BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies
https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Google Translate or ChatGPT-4? A Multi-Metric Evaluation of Chinese-to-English Technical Translation

Zhongming Zhang , ***Syed Nurulakla bin Syed Abdullah*** * , ***Muhammad Alif Redzuan Abdullah*** , ***Lina Zhou***

The Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Serdang 43400, Malaysia

ABSTRACT

The advent of large language models (LLMs), such as ChatGPT, has opened new avenues for machine translation (MT), particularly in specialised domains such as technical documentation. However, their performance, relative to neural MT systems like Google Neural Machine Translation (GNMT), lacks empirical validation for the Chinese-English language pair. This study aims to compare the Chinese-English translation quality of GNMT and ChatGPT-4 in technical manuals, evaluate the variability of six widely used automatic metrics, and examine their correlation with human assessment. A parallel bilingual corpus of eighty aligned segments from technical manuals was constructed. Translations generated by GNMT and ChatGPT-4 were evaluated using standard automatic lexical metrics (BLEU, METEOR, and CHRF), semantic metrics (BLEURT, BERTScore, and COMET-QE), and human assessments. Statistical analyses employed paired t-tests, Wilcoxon signed-rank tests, Friedman tests with Wilcoxon post hoc comparisons, and Spearman correlations. The results showed that human evaluators preferred ChatGPT-4 over GNMT for technical manual translation, whereas all automatic metrics favoured GNMT. Automatic evaluation revealed notable inconsistencies, with partial alignment observed in COMET-QE-related comparisons. Correlation patterns differed across systems: only semantic metrics exhibited limited correlations with human assessments for GNMT. In contrast, for ChatGPT-4, lexical metrics exhibited moderate to low correlations, whereas semantic metrics demonstrated no meaningful association. These findings highlight ChatGPT-4's advantage in human-judged translation quality, while also underscoring the misalignment between automatic metrics and

*CORRESPONDING AUTHOR:
Syed Nurulakla bin Syed Abdullah, The Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Serdang 43400, Malaysia; Email: syedakla@upm.edu.my

ARTICLE INFO
Received: 11 July 2025 | Revised: 21 July 2025 | Accepted: 28 July 2025 | Published Online: 15 September 2025
DOI: https://doi.org/10.30564/fls.v7i9.11014

CITATION
Zhang, Z., Syed Abdullah, S.N.b., Abdullah, M.A.R., et al., 2025. Google Translate or ChatGPT-4? A Multi-Metric Evaluation of Chinese-to-English Technical Translation. Forum for Linguistic Studies. 7(9): 770–788. DOI: https://doi.org/10.30564/fls.v7i9.11014

human assessments in LLM-based machine translation, thereby reinforcing the need for more context-sensitive and adaptive evaluation approaches.

Keywords: Automatic Evaluation Metrics; ChatGPT-4; Google Neural Machine Translation (GNMT); Technical Manual Translation

1. Introduction

Neural Machine Translation (NMT), as exemplified by Google Neural Machine Translation (GNMT), has led to notable improvements in translation quality. This progress stems from the use of neural networks and attention mechanisms, which support context-aware and semantically grounded translations[1,2]. More recently, large language models (LLMs), such as ChatGPT, have introduced a transformative translation paradigm shift[3], with some outputs approaching human-level quality[4]. Unlike traditional NMT systems, LLMs typically perform competitively in Chinese-English translation without extensive domain-specific fine-tuning, and sometimes even outperform specialised models[5,6].

Nonetheless, the quality of translations across both paradigms varies considerably depending on language pair and domain. While ChatGPT excels with high-resource European languages, its accuracy declines when handling low-resource or typologically distant languages[7,8]. Persistent challenges for ChatGPT translation, such as the rendering of cultural references[9] and domain-specific terminology[10], continue to hinder generalisability. In technical contexts, where translation errors may affect product functionality, legal clarity, or user safety, robust evaluation is vital. This underlines the importance of comparative research that assesses system performance under domain-specific constraints.

Technical translation differs from general-domain translation in its use of rigid syntax, a higher density of terminology, and its functional focus on facilitating product comprehension and usage[11]. Accuracy and consistency in rendering domain-specific collocations are essential to prevent ambiguity[12]. Even minor errors may lead to notable consequences, making technical translation a demanding benchmark for evaluating MT systems. Although recent studies[13–15] have confirmed the fluency and contextual adequacy of NMT and LLM systems, few have systematically evaluated their effectiveness in specialised Chinese-English technical translation tasks.

While lexical overlap metrics such as BLEU[16] and METEOR[17] remain widely used due to their efficiency and reproducibility, they are increasingly criticised for their limited ability to capture semantic adequacy and contextual nuance[18]. For instance, Chatzikoumi[19] highlighted that BLEU and related classic metrics often fail to adequately assess the semantic quality of NMT outputs, especially when translations diverge structurally from the source text yet preserve equivalent meaning.

Recent innovations have introduced semantically informed, pre-trained evaluation metrics: BERTScore[20] gauges contextual similarity using embeddings; BLEURT[21] leverages fine-tuned models aligned with human assessments; and COMET[22], developed by Unbabel, estimates quality without reference texts. However, few studies have employed both lexical and semantic metrics in tandem with human assessment to evaluate Chinese-English technical translation.

While recent developments in machine translation (MT) have yielded substantial progress, concerns persist regarding its consistency and effectiveness in domain-specific contexts. This study presents a comparative analysis of GNMT and ChatGPT-4, focusing on the Chinese-to-English translation of technical manuals. The study evaluates translation quality using both automatic metrics and human assessment, to elucidate system-level disparities, variations among metrics, and the extent to which metric scores align with human assessment. Specifically, this study seeks to address the following research questions:

RQ1. How does the Chinese-English translation quality of GNMT and ChatGPT-4 differ as evaluated by automatic metrics and human assessment?
RQ2. How do different automatic metrics vary in scoring translation quality?
RQ3. To what extent do automatic metrics correlate with human assessment across different systems?

The following section reviews relevant literature to contextualise the study within existing research and establish the analytical framework.

2. Literature Review

2.1. Technical Text Features and MT Systems Comparison

Unlike the translation of literary texts embedded with culture-specific elements by Zuo et al.[23] or the creative literary renderings explored by Zahrawi et al.[24], machine translation continues to exhibit limitations when processing culturally embedded discourse. By contrast, as Kaji[25] observed, controlled language texts in technical domains tend to be more concise and explicit, rendering them more amenable to machine translation.

Such texts typically employ passive constructions, nominalisations, and rigid syntactic patterns—particularly in formal English usage[26]. Additionally, they are marked by dense terminology, domain-specific abbreviations, and an objective, impersonal tone[27]. These features are especially prevalent in manuals, user guides, and procedural documentation, where clarity and functional precision are paramount[28]. Given these linguistic complexities, it is imperative to examine how MT systems respond to the specific demands of technical texts.

To address these challenges, researchers have increasingly applied NMT systems and LLMs to technical translation tasks involving diverse language pairs. Recent studies present a nuanced picture of their comparative strengths. For instance, Barák[29] found that ChatGPT-4 surpassed DeepL in semantic adequacy and error management when translating English-Slovak scientific texts, although DeepL demonstrated more consistent syntactic accuracy and fluency. In the English-Arabic context, Sadiq[30] reported that ChatGPT achieved the highest scores for fluency and semantic precision, while GNMT lagged in terminology rendering and cultural sensitivity. However, findings are not universally consistent. Alzain et al.[14] observed that GNMT produced fewer errors than ChatGPT when translating English scientific texts into Arabic. In contrast, Karim[31] emphasised ChatGPT's stronger ability to convey intended meaning, while DeepL outperformed both systems in terms of grammatical consistency.

These studies collectively underscore the variability in MT system performance depending on the language pair and specific textual genre. Notably, there has been scant attention to Chinese-English technical translation, a pairing characterised by significant structural divergence and high domain specificity. To address this gap, the present study conducts a domain-sensitive, multi-metric comparison of GNMT and ChatGPT-4 translation in Chinese-to-English technical texts.

2.2. Evaluation for Machine Translation Quality

Machine translation quality evaluation primarily comprises human assessment and automated metrics. Human assessment typically involves rating dimensions such as fluency and adequacy, often using Likert scales[32] or continuous scales within Direct Assessment frameworks[33]. More advanced approaches include semantically oriented methods such as HUME (Human UCCA-Based Machine Translation Evaluation), proposed by Birch et al.[34], which employs the UCCA framework to assess the extent to which source meaning is preserved in the translation. These methods offer practically meaningful insights, particularly useful in large-scale human assessment within specialised domains such as technical and scientific translation.

Another commonly employed approach to human evaluation involves categorising translation errors, ideally complemented by detailed error analysis using frameworks such as MQM (Multidimensional Quality Metrics)[35]. Its highly granular framework offers analytical precision but also poses practical challenges for annotators. Evaluating depends on expert judgement from trained annotators, yet distinctions between error types (e.g., mistranslation and terminology issues) and severity levels are often ambiguous, limiting reproducibility[19,36]. As a result, MQM proves less scalable for system-wide or cross-domain comparative evaluations. In this context, streamlined evaluation tools, such as five-point Likert scales that target key dimensions such as accuracy, fluency, terminology, and style, offer a pragmatic and scalable solution.

