Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9087

Comparing Chinese–English Translation Performance of Large Language Models: A Multidimensional Evaluation of Financial, Technological, and Political Texts

Yaowei Xu[1.a], Zhihan Fu[2.b]
[1]Foreign Language Experimental Teaching Center, Zhejiang Yuexiu University, Shaoxing 312000, China
[2]School of English Studies, Zhejiang Yuexiu University, Shaoxing 312000, China
Email: [a]20231027@zyufl.edu.cn. [b]fzh_0813@qq.com

How to cite this paper: Xu, Y., & Fu, Z. (2026). Comparing Chinese–English translation performance of large language models: A multidimensional evaluation of financial, technological, and political texts. Journal of Computer Science and Frontier Technologies, 3(3), 61–74. ISSN Print: 3104-4204, ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9087
Published: 2026-06-26

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Abstract

Large language models (LLMs) are increasingly used for machine translation; however, their reliability in professional Chinese–English translation, particularly for domain-specific texts, remains insufficiently examined. This challenge is especially evident in texts containing specialized terminology, numerical information, complex logical relations, and formal institutional phrasing. To address this issue, this study conducts a controlled comparison of six LLMs: ChatGPT-5.0, Claude Sonnet 4.5, DeepSeek-R1, GLM-4.6, Qwen3-MAX, and Grok 4.0. Three representative Chinese texts from the financial, technological, and political domains were selected and translated under three prompt conditions: zero-shot prompting, domain-enhanced prompting, and glossary-based prompting.

Translation quality was assessed through a multidimensional evaluation framework. Standard reference-based metrics, including BLEU, ChrF++, TER, ROUGE, BERTScore, and COMET, were used to measure automatic translation quality. In addition, lexical richness and syntactic-discourse indicators were incorporated to examine stylistic and structural differences across model outputs.

The results reveal a clear domain effect, indicating that financial, technological, and political texts present different levels and types of translation difficulty for LLMs. By contrast, prompt strategy did not produce significant main effects under the three tested conditions. At the model level, GLM-4.6 achieved the highest COMET-based ranking across all prompt settings, while DeepSeek-R1 showed comparatively stable second-place performance.

Overall, the findings suggest that, within this controlled exploratory dataset, model type and text domain exert a stronger influence on translation performance than prompt modification alone. Rather than offering a definitive ranking of LLM translation capability, this study provides empirical evidence on how selected LLMs handle specialized Chinese–English translation under comparable experimental conditions. Future research should expand the dataset, incorporate

human evaluation, and further investigate the role of terminology constraints in improving translation accuracy and stylistic appropriateness.

Keywords

Large Language Models; Professional Text Translation; Machine Translation Quality Evaluation; Multidimensional Evaluation; Prompt Strategy

1. Introduction

Machine translation has become an important tool in cross-lingual communication, especially in fields where large amounts of information need to be translated quickly. Earlier machine translation systems were mainly based on statistical or neural architectures, while recent translation practice has increasingly involved large language models (LLMs). Compared with traditional systems, LLMs are often better at producing fluent and contextually coherent texts. However, fluency alone does not guarantee reliable translation in professional domains, where terminology, numerical information, logical relations, and stylistic conventions may directly affect the quality of the translated text.

Professional Chinese-English translation is particularly challenging because different domains require different translation priorities. Financial texts often require accurate rendering of company names, figures, policy terms, and causal relations. Technological texts depend heavily on standardized terminology and clear procedural logic. Political texts, by contrast, are more sensitive to tone, institutional expressions, and pragmatic appropriateness. These differences suggest that LLM translation quality should not be evaluated only through general fluency or semantic similarity. Instead, domain-specific performance needs to be examined through a more detailed evaluation framework.

Previous studies have discussed the application of LLMs in financial, scientific, and political translation, but many of them focus on a single domain, a single model, or a limited set of evaluation metrics. For example, research on financial translation has shown that LLMs may still produce errors in terminology and logical coherence, while studies of scientific and political translation have emphasized problems related to technical terms, stylistic norms, and discourse appropriateness. These findings provide useful evidence, but they do not fully explain how different models behave across several professional domains under the same experimental conditions.

