Evaluating Google Neural Machine Translation from Chinese to English: Technical vs. Literary Texts

Zhongming Zhang [a]
809396037@qq.com
Faculty of Modern Languages and Communication
Universiti Putra Malaysia, Malaysia

Syed Nurulakla Syed Abdullah [b]
syedakla@gmail.com
Faculty of Modern Languages and Communication
Universiti Putra Malaysia, Malaysia

Muhammad Alif Redzuan Abdullah
muhammadalif@upm.edu.my
Faculty of Modern Languages and Communication
Universiti Putra Malaysia, Malaysia

Wenqi Duan
2352107020@qq.com
Faculty of Modern Languages and Communication
Universiti Putra Malaysia, Malaysia

ABSTRACT

As the global need for translation increases, machine translation (MT) has significantly enhanced the efficiency in facilitating information dissemination and cross-cultural communication. However, its quality remains bound by intrinsic limitations among language pairs and text genres. These discrepancies lead to distinct MT performance when processing technical and literary texts, forming the core gap and focus. This study aims to compare the quality of Google Neural Machine Translation (GNMT) in literary and technical texts, investigating error disparities and establishing the abilities and limits of MT across diverse linguistic contexts. The research was concerned with the English-Chinese language pair with the Multidimensional Quality Metrics (MQM) framework for manual annotation. The COMET automatic evaluation metric was also applied for validation and confirmation of quality differences observed. This study selected five excerpts from Apple product manuals (33 aligned units) and the novel, *the Old Man and Sea* (32 aligned units), respectively. Findings included (1) GNMT performed well with technical texts, but acted less effective with literary texts and technical texts exhibited notable terminology errors, whereas literary texts showed more stylistic inconsistencies; (2) MQM scores demonstrated a remarkable difference, with technical texts outperforming literary texts by 18.57%; and (3) COMET evaluation validated the above observations, confirming a significant difference between the two text styles. Although GNMT faced challenges with both texts, the quality remained acceptable within this study. Results recommend improving GNMT algorithms to enhance accuracy and remedy error patterns and distributions.

Keywords: Google Neural Machine Translation (GNMT); Translation Quality Evaluation; Technical and Literary Texts; Multidimensional Quality Metrics (MQM); COMET Metric

a Main author
b Corresponding author

INTRODUCTION

Neural network technology has greatly advanced machine translation (MT), leading to a significant improvement in the quality of neural machine translation (NMT) (Bahdanau et al., 2014). NMT, at this stage, has become the gold standard for practical applications (Tan et al., 2020). The Google Neural Machine Translation (GNMT) system, which relies on deep learning and large-scale corpus training, has become a widely used NMT system that demonstrates superiority over conventional statistical machine translation (SMT) models (Wu et al., 2016). Building on these developments, a further transformative milestone in NMT is marked by the introduction of the Transformer architecture by Vaswani et al. (2017). This model brings substantial improvements by effectively capturing long-range dependencies and enabling parallel processing through self-attention mechanisms.

While these advancements have considerably strengthened NMT systems, GNMT's performance remains highly inconsistent across various domains and language pairs, performing better on well-structured texts but worse on more flexible and context-dependent content (Shahnazaryan & Beloucif, 2024). To this end, investigating variations in MT quality across different genres holds considerable relevance for both theoretical inquiry and everyday translation practice.

Empirical studies on MT quality across various genres have been conducted for some time. Technical texts consistently yield higher translation quality and are well-aligned with MT system capabilities (Alenezi, 2024; Chéragui, 2012). However, the lack of domain-specific resources frequently results in errors in specialised terminology (Naveen & Trojovsky, 2024). On the other hand, MT of literary texts proves more challenging, as they rely heavily on rhetorical devices, wordplay, and nuanced expression (Chéragui, 2012). Nonetheless, Way et al. (2023) observed a growing openness among literary translators to employing machine translation as a draft-generation tool, particularly in facilitating creative translation strategies. As an increasing number of technologically adept translators enter the market, the use of MT tools in literary translation is likely to rise, with the potential to reshape conventional translation practices.

Recent advances in deep neural networks have substantially accelerated progress in machine translation. More recently, the field has shifted towards generative translation frameworks powered by large language models (LLMs) (Lyu et al., 2023). This transition marks more than a change in architecture—it even redefines translation as a facet of general-purpose language generation, exemplified by systems such as ChatGPT. Although both NMT and LLM-based translation share a foundation in deep learning, particularly the Transformer architecture, they differ markedly in training objectives and operational design. NMT systems, such as GNMT, are trained explicitly on large-scale parallel corpora, optimising for bilingual alignment. In contrast, LLMs are pre-trained on multilingual, multi-task corpora and generate translations through prompting or limited fine-tuning (Zhang et al., 2023). This enables zero-shot or few-shot translation, even without access to aligned bilingual data (Ji et al., 2024). Despite these innovations, LLM-based translation remains sensitive to prompt phrasing and often produces inconsistent outputs for same prompts (McIntosh et al., 2025). These inconsistencies pose significant challenges for evaluation and raise concerns about reliability in practical use. In light of these issues, this study focuses on Google Translate, a stable and widely adopted NMT system, to provide a consistent and interpretable basis for analysing translation quality across distinct text genres.

Although previous studies have offered valuable insights into the effectiveness of machine translation within specific domains, most have examined either technical or literary texts in isolation. In the context of English-Chinese literary translation, NMT systems continue to exhibit issues related to accuracy and fluency (Hu & Li, 2023), encounter difficulties with stylistic rendering (Zhao et al., 2024), and raise concerns about their potential to constrain creative expression (Guerberof-Arenas et al., 2022). By contrast, research on technical translation emphasises its formulaic and repetitive structure (Maxmudjanovna & Xamidjanovna, 2021), with terminology-related errors emerging as a prominent concern (Ying et al., 2021; Kostikova et al., 2019).

This study focuses on these two genres because they represent fundamentally distinct translation challenges. Literary texts require attention to accuracy, fluency, and stylistic fidelity, whereas technical texts demand terminological precision, consistency, and syntactic clarity. Comparing GNMT's performance across these genres enables a more comprehensive understanding of its capabilities and limitations in varied linguistic contexts. As MT tools become increasingly integrated into both professional and creative workflows, such genre-sensitive evaluation is relevant and essential.

This study aims to evaluate the performance of GNMT across technical and literary texts, focusing on translation quality and error patterns identified through automatic metrics and human annotation. To guide this investigation, the study addresses the following two research questions: (1) Do technical and literary texts exhibit significant differences in GNMT translation quality as measured through automatic evaluation metrics? (2) How do error categories and severity levels differ between the two text types as evaluated through human annotation? The structural framework is outlined as follows: the literature review analyses current research on NMT performance and error classification in technical and literary texts; the methodology details the text selection processes and both human and automatic evaluation methods; the analysis compares statistical scores and error types, thereby highlighting differences in translation quality; and the conclusion summarises the findings while offering recommendations for future research.

LITERATURE REVIEW

NMT FOR ENGLISH AND CHINESE LANGUAGE PAIRS

The concept of computer-based translation was first introduced in a memorandum by Warren Weaver in 1949 (Weaver, 1952). Since then, MT quality has faced persistent challenges in natural language processing (NLP), particularly in ensuring consistency across different language pairs and text genres. Hutchins (1986) was among the early researchers who described MT as the use of computers to enable translation between natural languages. As research in this field progressed, the MT paradigm has evolved from rule-based and transfer-based approaches to SMT, and more recently, to NMT.

Recent research highlights the advantages of NMT over SMT in Chinese and English translation, demonstrating considerable quality improvements. For example, Liu (2020) pointed out that NMT delivers better performance in handling cohesive devices, adverbs, and pronouns, leading to improved accuracy. Similarly, Hu and Li (2023) assessed DeepL's performance in Shakespearean drama translation, revealing over 80% accuracy and fluency, thereby highlighting its prospects for both literary translation and stylistic adaptation.

