International Journal of English Language Education
ISSN 2325-0887
2025, Vol. 13, No. 2

Translation Quality in an Evolving Paradigm: Neural Machine Translation and Large Language Models in Technical Domains

Zhongming Zhang
Universiti Putra Malaysia, Malaysia

Syed Nurulakla Syed Abdullah
Universiti Putra Malaysia, Malaysia

Muhammad Alif Redzuan Abdullah
Universiti Putra Malaysia, Malaysia

Wenqi Duan
Universiti Putra Malaysia, Malaysia

Received: July 10, 2025 Accepted: July 29, 2025 Published: July 31, 2025

doi:10.5296/ijele.v13i2.23057 URL: https://doi.org/10.5296/ijele.v13i2.23057

Abstract

The emergence of large language models (LLMs) has reshaped machine translation (MT). Although neural machine translation (NMT) systems like Google Translate (GT) remain dominant, systematic comparisons between LLMs and NMT systems across key quality dimensions are still limited, especially in specialised domains such as technical translation. This study aims to compare the translation quality and error subtypes of GT and ChatGPT-4 in Chinese-English technical manual translation. Eighty paragraph-level segments from Chinese product manuals were translated by both systems. Two trained annotators evaluated the outputs using a Likert scale across four MQM-based dimensions: accuracy, fluency, terminology, and style. Inter-rater agreement was tested and qualitative data analysis was conducted using NVivo. Results indicated that ChatGPT-4 outperformed GT across all dimensions, delivering higher quality translation, whereas GT frequently exhibited errors such as redundancy, stilted phrasing,

non-standard terminology, and formality mismatches. ChatGPT-4, however, occasionally produced over-translation and semantic overgeneralisation, compromising terminological precision. Despite the superior performance of ChatGPT-4, it still poses certain potential risks. Its context-driven outputs may introduce inferential or stylistic deviations, especially in specialised terminology. For high-stakes technical content, expert revision is recommended to ensure semantic fidelity and terminological consistency.

Keywords: ChatGPT-4, Google Translate (GT), human assessment, technical translation

1. Introduction

Driven by significant advancements in deep neural network architectures (Schmidhuber, 2015), Neural Machine Translation (NMT) has, over the past decade, become the prevailing paradigm in the field. In contrast to earlier approaches, NMT systems have delivered substantial improvement in both semantic adequacy and fluency, fundamentally reshaping the landscape of machine translation (MT) research and practice (Stahlberg, 2020). However, the rise of Large Language Models (LLMs), exemplified by ChatGPT, is heralding a new phase in the evolution of translation studies and its practical applications.

Although LLMs share the foundational Transformer architecture with NMT systems (Vaswani et al., 2017), they diverge significantly in terms of training goals, data dependencies, and operational methodologies. While NMT is built upon supervised learning from bilingual parallel corpora, LLMs are generally pre-trained on vast monolingual datasets and approach translation primarily through prompt-based generation (Wu & Hu, 2023). This shift signifies not only a structural divergence but also a redefinition of how translation is modelled and controlled in practice.

While NMT systems, such as Google Translate (GT), continue to dominate practical and academic translation applications, recent research suggests that LLMs may assume a central role in the future development of machine translation (MT) technologies (Lyu et al., 2023). Nonetheless, comparative performance between NMT and LLM systems varies substantially depending on the language pair and the nature of the source text. For example, Son and Kim (2023) found that NMT systems continue to outperform ChatGPT models in English-to-non-English translations across general-purpose genres such as news and reports. Conversely, ChatGPT-4 has demonstrated superior performance in select language pairs, such as German-English, indicating that the benefits of LLMs may be context-specific.

Recent findings further reveal that LLMs such as ChatGPT-4 can exceed traditional NMT systems in scientific and technical translation tasks, particularly when enhanced with domain-specific glossaries. However, their consistency remains an issue in low-resource language pairs, such as English-Slovak, where the variability of output poses challenges for terminological precision and coherence (Barák, 2024).

This evolving landscape underscores the need for a nuanced understanding of how different MT paradigms perform under domain-specific constraints. Technical translation, especially in

the context of user manuals and product documentation, requires not only linguistic accuracy but also high degrees of terminological consistency and stylistic appropriateness, which have been extensively problematised in translation studies (Radetska, 2024). Unlike general language texts, technical texts are characterised by high lexical density, a blend of specialised terminology, and a user-centric communicative purpose (Olohan, 2020). Therefore, evaluating MT performance in this domain requires more than surface-level metrics, incorporating qualitative analysis and structured error typologies to capture each system's strengths and weaknesses.

The translation of technical documentation, such as user manuals and product guides, plays a pivotal role in safeguarding user safety, facilitating functional comprehension, and supporting cross-market accessibility. The tekom Europe website (tekom Europe, 2018) offers a concise yet comprehensive definition of technical communication: It is the process of defining, creating, and delivering information products for use, with the aim of enabling the safe, efficient, effective, and sustainable operation of goods, technical systems, software, and services. Siikala (2018) further emphasises that technical texts must ensure product safety while assisting users in accurate understanding and proper use. In this context, poor-quality translations may result in user errors, safety hazards, and diminished consumer confidence.

Given the safety-critical and functional significance of technical documentation, it is crucial to assess the performance of current machine translation (MT) systems within this domain. Despite the rapid integration of large language models (LLMs) into translation workflows, their effectiveness in rendering technical content—particularly between Chinese and English—remains insufficiently investigated. This study seeks to address this gap through a qualitative comparison of ChatGPT-4 and GT in translating Chinese-English technical manuals. Drawing on the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014), the analysis focuses on four core quality dimensions (accuracy, fluency, terminology, and style) and adopts a data-driven approach to identify and categorise error subtypes.

Accordingly, this study aims to explore the extent to which GT and ChatGPT-4 differ in translation quality and typical error patterns when applied to Chinese-English technical manuals. The guiding research questions are as follows:

(a) How does the translation quality of GT and ChatGPT-4 differ in Chinese-English technical manuals?

(b) What error subtypes are typically produced by GT and ChatGPT-4 in technical manual translation?

2. Literature Review

2.1 NMT and LLMs: A Paradigm Shift in Machine Translation

Since the widespread adoption in 2014, NMT has marked a significant shift in the field of machine translation, transitioning from phrase-based statistical approaches to end-to-end modelling powered by deep learning. With the integration of the Transformer architecture into

NMT systems, leading platforms such as Google have gradually updated their underlying frameworks to enhance performance. Empirical research demonstrates that, compared with recurrent neural network (RNN) models, the Transformer architecture significantly improves translation quality and computational efficiency across multilingual tasks (Lakew et al., 2018). Its ability to capture sentence-level semantics makes it particularly effective in processing structured texts such as technical manuals. Nevertheless, when applied to longer documents, the Transformer still exhibits several limitations, including restricted input length, suboptimal computational efficiency, weakened handling of long-range dependencies, and insufficient support for complex hierarchical structures (Dong et al., 2023).

Concurrently, the emergence of LLMs, such as ChatGPT, has prompted a fundamental shift in how translation tasks are conceptualised and executed. In contrast to domain-specific systems, ChatGPT-4 leverages extensive pre-training on large-scale textual corpora and utilises an autoregressive generation mechanism. This allows it to approximate human-level translation quality across multi-domain evaluation datasets, including biomedical and technological texts, and high-resource language pairs such as Chinese-English (Yan et al., 2024). As highlighted by Lyu et al. (2023), LLMs demonstrate notable strengths in semantic generalisation—adapting to unseen tasks and stylistic variations—and in maintaining contextual coherence across long documents, multimodal content, and interactive exchanges. These capabilities grant LLMs a substantial advantage in MT contexts, positioning them not only as complements to traditional NMT systems but also as potential successors.

However, the increased generative flexibility of LLMs has also introduced novel challenges, most notably the phenomenon of "hallucination"—the production of outputs that deviate from factual content or misrepresent the input. Such issues include acronym ambiguity and numeric inaccuracy, collectively referred to as "numeric nuisance" (Rawte et al., 2023). These problems are particularly problematic in highly standardised technical documentation, where precision and factual reliability are critical to usability and safety. In response, recent studies have explored mitigation strategies such as prompt engineering and task-specific fine-tuning, aiming to enhance the stability and terminological consistency of LLM outputs in specialised technical domains (Wu et al., 2024).