Based on this gap, the present study compares the Chinese-English translation performance of six LLMs: ChatGPT-5.0, Claude Sonnet 4.5, DeepSeek-R1, GLM-4.6, Qwen3-MAX, and Grok 4.0. The study uses three representative texts from the financial, technological, and political domains and tests three prompt conditions: zero-shot prompting, domain-enhanced prompting, and glossary-based prompting. Translation quality is evaluated through automatic metrics, including BLEU, ChrF++,

TER, ROUGE, BERTScore, and COMET, together with lexical richness and syntactic-discourse indicators. The purpose is to examine whether model type, text domain, and prompt strategy lead to measurable differences in LLM-based professional translation.

2. Literature Review

2.1. The Evolution of Machine Translation Technologies and Changes in Domain-Specific Translation Capabilities

The evolution of machine translation has long been shaped by two related concerns: semantic representation and domain adaptation. From statistical machine translation to neural machine translation, and then to generative approaches represented by LLMs, each stage has changed how translation systems process meaning, terminology, and contextual information. Gu (2024) argues that statistical machine translation mainly depends on phrase-based probabilistic matching and local alignment, which makes it less effective in identifying terminological structures in specialized texts. This limitation is especially visible in professional translation, where meaning is often distributed across terminology, institutional expressions, and logical relations rather than contained in isolated phrases.

Neural machine translation changed this situation by introducing end-to-end modeling and Transformer-based architectures. These models greatly improved translation fluency and contextual coherence. Yet fluency alone does not solve the core problems of professional translation. Zheng (2025) observes that when explicit domain constraints are absent, NMT systems still show limited stability in translating texts with dense terminology and complex logical relations. In other words, the improvement from SMT to NMT does not remove the need for domain-sensitive evaluation. It only shifts the problem from basic fluency to terminological consistency, logical precision, and professional acceptability.

Representing a further paradigm shift, LLMs leverage large-scale pre-training to resolve long-range contextual relations and reorganize complex logical structures. Unlike earlier systems, these models not only generate highly fluent outputs but also dynamically adapt to specific task instructions and terminology constraints. However,Zhao et al. (2024) emphasize that the reliability and consistency of LLMs in professional translation still require systematic testing. This point is important because the advantages of LLMs are not equally stable across all domains. A model may reduce certain errors found in traditional NMT systems, yet still perform unevenly when financial, technological, and political texts impose different translation demands.

2.2. Domain-Specific Challenges in Specialized Text Translation and the Applicability of Large Language Models

Research on domain-specific translation shows that professional translation cannot

be treated as a homogeneous task. Fu et al. (2024) focus on financial-domain translation and point out that terminology, company names, and logical coherence remain difficult for LLMs. Their discussion suggests that financial translation requires more than readable output, because errors in figures, company names, or causal relations may directly change the meaning of the text. Gu and Leng (2024), by contrast, examine scientific and technological translation through mechanical engineering terminology. They show that term recognition and standardized expression are still unstable, even when the model output appears fluent. Wen and Tian (2024) shift the focus to political discourse, where lexical choice, tone, and institutionalized expressions become more important. Placed together, these three lines of research suggest that financial, technological, and political translation involve different risk points and should not be evaluated by a single general standard.

The distinction among these domains is important because professional translation quality is not determined by general fluency alone. Financial texts often require accurate rendering of company names, figures, policy terms, and causal relations. Technological texts depend heavily on standardized terminology and clear procedural logic. Political texts, by contrast, are more sensitive to tone, institutional expressions, and pragmatic appropriateness. A fluent translation may still fail if it distorts a financial relation, weakens a technical procedure, or uses an unsuitable political tone. For this reason, LLM translation quality should be evaluated through domain-sensitive criteria rather than through surface readability.