Despite these developments, NMT still faces persistent challenges. For instance, its performance degrades with longer sentences, and domain shifts significantly impact translation quality (Koehn, 2020). While NMT outperforms SMT in handling context, it may be less capable than LLMs in terms of contextual flexibility (Zhao et al., 2024). Lu (2023) evaluated GNMT's performance in Chinese-English translation, identifying consistent difficulties in processing texts with cultural nuances and Chinese-specific contexts, thereby reinforcing NMT's inability to fully replace human translators. Cai (2024) compared ChatGPT, DeepL, and GNMT, finding that NMT struggles with cultural references and idioms, often producing imprecise and incoherent translations. Additionally, the study found that NMT systems frequently generate grammatical errors and rigid source-language structures, limiting their flexibility and creativity. These limitations become particularly evident when comparing NMT performance across different text genres, as the following section explores.

NMT ERRORS IN TECHNICAL AND LITERARY TEXTS

The challenges in machine translation differ across text types, as each genre presents unique linguistic and stylistic difficulties. Recognising these genre-specific error patterns is crucial for evaluating MT performance and determining areas where human intervention is necessary. Kostikova et al. (2019) compared GNMT and Pragma Online for translating scientific and technological texts from English into Ukrainian in terms of vocabulary, grammatical correctness, terminology accuracy, and sentence structure. Their results revealed that GNMT performs better in vocabulary alignment, terminology translation accuracy, and phrasal cohesion. Ying et al. (2021) focused on terminology errors in English-Chinese patent translations, identifying major challenges including mislabelling terms as verbs or nouns, redundancy, homophonic ambiguity, and abbreviation-related errors. Alenezi (2024) explored GNMT errors in technical papers from Arabic to English, classifying the errors as comprehension-, linguistic-, or translation-related, and comparing these with human translation. The study revealed that, in some cases, GNMT performed better than human translators, highlighting its use as an efficient resource.

For literary texts, Kuzman et al. (2019) compared NMT systems for translating English novels into Slovenian, contrasting a novel-specific domain model with GNMT. Using human and automatic evaluations, they assessed errors in grammar, vocabulary, semantics, and cultural adaptation. While the custom model performed better in specific contexts, Google's model generally outperformed it. The study also indicated that NMT increases efficiency in literary translation by reducing post-editing demands. Long et al. (2023) tested MT effectiveness within literary contexts, identifying key hindrances in translating classical Chinese texts and poetry. Their findings showed that GNMT struggled with cultural vocabulary and accuracy, confirming the need for human intervention in the literary translation process. Recently, He et al. (2024) compared Bing and Youdao Translator for translating English literary texts into Chinese, measuring their accuracy, fluency, and contextual appropriateness. The study found that both NMT systems struggle to accurately convey literary nuances, especially in preserving stylistic and rhetorical complexities.

NMT QUALITY EVALUATION

The European Union-supported QTLaunchPad project created the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2013), a standardised tool for the assessment of human, machine, and AI-generated translation. One implementation of this framework is the DQF-

MQM error typology, which incorporates MQM into the Dynamic Quality Framework (DQF), developed by the Translation Automation User Society (TAUS), for practical application in both industry and research. For instance, Wang and Wang (2019) adopted MQM to assess post-edited MT output from English to Chinese. Their study compared human and MT performance in terms of speed, quality, and translator perceptions to evaluate GNMT's accuracy in technical texts. By focusing on accuracy, fluency, and their 13 sub-error categories, they found that accuracy issues were less prevalent compared to other error types, offering valuable insights for future MT improvements. Similarly, Liu et al. (2021) analysed four Chinese-to-English NMT systems in specialised scientific texts. Based on MQM, they further refined error types into six categories, including no translation, mistranslation, addition, repetition, grammar, and punctuation errors. They reported an average error rate exceeding 10%, highlighting persistent challenges in technical translation.

Dunder et al. (2021) applied the DQF-MQM model to examine Croatian-to-German poetry translation. Their analysis showed that accuracy-related errors accounted for the largest proportion, followed by issues related to language use, terminology, stylistic aspects, and localisation. In another study, Fakih et al. (2024) evaluated Instagram's NMT system for Arabic-to-English literary translation, finding that errors occurred in 90% of the samples examined. This study identified accuracy-related errors caused by contextual misunderstandings, fluency issues resulting from grammatical inconsistencies, and stylistic errors arising from literal translations that disregarded cultural nuances. In addition, Zhao et al. (2024) compared the use of traditional NMT systems with LLMs for Chinese-to-English translation and found accuracy-related errors. Their study demonstrated that the MQM framework adequately captured the weaknesses of both NMT systems and LLMs, thereby enabling a comprehensive evaluation of their respective performances.

While automatic evaluation metrics lack the holistic view typical of human judgement and often fail to address specific translation issues, they offer significant advantages in terms of speed, cost, and reduced dependence on human references. Ulitkin et al. (2021) demonstrated that modern automatic metrics are capable of capturing improvements within translation systems, thus providing a foundation for further quality enhancements. Classic surface metrics, such as BLEU (Papineni et al., 2002), may pose challenges in capturing linguistic nuances and aesthetic qualities in complex texts, such as literary works. Comparatively, COMET (Rei et al., 2020), a neural model created to evaluate multilingual machine translation, demonstrates a high correlation with human assessments. For instance, Peng (2023) analysed the performance of NMT in scientific texts and reported that while COMET performs well in sentence-level evaluation, it tends to overlook broader discourse-level issues. For example, even when only the first sentence in a document is accurately translated and the remainder contains substantial errors, COMET may still generate a high overall score. This raises concerns about its reliability in assessing coherence and consistency in document-level evaluation.

Toral et al. (2024) incorporated automatic metrics and human evaluation in literary translation and found that COMET performs better in capturing the nuances among various NMT systems. Similarly, COMET was used to evaluate English-to-Dutch novel translation by Ploeger et al. (2024), who found that while the baseline model of NMT achieved higher marks, excessive lexical diversity reduced COMET scores. It is worth noting that both Toral et al. (2024) and Ploeger et al. (2024) assert that automatic measures alone should not determine literary translation quality and emphasise the importance of human evaluation.

In summary, although prior research has shed light on NMT performance across various genres and evaluation approaches, direct comparisons between English-Chinese technical and literary texts remain limited, particularly those integrating both human annotation and automatic metrics. This study addresses this gap by conducting a focused, genre-specific evaluation of GNMT translations using a mixed-methods approach.

METHODOLOGY

In the process of human evaluation, assessment criteria are adapted to the text's intended function, intended user, and stylistic requirements, allowing identification of underlying linguistic difficulties (Siu, 2023). Accordingly, this study combined automatic evaluation with qualitative analysis to comprehensively assess GNMT performance on both text types at sentence level. Thus, the MQM framework was adopted for qualitative error analysis. Additionally, to increase objectivity and ensure cross-validation, the study adopted COMET metric, providing insights into overall translation patterns and differences.

As illustrated in Figure 1, this study employed a three-step workflow to evaluate the performance of GNMT across technical and literary texts. The process began with the selection of source materials, five excerpts each from Apple product manuals (technical) and *The Old Man and the Sea* (literary). These texts were translated into English using the GNMT engine. Following translation, evaluation was conducted through two complementary methods. First, automatic evaluation was carried out using the COMET metric. Human-translated Chinese-English texts served as reference translations: technical samples were drawn from publicly available Apple manuals, while the literary references were taken from Yu Guangzhong's published Chinese translation of *The Old Man and the Sea*. This yielded quantitative performance scores for each genre. In parallel, human evaluation was carried out using the MQM framework, which involves categorising and rating errors by type and severity. Inter-annotator agreement was assessed to ensure consistency across evaluators. Finally, results from both evaluation methods were compared to identify genre-specific translation challenges and performance patterns.

