation may benefit from the incorporation of more fine-grained cultural-semantic indicators, such as cultural reference retention and semantic hierarchy shift to better capture deviations in the rendering of culture-specific references. Overall, RQ2 highlights structural differences between LLM and NMT systems in the handling of cultural semantics and provides a more explanatory framework for understanding the relationship between error typology and patterns observed across different translation systems.

RQ3 aimed to examine whether and how different prompting strategies influence ChatGPT's performance across MQM quality dimensions in the English translation of the *Shan Hai Jing*. The results indicate that ChatGPT-min outperforms ChatGPT-enriched, suggesting that in the context of classical text translation, concise and direct instructions may contribute to greater stability in translation quality than more elaborate prompts. This finding further indicates that the effectiveness of prompt engineering does not necessarily increase with prompt complexity, but may instead depend on the characteristics of the source text and its semantic structure.

In the present study, the enriched prompting condition does not result in higher translation quality than the minimal prompting condition. This finding differs from the results reported by Yamada (2023), who suggested that incorporating translation-oriented elements—such as translation purpose and target audience—may help simulate real translation briefs and potentially improve translation quality. However, the present results are more consistent with Wang (2025), indicating that simple and direct translation instructions tend to perform better in both automatic and human evaluations, whereas more complex prompts that include explanatory or analytical requirements fail to produce consistent quality gains. This pattern also aligns with He (2024), whose findings suggest that increasing prompt elaboration does not necessarily improve translation quality.

eISSN: 2550-2131
ISSN: 1675-8021

Enriched prompting may involve additional steps of semantic inference, which may be associated with an increased tendency towards semantic generalisation and cultural shifts in the translation of classical texts. This phenomenon may be related to the task-dependent nature of prompt engineering (Lu et al., 2024; Zhang et al., 2023) and suggests that its effectiveness may be influenced by the semantic structure of the text. Therefore, for classical translation tasks, where meanings are highly compressed and culturally dense, increasing prompt complexity does not necessarily improve translation quality. The effectiveness of prompting strategies should thus be assessed with reference to specific text types and task objectives.

CONCLUSION

This study examined the translation performance of NMT and LLM systems in culturally dense classical texts using the *Shan Hai Jing* as the corpus, combining automatic evaluation metrics, MQM-based human assessment, and prompting strategies. The results tend to suggest that automatic metrics provide limited differentiation in system rankings, whereas MQM-based human evaluation appears to reveal clearer structural differences in error patterns. Deviations in NMT outputs appear to be largely concentrated in the Accuracy dimension, particularly in high-severity errors such as mistranslation and under-translation. By contrast, LLM outputs more frequently exhibit semantic generalisation and approximate conceptual substitution, often involving hierarchical shifts in the rendering of culture-specific references. In addition, prompt complexity does not appear to produce stable improvements in translation quality, suggesting that the effectiveness of prompting strategies may be strongly task-dependent.

Taken together, these findings indicate that translation quality in culturally dense texts may be more meaningfully interpreted through patterns of translation errors and cultural–semantic shifts across MQM dimensions, rather than solely through metric-based score comparisons. The results also suggest a possible sensitivity mismatch between translation generation processes and existing evaluation approaches, while the effectiveness of prompting strategies may be conditioned by the semantic structure of the source text. These patterns may further reflect how translation deviations influence the representation and transmission of culturally embedded meanings in classical texts.

The study is subject to certain limitations. The dataset is restricted to a small number of sentences drawn from a single classical text, and the results should therefore be interpreted as task-specific observations rather than generalisable properties of NMT or LLM translation behaviour. In addition, the enriched prompt may introduce a confounding effect by encouraging more interpretative processing, thereby influencing the results to some extent. Nevertheless, this study provides meaningful insights at both the theoretical and practical levels. Theoretically, it extends existing evaluation frameworks to some extent by suggesting that translation quality in culturally dense texts cannot be fully captured through metric-based approaches alone, but is more appropriately understood in terms of cultural–semantic accuracy and meaning representation, thereby highlighting the role of culture-specific meaning and MQM-based error analysis. Practically, the evaluation of such texts may benefit from combining MQM-based human assessment with automatic metrics to more effectively capture culturally embedded meanings and high-severity semantic deviations. At the same time, improvements in translation quality may be achieved by incorporating cultural contextualisation or explicit marking of CSIs, together with evaluation and post-editing strategies targeting cultural–semantic shifts.

eISSN: 2550-2131
ISSN: 1675-8021

This study is positioned as a preliminary empirical investigation, with its contribution lying primarily in analytical and evaluative insights. Accordingly, future research may expand the dataset to include a broader range of translation systems and text types, while further refining cultural–semantic evaluation dimensions and optimising experimental design to distinguish between contextual enhancement and interpretative guidance in prompting. Building on this, future work may explore the development of culturally aware evaluation models that integrate MQM-based error typologies with CSI-informed approaches to model culture-specific items and their semantic shifts in translation.