Against this backdrop, the contrast between different translation systems in terms of terminology control and contextual modelling has increasingly become a focal point of research. Some studies suggest that traditional NMT systems, such as GT, may outperform current LLMs in terminological precision and standardisation (Zayed, 2024). Similarly, in terms of grammatical consistency and structural alignment, as observed in Indonesian-English technical texts, NMT systems like DeepL have demonstrated a slight advantage (Karim, 2024). In contrast, LLMs tend to excel in maintaining contextual coherence and generating more naturally flowing language. Each system thus possesses distinct strengths. As such, comparative evaluations of GT and ChatGPT in technical translation scenarios not only carry significant practical relevance, but also reflect a broader shift in the field.

2.2 Translation Quality Assessment

As a central concern within translation studies, translation quality assessment has continually

evolved alongside advances in translation technologies. Traditionally, quality evaluation has been conceptualised across multiple dimensions, with accuracy and fluency widely recognised as the two core criteria (Koby et al., 2014; Chatzikoumi, 2020). However, the challenges posed by technical texts extend beyond purely linguistic factors. Terminology management, in particular, presents a persistent and critical obstacle to translation quality in this domain (Joshi, 2017). Moreover, in highly regulated genres stylistic consistency and register alignment are equally essential. Effective translations in this context must not only preserve the functional intent of the source text but also achieve naturalness and usability in the target language.

A common approach to human evaluation entails error categorisation, often supported by detailed analysis (Chatzikoumi, 2020), with the MQM framework (Lommel et al., 2014) emerging as a leading model in modern MT assessment. By employing a hierarchical taxonomy of error categories, MQM facilitates the consistent classification and quantification of a wide spectrum of translation problems. It is applicable across human, neural machine, and AI-generated translation outputs. The typology encompasses seven core error dimensions and can be tailored to accommodate specific assessment needs. These foundational categories serve as the theoretical underpinning for the present study.

In recent years, automatic evaluation metrics, including BLEU and more advanced pretrained models such as COMET and BERTScore, have been widely adopted for assessing machine translation. Nevertheless, human evaluation is still regarded as the gold standard for determining translation quality (Lommel et al., 2024). While automatic metrics offer advantages in terms of efficiency and cross-system comparability, they are often limited by their dependence on reference translations and struggle to capture more complex dimensions of translation quality, such as semantic equivalence and discourse coherence. Despite ongoing efforts to overcome these limitations through embedding-based approaches, challenges remain, particularly with respect to reliably judging semantic content (Freitag et al., 2021). These limitations become especially pronounced when evaluating outputs from Large Language Models (LLMs) such as ChatGPT, whose translations often diverge from reference texts in form while maintaining semantic fidelity, thus leading to potential misjudgements under reference-based automatic evaluation.

Consequently, human assessment is regarded as the most appropriate benchmark, as it can compensate for the limitations of automatic evaluation and is more sensitive to complex translation errors (Chatzikoumi, 2020). In comparative studies of GT and ChatGPT-4, it is essential to clarify the strengths and limitations of the chosen evaluation methods. Doing so not only provides a robust methodological foundation for empirical analysis but also facilitates the systematic identification of performance differentials across key quality dimensions.

2.3 NMT and LLMs: Performance and Error Analysis

Methodologically, some studies employ the MQM framework or its adapted variants for human annotation and scoring. For instance, Alzain et al. (2024) evaluated English-Arabic scientific texts using a Likert scale in conjunction with MQM dimensions, systematically annotating errors in terminology, fluency, and style. Their findings indicated that ChatGPT produced substantially more terminological errors than GT (119 vs. 70), and that both systems exhibited

notable issues in grammar and cohesion, reflecting the morphological complexity of Arabic as a significant challenge for MT systems. Similarly, Sanz-Valdivieso and López-Arroyo (2023) focused specifically on the terminological dimension, confirming ChatGPT's superiority over GT in this regard, though both systems demonstrated considerable limitations in handling specialised terminology.

Under the TAUS DQF framework, Barák (2024) conducted an evaluation of English-Slovak scientific text translations, comparing the performance of GT, DeepL, and ChatGPT-3.5/4.0. The findings revealed that ChatGPT-4.0 demonstrated the highest overall performance in terms of error rate and terminological consistency, whereas GT exhibited more issues in the dimensions of accuracy and style, with a total of 37 recorded errors. The study also emphasised that all systems continued to rely on post-editing by human experts, indicating that they remain insufficient as standalone replacements for professional translators.

Other studies have drawn upon the Adequacy-Fluency framework to establish evaluation dimensions. Brewster et al. (2024), for example, used professional human translations as a benchmark to assess the performance of GT and ChatGPT across Spanish, Portuguese, and Haitian Creole. Their results indicated that GT was more favourably evaluated in high-resource language contexts, while human translations remained clearly superior in low-resource settings. Similarly, Briva-Iglesias et al. (2024) combined the TAUS DQF framework with standardised scoring guidelines to assess GT and ChatGPT-4 in legal translation scenarios. They found that ChatGPT-4 performed better in maintaining terminological consistency and contextual coherence, whereas GT demonstrated comparatively greater stability in lower-resource languages such as Turkish.

Beyond general scientific texts, domain-specific investigations have also been conducted in fields such as medicine and law. Al-Maaytah and Almahasees (2024), for instance, evaluated medical terminology translation through manual annotation along orthographic, semantic, and grammatical dimensions, using reference translations for comparison. Their findings suggest that GT tends to favour literal rendering and exhibits weak cultural adaptability, whereas ChatGPT demonstrates stronger performance in interpreting terms and handling sensitive content, albeit with occasional grammatical instability. Similarly, Sadiq (2025) conducted a five-dimension blind evaluation of English-Arabic translations across multiple genres, including scientific and technical texts. The results ranked human translations highest in overall quality, followed by ChatGPT, while GT performed the worst, frequently exhibiting errors such as terminological mistranslations and awkward sentence structures.

While existing studies have achieved considerable granularity in multilingual translation quality assessment, systematic human-based comparisons between Chinese and English, particularly within the technical register, remain sparse. In particular, the classification and distribution patterns of error types have not been sufficiently explored, signalling the need for further empirical investigation in future research.

3. Methodology

3.1 Research Design

In recognition of the knowledge that human evaluation remains the gold standard in machine translation assessment (Lommel et al., 2024), this study adopted a qualitative, human-centred evaluation framework to compare the translation performance of GT and ChatGPT-4 in the context of Chinese-to-English technical manuals. The source material comprised paragraph-level segments extracted from product-oriented instructional texts, ensuring the inclusion of domain-specific terminologies and directive language structures that are typical of technical documentation.

Drawing on the Multidimensional Quality Metrics (MQM) framework, four core dimensions were used as top-level error categories. Within this structure, a theory-informed inductive thematic analysis was conducted to identify and categorise specific error subtypes emerging from human annotations. This approach allowed the study to combine the rigour of a predefined evaluative model with the flexibility to accommodate data-driven insights, ensuring both systematic comparability and contextual sensitivity in assessing translation quality.

3.2 Data Source

This study constructed a small-scale bilingual corpus by extracting paragraph-level segments (ranging from 80 to 120 words) from publicly available Chinese-English product manuals. The selection of texts adhered to four key criteria to ensure both methodological rigour and practical relevance. Firstly, the source materials were drawn from domains such as consumer electronics and automotive technologies, reflecting end-user scenarios with real-world applicability. Secondly, only materials featuring professionally produced English-Chinese parallel translations were included, serving as reference points for human evaluation. Thirdly, priority was given to open-access documentation in order to uphold research transparency, reproducibility, and ethical integrity. Lastly, the selected manuals were characterised by consumer-facing language, enhancing the societal relevance and practical utility of the study's findings.

The manuals exhibited hallmark features of technical communication, such as procedural clarity, terminological standardisation, and domain-specific linguistic conventions, which are widely recognised as persistent challenges for machine translation systems (Bowker & Ciro, 2019). These characteristics make such texts particularly suitable for evaluating the comparative performance of MT systems in high-stakes, domain-sensitive contexts.

3.3 Translation Evaluation Procedure

To ensure consistency in translation outputs, translations from GT were directly obtained via its official web interface, thereby ensuring standardised and reproducible system behaviour. In the case of ChatGPT-4, each source segment was translated using a deliberately minimal prompt: "Please translate the following text into English", to reduce prompt-related variability and ensure comparability across test items. All translation tasks were executed in May 2025, eliminating potential confounding effects from system updates and ensuring temporal control.