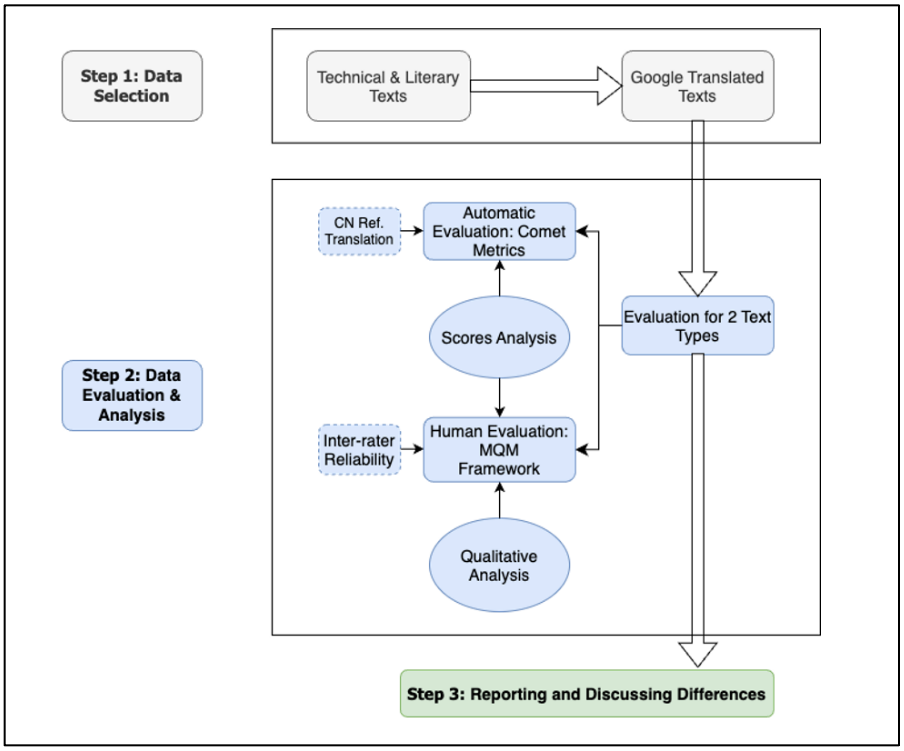
Step 1: Data Selection
Technical & Literary Texts
Google Translated Texts
CN Ref. Translation
Automatic Evaluation: Comet Metrics
Scores Analysis
Evaluation for 2 Text Types
Step 2: Data Evaluation & Analysis
Inter-rater Reliability
Human Evaluation: MQM Framework
Qualitative Analysis
Step 3: Reporting and Discussing Differences

FIGURE 1. GNMT comparative evaluation flowchart

DATA SELECTION

The technical and literary samples each comprised five excerpts, totalling 425 and 482 words, respectively. The technical excerpts were drawn from Apple product manuals[1], with their corresponding Chinese translations serving as reference texts for automatic evaluation. Each excerpt, approximately 80-100 words in length, represented a distinct Apple product: the Apple Vision Pro, HomePod, AirPods Max, Studio Display, and MacBook Air. Although the total word count is modest, the selection intentionally covers five product categories to ensure broad representation of Apple's consumer technology portfolio within the study's scope. The bilingual materials were sourced from publicly available documentation on Apple's official support website. As these documents are openly accessible and intended for general user reference, their inclusion in this study raises no privacy or ethical concerns.

The selected samples exhibit a wide range of terminology and complex syntactic structures, providing both linguistic variety and minimal redundancy. Multiple text types are also included, such as descriptive content, operational instructions, and regulatory information, thereby maintaining structural and functional diversity. Given the popularity and global reach of these products, the technical samples contribute to the study's practical value by offering insight into the translation challenges inherent in widely distributed, high-stakes technical materials.

The literary texts in this study come from Ernest Hemingway's 1952 novel *The Old Man and the Sea*, which is widely praised for its use of concise, conversational vocabulary and vivid imagery through the application of plain nouns and verbs to portray different scenes (Xie, 2008).

[1] *Apple Support Manuals available at: https://support.apple.com/zh-cn/docs (Chinese) and https://support.apple.com/en-gb/docs (English).*

Since the effectiveness of NMT models decreases as sentence length and syntactic difficulty increase (Koehn, 2020), the short sentence forms and plain narrative style of Hemingway make his novel particularly suitable for machine translation research. For COMET-based automatic evaluation, this study uses the Chinese translation of *The Old Man and the Sea* by Yu Guangzhong, a renowned Taiwanese scholar, translator, poet, and essayist. His translation is widely recognised for its linguistic fluency, literary depth, and fidelity to the original text (Ng Y. L. E., 2009). Yu's translation has been published in multiple widely circulated editions and remains publicly accessible. It is also frequently cited by Chinese readers and scholars as a representative version, praised for its expressive style and readability (Fang, 2022).

Using a single human reference translation may introduce stylistic bias, which constitutes a limitation of this study. However, employing multiple reference translations increases stylistic variability, potentially affecting the consistency and comparability of automatic evaluation results. To address this methodological trade-off, a single high-quality and widely recognised translation was adopted to reduce data variability and ensure greater uniformity in scoring (Freitag et al., 2021). Furthermore, as COMET (wmt20-COMET-da) evaluates semantic similarity and fluency based on reference translations, the use of a standard and reputable version enhances the validity and reliability of the assessment process (Rei et al., 2020).

Additionally, the sample selection prioritises excerpts with linguistic complexity and a higher likelihood of exhibiting typical translation errors, thereby enhancing the study's analytical depth. The selection strategy is guided by the following criteria: texts are chosen to reflect key linguistic features and thematic content, offering a representative overview of their stylistic and structural characteristics. Each excerpt includes four to eight complete sentences, providing sufficient structural complexity for meaningful analysis.

To address syntactic and structural differences between English and Chinese, the study employs aligned units—rather than individual sentences—as the basic unit of evaluation. In cross-linguistic translation, one sentence in the source language may correspond to multiple sentences in the target language. For instance, a long English sentence may be rendered into two shorter Chinese sentences, or vice versa, due to differences in syntactic conventions and information packaging. Each aligned unit in this study consists of a single English sentence and its full Chinese translation, regardless of sentence breaks. This approach ensures consistent and coherent alignment across both languages, thereby facilitating more accurate evaluation. As a result, the number of aligned units across text types remains comparable: 33 for the technical excerpts and 32 for the literary excerpts. This slight variation stems from differences in sentence segmentation across the selected passages.

EVALUATION METHODS

This study integrates human and automatic evaluation to comprehensively assess translation quality. The automatic evaluation employs COMET, an advanced evaluation metric and neural framework for translation assessment (Stewart et al., 2020), alongside the previously mentioned Chinese references to ensure a robust and balanced evaluation approach. The latest COMET algorithm complements human evaluation by extending its analytical scope and coverage. Scores are generated using the wmt20-COMET-da model, a reference-based regression approach built on XLM-R and trained on WMT17 to WMT20 direct assessments. COMET calculations[2] are executed via the Command Prompt on Windows 10 Professional.

[2] *GitHub: https://github.com/Unbabel/COMET*

Given that human evaluation is generally more reliable and sensitive to subtle translation errors (Stewart et al., 2020), this study adopts the MQM framework for qualitative analysis, as illustrated in Figure 2. The framework comprises over 100 error categories (Lommel et al., 2013), with detailed error classifications minimising subjectivity while enhancing evaluation efficiency and flexibility. The latest MQM taxonomy defines eight core error types, each further categorised into detailed subtypes, ensuring greater adaptability across diverse translation contexts (Lommel et al., 2014).