REFERENCES

Aixelá, F. (1996). Culture-specific items in translation. In R. Alvarez & M. C. Vidal (Eds.), *Translation, power, subversion* (pp. 52–78). Multilingual Matters.

Baker, M. (2011). *In other words: A coursebook on translation* (2nd ed.). Routledge.

Bennett, E., Han, H., Yang, X., Schonebaum, A., & Carpuat, M. (2025). Evaluating evaluation metrics for Ancient Chinese to English machine translation. *In Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025* (pp. 71–76). Association for Computational Linguistics. https://aclanthology.org/2025.alp-1.9/

Chen, A., Lou, L., Chen, K., Bai, X., Xiang, Y., Yang, M., Zhao, T., & Zhang, M. (2025). Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 33019–33036). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1678

Chen, S., & Lin, Y. (2025). A multidimensional comparison of ChatGPT, Google Translate, and DeepL in Chinese tourism texts translation: Fidelity, fluency, cultural sensitivity, and persuasiveness. *Frontiers in Artificial Intelligence, 8,* 1619489. https://doi.org/10.3389/frai.2025.1619489

Chow, R. C., Angeline, V., Jayata, G., Mujhid, A., & Hidayaturrahman. (2025). Comparing LLMs and NMTs performances in translating English–Indonesian texts. *Procedia Computer Science, 269*, 1455–1465. https://doi.org/10.1016/j.procs.2025.09.087

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20(1), 37–46.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches (4th ed.).* SAGE.

Dunder, I., Seljan, S., & Pavlovski, M. (2021). What makes machine-translated poetry look bad? A human error classification analysis. In *Proceedings of the Central European Conference on Information and Intelligent Systems* (pp. 183–191). Faculty of Organization and Informatics, Varaždin, Croatia.

Fakih, A., Ghassemiazghandi, M., Fakih, A. H., & Singh, M. K. (2024). Evaluation of Instagram's Neural Machine Translation for literary texts: An MQM-Based analysis. *GEMA Online Journal of Language Studies, 24*(1), 213–233. https://doi.org/10.17576/gema-2024-2401-13

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), 1-10. https://doi.org/10.1057/s41599-024-03363-0

Gao, Y., Wang, R., & Hou, F. (2023). How to design translation prompts for ChatGPT: An empirical study. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops* (pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3700410.3702123

He, S. (2024). Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation* (pp. 316–326). European Association for Machine Translation.

He, L., Ghassemiazghandi, M., & Subramaniam, I. (2024). Comparative assessment of Bing Translator and Youdao Machine Translation Systems in English-to-Chinese literary text translation. *Forum for Linguistic Studies, 6(2)*, 1189–1198.

Jiang, Z., Lv, Q., Zhang, Z., & Lei, L. (2024). Convergences and divergences between automatic assessment and human evaluation: Insights from comparing ChatGPT-Generated translation and Neural Machine Translation. *arXiv* preprint arXiv:2401.05176.

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. P. (2023). Is ChatGPT A Good Translator? Yes with GPT-4 as the Engine. *arXiv*. https://doi.org/10.48550/arXiv.2301.08745

Karabayeva, I., & Kalizhanova, A. (2024). Evaluating machine translation of literature through rhetorical analysis. *Journal of Translation and Language Studies, 5(1)*, 1–9. https://doi.org/10.48185/jtls.v5i1.962

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 33(1), 159–174. https://doi.org/10.2307/252931

Li, Z., & Chen, L. (2025). Mind vs. machine: Comparative analysis of metaphor-related word translation by human and AI systems. *Training, Language and Culture,* 9(1), 10–27. https://doi.org/10.22363/2521-442X-2025-9-1-10-27

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica,* 12, 455–463.

Lu, Q., Qiu, B., Ding, L., Zhang, K., Kocmi, T., & Tao, D. (2024). Error analysis prompting enables human-like translation evaluation in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 8801–8816). Association for Computational Linguistics.

Lyu, C., Du, Z., Xu, J., Duan, Y., Wu, M., Lynn, T., Aji, A. F., Wong, D. F., Liu, S., & Wang, L. (2023). A paradigm shift: The future of Machine Translation lies with Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 1339–1352). ELRA and ICCL.

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 4984–4997). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.448

Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience, 27*. (10). https://doi.org/10.1016/j.isci.2024.110878

Newmark, P. (1988). *A textbook of translation.* Prentice Hall.