In contrast to human assessment, automatic metrics offer superior efficiency, speed, consistency, and cost-effectiveness[19,37], making them widely adopted in both academic research and industrial applications. These metrics generally fall into two broad categories: lexical or string-based metrics (e.g., BLEU, CHRF, METEOR) and semantic or pre-trained metrics (e.g., BERTScore, BLEURT, COMET-QE)[7]. Most of these metrics rely on reference translations, whereas reference-free models such as COMET-QE evaluate output directly in relation to the source text, thereby enabling more independent and context-sensitive evaluation.

BLEU (Bilingual Evaluation Understudy)[16] calculates word-level n-gram precision between MT output and reference, applying a brevity penalty to penalise excessively short hypotheses. CHRF (Character n-gram F-score)[38] adopts the character-level n-gram, enabling finer-grained evaluation of morphological and orthographic variation through a balanced F-score of precision and recall. METEOR[17] extends n-gram matching by integrating stemming, synonymy, and word order penalties, thereby aiming to approximate human judgement more closely than surface-level metrics.

BERTScore[20] uses contextual embeddings from pre-trained transformer models to compute cosine similarity between tokens, generating soft alignments that underpin its calculation of precision, recall, and F1 scores. BLEURT[21] leverages pre-trained contextual representations fine-tuned on human-annotated data to predict semantic adequacy, showing promising performance in capturing semantic information[39]. COMET[18] similarly produces standardised quality estimates using both source and reference embeddings. Building upon COMET, COMET-QE[22] provides reference-free quality estimation based solely on source input and MT output. Although reference-free metrics may offer greater flexibility in application, they may exhibit bias against higher-quality outputs, including those produced by human translators[40].

As deep learning enables MT systems to move beyond surface-level lexical matches and capture richer semantic representations, evaluation frameworks must likewise evolve. As Ulitkin et al.[41] suggested, modern automatic evaluation tools not only facilitate error classification but also inform the continuous improvement of MT models. The integration of lexical and semantic metrics thus offers a more comprehensive and robust means of assessing translation quality. Nonetheless, the reliability of these metrics in evaluating translations produced by the latest AI-driven systems in scientific and technical domains remains to be substantiated.

2.3. Comparative Research on NMT and LLM-Based Translation Systems

This section reviews recent comparative studies on the translation performance of NMT systems and Large Language Models (LLMs) across various language pairs and text types. Particular attention is given to methodological differences that inform the design of the present study. For instance, Ding[42] compared ChatGPT and four NMT systems in the English-Chinese legal domain, finding that NMT systems generally produced fewer severe errors. Similarly, Briva-Iglesias[43] evaluated ChatGPT-4 and GNMT in multilingual legal translation. While NMT systems achieved higher scores on most automatic metrics (e.g., BLEU, CHRF), human evaluators preferred ChatGPT-4 for its consistent terminology usage and inter-sentential coherence in high-resource language pairs.

Evaluation methodologies across these studies vary considerably. Some employ purely qualitative human assessments, as in the works of Sanz-Valdivieso and López-Arroyo[44], AlAfnan[45], and Mohsen[46]. Others rely exclusively on automatic metrics. For example, Son and Kim[15] applied BLEU, TER, and CHRF to evaluate translation quality across eighteen language pairs in the context of news reporting. Hybrid approaches are also common: Brewster et al.[47] assessed ChatGPT and GNMT using human ratings of adequacy and fluency for paediatric health texts, revealing comparable performance in Spanish and Portuguese, but diminished quality for both systems in Haitian Creole. Notably, Sizov et al.[48] integrated BLEU and COMET with explainability tools to analyse linguistic features in translations produced by NMT, LLMs, and human translators. They found that LLM outputs more closely resembled human-authored language than those of NMT systems, although both remained distinguishable from native-authored texts.

Within the Chinese-English translation context, Cai[49] examined editorial texts and found that ChatGPT produced more fluent and grammatically coherent translations than GNMT. However, it was less effective in capturing cultural references and idiomatic expressions. Jiang et al.[50] noted that in stylistically sensitive genres such as diplomatic discourse, the correlation between automatic metrics and human assessment was relatively weak. Although ChatGPT's translation quality improved markedly with prompt engineering, automatic metrics failed to adequately reflect these improvements. This underscores the continued necessity of human assessment and highlights the need to develop automatic metrics capable of capturing the semantic diversity characteristic of generative models.

2.4. Research Gaps

While prior studies have explored the translation performance of NMT systems and LLMs across various domains, important gaps remain in the context of Chinese-English technical translation. First, existing research rarely focuses on system-level comparisons between NMT and LLMs within this specific language pair and domain; as a result, their relative strengths in technical contexts remain insufficiently examined. Second, although various automatic metrics have been widely adopted, little empirical attention has been given to how their scoring behaviours differ across technical domains. Third, the extent to which automatic metrics align with human assessments across systems remains insufficiently validated. These gaps underscore the need for a comprehensive evaluation framework that addresses system-level differences, metric-level variation, and the reliability of automatic evaluation.

3. Methodology

This study employed a quantitative design to evaluate the translation performance of GNMT and ChatGPT-4, using both automatic metrics and human assessments on Chinese-English technical manual texts. Six automatic metrics were employed, comprising lexical overlap-based measures (BLEU, METEOR, and CHRF); and semantically oriented metrics (BLEURT, BERTScore and COMET-QE) to capture both surface-level similarity and deeper semantic alignment. All translations were paragraph-level outputs, derived from a diverse range of product technical manuals. Evaluation scores were standardised through normalisation procedures to enhance the validity of cross-metric comparisons. This design facilitates the identification of potential performance differences between systems and allows for a structured examination of metric-level behaviour.

3.1. Research Hypotheses

Based on the preceding research questions and methodological framework, the following hypotheses were proposed to guide the quantitative analysis. They were designed to examine performance differentials between systems, the sensitivity of individual metrics to distinct aspects of translation quality, and the extent to which automatic metrics align with human assessment.

H1. *There is a statistically significant difference in translation quality between Google Translate and ChatGPT-4 across the four technical sub-domains, as evaluated by both automatic metrics and human assessment.*

H2. *Automatic metrics vary significantly in their sensitivity to lexical and semantic translation quality, depending on the translation systems.*

H3. *There is a statistically significant correlation between automatic metric scores and human assessment scores, with semantic metrics demonstrating stronger correlations than lexical metrics.*

To examine the proposed hypotheses, this study conducted a methodological framework that integrates a purpose-designed bilingual corpus with both automated and human evaluation procedures. The subsequent sections detailed the corpus construction process and articulated the underlying rationale for the key decisions made in data selection and assessment design. (Transition paragraph)

3.2. Corpus Selection and Rationale

The study constructed a small-scale bilingual corpus through purposive sampling of paragraph-level segments (80–120 words). These were drawn from publicly available Chinese-English product manuals, including those for the ThinkPad laptop, Apple devices, Toyota Highlander automobile, and Sony digital camera. The segment length was selected to provide adequate contextual information for automatic metrics. Excessive length was avoided, as longer inputs are associated with an increased risk of hallucinations in LLM-generated outputs[51].

Text selection was guided by four key criteria to ensure both methodological rigour and practical relevance. First, the materials were drawn from domains such as consumer electronics and automotive technology to reflect real-world, end-user contexts with practical significance. Second, only texts accompanied by professional parallel English-Chinese translations were included, facilitating the application of reference-based automatic metrics. Third, open-access documents were prioritised to maintain ethical transparency and support reproducibility. Finally, texts were chosen for their consumer-facing nature, thereby enhancing the usability and societal relevance of the research outcomes.

These manuals exemplify technical communication characterised by procedural clarity, standardised terminology, and domain-specific phraseology—linguistic features that continue to challenge MT systems[52]. The inclusion of authoritative, real-world parallel corpora enhances ecological validity. It also strengthens the correlation between automatic metrics and human assessments, as high-quality reference translations have been shown to increase metric reliability[53]. These texts were selected to provide a robust foundation for evaluating MT output in technical genres and to ensure consistency across automated and human evaluation.

3.3. Translation Generation

To ensure consistency in translation generation, outputs from GNMT were obtained directly via its web interface. For ChatGPT-4, a minimal standardised prompt, "Please translate the following text into English", was employed to minimise prompt-induced variability. Such minimal prompting reduces the risk of prompt-specific interference. It enables the model's output to reflect its default translational tendencies more accurately, rather than any stylistic bias introduced by elaborate instructions. This approach supports input uniformity across systems, thereby strengthening the internal validity of comparative analyses. According to Pourkamali and Sharifi[54], zero-shot prompting achieves greater accuracy and fluency in high-resource translation, with minimal prompts often surpassing n-shot configurations in both efficiency and quality. (add rationale for ChatGPT-4 prompting)

All translations were generated in May 2025, thereby avoiding confounding effects related to system updates and enabling a fair, synchronised comparison between the two systems. To evaluate translation quality, this study adopted both automatic and human assessment methods, selected for their complementary strengths and relevance to the research aims. (transition paragraph)

3.4. Automatic Metrics

Translation quality in this study was evaluated through a combination of automatic metrics and human assessment to ensure both depth and balance in the analysis. The automatic evaluation involved two complementary categories: lexical metrics and semantic metrics.