For the human assessment phase, this study adopted the four core dimensions from the MQM framework: accuracy, fluency, terminology, and style. Two trained annotators independently rated all translated outputs using a five-point Likert scale (1 = very poor, 5 = excellent) and provided qualitative error annotations based on MQM core dimensions. Mean scores for each system and dimension were calculated by averaging the individual ratings given by both annotators for each translation segment, followed by determining the overall mean for each of the four evaluation dimensions.

Inter-rater agreement was subsequently tested using the Intraclass Correlation Coefficient (ICC) (Shrout & Fleiss, 1979) to validate scoring reliability. This dual-layered procedure, combining scalar judgments with detailed error tagging, provided a robust basis for the subsequent comparative analysis.

All error types were given equal weight in the categorical, frequency-based analysis, with each instance counted equally irrespective of its specific category. However, the application of a five-point Likert scale in the quantitative assessment enabled a distinction to be made regarding the severity of errors: more serious errors were assigned lower scores, while less severe instances received higher ratings. Consequently, both the frequency of each error type and its average Likert score were considered in the interpretation of the comparative performance of the two systems.

Prior to the formal evaluation, both annotators underwent a structured training and calibration process. Several segments from technical manuals, excluded from the main evaluation, were selected as training material. Each annotator independently scored these samples, after which a focused discussion was held to review the rationale behind every judgement. Particular attention was paid to cases where discrepancies arose, and each assessment dimension was clarified with precise, operational definitions agreed upon by both annotators. Notably, both annotators possess more than five years of professional translation experience and hold a master's degree in translation, ensuring a robust foundation for expert evaluation.

During the official assessment phase, stringent measures were implemented to ensure both the rigour and consistency of the evaluation, especially given that GT and ChatGPT-4 represent advanced paradigms in machine translation. In situations where it was difficult to choose between two adjacent scores on the five-point Likert scale (e.g., 3 or 4), annotators were instructed to assign the lower score to maintain a conservative standard. In instances where 'accuracy errors' and 'terminology errors' overlapped, we explicitly stipulated that errors involving specialist vocabulary or terminology should be uniformly classified as 'terminology errors', in line with the distinctive characteristics of technical manual translation. Furthermore, a double-blind review was enforced: all source information was concealed, and annotators were only informed that the translations originated from different machine translation systems. These concrete measures collectively enhanced the consistency of the evaluation standards and contributed to the objectivity of the results.

3.4 Qualitative Data Analysis

Beyond Likert scalar ratings, this study conducted a qualitative exploration of translation errors

based on evaluators' annotations. To establish a robust professional baseline for comparison, this study utilised high-quality human translations, specifically the official bilingual parallel texts issued by recognised authorities, as reference material. Incorporating these authoritative translations enabled a more nuanced and comprehensive assessment of machine translation performance, as the outputs could be directly benchmarked against professional standards. This approach afforded a clearer contextualisation of the respective strengths and limitations of each system, thereby enhancing the rigour and validity of the comparative analysis.

All error comments generated during assessment were collated and imported into NVivo (version 15) for systematic coding according to MQM-defined core error categories. Through a process of iterative refinement, entailing repeated cycles of reviewing, adjusting, and reclassifying error annotations, this analysis enabled the identification of more precise error types and recurring patterns. This methodological approach enhanced the reliability of the coding process and provided deeper insight into the specific strengths and limitations of Google Translate and ChatGPT-4 in handling domain-specific translation challenges.

To maintain analytical consistency, annotations indicating only marginal concerns (e.g., signalled by terms like somewhat or slightly) were excluded from the coding procedure, as they were judged to reflect stylistic preferences rather than substantive errors. Similarly, comments affirming overall acceptable while noting only minor reservations were not treated as error instances. Annotations misaligned with their assigned categories were also excluded to prevent overlap; for instance, comments on conciseness mislabelled as accuracy were omitted.

The final analysis focused on identifying recurrent error subtypes under the four core MQM dimensions. This comparative approach allowed for the systematic examination of each system's translation tendencies, revealing the characteristic behaviours and potential weaknesses of GT and ChatGPT-4 in the technical translation context.

4. Results

4.1 Overview of Evaluation Results

This section provided an overview of evaluation results, including inter-rater agreement and average Likert scores across four dimensions. To evaluate the consistency between the two annotators, this study employed the Intraclass Correlation Coefficient (ICC), a widely acknowledged statistical measure used to assess inter-rater agreement in quantitative evaluations (see Table 1). The ICC is noted for its methodological robustness and broad applicability in contexts involving subjective judgement (Koo & Li, 2016). According to the interpretative thresholds proposed by Cicchetti (1994), values between .40 and .59 suggest moderate reliability, while scores exceeding .60 are generally regarded as substantial to excellent agreement.

Table 1. ICC reliability test

MT Systems	Evaluation Dimension	ICC(Average Measures)	Sig.
GT	Adequacy	0.766	< .001
	Fluency	0.615	< .001
	Terminology	0.558	< .001
	Style	0.625	< .001
ChatGPT-4	Adequacy	0.753	< .001
	Fluency	0.691	< .001
	Terminology	0.446	.002
	Style	0.431	.002

As illustrated in Table 1, ICC values for GT remained consistently above the .50 threshold across all four MQM dimensions, ranging from .558 for terminology to .766 for adequacy, denoting moderate to high levels of inter-rater agreement. ChatGPT-4 similarly achieved substantial agreement in adequacy (.753) and fluency (.691), yet demonstrated lower consistency in the dimensions of terminology (.446) and style (.431), both falling within the moderate reliability band. These discrepancies may be attributable to greater variation in ChatGPT-4's lexical choices and stylistic renderings in technical contexts, which may lead to more subjective divergence in rater interpretation.

Table 2. Mean Likert ratings for GT and ChatGPT-4

Dimension	GT (Mean)	ChatGPT-4 (Mean)
Accuracy	3.57	4.58
Fluency	3.26	4.71
Terminology	3.98	4.53
Style	3.31	4.25

As presented in Table 2, ChatGPT-4 demonstrated superior performance to GT across all four dimensions evaluated using the Likert scale. In terms of fluency and accuracy, ChatGPT-4 attained substantially higher mean scores (4.71 and 4.58, respectively) compared to GT (3.26 and 3.57), suggesting a higher degree of textual coherence and fidelity to the source content. Likewise, for terminology and style, ChatGPT-4 outperformed GT, scoring 4.53 and 4.25 versus 3.98 and 3.31, respectively, which reflects more precise lexical choices and improved

alignment with domain-specific stylistic conventions. These findings provided a direct response to research question 1, affirming that ChatGPT-4 consistently produces higher-quality translations than GT in the context of Chinese-English technical manuals.

4.2 Comparative Analysis of Error Subtypes in GT and ChatGPT-4

To further investigate differences in translation quality, all annotated errors were classified under four core dimensions, each comprising multiple subtypes derived from error annotations. To ensure clarity and consistency in the definition of each error type, a comprehensive codebook was developed (see Appendix A). Table 3 summarised the distribution and relative proportions of error types for both translation systems, providing a clear basis for comparison and further analysis.

Table 3. Distribution of Translation Error Types for GT and ChatGPT-4

Main Category	Subtype	GT Error	GT Rate (%)	ChatGPT-4 Error	ChatGPT-4 Rate (%)
Accuracy	Mistranslation	11	4.42%	1	2.17%
	Informational Redundancy	13	5.22%	0	0.00%
	Omission	10	4.02%	9	19.57%
	Structural Misalignment	11	4.42%	0	0.00%
	Ambiguity	10	4.02%	4	8.70%
	Overgeneralisation	0	0.00%	1	2.17%
	Over-translation	0	0.00%	4	8.70%
Fluency	Stilted Expression	17	6.83%	0	0.00%
	Disjointed Cohesion	11	4.42%	1	2.17%
	Uneven Rhythm	7	2.81%	3	6.52%
	Grammatical Error	13	5.22%	1	2.17%
	Inappropriate Word Order	5	2.01%	0	0.00%
	Mechanical Repetition	19	7.63%	3	6.52%
Terminology	Non-standard Terminology	12	4.82%	2	4.35%
	Terminological Inconsistency	7	2.81%	2	4.35%
	Non-technical Wording	8	3.21%	3	6.52%
	Abbreviation Misuse	2	0.80%	0	0.00%
	Terminology Formatting Error	2	0.80%	0	0.00%
	Terminological Vagueness	3	1.20%	3	6.52%
	Incorrect Term Selection	5	2.01%	2	4.35%
Style	Formality Mismatch	36	14.46%	1	2.17%
	Insufficient Technical Register	9	3.61%	2	4.35%
	Directive Intensity Deviation	1	0.40%	1	2.17%
	Translationese Style	18	7.23%	0	0.00%
	Verbal Redundancy	12	4.82%	2	4.35%
	Stylistic Inconsistency	7	2.81%	1	2.17%

According to Table 3, a total of 249 errors were identified in the Google Translate output, whereas 46 errors were observed in that of ChatGPT-4. These errors are distributed across the four principal dimensions and their respective subtypes. For Google Translate, the most prevalent error types were formality mismatch, mechanical repetition, and translationese style. In contrast, the errors found in ChatGPT-4's output were fewer in number and more evenly spread across the various categories.