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., ... & Tao, D. (2023). Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5622–5633). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.373

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., de Souza, J. G. C., Glushkova, T., Alves, D. M., Lavie, A., Coheur, L., & Martins, A. F. T. (2022). COMETKIWI: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 634–645). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.60

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation,* 56(2), 593–619. https://doi.org/10.1007/s10579-021-09537-5

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7881–7892). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.704/

Sharifian, F. (2017). *Cultural linguistics: Cultural conceptualisations and language*. John Benjamins.

Shen, S., Wang, W., & Birch, A. (2025). Liaozhai through the looking-glass: On paratextual explicitation of culture-bound terms in machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 34400–34416). Association for Computational Linguistics. https://aclanthology.org/2025.emnlp-main.1744/

Shi, Y., Xu, H., Kwok, H. L., & Liu, K. (2024). ChatGPT in professional translation: A double-edged sword – Insights from Chinese translators on capabilities, concerns, and future prospects. In *Translation Studies in the Age of Artificial Intelligence* (pp. 125–149). Routledge. https://doi.org/10.4324/9781003482369-7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems,* 30, 5998–6008.

Wang, Q., Amini, M., & Tan, D. A. L. (2025). Strategies, errors, and challenges in translating culture-specific items in Chinese-English literary works: A systematic review. *Jurnal Arbitrer,* 12(2), 259–273. https://doi.org/10.25077/ar.12.2.259-273.2025

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 16646–16661). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.1036

Wang, Q. (2025). Evaluating Uighur literary translation: A comparative study of ChatGPT, Google Translate, and Bing Translator. *PLOS ONE, 20*(10), e0335261.

Wang, J. (2024). Exploring the potential of ChatGPT-4o in Translation Quality Assessment. *Journal of Theory and Practice in Humanities and Social Sciences,* 1(3), 18–30.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* preprint arXiv:1609.08144.

Wu, S., Wieting, J., & Smith, D. A. (2025). Multiple references with meaningful variations improve literary machine translation. *arXiv preprint arXiv:2412.18707*.

Yamada, M. (2023). Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track* (pp. 195-204).

Yao, G., & Fan, L. (2025). An entropy-based study of simplification in ChatGPT translations compared to neural machine translation and human translation across genres. *PLOS ONE, 20*(12): e0339762. https://doi.org/10.1371/journal.pone.0339762

Zhang, B., Haddow, B., & Birch, A. (2023). Prompting large language models for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 41092–41110). https://proceedings.mlr.press/v202/zhang23m.html

Zhang, Z., Syed Abdullah, S. N., Abdullah, M. A. R., & Duan, W. (2025). Evaluating Google neural machine translation from Chinese to English: Technical vs. literary texts. *GEMA Online® Journal of Language Studies,* 25(3), 732–753. https://doi.org/10.17576/gema-2025-2503-09

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)* (pp. 1–43). https://openreview.net/forum?id=SkeHuCVFDr

Zhang, R., Zhao, W., & Eger, S. (2025). How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 10961–10988). Association for Computational Linguistics.

Zhang, R., Zhao, W., Macken, L., & Eger, S. (2025). LiTransProQA: An LLM-based literary translation evaluation metric with professional question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 29087-29109). Association for Computational Linguistics.

Zhao, Z., Sun, G., Liu, C., & Wang, D. (2025). Research on machine translation of ancient books in the era of large language models. *npj Heritage Science, 13,* Article 122. https://doi.org/10.1038/s40494-025-01697-9

ABOUT THE AUTHORS

Duan Wenqi is a PhD candidate in Translation Studies at Universiti Putra Malaysia (UPM) and a university teacher in China. Her research interests include translation studies, classical literature, culture, and applied linguistics.
Email: gs70887@student.upm.edu.my

Ng Chwee Fang (Ph.D) is a senior lecturer at the Department of Foreign Languages, Faculty of Modern Languages and Communication, Universiti Putra Malaysia. Her research interests include Chinese linguistics, comparative linguistics, psycholinguistics, and linguistic typology.
Email: chweefang@upm.edu.my

Hazlina Abdul Halim (Ph.D) is an associate professor at Universiti Putra Malaysia. Her research interests include applied linguistics, French studies, and translation studies.
Email: hazlina_ah@upm.edu.my

Zhang Zhongming is a PhD candidate in Translation Studies at Universiti Putra Malaysia (UPM) and a university lecturer in China. His research interests focus on translation assessment, machine translation, and translation education.
Email: gs66634@student.upm.edu.my

eISSN: 2550-2131
ISSN: 1675-8021