The lexical metrics—BLEU, METEOR, and CHRF—remain standard tools for assessing surface-level correspondence between machine-generated translations and reference texts. BLEU is the most established benchmark in machine translation research, facilitating comparability with prior studies and industry norms. METEOR extends BLEU's capabilities by incorporating synonym recognition and stemming, thereby offering greater sensitivity to lexical and morphological variation[17]. CHRF, which operates at the character level, captures fine-grained errors such as inflectional inconsistencies and spelling deviations[36]. This makes it particularly well-suited to technical translation, where precision at the subword level is often critical.

To complement these surface-oriented measures, the study also employed three semantically informed metrics. BLEURT[21], fine-tuned on human judgment data, is designed to approximate human perceptions of overall adequacy and quality. BERTScore[20], by contrast, computes token-level semantic similarity using contextual embeddings, providing a granular measure of segment-level meaning alignment. The combination of BLEURT and BERTScore ensures that both holistic quality and detailed semantic correspondence are robustly captured. In addition, COMET-QE[18] was incorporated as a reference-free quality estimation metric. This allows for evaluation in scenarios where gold-standard references are unavailable. Such conditions are common in real-world translation applications. (rationale for choosing these metrics)

BLEU and CHRF scores were computed using SACREBLEU[55], following the standard configuration adopted in WMT evaluations. METEOR scores were derived using the meteor_score function from the NLTK library. BLEURT evaluations employed the official bleurt-base-512 checkpoint via TensorFlow, while BERTScore was calculated using the RoBERTa-large model implemented through Hugging Face's Transformers library. For COMET-QE, reference-free quality estimation was performed using the Unbabel COMET framework with the wmt20-comet-qe-da model. This dual-strand metric design enabled a comprehensive evaluation. It supported system-level comparisons between GNMT and ChatGPT-4, while also facilitating inter-metric analysis to identify scoring discrepancies and potential alignment with human preferences.

3.5. Human Assessment

To assess translation quality, this study adopted a human evaluation framework grounded in the Multidimensional Quality Metrics (MQM) framework. The assessment focused on four principal error categories: accuracy, fluency, terminology, and style. Specifically, accuracy pertains to meaning shifts arising from omissions, additions, or distortions of the source content. fluency involves grammatical correctness, proper spelling, punctuation, and the overall linguistic coherence of the target text. terminology refers to the improper use of domain-specific terms or inconsistencies with established terminological conventions. Lastly, style concerns instances where the translation, despite being grammatically correct, fails to align with the expected technical or organisational style.

Two professional translators, each possessing more than five years of industry experience and holding Master's degrees in Translation Studies, participated as independent human evaluators. Before the assessment, both assessors underwent calibration training through simulated scoring sessions to ensure consistency and minimise potential bias in judgement.

All translation segments were evaluated using a five-point Likert scale (1 = very poor, 5 = excellent) across the four quality dimensions. Under the procedures outlined by Graham et al.[33] and Castilho and O'Brien[56], the final score for each segment was calculated as the average of the two evaluators' ratings. This averaging method serves to minimise inter-rater variability and provides a systematic foundation for the human evaluation process. The human assessments complemented the automatic metrics employed in this study, offering a comprehensive perspective on translation performance. (more details of human assessment)

3.6. Data Analysis

Data were analysed using SPSS (v27) and JASP (v0.17.2) for statistical testing and visualisation, with Microsoft Excel used for data organisation and normalisation. To ensure all evaluation scores were presented on a consistent percentage scale, they were converted to a 0-100 range. For metrics such as COMET-QE, which may produce negative values, Min-Max normalisation was applied to ensure consistent scaling[57]. This transformation preserves score distributions while enabling consistent interpretation and facilitating valid cross-metric comparison.

The Shapiro-Wilk test was used to assess data normality. Based on the results, either paired-samples t-tests or Wilcoxon signed-rank tests were employed to compare the performance of translation systems (RQ1). To examine scoring variation across metrics (RQ2), the Friedman test was conducted alongside post hoc analyses, supplemented by clustered boxplot visualisations. For RQ3, Spearman correlation was tested to assess relationships between automatic metrics and human ratings, owing to its robustness against non-normal distributions and outliers.

With the methodological framework in place, the following section reported the results obtained through both automated and human evaluation. These findings were based on standardised scoring and statistical procedures designed to ensure analytical rigour and enable valid cross-system comparison. (transition paragraph)

4. Results and Discussion

This section reports the results of statistical analyses and discusses the findings in relation to the study's three research questions. Given the relatively small sample size of translations generated by GNMT and ChatGPT-4, the Shapiro-Wilk test was first conducted to assess the normality of score distributions, thereby informing the choice of subsequent statistical tests.

4.1. RQ1: Translation Quality Comparison: GNMT vs. ChatGPT-4

4.1.1. Results

As shown in **Table 1**, a *p*-value above 0.05 was interpreted as evidence of normal distribution. To ensure consistency, a metric was deemed normally distributed only when both GNMT and ChatGPT-4 outputs passed the Shapiro-Wilk test. CHRF, METEOR, and BLEURT met this criterion and were therefore analysed using paired-samples t-tests. In contrast, for BLEU, BERTScore, COMET-QE, and human assessment, where normality was not confirmed across both systems, the Wilcoxon signed-rank test was employed.

Results from the paired-sample t-tests and effect sizes (**Table 2**) showed that all three automatic metrics yielded *p*-values below 0.001, indicating statistically significant differences in translation quality between GNMT and ChatGPT-4. GNMT consistently outperformed ChatGPT-4, with mean differences of 4.31, 6.13, and 3.64, respectively. Following Cohen's d[58] benchmarks, CHRF, METEOR, and BLEURT showed moderate effect sizes, indicating statistically significant yet modest differences. These distinctions were likely perceptible to human evaluators but insufficient to suggest major performance divergence.

Table 1. Shapiro-Wilk test results (bold indicates *p* > 0.05).

Metrics	System	W (Shapiro-Wilk)	*p*-Value	Normality
BLEU	GNMT	0.983	**0.388**	**Yes**
BLEU	ChatGPT-4	0.939	<0.001	No
CHRF	GNMT	0.989	**0.706**	**Yes**
CHRF	ChatGPT-4	0.973	**0.082**	**Yes**
METEOR	GNMT	0.984	**0.418**	**Yes**
METEOR	ChatGPT-4	0.978	**0.186**	**Yes**
BERTScore	GNMT	0.982	**0.301**	**Yes**
BERTScore	ChatGPT-4	0.965	0.029	No
BLEURT	GNMT	0.992	**0.899**	**Yes**
BLEURT	ChatGPT-4	0.978	**0.180**	**Yes**
COMET-QE	GNMT	0.944	0.002	No
COMET-QE	ChatGPT-4	0.951	0.004	No
Human Assessment	GNMT	0.967	0.038	No
Human Assessment	ChatGPT-4	0.957	0.009	No

Table 2. Paired-sample T-test and effective sizes.

Comparison	Mean Difference	t(df)	*p*	95% CI	Cohen's d	95% CI
CHRF(GNMT-ChatGPT-4)	4.31	5.96 (79)	**<0.001**	[2.87, 5.74]	0.666	[0.422, 0.907]
METEOR(GNMT -ChatGPT-4)	6.13	5.86 (79)	**<0.001**	[4.05, 8.21]	0.655	[0.412, 0.895]
BLEURT(GNMT -ChatGPT-4)	3.64	3.85 (79)	**<0.001**	[1.76, 5.52]	0.430	[0.200, 0.658]

Wilcoxon signed-rank tests (**Table 3**) revealed significant differences in system performance across all metrics. The automatic metrics—BLEU, BERTScore, and COMET-QE—consistently rated GNMT higher than ChatGPT-4 (Z = −4.988 to −3.976, *p* < 0.001), with effect sizes ranging from r = 0.445 to 0.558, indicating medium to large effects. In contrast, human assessments strongly favoured ChatGPT-4 (Z = −7.739, *p* < 0.001), with a large effect size (r = 0.865), suggesting a pronounced divergence in perceived translation quality.

Table 3. Wilcoxon test result.

Comparison	*Z*-Value	*p*-Value	Effect Size (r)
BLEU (GNMT & ChatGPT-4)	-4.988^b	< 0.001	0.558
BERTScore (GNMT & ChatGPT-4)	-4.931^b	< 0.001	0.551
COMET-QE (GNMT & ChatGPT-4)	-3.976^b	< 0.001	0.445
Human Assessment (GNMT & ChatGPT-4)	-7.739^c	< 0.001	0.865

* a. Wilcoxon signed ranks text; b. based on positive ranks; c. based on negative ranks.