However, the distribution and nature of these errors varied notably across dimensions and between systems. In order to further address the second research question, the following sections offered a detailed breakdown of error subtypes, providing a comparative perspective on the relative strengths and weaknesses of each system.

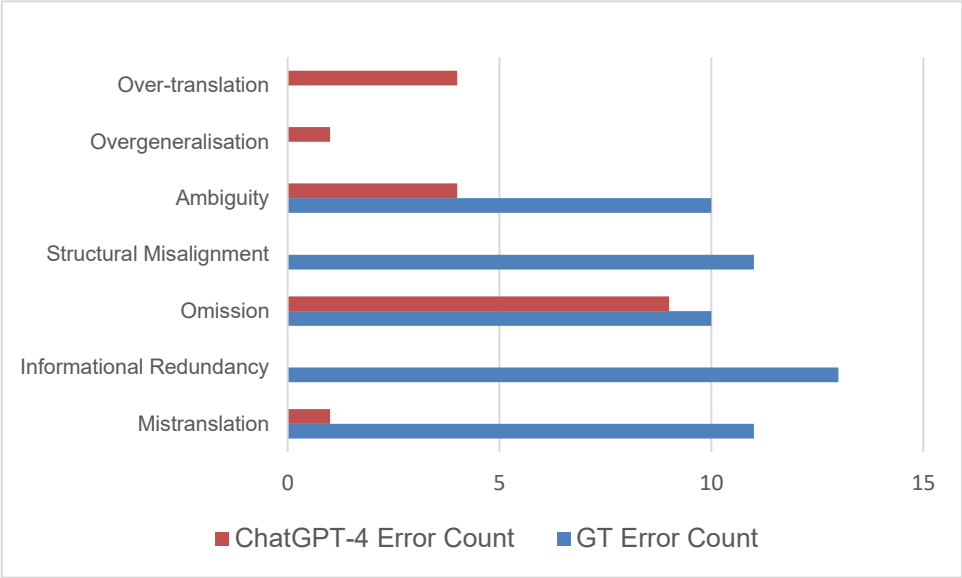

Figure 1. Accuracy error comparison: GT vs. ChatGPT-4

Figure 1 presented a comparative analysis of accuracy-related errors across seven subtypes in translations produced by GT and ChatGPT-4. Overall, GT generated a markedly higher number of errors, particularly in the categories of Informational Redundancy (13), Mistranslation (11), and Structural Misalignment (11). In contrast, ChatGPT-4, while producing fewer total errors, demonstrated a higher incidence of Omission (9) and exhibited error types not observed in GT output, such as Over-translation (4) and Overgeneralisation (1). These findings suggested that GT was more susceptible to issues involving semantic distortion, unnecessary repetition, and structural inconsistency. Meanwhile, ChatGPT-4, although generally more accurate, occasionally omitted essential source content or introduces unwarranted elaborations, reflecting a different set of accuracy-related challenges.

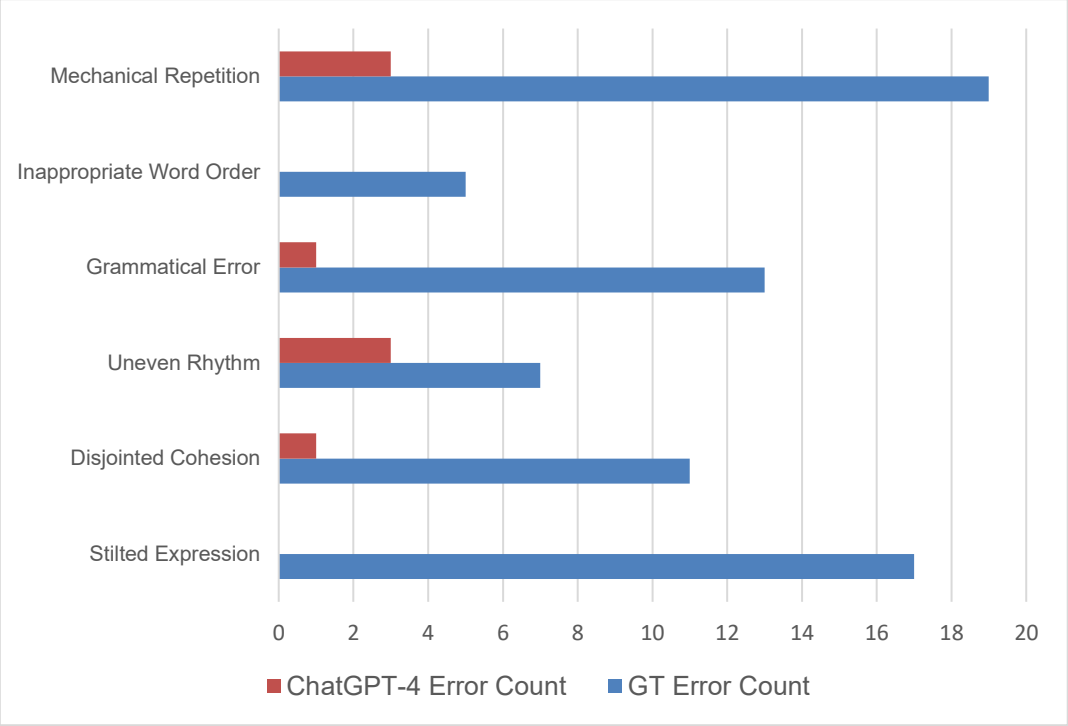

Figure 2. Fluency error comparison: GT vs. ChatGPT-4

Figure 2 compared fluency-related errors in the outputs of GT and ChatGPT-4 across six subtypes. The data clearly showed that GT produced markedly more errors in every category. The most frequent issues included Mechanical Repetition (19 instances), Stilted Expression (17), and Grammatical Error (13), suggesting a persistent tendency towards inflexible phrasing and syntactic awkwardness.

By contrast, ChatGPT-4 demonstrated substantially better fluency. It recorded no instances of Stilted Expression or Inappropriate Word Order, and only minimal occurrences of Uneven Rhythm (3), Mechanical Repetition (3), and Grammatical Error (1). The only dimension in which ChatGPT-4 exhibited a similar pattern was Disjointed Cohesion, although this was limited to a single case. These results suggested that ChatGPT-4 is considerably more capable of producing fluent and idiomatic translations, whereas GT remains susceptible to redundancy, unnatural phrasing, and grammatical inaccuracy.