The above research provides useful evidence, but its designs remain relatively limited. Many existing discussions examine one domain, one model, or one type of translation problem. Such designs can reveal local translation difficulties, yet they cannot fully explain how different LLMs behave across several professional domains under the same experimental conditions. This limitation makes a unified comparative design necessary. Only by comparing multiple models and multiple domains within the same evaluation framework can the influence of model type, text domain, and translation condition be more clearly identified.

2.3. Linguistic Features and Evaluation Methods: From Translation Quality to Translation Style

Translation evaluation has increasingly moved from single-score judgment toward a combined framework that includes automatic metrics and linguistic feature analysis. Reference-based metrics, including BLEU, ChrF++, TER, ROUGE, and BERTScore, measure translation quality from different angles, such as surface similarity, editing distance, lexical overlap, and semantic correspondence. These metrics are useful because they provide quantifiable evidence for comparing machine translation outputs. Their limitation, however, is also clear. They can show how close a translation is to a reference, but they do not fully explain how the translated text is

organized linguistically.

This problem becomes more obvious in LLM-generated translation. Two translations may receive similar automatic scores but differ greatly in lexical richness, syntactic structure, cohesion, and discourse organization. A translation with high lexical overlap may still sound unnatural or repetitive. Another translation may depart from the reference wording but better follow the stylistic norms of professional English. Therefore, automatic evaluation should be supplemented by linguistic feature analysis, especially when the research focus is not only translation accuracy but also translation style.

Rei et al. (2020) show that neural supervised evaluation methods, such as COMET, are trained on human assessment data and display strong consistency in detecting errors related to key information. This makes COMET useful for evaluating semantic adequacy and information transfer. At the same time, linguistic-feature-based research gives another perspective on LLM translation. Kwok et al. (2025) and Yu (2024) suggest that LLMs often perform better than traditional NMT systems in lexical density and coordination, but they do not always surpass human translations in the use of subordinate structures and discourse-level organization. These findings indicate that translation evaluation should not stop at whether the meaning is transferred. It should also examine how the target text constructs meaning through vocabulary choice, sentence structure, and discourse connection.

Although these studies provide useful evidence for understanding the linguistic style of LLM translations, most of them focus on general or literary texts and seldom adopt a systematic design involving multiple models, domains, and prompt strategies. This leaves room for a multidimensional evaluation framework that examines translation quality and linguistic characteristics at the same time.

2.4. Research Gap

Despite the development of machine translation and the growing interest in professional-domain translation, three gaps remain. First, most existing studies examine either a single domain or a single model. Few provide a unified comparative framework covering financial, technological, and political texts. As a result, it remains difficult to determine whether differences in translation quality are caused mainly by the model, the domain, or the interaction between them.

Secondly, the effectiveness and stability of LLMs under different prompt strategies in professional translation contexts have not been fully verified. Since LLMs respond sensitively to instructions, prompt design should not be treated as a neutral technical detail. A zero-shot prompt, a domain-enhanced prompt, and a glossary-based prompt may guide the model in different ways. Without controlled comparison, the role of prompt strategy in professional translation remains unclear.

Finally, limited attention has been paid to combining automatic evaluation with lexical and syntactic-discourse analysis. Automatic metrics can measure reference

similarity and semantic adequacy, but they cannot fully explain lexical variation, syntactic complexity, cohesion, or discourse organization. For professional translation, these features are directly related to readability, precision, and stylistic appropriateness.

Based on these gaps, the present study compares the Chinese-English translation performance of six LLMs across financial, technological, and political texts. It also examines three prompt conditions and combines automatic translation metrics with lexical and syntactic-discourse indicators. The purpose is to identify whether model type, text domain, and prompt strategy lead to measurable differences in LLM-based professional translation.

3. Research Design

3.1. Research Questions

This study evaluates the Chinese-English translation performance of different LLMs in professional text translation. The evaluation focuses on three major factors: model type, text domain, and prompt strategy. Rather than judging translation quality only through general fluency, the study combines automatic translation metrics with lexical and syntactic-discourse indicators. This design makes it possible to compare the outputs produced by different models under controlled experimental conditions.