This study supported Hypothesis H1, demonstrating statistically significant differences in translation quality between GNMT and ChatGPT-4 across technical manuals. Notably, a systematic divergence emerged between evaluation methods: while human annotators consistently favoured translations produced by ChatGPT-4, all automatic metrics ranked GNMT higher.

4.1.2. Discussion

These results revealed not only the relative performance disparity between the two systems in this study, but also a misalignment between human assessment and algorithmic scoring mechanisms. This pattern was similarly observed in Jiang et al.'s[50] comparative study on political and diplomatic discourse. Previous research has shown that automatic metrics tend to reward outputs that are structurally conventional, terminologically consistent, and closely aligned with expressions common in training data[59]. As GNMT tends to produce literal, syntax-preserving translations, it is more likely to score highly under such metrics.

This pattern mirrors the findings of Briva-Iglesias et al.[43] in the legal domain, where automatic metrics also favoured the output of NMT systems, while human evaluators preferred ChatGPT-4's more contextually adaptive style. This suggests that such evaluative bias may be consistent across domains, rather than confined to technical language.

The observed discrepancy largely stems from the design of current evaluation algorithms. Traditional surface-based metrics such as BLEU rely on n-gram overlap and fail to account for syntactic variation or discourse-level adequacy[19]. Even semantically informed models like COMET and BLEURT, which are trained on high-alignment reference corpora, may undervalue structurally divergent yet semantically accurate outputs[60,61].

Moreover, LLMs such as ChatGPT-4 often adopt rephrased or explicative strategies when handling infrequent or semantically ambiguous segments[62]. Such translations often appear contextually appropriate and stylistically natural. However, they may introduce greater semantic distance from reference texts, which can lead automatic metrics to underestimate their quality despite receiving favourable human assessments.

In summary, ChatGPT-4 tends to adopt contextually adaptive and interpretative translation strategies, which align more closely with human evaluators' preferences for semantic naturalness and contextual coherence. In contrast, GNMT prioritises syntactic fidelity and terminological consistency, thereby achieving higher scores under automatic metrics that favour surface-level alignment. The superior performance of ChatGPT-4 in human evaluation may stem from its enhanced capacity for contextual comprehension and natural language generation. This enables it to produce translations that are semantically complete, stylistically fluent, and more reflective of human linguistic patterns. Previous studies have likewise indicated that large language models, particularly in the case of extended texts and complex discourse scenarios, demonstrate greater linguistic adaptability and expressive naturalness[63], further substantiating the underlying mechanism behind the observed divergence between human and machine assessments.

4.2. RQ2: Variability across Automatic Metrics

4.2.1. Results

Although some automatic metrics met the normality assumption, others—namely BLEU, BERTScore, COMET-QE, and human assessments—did not. To ensure methodological consistency and analytical robustness, the Friedman test was conducted across all six automatic metrics to assess overall differences for ChatGPT-4 and GNMT, respectively.

Friedman test for the automatic metrics of ChatGPT-4 and GNMT were summarised in **Table 4**. The results showed that both systems demonstrate statistically significant differences across metrics ($\chi^2[5] > 268.5$, $p < 0.001$), with large effect sizes (Kendall's W > 0.67). Additionally, the mean rank suggested a high degree of concordance among evaluation metrics in ranking translation outputs within both ChatGPT-4 and GNMT.

Among all metrics, BLEURT consistently yielded the lowest mean scores (ChatGPT-4: M = 28.01, SD = 10.19; GNMT: M = 31.65, SD = 11.30) and the lowest mean ranks (1.23 and 1.18, respectively), followed by BLEU. In contrast, BERTScore produced the highest mean scores and ranks for both systems (ChatGPT-4: M = 71.04, SD = 8.01, Rank = 5.40; GNMT: M = 74.47, SD = 7.62, Rank = 5.23), indicating a relatively consistent scoring pattern. Thus, although absolute values differ between systems, the relative scores from automatic metrics across systems remained stable.

To visualise the distributional differences across evaluation metrics and between the two MT systems, a clustered boxplot was produced (**Figure 1**). Scores were normalised to a 0-100 scale on the x-axis, with individual metrics displayed along the y-axis for both GNMT and ChatGPT-4. This visual summary aligns with the Friedman test results (**Table 4**), which further identified statistically significant disparities in ranking patterns across the six metrics. To enable a more detailed comparison across individual automatic metrics, a Wilcoxon post hoc analysis (**Table 5**) was conducted to validate the preliminary distributional patterns.

Table 4. Friedman test for automatic metrics: GNMT vs. ChatGPT-4.

Metric	System	Mean Score	SD	Mean Rank
BLEU	ChatGPT-4	39.00	12.74	2.09
	GNMT	45.60	13.47	2.13
CHRF	ChatGPT-4	65.40	9.09	4.54
	GNMT	69.71	9.10	4.39
METEOR	ChatGPT-4	60.45	11.61	3.58
	GNMT	66.58	10.82	3.83
BERTScore	ChatGPT-4	71.04	8.01	5.40
	GNMT	74.47	7.62	5.23
BLEURT	ChatGPT-4	28.01	10.19	1.23
	GNMT	31.65	11.30	1.18
COMET-QE	ChatGPT-4	61.89	22.28	4.18
	GNMT	66.57	22.44	4.25

* Friedman test results indicated significant differences across metrics for both systems: ChatGPT-4, $\chi^2(5) = 281.564$, $p < 0.001$, Kendall's W = 0.704; GNMT, $\chi^2(5) = 268.510$, $p < 0.001$, Kendall's W = 0.671.

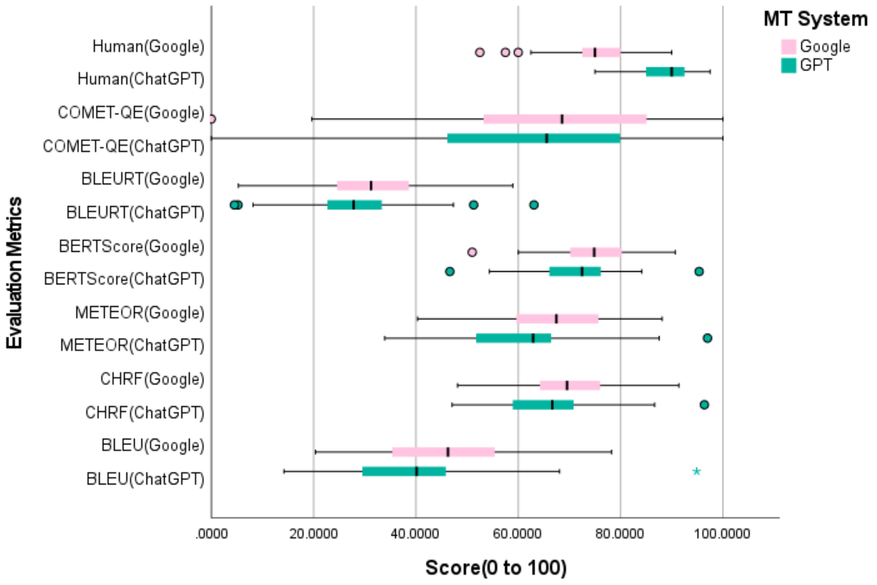


Figure 1. Clustered boxplot of metric scores for GNMT and ChatGPT-4.

Table 5. Wilcoxon post hoc tests for automatic metrics.

Comparison	Z (ChatGPT-4)	*p* (ChatGPT-4)	Z (GNMT)	*p* (GNMT)
CHRF vs. BLEU	−7.770[b]	<0.001	−7.770[b]	<0.001
BERTScore vs. CHRF	−6.823[b]	<0.001	−6.034[b]	<0.001
METEOR vs. CHRF	−7.005[c]	<0.001	−5.498[c]	<0.001
BLEURT vs. CHRF	−7.770[c]	<0.001	−7.770[c]	<0.001
COMET-QE vs. CHRF	−0.609[c]	**0.542**	−0.470[c]	**0.638**
BERTScore vs. BLEU	−7.770[b]	<0.001	−7.770[b]	<0.001
METEOR vs. BLEU	−7.770[b]	<0.001	−7.770[b]	<0.001
BLEURT vs. BLEU	−7.101[b]	<0.001	−7.581[c]	<0.001
COMET-QE vs. BLEU	−5.875[b]	<0.001	−5.482[b]	<0.001
BERTScore vs. METEOR	−7.573[b]	<0.001	−7.046[b]	<0.001
BLEURT vs. BERTScore	−7.770[c]	<0.001	−7.770[c]	<0.001
COMET-QE vs. BERTScore	−2.878[c]	**0.004**	−2.139[c]	**0.032**
BLEURT vs. METEOR	−7.770[c]	<0.001	−7.770[c]	<0.001
COMET-QE vs. METEOR	−1.050[b]	**0.294**	−0.350[b]	**<0.726**
COMET-QE vs. BLEURT	−7.266[b]	<0.001	−7.262[b]	<0.001

* a. Wilcoxon signed-rank test; b. based on negative ranks; c. based on positive ranks.