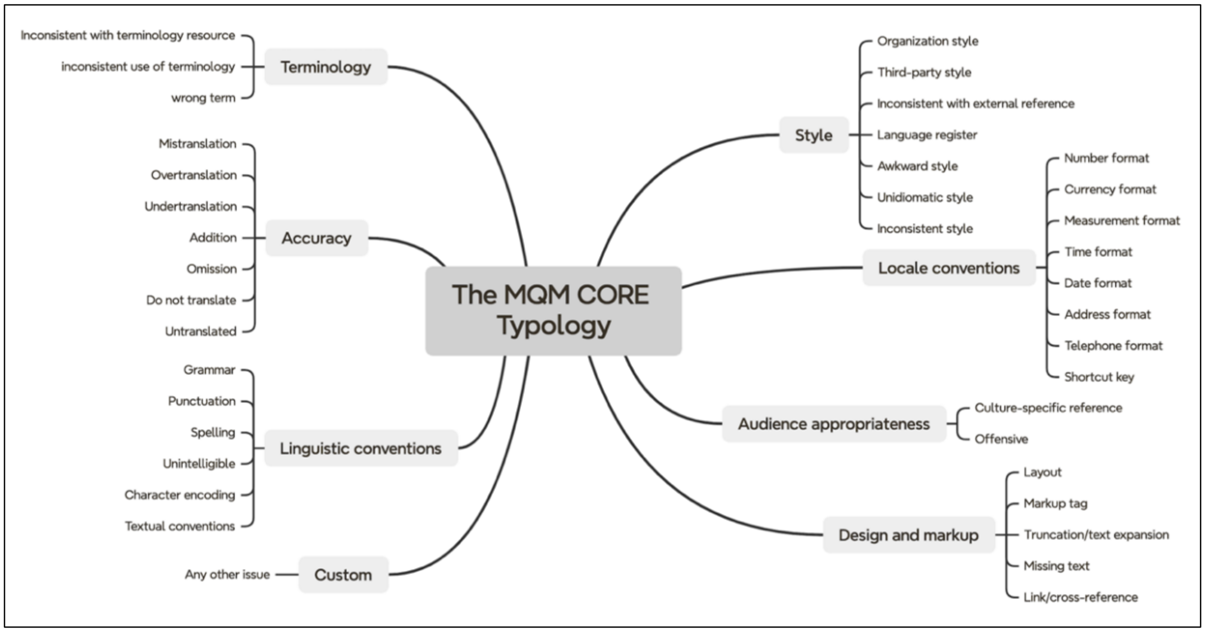

FIGURE 2. The MQM-core typology (https://themqm.org/the-mqm-typology/)

This study employs four core error categories from the MQM framework—accuracy, linguistic conventions (formerly referred to as "fluency" in earlier MQM versions), terminology, and style. Accuracy and fluency are critical components of high-quality translation (Lommel et al., 2014). Meanwhile, terminology and style errors also significantly impact both technical and literary texts. Technical translation depends on precise terminology usage to prevent ambiguity and ensure clarity (Mohsen, 2024), while literary translation requires the preservation of stylistic features to maintain the original work's essence (Jinfang et al., 2025).

Annotators are required to categorise error types and severity levels, and errors are classified as minor, major, or critical, with penalty weights of 1×, 5×, and 10×, respectively. According to MQM guidelines, error severity reflects the extent to which an issue impacts the usability and communicative effectiveness of the text. Minor errors may slightly affect accuracy, style, or clarity but do not impede overall comprehension. Major errors substantially alter meaning or undermine reliability, potentially interfering with proper use or interpretation. Critical errors render the content unusable for its intended function or may result in significant consequences, including physical, financial, or reputational harm. To maintain consistency, all error types are assigned equal weighting (1), while additional values, such as the Max Score Value, are calculated by default. Two professional translators perform manual annotation, with Cohen's Kappa coefficient (Cohen, 1960) applied to measure inter-annotator agreement. Any discrepancies are resolved through researcher consensus to ensure evaluation reliability.

RESULTS AND ANALYSIS

COMET EVALUATION

The study began by evaluating the translation quality of both text types using the COMET metric, which provided a statistical measure of performance based on aligned units. Each aligned unit comprised a sentence from the source text and its complete GNMT translation, accommodating structural differences between English and Chinese. In COMET, scores approaching one indicated high translation quality, whereas those closer to zero suggested random or unreliable translations. While raw COMET scores lacked direct interpretability, comparative analysis offered valuable insights into quality variations. As shown in Table 1, both text types achieved relatively high scores (>0.6), with technical texts consistently outperforming literary texts, underscoring a distinct advantage in structural and linguistic regularity.

Technical texts demonstrated average COMET scores ranging from 0.8971 to 0.9387, compared to 0.8062 to 0.8557 for literary texts. This disparity highlighted the greater syntactic consistency and terminological precision of the former, which facilitated higher translation quality. A closer analysis of individual units revealed stable performance in technical texts, with scores exceeding 0.9 in some segments. Conversely, literary texts exhibited greater variability, with notably lower scores in Unit 6, No. 2 (0.6663) and Unit 2, No. 3 (0.6748). These fluctuations likely resulted from the inherent complexities of literary language, including nuanced expressions and layered meanings, which posed significant challenges for GNMT.

Interestingly, the literary text in Unit 3, No. 2 achieved a score of 0.9579, surpassing certain technical text segments, suggesting that literary translations can occasionally attain comparable quality levels. Notably, minor declines in technical text scores, such as in Unit 1, No. 5 (0.8847) and Unit 4, No. 3 (0.8844), indicated challenges in translating specialised terminology or complex conceptual structures. To substantiate these findings, further qualitative analysis through human evaluation was necessary.

TABLE 1. COMET scores for literary and technical excerpts

	No.	Aligned unit 1	Aligned unit 2	Aligned unit 3	Aligned unit 4	Aligned unit 5	Aligned unit 6	Aligned unit 7	Aligned unit 8	Avg. Score
Literary Texts	1	0.9159	0.8513	0.7169	0.9312	0.7790	0.8965	0.8992		0.8557
	2	0.9287	0.9470	0.9579	0.7570	0.8714	0.6663	0.8263	0.8777	0.8540
	3	0.9429	0.6748	0.9429	0.7859	0.7062	0.8124	0.8707		0.8194
	4	0.8264	0.7237	0.7662	0.9083					0.8062
	5	0.9273	0.7850	0.9329	0.8202	0.7427	0.7165			0.8208
Technical Texts	1	0.9821	0.9130	0.9362	0.9389	0.9514	0.9167	0.9197		0.9369
	2	0.9733	0.9424	0.8564	0.9171	0.8913	0.9496			0.9217
	3	0.9186	0.9080	0.8817	0.8844	0.8498	0.9830	0.9671	0.9251	0.9147
	4	0.9758	0.8923	0.9772	0.9828	0.8881	0.8946	0.9602		0.9387
	5	0.8847	0.9327	0.9087	0.8871	0.8723				0.8971

Due to the small sample size (technical text: n = 33; literary text: n = 32), the Shapiro-Wilk test was used to assess normality in both groups. Results (Table 2a) indicated that the technical text group did not differ significantly from normality (W = 0.956, p = 0.192), whereas the literary text group showed a significant departure (W = 0.930, p = 0.038), violating the assumption of normality. The non-parametric Mann-Whitney U test was therefore used to compare translation quality scores. The findings (Table 2b) revealed a statistically significant difference (U = 227.000, W = 755.000, p < 0.001), demonstrating that the GNMT quality of technical texts differed significantly from that of literary texts.

TABLE 2a. Normality test results for COMET scores

Group	Kolmogorov-Smirnov	Sig.	Shapiro-Wilk	Sig.
Technical text	0.103	0.200	0.956	0.192
Literary text	0.129	0.189	0.930	0.038

TABLE 2b. Mann-Whitney U test for translation quality comparison

Test	Score	p-value (2-tailed)
Mann-Whitney U	227.000	<0.001
Wilcoxon W	755.000	<0.001
Z	-3.950	-

The grouped scatter plot (Figure 3) showed the data distribution across excerpts, revealing differences and inconsistencies. Technical reports consistently had higher median COMET scores for all five excerpts, implying that machine translation performed better with technical texts. In Excerpt 1, technical text scores were closely clustered around a high median, reflecting low variability and consistent performance. In contrast, literary text scores for the same passage were more dispersed, indicating translation inconsistencies.

Overall, literary texts had lower median values and broader interquartile ranges than technical texts, highlighting greater variability and challenges in achieving comparable translation quality. Excerpt 2 exemplified this pattern, as technical text scores showed higher medians and a narrower range, whereas literary text scores are more widely spread. This discrepancy underscored GNMT's difficulty in translating the nuanced language of literary works, necessitating further qualitative analysis to identify specific challenges.

These results illustrated the impact of text type on MT performance. Technical texts, with standardised vocabulary and straightforward structures, aligned better with MT systems, ensuring greater consistency in translation quality. By contrast, although literary texts may have achieved comparable COMET scores, their translation reliability was lower, with greater fluctuations across segments. This volatility highlighted GNMT's difficulty in handling literary texts, emphasising the complexities of literary translation.