Specifically, this study addresses the following research questions:

RQ1: How do different LLMs perform in Chinese-English professional text translation when evaluated by automatic translation quality metrics?

RQ2: Does text domain significantly affect LLM translation quality across financial, technological, and political texts?

RQ3: Does prompt strategy significantly influence LLM translation quality under zero-shot, domain-enhanced, and glossary-based prompting conditions?

RQ4: Do different LLMs show measurable differences in lexical richness, syntactic features, and discourse-level characteristics?

By integrating multi-model comparison, multi-domain analysis, prompt-condition testing, and multidimensional evaluation metrics, this study contributes to a more systematic understanding of LLM performance in professional Chinese–English translation. The findings are expected to provide methodological insights for machine translation evaluation and practical implications for the use of LLMs in domain-specific translation tasks.

3.2. Source Text Collection and Preprocessing

This study focuses on the Chinese-English translation performance of different LLMs in finance, technology, and politics. It also examines how prompt conditions affect translation quality. To support this design, three Chinese source texts were collected from three professional domains: finance, technology, and politics. One

representative text was selected from each domain so that the experimental materials could cover different types of professional language.The three selected source texts were controlled at a comparable scale, with approximately 800–1,100 Chinese characters and an average of about eight paragraphs per text. This design helps reduce the possible influence of text length and text structure on automatic evaluation results, thereby improving the reliability of cross-domain comparison.

The financial text contains business and policy-related expressions, numerical information, and financial terminology. The technological text contains technical terms, procedural descriptions, and system-related information. The political text contains institutional expressions, formal discourse, and policy-related language. These three types of texts were selected because they represent different professional translation requirements and create different translation difficulties.

The collected texts were cleaned before translation and evaluation. The preprocessing mainly included removing irrelevant formatting marks, unifying paragraph structure, checking sentence boundaries, and ensuring that each source text could be translated as a complete text. This step was necessary because inconsistent formatting, incomplete sentence boundaries, or unnecessary symbols could affect the calculation of automatic evaluation metrics.

Each Chinese source text was paired with a human English reference translation. The reference translation was used as the comparison standard for automatic evaluation. The LLM-generated translations were compared with the corresponding human reference translations through BLEU, ChrF++, TER, ROUGE, BERTScore, and COMET.

It should be noted that this study does not construct an annotated corpus in the strict linguistic sense. The collected texts are used as controlled experimental materials for LLM translation evaluation. Therefore, this section uses the term “source text collection and preprocessing” rather than “corpus construction.” Since the paragraph-level scoring part has been removed from the present study, the evaluation is conducted mainly at the full-text level rather than through paragraph-level grading.

3.3. Translation Systems and Testing Time

This study selected representative LLMs from domestic and international AI platforms to ensure comprehensive evaluation results. For domestic models, DeepSeek-R1, GLM-4.6, and Qwen3-MAX were selected. For international models, ChatGPT-5.0, Grok 4.0, and Claude Sonnet 4.5 were tested to compare the overall translation performance of different models.

Because LLMs are frequently updated, the findings of this study should be understood within a specific time frame. The model outputs used in this study were generated around November 2025. Therefore, the results reflect the translation performance of the selected model versions during that period. Later updates of

these models may lead to changes in translation quality, terminology handling, prompt responsiveness, and stylistic features.

All models were tested using the same Chinese source texts and the same prompt conditions. No manual post-editing was applied to the model outputs, so the evaluation results reflect the original translation performance of each model under the given experimental settings.

3.4. Prompts

This study investigates the impact of prompts on the professional text translation performance of LLMs. The focus is placed on translation quality and stylistic differences under different prompt conditions across different fields. The prompt design is categorized into three types: zero-shot baseline prompts, domain-enhanced prompts, and glossary-based prompts. These three types correspond to unguided, weakly guided, and explicitly constrained translation scenarios.