Wilcoxon signed-rank tests were conducted across all 15 pairwise metric comparisons within each system, with Bonferroni correction applied ($\alpha = 0.0033$). As shown in **Table 5**, 12 comparisons yielded statistically significant results ($p < 0.001$) across both systems, indicating substantial divergences in how different metrics evaluate translation quality. These results provided partial support for Hypothesis 2, suggesting that most automatic metrics exhibit statistically significant differences in their evaluation patterns.

However, the expected lexical-semantic distinction was not clearly reflected in the observed results. Notably, significant differences were also observed within both lexical and semantic metric groups. This indicated that even metrics within the same category may display divergent scoring behaviours, thereby challenging the assumed consistency between metric typology and actual evaluative patterns. Furthermore, despite these divergences, the scoring patterns remained highly consistent across systems. This suggests that the observed differences arise more from the metrics themselves than from the characteristics of the translation systems.

4.2.2. Discussion

According to the results of **Table 4** and **Figure 1**, considerable variability was observed across the metrics in both score magnitude and distribution. No consistent distinction emerged between lexical and semantic metrics in terms of stability; however, COMET-QE exhibited the widest dispersion. Notably, before normalisation, some of COMET-QE absolute scores were negative and not corroborated by human assessments. This volatility suggested potential limitations in the robustness of COMET-QE for technical-domain applications, echoing concerns raised by Marie[64] and He et al.[65] regarding the interpretability and reliability of reference-free evaluation systems. Similarly, Deutsch et al.[40] identified evaluation biases for higher-quality outputs, potentially linked to the constrained domain scope of the training data. These findings suggest that further recalibration or threshold tuning may be necessary when applying such models in evaluative contexts where accuracy and reliability are critical.

BLEU exhibited consistently low and tightly clustered scores, with medians below 50. This outcome likely reflected BLEU's strict dependence on n-gram overlap and brevity penalties, which penalise syntactic and lexical variation—particularly characteristic of LLM-generated translations. These findings support previous research[7,64], which identified BLEU as a conservative metric that may underrepresent translation quality in certain contexts.

Similarly, BLEURT, despite being semantically oriented, returned the lowest absolute scores across all metrics. Its reliance on large-scale reference-aligned training data would compromise its responsiveness to paraphrastic or stylistically nuanced outputs, particularly in specialised domains. This corroborates prior findings by Yan et al.[61], who noted BLEURT's tendency to undervalue linguistically varied LLM outputs.

By contrast, BERTScore, METEOR, and CHRF demonstrated higher median scores and narrower distributions, indicating better internal consistency. While CHRF is also based on n-gram matching, it operates at the character level[38], enabling it to capture morphological variants and symbol-level alignment. METEOR, though surface-level in nature, incorporates stemming and synonym matching[17], which may afford it greater tolerance for terminological variation. Overall, these results indicated that lexical metrics should not be dismissed as unsuitable for technical MT evaluation. Semantic metrics could not categorically outperform surface-level counterparts, and traditional measures might offer distinctive strengths in domain-specific contexts.

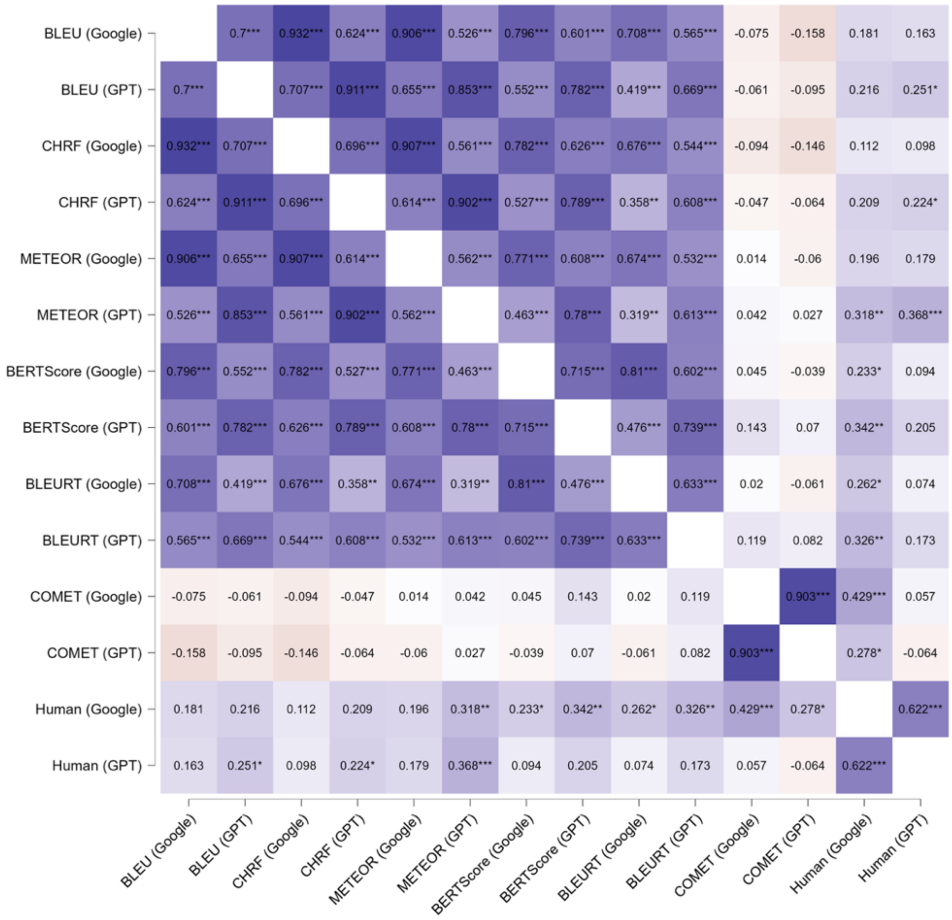
As indicated in **Table 5**, the Wilcoxon post hoc tests reaffirmed that most evaluation metrics exhibited statistically significant differences when assessing the two systems' translations of technical manual texts. However, the resulting pattern did not conform to the anticipated lexical-semantic distinction, suggesting that the boundaries between metric categories may be more fluid than previously assumed. Differences were found both across and within metric categories, indicating that sensitivity variations are largely attributable to metric-specific characteristics rather than to group-level distinctions. These findings underscore the need to treat evaluation metrics as individually calibrated tools rather than members of rigid lexical or semantic categories. This aligns with Sai et al.[66] who emphasised that variation in evaluation outcomes often arises from differences in metrics' underlying architectures, modelling strategies, and parameter settings. These distinctions highlighted the individuality of metric behaviour in translation evaluation.

Notably, three COMET-QE-related comparisons—COMET-QE vs. CHRF, METEOR, and BERTScore—did not reach statistical significance ($p > 0.0033$), suggesting an apparent convergence. However, this convergence should be interpreted with caution. The COMET-QE scores, before normalisation, exhibited high dispersion and frequent outliers, potentially weakening the statistical power of the test. Similar findings were raised by Marie[64] in the context of WMT evaluations, where COMET-QE was found to be prone to scoring instability and inconsistent sensitivity.

In summary, the results highlighted substantive divergences among metrics in terms of scoring logic and evaluative focus, underscoring the importance of metric triangulation in translation quality assessment. For contexts requiring scoring stability and responsiveness, CHRF, METEOR, and BERTScore might serve as more robust alternatives to COMET-QE in this context.

4.3. RQ3: Correlation between Automatic Metrics and Human Assessment

Although automatic metrics tended to favour GNMT, their alignment with human assessments at the paragraph level remains inconclusive. To explore this further, correlation analyses were undertaken to assess the extent to which each metric reflects human preferences at the item level. Owing to the non-normal distribution of the data, Spearman's rank-order correlation was deemed a suitable method. While some evaluation scores appeared unusually low and might be considered outliers, they were retained to reflect potential limitations of the metrics themselves. This decision allows for a more nuanced account of variability and evaluative bias. To examine H3, Spearman's rho (**Figure 2**) was computed between six metrics and human scores for both GNMT and ChatGPT-4 outputs.