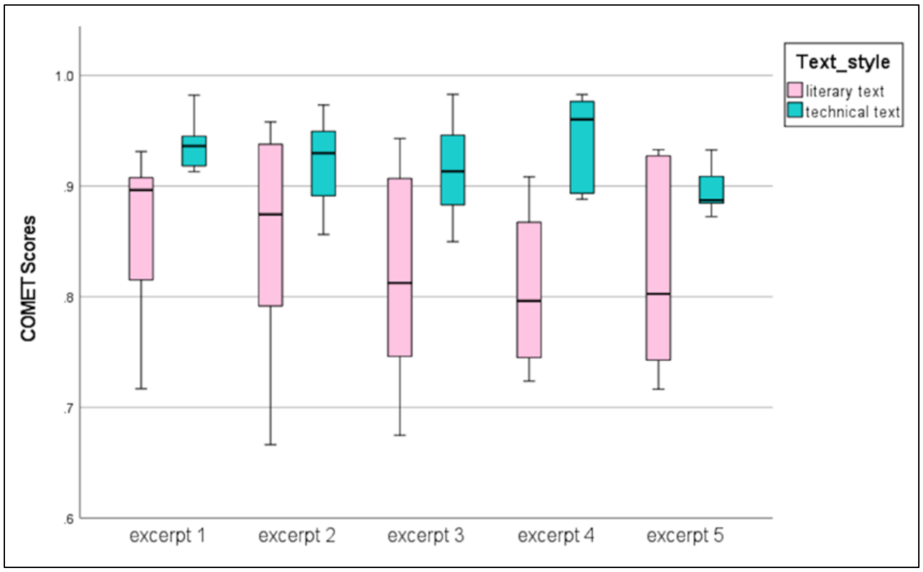

FIGURE 3. The grouped scatter plot of COMET scores

INTER-ANNOTATOR AGREEMENT

Assessing agreement between annotators is vital for evaluating the reliability of results. This study employed Cohen’s Kappa coefficient (Cohen, 1960) to measure annotation consistency across translated texts. As shown in Table 3, the p-values for both text categories and the overall Kappa coefficient were below 0.001, confirming statistical significance and reducing the likelihood of chance agreement. The Kappa values for technical and literary texts were 0.669 and 0.658, respectively, with an overall coefficient of 0.669. Landis and Koch (1977) considered Kappa values between 0.6 and 0.7 to indicate substantial agreement. The slightly lower Kappa value for literary texts suggested greater variation in error classification, likely due to the complexity of literary translation assessment.

The higher frequency of annotation errors for literary texts (23) compared with technical texts (15) supported this finding. The complexity of annotating literary texts likely stems from their linguistic nuances, interpretative openness, and subjective nature, posing additional challenges for annotators. These factors contribute to differences in agreement, highlighting the difficulty of consistently evaluating literary translations.

TABLE 3. Kappa coefficient by two annotators

Text styles	Cohen’s Kappa	p-value	Number of Errors
Technical texts	0.669	<.001	15
Literary texts	0.658	<.001	23
Total texts	0.669	<.001	38

MQM ERROR ANALYSIS

According to the MQM framework, the study categorised errors into four primary types: style, linguistic conventions, accuracy, and terminology. Errors were further classified by severity as minor, major, or critical, as shown in Table 4. A total of 15 errors were identified in technical texts and 21 in literary texts, thus indicating a higher frequency and severity of errors in literary translation.

TABLE 4. Error types and severity in technical and literary texts

Text Type	Error Type	Sub-Error Type	Minor	Major	Critical
Technical Text	**Terminology**	Inconsistent with terminology resource	2	0	0
	Accuracy	Mistranslation	2	3	0
		Over-translation	1	0	0
		Under-translation	1	0	0
		Omission	0	1	0
		MT hallucination	0	1	0
	Linguistic Conventions	Grammar	1	0	0
	Style	Awkward style	2	0	0
		Inconsistent style	1	0	0
Literary Text	**Accuracy**	Mistranslation	6	1	2
		Untranslated	0	0	1
		Under-translation	0	1	0
	Linguistic Conventions	Grammar	1	0	0
	Style	Awkward style	3	1	0
		Inconsistent style	1	0	0
		Unidiomatic style	3	1	0

Technical translations exhibited a structured error distribution across all four categories. Accuracy errors were the most frequent, with nine cases (four minor, five major), followed by two minor terminological inconsistencies. Style errors in technical texts were fewer (three minor). Notably, technical texts contained no critical errors, and most inaccuracies involved minor lexical or grammatical deviations rather than major meaning distortions.

In contrast, literary translations showed greater variability and more severe errors, particularly in accuracy and stylistic features. Accuracy problems dominated, including six minor mistranslations, two major errors, and three critical errors, where the untranslated segment (critical error) and under-translation (major error) highlighted GNMT's difficulty in capturing nuanced and implicit meanings in literary texts. Compared with technical texts, style-related errors were more frequent and severe (seven minor and two major). These patterns indicated that GNMT struggled to maintain the stylistic and idiomatic coherence necessary for literary translation.

Both genres contained only one minor grammatical error, indicating GNMT's fluency. However, the absence of critical errors in technical translations, contrasted with their presence in literary texts, underscored the disparity. Thus, the observation confirmed that GNMT performed better on technical texts. These findings highlighted the need for advanced MT methods, particularly for literary translation, to improve semantic accuracy and maintain stylistic fidelity. The following sections provide example analyses, offering insights into key translation challenges and potential improvements in MT technology.

TERMINOLOGY

Terminology errors occur when target terms deviate from their expected or conventional counterparts in the source text. In this study, two minor terminology errors were identified in technical texts, whereas no such errors were observed in literary texts. The finding suggests that literary texts are less susceptible to terminology inconsistencies, while the specialized terminology in technical texts poses a greater challenge for NMT systems. Table 5 provides a case analysis of terminology errors.

TABLE 5. Examples of evaluation results on terminology errors

ST	This equipment should be **serviced** by Apple or an authorised service provider.
GNMT	本设备应由Apple或授权服务提供商**进行维修**。
Ref. T	此设备应由Apple或授权服务提供商**提供相关服务**。
ST	Discoloration of the **Lightning connector** after regular use is normal.
GNMT	定期使用后，**Lightning接口**变色是正常现象。
Ref. T	**闪电接头**随着正常使用而褪色是正常现象。

The translation of the term “service” as “维修” (maintenance) is unduly restrictive. Though “维修” conveys repair, in technical contexts the English term “service” refers to a wide spectrum of activities, including maintenance, repair, upgrading, and general upkeep. By limiting the translation to “maintenance”, the translated text fails to capture the full range of services intended in the original, potentially leading to ambiguity—particularly in maintenance or product support contracts, where service coverage is a key factor.

In addition, the word “lightning”, left untranslated, presents a particular challenge. Though its presence is unlikely to cause misunderstanding due to its established use in Chinese technical nomenclature, its untranslated state—rather than being rendered with an accurate Chinese equivalent—constitutes a minor terminological anomaly. Although such non-translation does not necessarily impede comprehension, it diverges from standard translation practice and may affect the formality and coherence of the text. In formal or official contexts, such inconsistencies could be interpreted as a lack of precision or meticulousness.

These examples highlight the importance of terminological precision and contextual awareness in technical translation. Maintaining terms that are not only precise but also contextually comprehensive and aligned with the expectations of the target audience is crucial to ensuring the clarity, professionalism, and coherence of technical documentation.

ACCURACY

TABLE 6. Examples of evaluation results on mistranslation errors

	ST	GNMT	Ref. T
Technical text	It’s important to keep your MacBook Air on a **hard**, stable, and well-ventilated work surface when in use or charging.	使用或充电时，务必将MacBook Air放在**坚硬**、稳定且通风良好的工作台面上。	使用 MacBook Air 或给其充电时，请注意将设备放置在**牢固**、平稳且通风良好的工作平台上。
Literary text	Five, and **you nearly were killed** when I brought the fish in **too green**, and he nearly tore the boat to pieces.	五岁，**我**把一条鱼带上船时**差点把你弄死**，它差点把船撕成碎片。	五岁。**你差点送了命**，当时我**太早**把鱼拉上来了，它几乎把船撞碎。

Accuracy errors are pronounced in both text categories, with mistranslations being particularly evident. As shown in Table 6, the technical text contains a basic error where the word "hard" is incorrectly translated as "坚硬的" (hard, physically) rather than the more appropriate "牢固的" (secure) or "安全的" (safe). This misinterpretation stems from the polysemous nature of the word "hard", leading to an erroneous lexical choice that compromises clarity in a technical context, where accuracy is paramount.