First, zero-shot baseline prompts only include basic translation instructions. They are used to assess the model's original translation ability without additional information. A typical prompt is: "Please translate the following text into English: [source text]."

Second, domain-enhanced prompts are based on zero-shot prompts, but role descriptions, domain contexts, and quality requirements are added. For example, the model is instructed to act as a professional translator in a specific field and to produce a translation that conforms to professional norms.

Third, glossary-based prompts are based on domain-enhanced prompts, but core terms are explicitly provided to directly constrain translation behavior. This study limits each text to five terms to balance constraint intensity and linguistic naturalness.

By comparing these three prompt conditions, the study examines whether prompt strategy produces measurable changes in LLM translation quality and linguistic features.

3.5. Automatic Evaluation Metrics

This study establishes an automatic evaluation framework that combines reference-aligned quality indicators and linguistic feature indicators. The quality indicators include BLEU, ChrF++, TER, BERTScore, ROUGE, and COMET. Among them, BLEU focuses on phrase-level overlap; ChrF++ is sensitive to word order and collocation; TER measures the minimum editing operations; BERTScore evaluates semantic similarity; ROUGE evaluates lexical overlap and recall-oriented similarity; and COMET exhibits high consistency with human evaluation (Rei et al., 2020).

In addition, lexical richness indicators, including TTR, MSTTR, MATTR, Herdan's C, Root TTR, and CTTR, and syntactic-discourse indicators, including Syntactic

Simplicity, Word Concreteness, Referential Cohesion, and Deep Cohesion, were introduced to describe the linguistic features of the translations.

The combination of automatic metrics and linguistic feature indicators allows this study to evaluate translation quality from multiple perspectives. Automatic metrics provide quantitative evidence for translation adequacy and similarity to the reference translation, while linguistic indicators help explain the stylistic and structural characteristics of different model outputs.

3.6. Experimental Procedure

The experiment commenced with the selection and formatting of three representative source texts, which were immediately paired with human English references. Subsequently, the six selected LLMs processed these texts under the varying prompt conditions, yielding a total corpus of 54 unedited English translations. These raw outputs were then subjected to multidimensional evaluation: automatic metrics (e.g., BLEU, COMET) benchmarked reference similarity, while parallel computational analysis extracted key lexical and syntactic-discourse features. Finally, ANOVA was employed to isolate the specific effects of model type, prompt strategy, and text domain on the translation outputs.

4. Data Analysis

4.1. Machine Evaluation Rankings

4.1.1. COMET-Based Machine Evaluation Ranking Method

COMET is used as the main criterion for machine evaluation ranking because of its strong alignment with human judgments. Based on multilingual pre-trained encoders, it jointly models the source text, machine translation output, and reference translation, and is trained under the supervision of human quality assessments (Rei et al., 2020). Therefore, it shows considerable stability in evaluating semantic completeness and logical coherence.

Table 1. Internal rankings under zero-shot prompting.

Rank	Model	COMET
1	GLM-4.6	0.4067
2	DeepSeek-R1	0.4000
3	ChatGPT-5.0	0.3867
3	Qwen3-MAX	0.3867
5	Claude Sonnet 4.5	0.3833
5	Grok 4.0	0.3833

Table 2. Internal rankings under domain-enhanced prompting.

Rank	Model	COMET
1	GLM-4.6	0.4100
2	DeepSeek-R1	0.3867
2	Claude Sonnet 4.5	0.3867
4	ChatGPT-5.0	0.3833

4	Grok 4.0	0.3833
6	Qwen3-MAX	0.3633

Table 3. Internal rankings under glossary-based prompting.