	BLEU (Google)	BLEU (GPT)	CHRF (Google)	CHRF (GPT)	METEOR (Google)	METEOR (GPT)	BERTScore (Google)	BERTScore (GPT)	BLEURT (Google)	BLEURT (GPT)	COMET (Google)	COMET (GPT)	Human (Google)	Human (GPT)
BLEU (Google)		0.7***	0.932***	0.624***	0.906***	0.526***	0.796***	0.601***	0.708***	0.565***	-0.075	-0.158	0.181	0.163
BLEU (GPT)	0.7***		0.707***	0.911***	0.655***	0.853***	0.552***	0.782***	0.419***	0.669***	-0.061	-0.095	0.216	0.251*
CHRF (Google)	0.932***	0.707***		0.696***	0.907***	0.561***	0.782***	0.626***	0.676***	0.544***	-0.094	-0.146	0.112	0.098
CHRF (GPT)	0.624***	0.911***	0.696***		0.614***	0.902***	0.527***	0.789***	0.358**	0.608***	-0.047	-0.064	0.209	0.224*
METEOR (Google)	0.906***	0.655***	0.907***	0.614***		0.562***	0.771***	0.608***	0.674***	0.532***	0.014	-0.06	0.196	0.179
METEOR (GPT)	0.526***	0.853***	0.561***	0.902***	0.562***		0.463***	0.78***	0.319**	0.613***	0.042	0.027	0.318**	0.368***
BERTScore (Google)	0.796***	0.552***	0.782***	0.527***	0.771***	0.463***		0.715***	0.81***	0.602***	0.045	-0.039	0.233*	0.094
BERTScore (GPT)	0.601***	0.782***	0.626***	0.789***	0.608***	0.78***	0.715***		0.476***	0.739***	0.143	0.07	0.342**	0.205
BLEURT (Google)	0.708***	0.419***	0.676***	0.358**	0.674***	0.319**	0.81***	0.476***		0.633***	0.02	-0.061	0.262*	0.074
BLEURT (GPT)	0.565***	0.669***	0.544***	0.608***	0.532***	0.613***	0.602***	0.739***	0.633***		0.119	0.082	0.326**	0.173
COMET (Google)	-0.075	-0.061	-0.094	-0.047	0.014	0.042	0.045	0.143	0.02	0.119		0.903***	0.429***	0.057
COMET (GPT)	-0.158	-0.095	-0.146	-0.064	-0.06	0.027	-0.039	0.07	-0.061	0.082	0.903***		0.278*	-0.064
Human (Google)	0.181	0.216	0.112	0.209	0.196	0.318**	0.233*	0.342**	0.262*	0.326**	0.429***	0.278*		0.622***
Human (GPT)	0.163	0.251*	0.098	0.224*	0.179	0.368***	0.094	0.205	0.074	0.173	0.057	-0.064	0.622***	

Figure 2. Spearman Correlations Between Human Assessment and Automatic Metrics.

(* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Colour intensity denotes correlation strength: darker shades indicate stronger positive relationships; lighter tones denote weaker or negative ones.)

4.3.1. Results

The analysis first examined correlations under the GNMT condition. **Figure 2** and the results presented in Appendix A revealed notable variation in the correlations between automatic metrics and human assessments. Overall, semantically-oriented metrics demonstrated statistically significant positive correlations with human scores. Specifically, COMET-QE showed a moderate correlation (ρ = 0.429, *p* < 0.001), while BERTScore (ρ = 0.233, *p* = 0.037) and BLEURT (ρ = 0.262, *p* = 0.019) exhibited weaker but still significant associations. By contrast, traditional lexical metrics such as BLEU (ρ = 0.181, *p* = 0.108), CHRF (ρ = 0.112, *p* = 0.322), and METEOR (ρ = 0.196, *p* = 0.082) failed to reach statistical significance (*p* > 0.05), and showed only weak correlations in this domain-specific context. These results partially supported for Hypothesis 3 under the GNMT condition, as semantic metrics correlated more strongly and significantly with human evaluations than their lexical counterparts.

A similar analysis was conducted for ChatGPT-4 outputs. Significant differences in correlation were also observed between automatic metrics and human assessments (refer to **Figure 2** and **Appendix A** for details). Traditional lexical metrics, including METEOR (ρ = 0.368, *p* < 0.001), BLEU (ρ = 0.251, *p* = 0.025), and CHRF (ρ = 0.224, *p* = 0.046), exhibited weak yet statistically significant positive correlations with human assessments. In contrast, the performance of semantic metrics was unexpectedly poor. COMET-QE even exhibited a non-significant negative correlation with human assessments (ρ = −0.064, *p* = 0.570), while BLEURT and BERTScore similarly failed to yield significant results. Therefore, Hypothesis 3 was not supported in the case of ChatGPT-4, as semantic metrics did not demonstrate stronger or more consistent correlations with human assessments than lexical metrics.

4.3.2. Discussion

Building on the Spearman correlation results presented above, the following discussion examines the observed patterns in relation to prior literature and the underlying properties of each metric. For GNMT, these findings, consistent with Glushkova et al.[67], indicated that semantic metrics better reflect human evaluative preferences. This limited alignment for lexical metrics may be attributable to their reliance on surface-level n-gram overlap, which constrains their ability to accommodate valid lexical and structural variation—features especially salient in technical translation tasks.

These results provided support for Hypothesis 3 within the GNMT condition. While significant correlations were observed between some semantic metrics and human assessments, the correlation strength varied. Importantly, semantic metrics tended to outperform lexical ones. The finding echoes recent research, including WMT22 metrics shared task from Freitag et al.[68] and the meta-review by Lee et al.[69], which highlight the superior adaptability of neural evaluation models in handling paraphrasing and syntactic variation. While the observed correlations were not strong, semantic metrics showed a more consistent alignment with human preferences in evaluating technical manual translations, suggesting their relative suitability for GNMT tasks.

For ChatGPT-4, these findings contradicted the expected pattern posited in Hypothesis 3, which predicted stronger correlations for pre-trained semantic metrics. The results suggested that, in the context of technical manual translation, existing semantic evaluation models may not fully capture the quality dimensions emphasised by human annotators in evaluating ChatGPT-generated outputs.

One possible explanation lies in the training data of these metrics: models such as COMET and BLEURT are typically trained on reference translations characterised by structural uniformity and stylistic conservatism[18,21]. Compared with NMT, ChatGPT outputs tend to exhibit greater stylistic diversity and syntactic flexibility[3]. Moreover, the specialised nature of technical manuals demands high precision and consistency, which may not be adequately captured by semantic similarity models[20].

This interpretation aligns with recent findings. Mukherjee and Shrivastava[59] noted that many modern evaluation tools fail to accommodate the stylistic and structural characteristics of LLM outputs. Jiang et al.[50] also observed weak automatic-human alignment in a multi-genre evaluation of generative systems. However, Qian et al.[70], in the context of the WMT22 shared task on quality estimation, reported encouraging results. Pre-trained semantic metrics such as COMET and TransQuest showed high correlations with human assessments when applied to LLM-generated translations. This may be attributed to their use of reference-based COMET models and the focus on general-domain language, rather than technical or domain-specific content.

In summary, this section explored how automatic metrics align with human assessments in translating Chinese technical manuals. For GNMT, only semantic metrics showed moderate or weak correlations, indicating that existing tools remain reasonably calibrated for NMT outputs, while lexical metrics performed poorly. For ChatGPT-4, only some lexical metrics showed modest or weak correlations, whereas semantic metrics correlated poorly and non-significantly with human judgements. These findings provided partial support for Hypothesis 3 and highlight the need for domain-adaptive, LLM-based evaluation metrics.

5. Conclusions

This study conducted a quantitative comparison of Chinese-English translations produced by GNMT and ChatGPT-4, using a self-constructed corpus of 80 segments drawn from technical manuals. Six commonly adopted automatic metrics—BLEU, METEOR, CHRF, BERTScore, BLEURT, and COMET-QE—were applied as well as human assessments based on a five-point Likert scale evaluating accuracy, fluency, terminology, and style. The investigation addressed three key dimensions: system-level translation differences, inter-metric variation, and the degree of alignment between automatic and human assessments.

The results revealed significant differences between human and automatic evaluation. While human annotators consistently rated ChatGPT-4 significantly higher than GNMT, all six automatic metrics favoured GNMT. Statistically significant variation was also observed among the metrics themselves. Post hoc analysis showed that COMET-QE's comparisons with CHRF, METEOR, and BERTScore were non-significant across both systems, likely due to its broad score range after normalisation, which may have reduced statistical sensitivity. Correlation analyses yielded divergent trends. For GNMT, COMET-QE showed a moderate, statistically significant correlation with human scores ($\rho = 0.429$, $p < 0.001$), while other semantic metrics produced weaker associations. For ChatGPT-4, however, none of the semantic metrics correlated significantly with human assessments. Conversely, lexical metrics, especially METEOR ($\rho = 0.368$, $p < 0.001$), exhibited moderate or weak alignment with human assessments.

Divergent evaluations from human annotators and automatic metrics underscore a misalignment between computational outputs and human judgement. This highlights fundamental shortcomings in current evaluation metrics, particularly their limited capacity to capture the nuanced strengths of LLM-generated translations in specialised domains such as technical manuals. It may be beneficial for future evaluation approaches to incorporate contextual semantic representations and consider LLM-generated content in training data, particularly for specialised translation domains. Expanding evaluation frameworks to incorporate stylistic features and domain-specific terminology may further enhance their sensitivity and reliability in assessing LLM-generated translations. Collectively, these findings provide constructive insight for refining translation evaluation practices in line with ongoing advances in machine translation technologies.

6. Limitations

This study is subject to several limitations. First, the translation prompts employed for ChatGPT-4 were concise and basic; future research could explore more sophisticated prompt engineering techniques to potentially improve translation quality. Second, human assessment was conducted by only two professional annotators. While a standardised assessment rubric was employed to enhance scoring consistency, some degree of subjective variability remains inevitable. Third, the study relied on a relatively small corpus of technical manual texts. Future studies should consider expanding both the scale and domain range of the dataset to enhance the generalisability of the findings.