In the given literary text, a far more severe mistranslation occurs. The first passage depicts a dangerous fishing incident in which the old man's life is endangered, and the young boy is nearly killed. The machine translation, however, misrepresents the original sentiment and distorts it as "老人差点把小男孩弄死了" (The old man nearly killed the boy), fundamentally altering the nature of the relationship between the characters and the emotional atmosphere of the scene. Moreover, the omission of "too green", which conveys a lack of maturity and experience, diminishes the thematic complexity of the literary work. This significant deletion removes a key component of meaning, thereby reducing insight into vulnerability and character development.

In summary, the accuracy issues observed, particularly instances of mistranslation and information loss, highlight the limitations of GNMT in achieving literal accuracy and contextual integrity. These inaccuracies are especially problematic in technical reporting, where precise terminology is essential, and in literary works, where emotional and thematic consistency must be preserved. The results of this study demonstrate the ongoing need for improvements in GNMT to better handle semantic subtleties and contextual complexities.

LINGUISTIC CONVENTIONS

TABLE 7. Examples of evaluation results on linguistic conventions errors

	ST	GNMT	Ref. T
Technical text	HomePod contains a radio and other components **that emit electromagnetic fields.**	HomePod 包含无线电和其他**发射电磁场的**组件。	HomePod 包含的无线电和其他组件**会发射电磁场**。
Literary text	…and the noise of you clubbing him **like chopping a tree down**...	你**像砍树一样**用棍棒打它的声音	你用棍子打它的声音**就像砍倒了一棵树**

As Table 7 illustrates, both text types exhibit minor errors in linguistic conventions which, though not greatly affecting fluency, still require correction, emphasising the need for precision in translation. In the technical text, GNMT constructs a "that" clause preceding "other components" as an attributive, which, although grammatically acceptable, shifts the original focus. This change weakens the emphasis on "Medical Device Interference", diverting attention from a critical warning. While not grammatically incorrect, annotators observe that it fails to adequately highlight key information in this context. They suggest adding commas to mark the clause, thereby enhancing clarity and ensuring the warning is more effectively conveyed to the reader. This modification not only improves readability but also aligns more closely with the intended grammatical structure.

In contrast, the linguistic convention error in the literary passage is more pronounced due to unnatural word order. Although the phrase "…like chopping a tree" remains understandable, its premature placement disrupts the sentence's natural flow and rhythm. A better structure would move this adjectival clause to the end of the sentence to preserve stylistic harmony. This minor word order issue does not constitute a major mistranslation but does affect the text's expressiveness. In literary translation, word order adjustments are crucial for retaining emotional depth and stylistic consistency, making such corrections particularly important.

Overall, while linguistic convention errors do not hinder comprehension, they undermine fluency. The structural issue in the technical document underscores the need for clarity in conveying critical information, whereas the altered word order in the literary text highlights the importance of smoothness and naturalness in creative writing.

STYLE

TABLE 8. Examples of evaluation results on style errors

	ST	GNMT	Ref. T
Technical text	**Use common sense to** avoid situations...	**使用常识时**，应避免…	**应当使用常识**以避免…
Literary text	…**in the stupidity of their great hunger** they were losing and finding the scent in their excitement.	…**在极度饥饿的愚蠢中**，它们兴奋地寻找着气味，又迷失了方向。	…**由于十分饥饿…它们昏头昏脑地**，一会儿追丢了腥味，一会儿又再找到。

While both translation types exhibit style issues, they are more pronounced in literary translations (Table 8). Technical translations contain stylistic faults that are generally less severe, whereas literary translations present them more frequently and with greater intensity. This difference underscores the heightened importance of style in literary translation, where linguistic fluidity and artistic coherence are paramount.

For example, in the technical text, "使用常识时" ("use common sense") is somewhat abrupt, affecting overall fluency; however, this error is relatively minor and does not significantly impair comprehension. By contrast, the literary text presents a more severe issue: the word-for-word translation of "在极度饥饿的愚蠢中" ("in the extreme hunger of stupidity") is awkward and unnatural, markedly obstructing both fluency and readability. Identified as a serious unidiomatic translation error, this version fails to convey the emotional depth and literary subtleties of the original, stripping it of its aesthetic appeal. The contrast highlights the critical role of style in literary translation, where tone, atmosphere, and expressiveness are integral to preserving the author's intent. Technical texts, by comparison, prioritise clarity and functionality, allowing more leeway for stylistic infelicities.

COMPARISON OF MQM SCORECARDS

The MQM scorecards categorised errors into four broad types to calculate quality scores for different text genres. As shown in Table 9, the technical text (425 words) was paired with a reference translation of 733 words, accumulating a total penalty of 35 points. With a word penalty of 0.0824, this yielded a quality score of 91.76. In contrast, the literary text (482 words) was matched with a reference translation of 792 words, incurring a total penalty of 109 points. This resulted in a word penalty of 0.2261 and a final quality score of 77.39.

The 18.57% higher quality score in technical translation further illustrated the substantial gap in translation quality between the two genres, reflecting GNMT's superior performance in handling structured, clearly expressed technical content. Conversely, the significantly lower score in literary translation highlighted the challenges NMT faced in processing complex, context-dependent language. Additionally, the consistency between COMET scores and human evaluation reinforced the reliability of automatic evaluation tools, further validated their applicability in translation quality assessment.

TABLE 9. Comparative metrics for GNMT quality

Metric	Technical texts	Literary texts
Evaluation Word Count	425	482
Reference Word Count	733	792
ET Weight	1	1
Absolute Penalty Total	35	109
Per-Word Penalty Total	0.0824	0.2261
Overall Quality Score	91.76	77.39

DISSCUSSION

The automatic assessment revealed notable disparities in translation quality and consistency between the two text types. Technical texts consistently scored higher, enabling more accurate machine translation outcomes. They also exhibited greater score consistency, as reflected by narrower interquartile ranges in the boxplot. Conversely, literary texts showed greater score variability, reflecting differences in quality. Notably, some alignment units in Table 1 had similar scores for both literary and technical texts, suggesting that simpler literary segments achieved comparable translation quality. However, this does not invalidate the overall trend of lower and more variable scores for literary texts, highlighting GNMT's limitations in capturing the stylistic and interpretative subtleties essential for high-quality literary translation. Previous studies, such as Peng (2023), reported high COMET scores (>0.6) for English-French technical texts, while Toral et al. (2024) found lower scores (<0.6) for English-Dutch literary texts. Given the limited COMET-based comparisons in English-Chinese NMT, this study examined the effectiveness of COMET in distinguishing between technical and literary texts.

The qualitative findings from human evaluation also indicated that GNMT performed well in translating technical texts with fewer annotated errors, which aligned with the findings of Alenezi (2024), who highlighted NMT's extensive application in technical fields. Similarly, Kostikova et al. (2019) noted that while GNMT may introduce some grammatical and lexical errors, it demonstrated fewer alignment errors in technical texts, making it suitable for this type of translation. Moreover, Ulitkin et al. (2021) found that GNMT achieved a high rate of lexical coherence compared to reference translations under technical conditions, further supporting its usability in this domain. This study found a higher number of accuracy-related errors in GNMT, consistent with the findings of Wang and Wang (2019). It confirmed that GNMT struggled with terminology, particularly in technical texts, in line with the findings of Ying et al. (2021). Inconsistent terminological accuracy often made its translations appear less professional. Furthermore, MT hallucination, a major error type in this study, reflected GNMT's challenges in processing complex sentence structures, resulting in output uncertainty.

Conversely, GNMT's primary challenges in literary translation lay in accuracy and stylistic fidelity. Accuracy errors were the most frequent, as noted by Zhao et al. (2024) and Dunder et al. (2021), while stylistic inconsistencies were also prominent, aligning with the findings of He et al. (2024) on NMT's struggles with literary style. Similarly, Fakih et al. (2024) found that Instagram's

NMT system exhibited major deficiencies in accuracy, fluency, and stylistic adaptation. Style errors were particularly evident in literary works, corroborating earlier findings by Dunder et al. (2021) and Fakih et al. (2024).