Rank	Model	COMET
1	GLM-4.6	0.4067
2	DeepSeek-R1	0.3967
3	ChatGPT-5.0	0.3933
4	Grok 4.0	0.3900
5	Claude Sonnet 4.5	0.3833
6	Qwen3-MAX	0.3600

4.1.2. Ranking and Analysis of 18 LLM Translated Text Patterns Based on COMET

The COMET scores of the six LLMs under the three prompt conditions were calculated respectively. The results reveal that the COMET scores of the different models are distinctly stratified, and the relative ranking of the models changes minimally across the three prompt strategies, indicating that the cross-domain translation quality differences between models are relatively stable.Specifically, GLM-4.6 consistently ranks first under all three prompt conditions; DeepSeek-R1 ranks second with marginally lower scores; ChatGPT-5.0 and Claude-Sonnet 4.5 fall within the middle range; and Qwen3-MAX and Grok-4.0 rank lower overall. The prompt strategy has a limited impact on the overall ranking structure, and the system-level COMET ranking is predominantly determined by the model itself.

4.2. Significance Analysis of Six Automatic Evaluation Metrics (Full-Text Level)

The results in Table 4 show that domain has a significant effect on all six automatic evaluation metrics. The strongest effect appears in TER, $F = 127.84$, $p = 1.36 \times 10^{-20}$, followed by ROUGE, $F = 52.86$, $p = 3.69 \times 10^{-13}$. This means that the differences among financial, technological, and political texts are not only reflected in semantic similarity, but also in surface matching and editing distance. By comparison, the prompt factor remains non-significant across all six metrics in Table 6. For example, the p-values for COMET, BERTScore F1, and ChrF++ are 0.812, 0.937, and 0.950, respectively. In this dataset, simple prompt modification therefore does not substantially change the full-text translation scores.

Table 4. Main effects of domain on six automatic evaluation metrics.

Metric	df (Domain)	F-value	p-value
BLEU	2	23.81	4.98×10^{-8}
ChrF++	2	7.46	1.44×10^{-3}
TER	2	127.84	1.36×10^{-20}
ROUGE	2	52.86	3.69×10^{-13}
BERTScore F1	2	6.60	2.83×10^{-3}
COMET	2	12.80	3.13×10^{-5}

Table 5. Main effects of model on six automatic evaluation metrics.

Metric	df (Model)	F-value	p-value
BLEU	5	1.96	0.1020
ChrF++	5	2.51	0.0427
TER	5	0.20	0.9620
ROUGE	5	1.11	0.3690
BERTScore F1	5	2.79	0.0273
COMET	5	2.46	0.0459

Table 6. Main effects of prompt strategy on six automatic evaluation metrics.

Metric	df (Prompt)	F-value	p-value
BLEU	2	0.38	0.688
ChrF++	2	0.05	0.950
TER	2	0.07	0.929
ROUGE	2	0.40	0.669
BERTScore F1	2	0.07	0.937
COMET	2	0.21	0.812

4.3. Full-Text Level Analysis of Lexical Richness and Syntactic Complexity

The lexical-level analysis indicates that the model factor exhibits significant differences in Herdan's C and Root TTR, suggesting that different LLMs possess varying lexical coverage and type expansion rates at the full-text scale, while their local lexical variation patterns remain similar. The syntactic discourse-level analysis shows that the model factor has significant effects on Syntactic Simplicity, Word Concreteness, Referential Cohesion, and Deep Cohesion, indicating that the models have stable differences in syntactic preferences, vocabulary selection, and the maintenance of discourse coherence. The prompt strategy has no significant effect on these indicators, meaning that the syntactic structure and discourse coherence of the translations are mainly determined by the model's inherent language modeling characteristics.

Table 7. ANOVA Results of Full-Text Level Lexical Richness and Syntactic Complexity Metrics

Metric	Model F	Model p	Prompt F	Prompt p	Domain F	Domain p
Syntactic Simplicity	3.28	0.0126	0.28	0.75	29.48	0.0001
Word Concreteness	0.92	0.48	0.003	0.997	119.2	0.0001
Referential Cohesion	1.85	0.12	0.16	0.85	37.35	0.0001
Deep Cohesion	0.4	0.85	0.18	0.84	65.29	0.0001
TTR	4.98	0.00094	0.94	0.4	23.13	0.0001
MSTTR	1.18	0.33	0.28	0.76	74.74	0.0001
MATTR	1.5	0.21	0.28	0.76	63.2	0.0001
Herdan's C	6.23	0.00016	0.49	0.62	18.21	0.0001
Root TTR	5.26	0.00063	0.91	0.41	23.11	0.0001
CTTR	5.23	0.00065	0.92	0.41	23.14	0.0001