Author Contributions

Z.Z. was responsible for the main research design, data collection, and manuscript preparation. S.N.b.S.A. provided overall supervision and methodological guidance. M.A.R.A. contributed to quantitative analysis and critically reviewed and edited the manuscript. L.Z. assisted with data collection, data organization, and language polishing. All authors have read and approved the final version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

As the study did not involve human or animal subjects, no ethical approval was required. The manuscript adheres to ethical academic standards for theoretical and conceptual research.

Informed Consent Statement

Informed consent was obtained from all trained annotators who participated in the manual error analysis. Their involvement was entirely voluntary, and they were fully informed of the study's aims and procedures beforehand.

Data Availability Statement

Data is available upon request.

Acknowledgments

The authors would like to thank the administrative staff and colleagues at the Faculty of Modern Languages and Communication, Universiti Putra Malaysia, for their helpful support during the preparation of this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest related to the content of this article. The research was conducted independently, without any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

Table A1. Spearman Correlations between Automatic Metrics and Human Scores.

Variable		**BLEU (Google)**	**BLEU (GPT)**	**CHRF (Google)**	**CHRF (GPT)**	**METEOR (Google)**	**METEOR (GPT)**	**BERTScore (Google)**	**BERTScore (GPT)**	**BLEURT (Google)**	**BLEURT (GPT)**	**COMET (Google)**	**COMET (GPT)**	**Human (Google)**	**Human (GPT)**
BLEU (Google)	Spearman's rho	—													
	p-value	—													
BLEU (GPT)	Spearman's rho	0.700 ***	—												
	p-value	<0.001	—												
CHRF (Google)	Spearman's rho	0.932 ***	0.707 ***	—											
	p-value	<0.001	<0.001	—											
CHRF (GPT)	Spearman's rho	0.624 ***	0.911 ***	0.696 ***	—										
	p-value	<0.001	<0.001	<0.001	—										
METEOR (Google)	Spearman's rho	0.906 ***	0.655 ***	0.907 ***	0.614 ***	—									
	p-value	<0.001	<0.001	<0.001	<0.001	—									
METEOR (GPT)	Spearman's rho	0.526 ***	0.853 ***	0.561 ***	0.902 ***	0.562 ***	—								
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001	—								
BERTScore (Google)	Spearman's rho	0.796 ***	0.552 ***	0.782 ***	0.527 ***	0.771 ***	0.463 ***	—							
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	—							
BERTScore (GPT)	Spearman's rho	0.601 ***	0.782 ***	0.626 ***	0.789 ***	0.608 ***	0.780 ***	0.715 ***	—						
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	—						
BLEURT (Google)	Spearman's rho	0.708 ***	0.419 ***	0.676 ***	0.358 **	0.674 ***	0.319 **	0.810 ***	0.476 ***	—					
	p-value	<0.001	<0.001	<0.001	0.001	<0.001	0.004	<0.001	<0.001	—					
BLEURT (GPT)	Spearman's rho	0.565 ***	0.669 ***	0.544 ***	0.608 ***	0.532 ***	0.613 ***	0.602 ***	0.739 ***	0.633 ***	—				
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	—				
COMET (Google)	Spearman's rho	−0.075	−0.061	−0.094	−0.047	0.014	0.042	0.045	0.143	0.020	0.119	—			
	p-value	0.509	0.591	0.409	0.679	0.899	0.712	0.695	0.207	0.858	0.295	—			
COMET (GPT)	Spearman's rho	−0.158	−0.095	−0.146	−0.064	−0.060	0.027	−0.039	0.070	−0.061	0.082	0.903 ***	—		
	p-value	0.162	0.402	0.195	0.573	0.599	0.812	0.733	0.535	0.593	0.471	<0.001	—		
Human (Google)	Spearman's rho	0.181	0.216	0.112	0.209	0.196	0.318 **	**0.233** *	0.342 **	**0.262** *	0.326 **	**0.429** ***	0.278 *	—	
	p-value	0.108	0.054	0.322	0.063	0.082	0.004	0.037	0.002	0.019	0.003	<0.001	0.012	—	
Human (GPT)	Spearman's rho	0.163	**0.251** *	0.098	**0.224** *	0.179	**0.368** ***	0.094	0.205	0.074	0.173	0.057	-0.064	0.622 ***	—
	p-value	0.149	0.025	0.386	0.046	0.112	<0.001	0.409	0.068	0.514	0.125	0.613	0.570	<0.001	—

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

References

[1] Ahammad, S.H., Kalangi, R.R., Nagendram, S., et al., 2024. Improved neural machine translation using Natural Language Processing (NLP). Multimedia Tools and Applications. 83(13), 39335–39348. DOI: https://doi.org/10.1007/s11042-023-17207-7

[2] Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. Preprint. arXiv:1409.0473.

[3] Lyu, C., Du, Z., Xu, J., et al., 2023. A paradigm shift: The future of machine translation lies with large language models. Preprint. arXiv:2305.01181. DOI: https://doi.org/10.48550/arXiv.2305.01181

[4] Chan, V., Tang, W.K.W., 2024. GPT for Translation: A systematic literature review. SN Computer Science. 5(8), 1–9. DOI: https://doi.org/10.1007/s42979-024-03340-z

[5] Zhu, S., Xu, S., Sun, H., et al., 2024. Multilingual large language models: A systematic survey. Preprint. arXiv:2411.11072. DOI: https://doi.org/10.48550/arXiv.2411.11072

[6] Ouyang, L., Wu, J., Jiang, X., et al., 2022. Training language models to follow instructions with human feedback. Preprint. arXiv:2203.02155. DOI: https://doi.org/10.48550/arXiv.2203.02155

[7] Kocmi, T., Federmann, C., Grundkiewicz, R., et al., 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. Preprint. arXiv:2107.10821. DOI: https://doi.org/10.48550/arXiv.2107.10821

[8] Jiao, W., Wang, W., Huang, J.T., et al., 2023. Is ChatGPT a good translator? A preliminary study. Preprint. arXiv:2301.08745.

[9] Obeidat, M.M., Haider, A.S., Tair, S.A., et al., 2024. Analyzing the performance of Gemini, ChatGPT, and Google Translate in rendering English idioms into Arabic. FWU Journal of Social Sciences. 18(4).

[10] Al-Maaytah, M., Almahasees, Z., 2024. A linguistic investigation for a case study of ChatGPT and Google Translate in rendering special needs texts from English into Arabic: A synchronic case study. Pakistan Journal of Life & Social Sciences. 22(2).

[11] Byrne, J., 2006. Technical Translation: Usability Strategies for Translating Technical Documentation. Springer: Dordrecht, Netherlands.

[12] Olohan, M., 2015. Scientific and Technical Translation. Routledge: Abingdon, UK.

[13] 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), 63–87.

[14] 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. Forum for Linguistic Studies. 6(3), 837–849. DOI: https://doi.org/10.30564/fls.v6i3.6799

[15] 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. DOI: https://doi.org/10.3390/info14100574

[16] Papineni, K., Roukos, S., Ward, T., et al., 2002. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.

[17] Banerjee, S., Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.

[18] Rei, R., Stewart, C., Farinha, A.C., et al., 2020. COMET: A neural framework for MT evaluation. Preprint. arXiv:2009.09025. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.213

[19] Chatzikoumi, E., 2020. How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering. 26(2), 137–161. DOI: https://doi.org/10.1017/S1351324919000469

[20] Zhang, T., Kishore, V., Wu, F., et al., 2019. BERTScore: Evaluating text generation with BERT. Preprint. arXiv:1904.09675.

[21] Sellam, T., Das, D., Parikh, A.P., 2020. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7881–7892. DOI: https://doi.org/10.18653/v1/2020.acl-main.704

[22] Rei, R., Farinha, A.C., Zerva, C., et al., 2021. Are references really needed? UNBABEL-IST 2021 submission for the metrics shared task. In Proceedings of the Sixth Conference on Machine Translation, Online, 10–11 November 2021; pp. 1030–1040.

[23] Zuo, Y., Abdullah, S.S., Toh, F.H.C., 2023. Strategies for translating culture-specific items from Chinese into English. World Journal of English Language. 13(7), 27–38. DOI: https://doi.org/10.5430/wjel.v13n7p27

[24] Zahrawi, R.M.T., Abdullah, S.N.S., Mustapha, N.F., et al., 2024. Strategies for translating Arabic similes in Al-Manfaluti's Al-Abrat into English. International Journal of Academic Research in Progressive Education and Development. 13(1), 223–239. DOI: https://doi.org/10.6007/IJARPED/v13-i1/20002

[25] Kaji, H., 1999. Controlled languages for machine translation: State of the art. In Proceedings of Machine Translation Summit VII, Singapore, 13–17 September 1999; pp. 37–39.