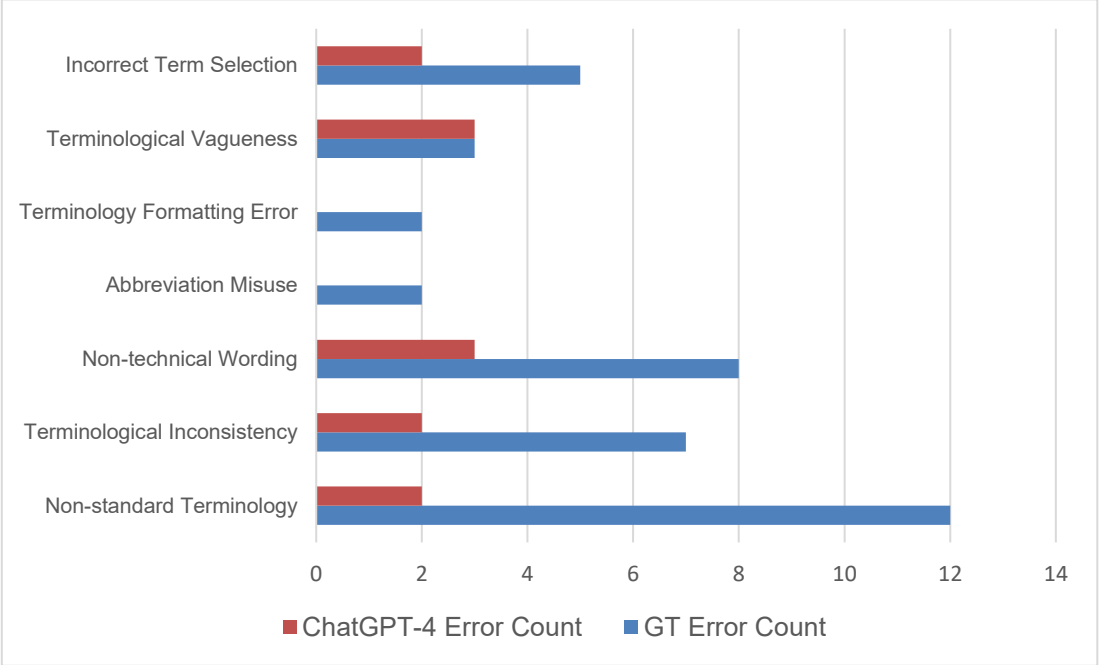

Figure 3. Terminology error comparison: GT vs. ChatGPT-4

Figure 3 compared terminology-related translation errors between GT and ChatGPT-4, distributed across seven subtypes. The data revealed a substantial disparity in overall performance, with GT generating a total of 39 errors, compared to just 13 in the output of ChatGPT-4. Among GT's most frequent issues were Non-standard Terminology (12 instances), Terminological Inconsistency (7), Non-technical Wording (8), and Incorrect Term Selection (5). By contrast, ChatGPT-4 recorded significantly fewer errors in these areas, typically between two and three occurrences per category, and exhibited no instances of Abbreviation Misuse or Terminology Formatting Errors.

Although both systems demonstrated comparable difficulty with Terminological Vagueness (three instances each), GT consistently deviated more from standardised and contextually appropriate terminology. Errors produced by ChatGPT-4, while fewer in number, often involved more nuanced challenges relating to lexical precision and domain register, rather than systematic misuse or inconsistency.

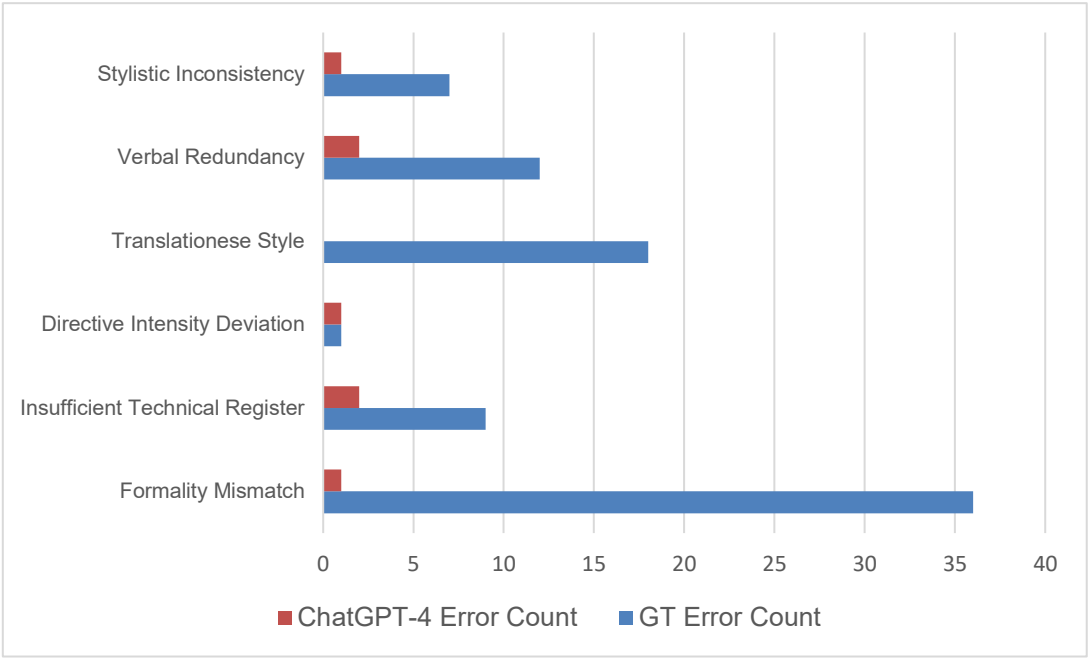

Figure 4. Style error comparison: GT vs. ChatGPT-4

Figure 4 provided a comparative analysis of style-related translation errors for GT and ChatGPT-4, categorised across six subdimensions. The data revealed that GT produced a significantly greater number of stylistic errors overall. The most prevalent issue in GT's output was Formality Mismatch, with 36 occurrences, compared to just one instance observed in ChatGPT-4. Likewise, GT exhibited 18 instances of Translationese Style and 12 of Verbal Redundancy, whereas ChatGPT-4 demonstrated no issues with translationese and only two cases of redundancy.

GT also recorded higher frequencies of errors in Stylistic Inconsistency and Insufficient Technical Register. These findings suggested that ChatGPT-4 maintains a higher degree of stylistic appropriateness and consistency, while GT was more prone to overly literal renditions, misaligned formality, and stylistic irregularities.

4.3 Case Study

This section presented qualitative insights derived from translation outputs, emphasising error types identified during the human assessment process. To facilitate a more intuitive understanding of these patterns, four representative cases were selected from diverse subdomains of technical translation, namely Sony, Apple, Toyota, and ThinkPad. The analysis of these examples revealed four principal categories of MT errors, frequently observed across both systems.

Table 4. Accuracy error from ThinkPad dataset

Example one
Source: （直接关闭计算机可能导致部分数据或进度丢失）但是， 如果不立即关闭，可能导致**计算机**完全损毁。
Reference: …leaving the computer on might make your **computer** unusable.
ChatGPT-4: …however, failure to shut it down immediately could lead to complete **hardware** failure.
GT: …however, if it is not turned off immediately, it may cause the **computer** to be completely damaged.

As illustrated in Table 4, ChatGPT-4's rendering of the "计算机" as "hardware failure" constitutes a clear instance of Over-translation. The source term encompasses the entire computer system, including both physical hardware and data integrity. By translating it solely as "hardware," the output unduly restricts the semantic scope, omitting potential references to data loss or broader system malfunctions that are implicitly conveyed in the original. In contrast, GT retains this broader interpretive range, more accurately reflecting the intended meaning of the source text.

Table 5. Fluency Error from Sony Dataset

Example two
Source: 如果几分钟后仍未显示定位图标,则**可能**是信号接收有问题。
Reference: …if a positioning icon is not displayed after several minutes, **there may be** a problem with signal reception.
ChatGPT-4: …if the location icon does not appear after a few minutes, **there may be** a problem with signal reception.
GT: …if the positioning icon is not displayed after a few minutes, **it may be there is** a problem with signal reception.

As illustrated in Table 5, the sentence generated by GT— "it may be (that) there is a problem with signal reception"—contains a clear Grammatical Error. This construction deviates from standard English syntactic conventions, thereby compromising the sentence's grammatical integrity and logical coherence. In contrast, both the reference translation and ChatGPT-4's output, "there may be a problem...", employ a grammatically standard and idiomatic structure, which aligns more closely with the expectations of accuracy and readability in technical text.

Table 6. Terminology error from the Toyota dataset

Example three
Source: 在下列情况下，应更换轮胎：轮胎上显示外胎磨损标记。轮胎有诸如割伤、断裂、露出**帘布层**的较深裂缝或出现表示轮胎内部有损伤的凸起时。
Reference: Tires should be replaced if: the **treadwear indicators** are showing on a tire. You have tire damage such as cuts, splits, cracks deep enough to expose the **fabric**, and bulges indicating internal damage.
ChatGPT-4: Tires should be replaced under the following conditions: when the **tread wear indicators** appear on the tire. When the tire has deep cuts, cracks, breaks, exposed **cords**, or bulges indicating internal damage.
GT: Tires should be replaced in the following cases: the tire shows signs of **outer wear**. When the tire has cuts, breaks, deep cracks that expose the **ply**, or bulges that indicate damage inside the tire.