To avoid redundancy and maintain analytical focus, this study emphasizes three representative full-text linguistic indicators: word concreteness, MATTR, and deep

cohesion. Word concreteness captures the degree of lexical specificity, MATTR represents lexical diversity while reducing the influence of text length, and deep cohesion reflects discourse-level organization. The ANOVA results show that the domain factor had significant effects on all three indicators: word concreteness ($F = 119.20$, $p = 0.0001$), MATTR ($F = 63.20$, $p = 0.0001$), and deep cohesion ($F = 65.29$, $p = 0.0001$). These results indicate that financial, technological, and political texts differ substantially in lexical specificity, vocabulary distribution, and discourse organization. In contrast, prompt strategy showed no significant effect on these indicators, with non-significant results for word concreteness ($F = 0.003$, $p = 0.997$), MATTR ($F = 0.28$, $p = 0.76$), and deep cohesion ($F = 0.18$, $p = 0.84$). This suggests that zero-shot, domain-enhanced, and glossary-based prompting did not substantially reshape the full-text linguistic profile of the translations. Overall, the findings indicate that the variation captured by these representative indicators primarily reflects broad differences among text domains rather than localized adjustments triggered by prompt strategies. Thus, the inherent linguistic and discourse characteristics of the source texts appear to play a more decisive role than prompt-level intervention in shaping LLM-generated Chinese–English translations.

5. Conclusion

This study systematically examines the Chinese-to-English translation performance of multiple large language models (LLMs) on professional texts in the financial, technological, and political domains, with a primary focus on the impacts of model type, text domain, and prompt strategies. By integrating automatic evaluation metrics, linguistic feature analysis, and benchmarking against human-translated references, the research constructs a multi-layered analytical framework to assess translation quality and stylistic features, providing a comprehensive empirical perspective for evaluating LLM-based professional translation.

Text domain serves as the dominant driver of translation performance differences, exerting statistically significant effects at the macro document level; the unique discourse structures, syntactic norms, and functional constraints of each professional domain are explicitly reflected in the model outputs. Inherent model characteristics shape distinct linguistic features: while overall automatic evaluation scores vary moderately across models, stable divergences emerge in lexical diversity, syntactic complexity, and discourse coherence, marking consistent stylistic and structural traits for each model. Conventional prompt strategies exhibit no significant influence on overall automatic evaluation scores or linguistic feature measurements, demonstrating that model architecture and domain-adapted training data possess far greater decisive power over translation results.

The results suggest that LLMs can produce relatively reliable translations in the selected professional texts, but their performance still varies across domains and models. These results may be useful for selecting LLMs in professional translation

tasks and for designing future evaluation frameworks. Future research can explore adaptive prompt engineering, interactive human-machine post-editing, and cross-lingual style transfer techniques to further elevate the translation performance and stylistic appropriateness of LLMs in specialized professional texts. However, this study is limited by the size of the corpus, as only one representative text was selected from each domain. Therefore, the findings should be interpreted as evidence from a controlled exploratory comparison rather than as a general conclusion about all professional translation scenarios. In addition, the study relies mainly on automatic metrics and linguistic indicators, while independent human evaluation is not included. Future research may expand the corpus, include human rating or post-editing analysis, and test more adaptive prompt designs.