Parallel structure-based translations tend to simplify the original text, diminishing stylistic richness. For instance, in the sentence: “I can remember the tail slapping and banging and the thwart breaking and the noise of the clubbing,” the translation “我记得尾巴拍打和撞击的声音，座板断裂的声音和棍棒的声音。” disrupts the rhythm and introduces redundancy with “声音” (sound), weakening variation and expressiveness. This tendency aligns with Baker’s (1993) translation universals, particularly simplification and normalisation—features that prioritise clarity and regularity over stylistic nuance. Additionally, errors in proper noun translation reveal GNMT’s insensitivity to capitalisation and contextual implications. These findings highlight that while GNMT can achieve functional correctness, it lacks the complexity required for literary style and rhetorical precision, thereby necessitating human intervention.

From a holistic perspective, GNMT’s translation of *The Old Man and the Sea* demonstrated potential and maintained readability. This study aligns with the view of Kuzman et al. (2019), who suggest that NMT systems serve as useful tools for human review or post-editing. However, this contrasts with Tahseen and Hussein’s (2024) claim that MT renders literary texts “meaningless and ambiguous” because machines can neither “think nor feel”. The variation in findings may stem from differences in test text selection, particularly the greater challenges posed by classical texts or poetry translation (Long et al., 2023). Hence, while the limitations of NMT in literary translation remain evident, its output is acceptable under certain conditions.

Human evaluation and automatic scores reveal a positive correlation, indicating that the COMET metric is an effective means of validation. For instance, in the second literary text excerpt, aligned unit 3 received the highest COMET score (0.9579). The source sentence, “He hasn’t much faith,” is a simple structure, allowing GNMT to generate a correct translation. Human annotators also deem the translation accurate, noting no significant errors. However, in the same excerpt, aligned unit 6 has a considerably lower COMET score (0.6663). The first sentence, “‘Why not?’ the old man said. ‘Between fishermen.’” is translated by GNMT as “‘为什么不呢？’老人说, ‘渔民之间。’”. Overreliance on the syntactic structure of the source text highlights GNMT’s inadequate contextual comprehension (Cai, 2024). Consequently, human annotators identify stylistic deficiencies (awkward style, major error) and accuracy-related issues (overly literal translation, major error). Two major errors in this unit indicate that GNMT failed to accurately convey the intended meaning and required significant revision.

According to the findings, GNMT-translated technical texts tend to achieve high quality, as indicated by strong automatic evaluation scores and few errors. GNMT also performs well in technical translation, significantly reducing time and cost while delivering high-quality output. However, literary translation tends to receive lower evaluation scores and produce more major and critical errors. Nonetheless, for relatively simple literary texts from *The Old Man and the Sea*, machine translation remains within an acceptable quality range.

CONCLUSION

This study conducted a comparative evaluation of GNMT's translation quality for English-Chinese technical and literary texts, using the Apple Product Manual and *The Old Man and the Sea* as sample texts. The evaluation combined qualitative analysis under the MQM framework with the COMET metric, categorising errors into four main types: terminology, accuracy, linguistic conventions, and style. The results revealed significant differences between the two genres in GNMT quality, as reflected in both evaluation methods. Technical translations demonstrated greater consistency and accuracy, whereas literary translations contained more major and critical errors, particularly in accuracy and stylistic adaptation, which affected meaning conveyance. Despite these challenges, GNMT's literary translations remained within an acceptable quality range.

Moreover, Apple product manuals exhibited higher quality but included terminological inconsistencies, reflecting a lack of technical precision. Literary translation showed stylistic differences, where parallel structures tended to oversimplify the source text, diminishing its depth. These findings suggest that GNMT may be suitable for technical drafts with light post-editing but remains inadequate for literary texts requiring stylistic fidelity and cultural nuance. Future studies should expand empirical evaluations of English-Chinese machine translation across various text types and incorporate post-editing strategies to enhance translation quality. Developers should also refine GNMT algorithms to improve contextual understanding and stylistic sensitivity in literary translation.

While the findings offer insight into GNMT's genre-specific performance, some limitations remain. The small sample size restricts the generalisability of the results, and the selected literary texts—with their relatively simple narrative style—may not reflect the challenges posed by more complex works. Further research with a broader and more diverse corpus is needed to enhance the robustness of the conclusions. The analysis also confirms that GNMT output still falls short of high-quality translation benchmarks, underscoring the need for continued algorithm refinement and post-editing strategies.

REFERENCES

Alenezi, A. M. (2024). Error analysis of neural machine translation in technical texts: Google Translate as a case study. *Journal of the North for Humanities, 9*(2, Part 1), 167–181. https://doi.org/10.12816/0061799

Alzain, E., Nagi, K. A., & Algobaei, F. (2024). The Quality of Google Translate and ChatGPT English to Arabic Translation: The Case of Scientific Text Translation. In *Forum for Linguistic Studies* (Vol. 6, No. 4, pp. 837-849). http://dx.doi.org/10.30564/fls.v6i3.6799

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baker, M. (2011). Corpus linguistics and translation studies—implications and applications. In *Text and technology: In honour of John Sinclair* (pp. 233-250). John Benjamins Publishing Company.

Cai, L. (2024). How does ChatGPT Compare with Conventional Neural Machine Translation Systems in Performing a Chinese to English Translation Task?. *Journal of Translation Studies, 4*(1), 25-45. http://dx.doi.org/10.3726/JTS012024.02

Chéragui, M. A. (2012). Theoretical Overview of Machine translation. *ICWIT*, 160-169.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Dunder, I., Seljan, S., & Pavlovski, M. (2021). What Makes Machine-Translated Poetry Look Bad? A Human Error Classification Analysis. In *Central European conference on information and intelligent systems* (pp. 183-191). Faculty of Organization and Informatics Varazdin.

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). http://dx.doi.org/10.17576/gema-2024-2401-13

Fang, Q. A Comparative Analysis on Wu Lao's and Yu Guangzhong's Chinese Versions of The Old Man and the Sea. *Journal of Innovation and Social Science Research, 9*(9), 504–507.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics,9*, 1460-1474. http://dx.doi.org/10.1162/tacl_a_00437

Guerberof-Arenas, A., & Toral, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces, 11*(2), 184-212. http://dx.doi.org/10.1075/ts.21025.gue

He, L., Ghassemiazghandi, M., & Subramaniam, I. (2024). Comparative assessment of Bing Translator and Youdao Machine Translation Systems in English-to-Chinese literary text translation. *In Forum for Linguistic Studies (Transferred)* (Vol. 6, No. 2, pp. 1189-1189). http://dx.doi.org/10.59400/fls.v6i2.1189

Hu, K., & Li, X. (2023). The creativity and limitations of AI neural machine translation: A corpus-based study of DeepL's English-to-Chinese translation of Shakespeare's plays. *Babel, 69(4)*, 546-563. http://dx.doi.org/10.1075/babel.00331.hu

Hutchins, W. J. (1986). *Machine translation: past, present, future* (p. 66). Chichester: Ellis Horwood.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174

Ji, B., Duan, X., Zhang, Y., Wu, K., & Zhang, M. (2024). Zero-shot prompting for llm-based machine translation using in-domain target sentences. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. http://dx.doi.org/10.1109/TASLP.2024.3519814

Jinfang, Y., Kasuma, S. A., & Moindjie, M. A. (2025). Translator's Style in Fiction Translation: A Ten-Year Systematic Literature Review. *Journal of Language Teaching and Research, 16(1)*, 125-133. http://dx.doi.org/10.17507/jltr.1601.14

Koehn, P. *Neural Machine Translation*. Cambridge University Press: Cambridge, UK, 2020.

Kostikova, I., Shevchenko, A., Holubnycha, L., Popova, N., & Budianska, V. (2019). Use of machine translation technology for understanding scientific and technical texts. *Journal of Theoretical and Applied Information Technology, 97*(4), 1350-1361.

Kuzman, T., Vintar, Š., & Arcan, M. (2019, August). Neural machine translation of literary texts from English to Slovene. In *Proceedings of the qualities of literary machine translation* (pp. 1-9).