As illustrated in Table 6, GT renders the precise technical term “treadwear indicators” as the vague and inaccurate expression “outer wear.” This constitutes a case of Terminological Vagueness, as it obscures the intended meaning by omitting reference to the specific safety feature, raised rubber bars embedded in the tyre tread, that signal the need for replacement. Such terminological imprecision undermines both the clarity and functional utility of the translation. Furthermore, both GT and ChatGPT-4 fail to render the term “fabric” correctly, which refers to the tyre’s internal reinforcement layer. This error falls under Non-standard Terminology. While ChatGPT-4’s use of “cords” more closely approximates the intended concept, it remains technically inaccurate and may still lead to misunderstanding in domain-specific contexts.

Table 7. Style error from Apple dataset:

Example four
Source: 如果 **studio display** 使用 nano-texture 纳米纹理玻璃,请**仅使用**随附的抛光布进行清洁。
Reference: if **your studio display** has nano-texture glass, **clean it only** using the included polishing cloth.
ChatGPT-4: if **your studio display** uses nano-texture glass, **clean it only** with the included polishing cloth.
GT: if **the studio display** uses nano-texture glass, **use only** the included polishing cloth **to clean** it.

As shown in Table 7, the GT output exhibits a Directive Intensity Deviation by rendering the personalised "your studio display" as the neutral "the studio display". This shift weakens the directive tone typical of technical manuals, where direct user engagement is standard practice. While grammatically correct, the impersonal phrasing reduces clarity and user orientation, resulting in a stylistic mismatch that may undermine the communicative intent of the instruction.

Moreover, the phrase "use only... to clean it" introduces unnecessary repetition and lacks the conciseness characteristic of effective directive language, thereby constituting a case of Verbal Redundancy. This rigid adherence to source syntax leads to a stylistic mismatch, ultimately reducing the communicative efficacy of the instruction.

4.4 Summary of Findings

This study reveals a consistent advantage for ChatGPT-4 over GT in the translation of Chinese-English technical manuals. Likert-scale evaluations show that ChatGPT-4 outperforms across all dimensions—fluency, accuracy, terminology, and style. Qualitative analysis further supports this, with GT exhibiting more grammatical and structural errors, often resulting in stylistic mismatches and translationese, particularly due to its reliance on source-language alignment (Ni & Jin, 2022).

By contrast, ChatGPT-4 generates outputs that are more fluent and stylistically attuned to target-language norms, benefitting from its superior contextual modelling. However, this strength occasionally leads to semantic overgeneralisation and inferential embellishment, which may undermine precision and domain-specific terminological fidelity (Allaway et al., 2024). On balance, ChatGPT-4 appears better suited for Chinese-English translation tasks in technical manuals.

5. Discussion

While the overall findings reaffirm ChatGPT-4's superiority over GT across key quality dimensions, a more granular analysis of the error distributions uncovers subtler distinctions. This section investigates the divergent error patterns and their underlying causes in relation to each system's translation architecture. By situating these results within the context of existing scholarship, the discussion seeks to elucidate not only the scope of the observed differences, but also their implications for domain-specific machine translation practices.

GT's errors were primarily concentrated in areas such as structural misalignment, informational redundancy, and terminological inaccuracies, issues that are closely tied to its traditional NMT architecture, which tends to align rigidly with source-language structures. Such alignment often results in mechanical repetition and syntactic rigidity, particularly in highly formulaic texts like technical manuals. These findings closely echo those of Sadiq (2025), who reported that GT performs poorly in professional registers, frequently exhibiting terminological confusion and unnatural syntax that impairs readability.

By contrast, ChatGPT-4 produced translations that were generally more natural and coherent.

Its advanced contextual modelling capabilities enhanced overall textual cohesion, but simultaneously introduced new types of errors. In this study, these took the form of "over-translation" and "overgeneralisation", which are not strictly mistranslations, but rather instances where the model extended beyond the source content by making inferential additions. Such behaviour reflects the generative nature of LLMs, which may incorporate plausible but unsupported elaborations. Alzain et al. (2024) similarly observed phenomena such as content addition and hallucination in ChatGPT's scientific text translations. Although no hallucinations were detected in the present study, this may be attributed to the relatively short input texts and the use of a high-resource language pair (Chinese-English), suggesting that text length and training resource density are probably important variables influencing translation stability.

It is also worth noting that Barák (2024), in a study on English-Slovak translation, found that while ChatGPT-4 performed well in terminological accuracy, its stylistic control remained problematic, particularly with regard to sentence structure and lexical tone. This aligns with the present study's observation that, despite ChatGPT-4's overall superiority, its handling of stylistic nuance can be inconsistent in certain sentence constructions.

Terminology translation and stylistic appropriateness are particularly critical dimensions in the context of technical texts. In the area of terminology, this study identified novel error categories in GT's output, such as terminology formatting errors and abbreviation misuse, suggesting systemic shortcomings in its ability to enforce terminological consistency and formatting norms. This finding is consistent with the conclusions of Mohsen (2024), who observed that GT frequently resorts to literal or incorrect term translation, thereby compromising both terminological consistency and domain-specific accuracy.

By contrast, ChatGPT-4 exhibited significantly fewer terminological errors. This aligns with Wu's (2023) findings that ChatGPT handles specialised terminology with greater precision, and also supports the MQM-based evaluation by Sanz-Valdivieso and López-Arroyo (2023), which showed that ChatGPT produced 21.57% fewer terminological errors than GT. However, it must be emphasised that fewer errors do not imply error-free output. As the present study demonstrates, ChatGPT occasionally engages in semantic overgeneralisation or substitution with near-synonyms, suggesting limitations in maintaining the strict terminological boundaries often required in technical documentation.

In terms of style, GT's issues were primarily related to mismatches in register (e.g., inappropriate levels of formality and errors of directive intensity deviation), syntactic rigidity, and generally unnatural phrasing. These issues can impair the clarity or coherence of instructional content, especially in user-facing materials. Similar stylistic shortcomings have been noted by Cai (2024) and Karim (2024), both of whom reported that GT's tendency to closely replicate source syntax results in a lack of flexibility and stylistic nuance—an issue particularly salient in instructional genres.

Although ChatGPT's outputs are generally more idiomatic and better aligned with the stylistic conventions of the target language, there is evidence of "over-optimisation" in tone, where efforts to enhance naturalness occasionally lead to diminished terminological precision or divergence from the expected stylistic register of the source domain. As noted by Alzain et al.

(2024), such stylistic deviations occur more frequently in scientific texts translated by ChatGPT. Barák (2024) also observed that despite ChatGPT's gains in fluency, it continues to face challenges in balancing naturalness of style with terminological accuracy.

This study adopted a deliberately minimalist prompt for ChatGPT-4. Pourkamali and Sharifi (2024) demonstrated that zero-shot prompting achieves greater accuracy and fluency in high-resource language translation, with minimal prompts often surpassing n-shot configurations in both efficiency and output quality. Nonetheless, emerging research indicates that more complex or tailored prompts can elicit alternative and sometimes superior results, highlighting the inherent flexibility and adaptability of large language models. (Yamada, 2023). Nair et al. (2025) further observed that extended prompts may increase the depth and creativity of responses, but they also bring additional challenges, such as greater cognitive demands on both the model and the user, as well as higher expectations for user proficiency in prompt design. Chen et al. (2024) also observed that excessively long prompts may cause LLMs to generate content unrelated to the target language, thereby increasing the risk of hallucination. As prompt engineering continues to develop, future research should systematically evaluate the merits and limitations of both minimalist and advanced prompting strategies in specific translation practices.

In conclusion, the findings of this study offer several practical implications for current translation practice. First, in the context of Chinese-English translation of technical manuals, ChatGPT-4 demonstrates superior linguistic quality. For texts that require a relatively high degree of linguistic fluency, yet involve standardised or moderately demanding terminology, ChatGPT-4 presents itself as a viable and effective automated translation solution.

Secondly, while GT exhibits a higher frequency of lower-level errors, its outputs tend to be more conservative and, consequently, easier to detect and amend. By contrast, although ChatGPT-4 produces more fluent and contextually appropriate renderings, its tendency to introduce additional content or semantically extended interpretations poses a potential risk in professional contexts. Such errors are less readily apparent and, if left unchecked in sensitive domains, such as medical instructions or legal provisions, may mislead users or result in operational misjudgements. Thus, caution remains necessary when deploying ChatGPT-4 in high-stakes technical settings without subsequent human revision.