Acknowledgements

This article is a phased outcome of the following research projects: the 2025 Zhejiang Provincial Philosophy and Social Sciences Planning Project, under the interdisciplinary and niche "Juexue" research category, titled "Emotional Narrative in Precision International Communication and Translation in the Context of Large Language Models" (Grant No. 25JCXK004YB); the Ministry of Education–Transn Industry–University Collaborative Education Project, titled "Practical Applications of Artificial Intelligence Large Language Models in Translation Education" (Project No. 241000630161313); the 2024 commissioned project of the Research Base of Zhejiang Yuexiu University, titled "Innovative Pathways and Effects of Large Language Models in Translation Practice Teaching" (Grant No. 2024WGYYWH04); and the 2024 university-level Higher Education Teaching Reform Research Project of Zhejiang Yuexiu University, titled "A Study of the Training Model for Post-editing Competence in Machine Translation in the AIGC Era" (Grant No. 2405160003).

References

[1] Chan, S.W. (2004) Machine Translation. In: Baker, M. and Saldanha, G., Eds., Routledge Encyclopedia of Translation Studies, Routledge, London, 135.

[2] Fu, Y., Si, S., Mai, L. and Li, X. (2024) FFN: A Fine-Grained Chinese-English Financial Domain Parallel Corpus. arXiv Preprint, arXiv:2406.18856.

[3] Zhao, S., Qiao, L., Luo, K., Zhang, Q.-W., Lu, J. and Yin, D. (2024) SNFinLLM: Systematic and Nuanced Financial Domain Adaptation of Chinese Large Language Models. arXiv Preprint, arXiv:2408.02302.

[4] Kwok, H.L., Shi, Y., Xu, H., Li, D. and Liu, K. (2025) GenAI as a Translation Assistant? A Corpus-Based Study on Lexical and Syntactic Complexity of GPT-Post-Edited Learner Translation. System, 130, Article ID: 103618.

[5] Rei, R., Stewart, C., Farinha, A.C. and Lavie, A. (2020) COMET: A Neural Framework for MT Evaluation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16-20 November 2020, 2685-2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

[6] Zheng, J. (2021) The Importance of a Glossary for Translation Services. Globalization Partners.

[7] Semenov, K., Ailem, M., Jotti, A., España-Bonet, C., Semmar, N., Wanner, L. and

Wisniewski, G. (2023) Findings of the WMT 2023 Terminology Shared Task on Machine Translation with Terminologies. Proceedings of the Eighth Conference on Machine Translation.

[8] Gu, W.H. and Leng, B.B. (2024) Four Types of Terminology Mistranslation in the Application of ChatGPT to Scientific and Technological Translation: A Case Study of Mechanical Engineering Terminology. Chinese Science & Technology Translators Journal, 37, 1.

[9] Zheng, X. (2025) A Comparative Study of Translation Quality of Scientific and Technological Texts under Large Language Models: A Case Study of Four Typical Texts. Modern Linguistics.

[10] Yu, L. (2024) A Study on Lexical Diversity and Syntactic Complexity in ChatGPT Translation. Foreign Language Teaching and Research, 56, 2.

[11] Yu, J. (2025) Exploring DeepSeek's Translation Capability: A Case Study of Literary and Financial Texts. Chinese Translators Journal, 2025, 172-179.

[12] Wen, X. and Tian, Y.L. (2024) The Effectiveness of ChatGPT in Translating Discourse with Chinese Characteristics. Shanghai Journal of Translators, 2024, 27-34.

[13] Zhou, Z.L. (2024) Generative-AI-Based Maritime Translation: Advantages, Challenges and Prospects. Research Report.

[14] Zhao, Y., Zhang, H. and Yang, Y.C. (2024) A Comparative Study of Large Language Models in Text Translation Quality: A Case Study of the Translation of Blossoms. Technology Enhanced Foreign Language Education, 2024, 60-66.

[15] Feng, Q.H. (2025) Innovative Applications of DeepSeek in Translation Teaching and Research. Chinese Translators Journal, 2025, 58-67.

[16] Zhang, S.K. and Zhao, C.Y. (2024) The Applicability of Large Language Models to Literary Translation: A Multi-Metric Evaluation of Multi-Model Translations of Border Town. Research Report.

[17] Liu, S.J. (2024) The Application Effectiveness of Machine Translation in Maritime Translation: A Comprehensive Evaluation Based on BLEU, chrF++ and BERTScore. Research Report.