[26] Wright, S.E., 2011. Scientific, technical, and medical translation. In: Malmkjær, K., Windle, K. (eds.). The

Oxford Handbook of Translation Studies (online ed.). Oxford University Press: Oxford, UK. DOI: https://doi.org/10.1093/oxfordhb/9780199239306.013.0018

[27] Suima, I., 2024. Scientific and technical texts: Translation aspects in electrical and computer engineering. Challenges and Issues of Modern Science. 3, 74–82.

[28] Axunbabayeva, N., Yunusova, N., 2020. The importance of consistent terminology in technical translation. The Scientific Heritage. (49-3), 31–33.

[29] 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.

[30] Sadiq, S., 2025. Evaluating English–Arabic translation: Human translators vs. Google Translate and ChatGPT. Journal of Languages and Translation. 12(1), 67–95. DOI: https://doi.org/10.21608/jltmin.2025.423147

[31] 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. DOI: https://doi.org/10.23971/jobit.v1i2.297

[32] Callison-Burch, C., Fordyce, C.S., Koehn, P., et al., 2007. (Meta-) evaluation of machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007; pp. 136–158. DOI: https://doi.org/10.3115/1626355.1626373

[33] Graham, Y., Baldwin, T., Moffat, A., et al., 2013. Continuous measurement scales in human evaluation of machine translation. In Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, Sofia, Bulgaria, 8–9 August 2013; pp. 33–41.

[34] Birch, A., Abend, O., Bojar, O., et al., 2016. HUME: Human UCCA-based evaluation of machine translation. Preprint. arXiv:1607.00030. DOI: https://doi.org/10.18653/v1/D16-1134

[35] Lommel, A., Popovic, M., Burchardt, A., 2014. Assessing inter-annotator agreement for translation error annotation. In MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation, Reykjavik, Iceland, 26–31 May 2014; pp. 31–37.

[36] Popović, M., 2018. Error classification and analysis for machine translation quality assessment. In: Translation Quality Assessment: From Principles to Practice. Springer: Cham, Switzerland. pp. 129–158. DOI: https://doi.org/10.1007/978-3-319-91241-7_7

[37] Reiter, E., Belz, A., 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Computational Linguistics. 35(4), 529–558. DOI: https://doi.org/10.1162/coli.2009.35.4.35405

[38] Popović, M., 2015. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 11–12 September 2015; pp. 392–395. DOI: https://doi.org/10.18653/v1/W15-3049

[39] Ghosh, S., Ghose, A., Chattopadhya, R., et al., 2024. A study on evaluation techniques for machine translation. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 3–5 July 2024; pp. 1–7. DOI: https://doi.org/10.1109/ICCCNT61001.2024.10723917

[40] Deutsch, D., Dror, R., Roth, D., 2022. On the limitations of reference-free evaluations of generated text. Preprint. arXiv:2210.12563. DOI: https://doi.org/10.18653/v1/2022.emnlp-main.753

[41] Ulitkin, I., Filippova, I., Ivanova, N., et al., 2021. Automatic evaluation of the quality of machine translation of a scientific text: the results of a five-year-long experiment. E3S Web of Conferences 284. 08001. DOI: https://doi.org/10.1051/e3sconf/202128408001

[42] Ding, L., 2024. A comparative study on the quality of English-Chinese translation of legal texts between ChatGPT and neural machine translation systems. Theory and Practice in Language Studies. 14(9), 2823–2833. DOI: https://doi.org/10.17507/tpls.1409.18

[43] 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?. MonTI. 16. DOI: https://doi.org/10.6035/MonTI.2024.16.02

[44] Sanz-Valdivieso, L., López-Arroyo, B., 2023. Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation?. In Proceedings of the International Conference Human-informed Translation and Interpreting Technology (HiT-IT 2023); pp. 97–107. DOI: https://doi.org/10.26615/issn.2683-0078.2023_008

[45] AlAfnan, M.A., 2025. Large language models as computational linguistics tools: A comparative analysis of ChatGPT and Google machine translations. Journal of Artificial Intelligence and Technology. 5, 20–32. DOI: https://doi.org/10.37965/jait.2024.0549

[46] Mohsen, M., 2024. Artificial intelligence in academic translation: A comparative study of large language models and Google Translate. Psycholinguistics. 35(2), 134–156. DOI: https://doi.org/10.31470/2309-1797-2024-35-2-134-156

[47] Brewster, R.C., Gonzalez, P., Khazanchi, R., et al., 2024. Performance of ChatGPT and Google Translate for pediatric discharge instruction translation. Pediatrics. 154(1), e2023065573. DOI: https://doi.org/10.1542/peds.2023-065573

[48] Sizov, F., España-Bonet, C., van Genabith, J., et al., 2024. Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations. In Proceedings of the Ninth Conference on Machine Translation; pp. 1183–1199. DOI: https://doi.org/10.18653/v1/2024.wmt-1.116

[49] 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. DOI: https://doi.org/10.3726/JTS012024.02

[50] Jiang, Z., Lv, Q., Zhang, Z., et al., 2024. Convergences and divergences between automatic assessment and human evaluation: Insights from comparing ChatGPT-generated translation and neural machine translation. Preprint. arXiv:2401.05176. DOI: https://doi.org/10.48550/arXiv.2401.05176

[51] Emery, D., Goitia, M., Vargus, F., et al., 2025. HalluMix: A task-agnostic, multi-domain benchmark for real-world hallucination detection. Preprint. arXiv:2505.00506.

[52] Bowker, L., Ciro, J.B., 2019. Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community. Emerald Publishing Limited: Leeds, England. DOI: https://doi.org/10.1108/9781787567214

[53] Zouhar, V., Bojar, O., 2024. Quality and quantity of machine translation references for automatic metrics. Preprint. arXiv:2401.01283. DOI: https://doi.org/10.48550/arXiv.2401.01283

[54] Pourkamali, N., Sharifi, S.E., 2024. Machine translation with large language models: Prompt engineering for Persian, English, and Russian directions. Preprint. arXiv:2401.08429.

[55] Post, M., 2018. A call for clarity in reporting BLEU scores. Preprint. arXiv:1804.08771. DOI: https://doi.org/10.18653/v1/W18-6319

[56] Castilho, S., O'Brien, S., 2017. Acceptability of machine-translated content: A multi-language evaluation by translators and end-users. Linguistica Antverpiensia, New Series – Themes in Translation Studies. 16. DOI: https://doi.org/10.52034/lanstts.v16i0.430

[57] Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann: Burlington, MA, USA.

[58] Cohen, J., 1988. The effect size. In: Statistical Power Analysis for the Behavioral Sciences. Routledge: Abingdon, UK. pp. 77–83.

[59] Mukherjee, A., Shrivastava, M., 2025. Lost in translation? Found in evaluation: A comprehensive survey on sentence-level translation evaluation. ACM Computing Surveys. DOI: https://doi.org/10.1145/3735970

[60] Glushkova, T., Zerva, C., Rei, R., et al., 2021. Uncertainty-aware machine translation evaluation. Preprint. arXiv:2109.06352. DOI: https://doi.org/10.18653/v1/2021.findings-emnlp.330

[61] Yan, Y., Wang, T., Zhao, C., et al., 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. Preprint. arXiv:2307.03131. DOI: https://doi.org/10.18653/v1/2023.acl-long.297

[62] Balashov, Y., 2025. Translation in the wild. Preprint. arXiv:2505.23548. DOI: https://doi.org/10.48550/arXiv.2505.23548

[63] Wang, L., Lyu, C., Ji, T., et al., 2023. Document-level machine translation with large language models. Preprint. arXiv:2304.02210. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.1036

[64] Marie, B., 2022. An automatic evaluation of the WMT22 general machine translation task. Preprint. arXiv:2209.14172. DOI: https://doi.org/10.48550/arXiv.2209.14172

[65] He, T., Zhang, J., Wang, T., et al., 2022. On the blind spots of model-based evaluation metrics for text generation. Preprint. arXiv:2212.10020. DOI: https://doi.org/10.18653/v1/2023.acl-long.674

[66] Sai, A.B., Mohankumar, A.K., Khapra, M.M., 2022. A survey of evaluation metrics used for NLG systems. ACM Computing Surveys. 55(2), 1–39. DOI: https://doi.org/10.1145/3485766

[67] Glushkova, T., Zerva, C., Martins, A.F.T., 2023. BLEU meets COMET: Combining lexical and neural metrics towards robust machine translation evaluation. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation; pp. 47–58.

[68] Freitag, M., Rei, R., Mathur, N., et al., 2022. Results of WMT22 metrics shared task: Stop using BLEU—neural metrics are better and more robust. In Proceedings of the Seventh Conference on Machine Translation (WMT), Abu Dhabi, United Arab Emirates, 7–8 December 2022; pp. 46–68.

[69] Lee, S., Lee, J., Moon, H., et al., 2023. A survey on evaluation metrics for machine translation. Mathematics. 11(4), 1006. DOI: https://doi.org/10.3390/math11041006

[70] Qian, S., Sindhujan, A., Kabra, M., et al., 2024. What do large language models need for machine translation evaluation? Preprint. arXiv:2410.03278. DOI: https://doi.org/10.18653/v1/2024.emnlp-main.214