Moreover, during manual annotation, some disagreement was observed between the two professional annotators regarding ChatGPT-4's output, particularly in the dimensions of terminology and style. This may be attributed to the model's flexible rendering tendencies: its terminological choices often reflect contextual generalisation rather than strict lexical equivalence, which, while semantically acceptable, can lead to divergent judgements on whether an error has occurred. Similarly, in terms of stylistic register, ChatGPT-4 often generates fluent but slightly colloquial expressions; while one annotator considered these as stylistic deviations, another deemed them acceptable variants. Such subjectivity underscores the need for more clearly defined evaluation criteria or the inclusion of a broader annotator pool to ensure inter-rater agreement in future assessments.

Finally, when compared to findings from other studies, it is evident that ChatGPT's strengths are most pronounced in high-resource language pairs such as Chinese-English or Spanish-

English. Its performance remains inconsistent in low-resource contexts or when handling extended texts. As Son and Kim (2023) observe, traditional NMT systems still outperform current large language models in multi-language environments and structurally complex inputs. Therefore, future research and practice should avoid a one-size-fits-all reliance on ChatGPT, and instead adopt a more nuanced, context-sensitive approach. It is essential to select the most appropriate tool based on the language pair, text genre, and intended use, while also incorporating human revision to ensure both translation quality and communicative reliability.

6. Conclusion and Limitations

This study employed qualitative analysis to compare the translation quality of GT and ChatGPT-4 across eighty Chinese-English technical manual segments. Two evaluators assessed the outputs using a five-point Likert scale based on the four core MQM dimensions: accuracy, fluency, terminology and style, and annotated the errors for systematic classification.

The comprehensive analysis demonstrated that ChatGPT-4 consistently outperformed GT in this domain, corroborating the findings of Chan and Tang (2024) regarding the advantages of LLMs in translation. In terms of accuracy, GT frequently exhibited Structural Misalignment, Information Redundancy, and Mistranslation, whereas ChatGPT-4, though generally more accurate, occasionally introduced unique error types characterised by Over-translation and semantic Over-generalisation. Regarding fluency and style, ChatGPT-4 produced smoother and more natural output, while GT was prone to Mechanical Repetition, Stilted Expression, and Formality Mismatches. On terminology, GT struggled with Formatting Errors and Abbreviation Misuse, whereas ChatGPT-4 delivered more consistent performance across all sub-dimensions.

Overall, ChatGPT-4 proves better suited to Chinese–English technical translation tasks, particularly due to its enhanced fluency and greater adaptability to stylistic conventions. However, its tendency towards semantic inference and contextual generalisation introduces subtle risks of diminished terminological precision and semantic fidelity. It is therefore essential to carefully balance these benefits and limitations when deploying ChatGPT-4 in technical translation settings.

Although the segments of product manuals provide some degree of domain representativeness, the sample size and textual diversity remain insufficient to capture the full complexity and stylistic variation inherent in technical translation tasks. The evaluation process relied on human judgement for both scoring and error annotation. While efforts were made to enhance consistency, individual preferences in language style and sensitivity to nuance may still have influenced the assessments. Furthermore, the study employed only a minimal prompt when generating translations with ChatGPT-4 and did not systematically examine the impact of prompt design on output quality, highlighting a potential avenue for future research.

References

Al-Maaytah, M., & Almahasees, Z. (2024). A Linguistic Investigation for a Case Study of Chat GPT and Google Translate in Rendering Special Needs Texts from English into Arabic: A Synchronic Case Study. *Pakistan Journal of Life & Social Sciences*, *22*(2).

https://doi.org/10.57239/PJLSS-2024-22.2.00812

Allaway, E., Bhagavatula, C., Hwang, J. D., McKeown, K., & Leslie, S. J. (2024). Exceptions, instantiations, and overgeneralization: Insights into how language models process generics. *Computational Linguistics,* 50(4), 1211-1275.

http://dx.doi.org/10.1162/coli_a_00530

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. 3, pp. 837-849).

https://doi.org/10.30564/fls.v6i3.6799

Barák, A. (2024). Comparing Machine Translation Effectivity of Selected Engines from English into Slovak on the Example of a Scientific Text. *L10N Journal*, 3(2), 7-28.

https://l10njournal.net/index.php/home/article/view/40/42

Bowker, L., & Ciro, J. B. (2019). *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Emerald Publishing Limited.

http://dx.doi.org/10.1108/9781787567214

Brewster, R. C., Gonzalez, P., Khazanchi, R., Butler, A., Selcer, R., Chu, D., ... & Hron, J. D. (2024). Performance of ChatGPT and Google Translate for pediatric discharge instruction translation. *Pediatrics*, *154*(1), e2023065573.

http://dx.doi.org/10.1542/peds.2023-065573

Briva-Iglesias, V., Camargo, J. L. C., & Dogru, G. (2024). Large language models "ad referendum": How good are they at machine translation in the legal domain?. *arXiv preprint arXiv:2402.07681.* http://dx.doi.org/10.6035/MonTI.2024.16.02

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

Chan, V., & Tang, W. K. W. (2024). GPT for Translation: A Systematic Literature Review. *SN Computer Science, 5*(8), 1-9. http://dx.doi.org/10.1007/s42979-024-03340-z

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering, 26*(2), 137-161.

http://dx.doi.org/10.1017/S1351324919000469

Chen, S., Shi, X., Li, P., Li, Y., & Liu, J. (2024). Refining translations with llms: A constraint-aware iterative prompting approach. *arXiv preprint arXiv:2411.08348*.

https://doi.org/10.48550/arXiv.2411.08348

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, *6*(4), 284. http://dx.doi.org/10.1037/1040-3590.6.4.284

Dong, Z., Tang, T., Li, L., & Zhao, W. X. (2023). A survey on long text modeling with transformers. *arXiv preprint arXiv:2302.14502.*

https://doi.org/10.48550/arXiv.2302.14502

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.

https://doi.org/10.1162/tacl_a_00437

Joshi, S. K. (2017). Challenges in translating technical texts. *Nepalese translation*, 1, 49-54.

Karim, H. A. (2024). ChatGPT vs DeepL: Comparing the English Translation Quality of Digital Business and Information Technology Texts Using BLEU Metric: ChatGPT vs DeepL: Membandingkan Kualitas Terjemahan Bahasa Inggris Teks Bisnis Digital dan Teknologi Informasi Menggunakan Metrik BLEU. *Journal of Digital Business and Information Technology, 1*(2), 50-60. https://doi.org/10.23971/jobit.v1i2.297

Karim, H. A. (2024). ChatGPT vs DeepL: Comparing the English Translation Quality of Digital Business and Information Technology Texts Using BLEU Metric. *Journal of Digital Business and Information Technology, 1*(2), 50-60.

http://dx.doi.org/10.23971/jobit.v1i2.297

Koby, G. S., Fields, P., Hague, D. R., Lommel, A., & Melby, A. (2014). Defining translation quality. *Tradumàtica*, (12), 0413-420.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine, 15*(2), 155-163.

http://dx.doi.org/10.1016/j.jcm.2016.02.012

Lakew, S. M., Cettolo, M., & Federico, M. (2018). A comparison of transformer and recurrent neural networks on multilingual neural machine translation. *arXiv preprint arXiv:1806.06957*. https://arxiv.org/pdf/1806.06957

Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation, 23*, 105-115.

Lommel, A., Gladkoff, S., Melby, A., Wright, S. E., Strandvik, I., Gasova, K., ... & Nenadic, G. (2024). The multi-range theory of translation quality measurement: Mqm scoring models

and statistical quality control. *arXiv preprint arXiv:2405.16969*.

https://arxiv.org/pdf/2405.16969

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12), 0455-463.