Liu, J. (2020). Comparing and analyzing cohesive devices of SMT and NMT from Chinese to English: a diachronic approach. *Open Journal of Modern Linguistics, 10*(06),765. http://dx.doi.org/10.4236/ojml.2020.106046

Liu, M., Zhang, H., & Wu, G. (2021). Fine grained human evaluation for English-to-Chinese machine translation: A case study on scientific text. *arXiv preprint arXiv:2110.14766.*

Lommel, A. (2013). Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*.

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

Long, X., Chen, K., Bamigbade, O. O., & Swenson, D. L. (2023, September). In-Depth Analysis of Machine Translation and Human Translation of Literary Book Chinese Traditional Culture and a Community with a Shared Future for Mankind. In *3rd International Conference on Internet, Education and Information Technology (IEIT 2023)* (pp. 1163-1170). Atlantis Press. http://dx.doi.org/10.2991/978-94-6463-230-9_139

Lu, Y. (2023, July). An Analysis of Error Types in Chinese to English Translation by Google Neural Machine Translation. *In Proceedings of the 2023 International Joint Conference on Robotics and Artificial Intelligence* (pp. 148-154).

Lyu, C., Du, Z., Xu, J., Duan, Y., Wu, M., Lynn, T., ... & Wang, L. (2023). A paradigm shift: The future of machine translation lies with large language models. *arXiv preprint arXiv:2305.01181*.

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., & Halgamuge, M. N. (2025). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence*. http://dx.doi.org/10.1109/TAI.2025.3569516

Maxmudjanovna, Y. N., & Xamidjanovna, A. N. (2021). Technical translation as a type of specialized translation. *Central Asian Journal of Literature, Philosophy and Culture*.

Mohsen, M. (2024). Artificial intelligence in academic translation: A comparative study of large language models and google translate. *PSYCHOLINGUISTICS, 35*(2), 134-156. http://dx.doi.org/10.31470/2309-1797-2024-35-2-134-156

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

Ng, Y. L. E. (2009). *A Systemic Approach to Translating Style: A Comparative Study of Four Chinese Translations of Hemingway's The Old Man and the Sea.* (Doctoral dissertation, University College London). UCL Discovery.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). 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).

Peng, Z., & Yvon, F. (2023). *Document-level Machine Translation for Scientific Texts* (Doctoral dissertation, ISIR, Université Pierre et Marie Curie UMR CNRS 7222).

Ploeger, E., Lai, H., Van Noord, R., & Toral, A. (2024). Towards Tailored Recovery of Lexical Diversity in Literary Machine Translation. *arXiv preprint arXiv:2408.17308.*

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. *arXiv preprint arXiv:2009.09025*. http://dx.doi.org/10.18653/v1/2020.emnlp-main.213

Shahnazaryan, L., & Beloucif, M. (2024). Defining Boundaries: The Impact of Domain Specification on Cross-Language and Cross-Domain Transfer in Machine Translation. *arXiv preprint arXiv:2408.11926.*

Siu, S. C. (2023). ChatGPT and GPT-4 for Professional Translators: Exploring the Potential of Large Language Models in Translation. *Available at SSRN 4448091.*

http://dx.doi.org/10.2139/ssrn.4448091

Stewart, C., Rei, R., Farinha, C., & Lavie, A. (2020, October). COMET-Deploying a New State-of-the-art MT Evaluation Metric in Production. *In AMTA (2)* (pp. 78-109).

Tahseen, W., & Hussein, S. H. (2024). Investigating Machine translation errors in rendering English literary texts into Arabic. *Integrated Journal for Research in Arts and Humanities, 4*(1), 68-81. http://dx.doi.org/10.55544/ijrah.4.1.11

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1, 5-21. http://dx.doi.org/10.1016/j.aiopen.2020.11.001

Toral, Antonio, Andreas Van Cranenburgh, and Tia Nutters. “Literary-adapted machine translation in a well-resourced language pair: Explorations with More Data and Wider Contexts.” *Computer-Assisted Literary Translation*. Routledge, 2023. 27-52. http://dx.doi.org/10.4324/9781003357391-3

Ulitkin, I., Filippova, I., Ivanova, N., & Poroykov, A. (2021). Automatic evaluation of the quality of machine translation of a scientific text: the results of a five-year-long experiment. In *E3S Web of Conferences* (Vol. 284, p. 08001). EDP Sciences. http://dx.doi.org/10.1051/e3sconf/202128408001

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, X., & Wang, T. (2019). A comparative study of human translation and machine translation post-editing in EC Translation: Translation speed, quality and translators’ attitude. *Foreign Languages and Cultures, 3*(4), 83-93.

Way, A., Youdale, R., & Rothwell, A. (2023). Why more literary translators should embrace translation technology. *Revista Tradumática, 21*, 87-102. https://doi.org/10.5565/rev/tradumatica.344

Weaver, W. (1952). Translation. In *Proceedings of the Conference on Mechanical Translation*.

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., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J.R., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G.S., Hughes, M., & Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv, abs/1609.08144*.

Xie, Y. (2008). Hemingway’s Language Style and Writing Techniques in “The Old Man and the Sea”. *English language teaching, 1*(2), 156-158. http://dx.doi.org/10.5539/elt.v1n2p156

Ying, C., Shuyu, Y., Jing, L., Lin, D., & Qi, Q. (2021). Errors of machine translation of terminology in the patent text from English into Chinese. *ASP Transactions on Computers, 1*(1), 12-17.

Zhang, B., Haddow, B., & Birch, A. (2023, July). Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning* (pp. 41092-41110). PMLR.

Zhao, Y, Zhang, H &Yang, Y. (2024). Comparative Study on the Translation Quality of Large Language Models—Taking the Translation of “Fan Hua” as an Example. *Technology Enhanced Foreign Language Education, 4*(109), 60-66.

ABOUT THE AUTHORS

Zhang Zhongming is a PhD candidate in Translation Studies at UPM, Malaysia, and a university lecturer in China, where he teaches courses in translation and language education. His research interests focus on translation assessment, particularly machine translation, and its applications in education. He also explores innovative teaching methods to enhance students' practical and academic skills. With a strong interdisciplinary background, his work aims to improve translation quality and modernize teaching approaches, bridging theory and practice in both fields.

Syed Nurulakla Syed Abdullah is an Assistant Professor at Universiti Putra Malaysia (UPM) and a Senior Lecturer at the Department of Foreign Languages, Faculty of Modern Languages and Communication, Universiti Putra Malaysia. He is renowned for translating the world masterpiece Rihlah Ibn Battutah into Malay as Pengembaraan Ibn Battutah: Pengembara Agung, Karya Terulung, Menyingkap Wajah Dunia, launched by Sultan Selangor in 2004. He earned a Ph.D. from the University of Malaya in 2015, specializing in translation. Widely recognized as an instructor by language institutions, he translated Roger T. Bell's Translation and Translating into Malay in 2012 and recently translated Iktibar daripada Kehidupanku (2021). Actively involved in national translation activities, he contributes to organizations like Bank Negara Malaysia, RMK-12, and PR agencies. His extensive publications encompass journal articles, book chapters, books, and conference proceedings. At UPM, he strengthens translation initiatives through teaching, research, supervision, and publication while guiding international Ph.D. students from the Middle East and China.

Muhammad Alif Redzuan Abdullah is currently a Senior Lecturer in the Faculty of Modern Languages and Communication at Universiti Putra Malaysia (UPM). He has published articles in research journals in the area of his studies. His research interests include Translation, Interpretation, and Comparative Applied Linguistics. Furthermore, he is actively shaping the next generation as he supervises a cadre of Chinese Ph.D. students in Translation and Interpreting at UPM, showcasing his commitment to cross-cultural communication and academic mentorship.

Duan Wenqi is a PhD candidate in Translation Studies at Universiti Putra Malaysia (UPM) and a university teacher in China, where she teaches English Interpretation, College English, and Academic Writing. Her research interests include translation and culture, classical literature, and applied linguistics. She also has extensive experience in education and teaching. Her work aims to explore the intersection of translation and cultural studies, particularly in the context of classical literature, while enhancing translation strategies that bridge linguistic and cultural gaps and improving pedagogical approaches in language and translation education.