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. https://aclanthology.org/2024.lrec-main.120/

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

Nair, A. B., Gafoor, F. A., Reddy, B. V., Sivasakthivel, R., RajaGopal, M., & Ramar, G. (2025, March). Strategical Framework Design for Prompt Engineering Analysis-ChatGPT. In *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)* (pp. 1-8). IEEE. http://dx.doi.org/10.1109/COMP-SIF65618.2025.10969948

Ni, J., Jin, Z., Freitag, M., Sachan, M., & Schölkopf, B. (2022). Original or translated? a causal analysis of the impact of translationese on machine translation performance. *arXiv preprint arXiv:2205*.02293. http://dx.doi.org/10.18653/v1/2022.naacl-main.389

Olohan, M. (2020). Translating Technical Texts. In *Cambridge Handbook of Translation*. Cambridge University Press. http://dx.doi.org/10.1017/9781108616119.017

Pourkamali, N., & Sharifi, S. E. (2024). Machine translation with large language models: Prompt engineering for persian, english, and russian directions. arXiv preprint arXiv:2401.08429.

Radetska, S. (2024). Challenges and innovations in scientific and technical translation: Terminological complexities and 'false friends'. *The Modern Higher Education Review*, (9), 119–131. https://doi.org/10.28925/261

Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. I., Chadha, A., ... & Das, A. (2023, January). The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. *Association for Computational Linguistics*. http://dx.doi.org/10.18653/v1/2023.emnlp-main.155

Sadiq, S. (2025). Evaluating English-Arabic translation: Human translators vs. Google Translate and ChatGPT. *Journal of Languages and Translation, 12*(1), 67-95.

http://dx.doi.org/10.21608/jltmin.2025.423147

Sanz-Valdivieso, L., & López-Arroyo, B. (2023). Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation. In *International Conference Human-informed Translation and Interpreting Technology (HiT-IT 2023)* (pp. 97-107).

http://dx.doi.org/10.26615/issn.2683-0078.2023_008

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117. https://arxiv.org/pdf/1404.7828

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Siikala, T. (2018). *Value and Benefits of Technical Documentation Services. An Analysis of Corporate Websites* (Doctoral dissertation, Master Thesis). University of Tampere).

Son, J., & Kim, B. (2023). Translation performance from the user's perspective of large language models and neural machine translation systems. *Information, 14*(10), 574.

http://dx.doi.org/10.3390/info14100574

Stahlberg, F. (2020). Neural machine translation: A review. Journal of Artificial Intelligence Research, 69, 343-418. http://dx.doi.org/10.1613/jair.1.12007

Tekom Europe 2018. Defining Technical Communication. Available from http://www.technical-communication.org/technical-communication/defining-technical-communication.html

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*. https://arxiv.org/pdf/1706.03762

Wu, Y., & Hu, G. (2023, December). Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 166-169).

http://dx.doi.org/10.18653/v1/2023.wmt-1.15

Wu, M., Vu, T. T., Qu, L., Foster, G., & Haffari, G. (2024). Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401*.06468.

https://arxiv.org/pdf/2401.06468

Wu, J. (2023). A comparative analysis of Chinese-English translation quality based on ChatGPT: A case study of Chinese characteristic words. *Journal of Social Science Humanities and Literature, 6*(5), 53-58. http://dx.doi.org/10.53469/jsshl.2023.06(05).08

Yan, J., Yan, P., Chen, Y., Li, J., Zhu, X., & Zhang, Y. (2024). Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*. https://doi.org/10.48550/arXiv.2407.03658

Yamada, M. (2023). Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. *arXiv preprint arXiv:2308.01391.*

https://doi.org/10.48550/arXiv.2308.01391

Zayed, A. B. (2024). Evaluating the Fidelity and Accuracy of ChatGPT 4 and Google Translate

in Translating Legal English Documents into Arabic–and Vice Versa. *Faculty of Languages Journal-Tripoli-Libya*, *1*(29), 87-63.

Privacy Statement

All data were collected anonymously and used solely for academic research purposes. No personally identifiable information was recorded or disclosed.

Acknowledgments

Not Applicable.

Funding

Not Applicable.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Macrothink Institute.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Appendix A. Codebook of Translation Errors and Sample Annotations

Codebook A: Accuracy

Accuracy-errors occurring when the target text does not accurately correspond to the propositional content of the source text, introduced by distorting, omitting, or adding to the message.

a. Mistranslation: The translated text misrepresents the intended meaning of the source, causing distortion or misunderstanding.
"Mistranslated 'charged to 80%' as '100%', distorting key information."
b. Informational Redundancy: The translation contains unnecessary repetition of content, leading to excessive or redundant information.
"Repetition of 'at that point in time' weakens logical coherence."
c. Omission: Information explicitly present in the source text is partially or entirely missing in the translation.
"Omitted the meaning of 'positioning'."
Structural Misalignment: The grammatical or logical structure of the translation does not correspond to that of the source text.
"Misplaced 'work', affecting clarity."
Ambiguity: The translated text is vague, unclear, or open to multiple interpretations, reducing comprehension accuracy.
"Ambiguous reference of 'it' may lead to misunderstanding."
Overgeneralisation: Specific concepts in the source are translated too broadly, resulting in a loss of detail or precision.
"Weakened and generalised the meaning of 'as shown in the figure'."
Over-translation: The translation includes extra explanations or added information that are not present in the source text.
"Over-explained."

Codebook B: Fluency

Fluency-errors related to the linguistic well-formedness of the text, including problems with grammaticality, spelling, punctuation, and mechanical correctness.

Stilted Expression: Unnatural or awkward phrasing that lacks idiomatic fluency; often influenced by the source language or overly formal tone.
"Repetitive sentence structure and rigid word order identified."
Disjointed Cohesion: Poor connection between clauses or sentences, causing the text to read as fragmented or lacking in flow.
"Logical cohesion appears rigid."
Uneven Rhythm: The pacing of the sentence feels sluggish, verbose, or unbalanced, hindering smooth progression.
"The first clause is overly long."

Grammatical Error: Violations of English grammar rules, such as incorrect verb tense, subject–verb disagreement, or article misuse.
"Significant grammatical errors observed."
Inappropriate Word Order: Word or phrase sequence does not follow natural English syntax, resulting in confusion or awkwardness.
"Unnatural word order with awkward phrasing (e.g., 'cloth max')."
Mechanical Repetition: Repetitive use of similar structures or vocabulary in a rigid, formulaic manner that reduces stylistic variety.
"Repetition and mechanical wording present."

Codebook C: Terminology

Terminology-errors arising when a term does not conform to normative domain or organizational terminology standards or when a term in the target text is not the correct, normative equivalent of the corresponding term in the source text.

Non-standard Terminology: The term used does not conform to established standards, industry glossaries, or official terminology norms.
"The term 'middle dot button' is non-standard."
Terminological Inconsistency: The same term appears inconsistently across the text, causing confusion or lack of cohesion.
"Terminology is inconsistent or vague (e.g., 'AC power source')."
Non-technical Wording: The translation uses lay or colloquial expressions instead of domain-appropriate technical terms.
"The translated term 'bright spots' lacks technical precision."
Abbreviation Misuse: Abbreviations are incorrect, inappropriate, or undefined upon first use.
"Incorrect use of singular/plural forms and abbreviation errors in terminology."
Terminology Formatting Error: Terminological formatting (e.g., capitalisation, punctuation, italics) does not follow professional standards.
"Spacing error in the term 'time code'."
Terminological Vagueness: The term is overly vague, general, or lacks conceptual clarity in the technical context.
"The phrase 'loss of profit, etc.' is unclear and terminologically imprecise."
Incorrect Term Selection: The selected term is factually incorrect or misrepresents the intended concept of the source.
"Using 'lever' instead of "control arm" indicates a terminological mismatch."

Codebook D: Style

Style – errors occurring in a text that are grammatically acceptable but are inappropriate because they deviate from organizational style guides or exhibit inappropriate language style.

Formality Mismatch: The level of formality is either too casual or overly formal compared to the expected tone of the target text type.

"Some expressions are slightly colloquial."

Insufficient Technical Register: The translation lacks domain-specific tone or technical precision expected in professional or instructional contexts.

"Lacks a professional tone typical of technical documentation."

Directive Intensity Deviation: The degree of command or warning is too weak or too strong, failing to align with the communicative intent.

"Some parts are overly narrative."

Translationese Style: The text displays source-language interference or literal rendering, lacking naturalness and idiomatic expression.

"Tone is mechanical and repetitive."

Verbal Redundancy: The text contains wordy constructions, repeated phrases, or unnecessarily elaborate sentence structures.

"Style is mechanical and less concise"

Stylistic Inconsistency: The tone, punctuation, or formatting shifts inconsistently across or within sentences, reducing coherence.

"Style does not align with technical documentation; structure is disorganised."