Education, Language and Sociology Research
ISSN 2690-3644 (Print) ISSN 2690-3652 (Online)
Vol. 7, No. 1, 2026
www.scholink.org/ojs/index.php/elsr

Original Paper

Quality Assessment of Large Language Models Empowering Traditional Chinese Medicine English Translation: A Comparative Study with Online Translation Tools

Li Xu[1] & Zhe Zhang[1*]

[1] Shandong University of Traditional Chinese Medicine, Jinan, China

[*] Corresponding author: Zhe Zhang, zhe_zhang@163.com

Received: February 8, 2026 Accepted: February 22, 2026 Online Published: March 8, 2026

doi:10.22158/elsr.v7n1p110 URL: http://dx.doi.org/10.22158/elsr.v7n1p110

Abstract

Objective: To compare the quality differences in English translations of modern traditional Chinese Medicine (TCM) texts between artificial intelligence (AI)-driven large language models (LLMs) and mainstream online translation platforms, providing empirical evidence for TCM translation practice and research in the AI era. Methods: A representative passage from the TCM textbook Fundamentals of Chinese Medicine was selected as the source text. The text was input into online translation services to generate English translations. The multidimensional quality metrics (MQM) framework was adopted to construct a scoring card covering dimensions such as terminology accuracy, omission, mistranslation, and grammar. Two TCM translation experts independently scored the translations, and the average total penalty points were calculated to measure translation quality. Results: Advanced international LLMs (Claude 3.5 Sonnet, ChatGPT-4o mini, Gemini 1.5 Flash) produced the highest quality translations, with a total penalty score of only 2, significantly outperforming traditional online translation tools. The domestic LLM (ERNIE 4.0) performed next best but was still markedly superior to all conventional online platforms. Translations from traditional online tools commonly suffered from terminological inaccuracies, mistranslations, and grammatical awkwardness. Conclusion: Current mainstream LLMs demonstrate a significant advantage in translating modern TCM texts regarding terminological accuracy and linguistic fluency, achieving a basically usable level. However, LLMs still exhibit limitations when processing specific TCM terms with abstract meanings. Future research should further explore their application potential in translating ancient TCM classics and develop optimized prompt strategies integrated with expert knowledge.

Keywords

Large Language Model, Traditional Chinese Medicine English Translation, Online Translation, Translation Quality Assessment, Multidimensional Quality Metrics

1. Introduction

With the accelerative internationalization of traditional Chinese Medicine (TCM), TCM English translation, serving as a bridge for disseminating TCM science and culture, has become increasingly more important. Traditional translation models primarily rely on manual translation, which is relatively inefficient. In recent years, the proliferation of computer-aided translation tools, especially the breakthrough advancements in artificial intelligence (AI) technology represented by large language models (LLMs), has brought profound changes to the field of TCM translation (Wu, 2021). LLMs represented by ChatGPT, Claude, Gemini, and domestic products like ERNIE Bot, with its powerful natural language understanding and generation capabilities, are increasingly being applied in text translation practice, and users often use them as convenient, low-cost online translation tools (Wang & Zhang, 2024).

Currently, the online translation tools available to translators and researchers can be broadly categorized into two types: first, traditional statistical machine translation or neural machine translation tools such as Google Translate, Baidu Translate, and DeepL; second, conversational AI agents based on LLMs, which can not only translate but also understand context and generate more natural text. How these two technological approaches perform in the highly specialized field of TCM, which is rich in culturally loaded terms, still lacks systematic comparative research. In particular, the accuracy of LLMs' understanding of TCM terminology, the fluency of their translations, and their advantages and disadvantages compared to traditional tools urgently need to be revealed through rigorous empirical studies.

This study aims to fill this gap. By selecting typical modern TCM texts and applying the multidimensional quality metrics (MQM) framework, it conducts a quantitative comparison and qualitative analysis of the English translation quality of mainstream domestic and international LLMs versus traditional online translation platforms. The research results can not only provide data support for TCM translation practitioners in selecting and using translation tools but also offer a theoretical basis for future optimization of TCM translation processes and the formulation of more scientific translation strategies by integrating AI technology.

2. Method

2.1 Source Text Selection

To ensure the representativeness and relevance of the research, a passage from page 135 of the national planning textbook for higher education in the TCM field during the "Twelfth Five-Year" plan period, *Fundamentals of Chinese Medicine*, was selected as the source text (Sun & Zheng, 2012). This passage

covers several core TCM concepts such as “the spleen governs transportation and transformation,” “source of qi and blood production,” and “heart-spleen deficiency pattern,” as well as typical clinical symptom descriptions like “lusterless complexion.” This effectively tests the translation tools’ mastery of fundamental TCM theories and terminology. The specific text is as follows:

“脾主运化而为气血生化之源，水谷精微经脾转输至心肺，贯注于心脉而化赤为血；心主血脉，心生血养脾以维持其运化功能。若脾失健运，化源不足，可导致血虚而心失所养。劳神思虑过度，不仅暗耗心血，又可损伤脾气，形成心脾两虚证。临床常见眩晕，心悸，失眠多梦，腹胀食少，体倦乏力，精神萎靡，面色无华等症。”

2.2 Selection of Translation Tools

This study selected 14 mainstream online translation tools, categorized into three groups based on their technological background: traditional online translation platforms (Google Translate, Baidu Translate, Youdao Translate, Bing Translate, DeepL), Domestic LLMs (ERNIE Bot 3.5, ERNIE Bot 4.0), International LLMs (ChatGPT-4o, ChatGPT-4o mini, Claude 3 Haiku, Claude 3.5 Sonnet, Gemini 1.5 Pro, Gemini 1.5 Flash). The source text was input into each of these tools to generate English versions. All translations were completed within the same time-period to ensure consistency in tool versions.

2.3 Evaluation Framework and Criteria

This study adopted the MQM framework as the evaluation foundation and customized it according to the characteristics of TCM translation. MQM is a theoretical framework for fine-grained translation quality assessment, enabling the classification of different error types and severity levels (Tian, 2020). Within MQM, translation errors are primarily divided into the following seven dimensions: 1. Terminology; 2. Accuracy; 3. Linguistic conventions; 4. Style; 5. Locale conventions; 6. Audience appropriateness; 7. Design and markup. Referring to the International Standard Nomenclature of Traditional Chinese Medicine (ISNTCM) (World Federation of Chinese Medicine Societies, 2007) and the World Health Organization’s WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region (ISTTM) (WHO Regional Office for the Western Pacific, 2007), this study formulated the MQM scorecard shown in Table 1. Recent comparative studies by scholars on TCM terminology translation standards also provide important references for the evaluation in this study.

Table 1. MQM Scorecard for Evaluating TCM Text English Translation Quality

Error Dimension	Error Type	Error Level	Penalty Points	Error Count	Penalty Subtotal
Terminology	Inaccurate meaning	Minor	1		
	Term inconsistency	Major	5		
	Term error	Critical	25		
Accuracy	Omission	Major	5		
	Over/Under-translation	Major	5		
	Mistranslation	Critical	25		
Linguistic conventions	Spelling error	Minor	1		
	Gramma error	Critical	25		
Style	Inconsistent with source	Major	10		
Audience appropriateness	Unsuitable	Major	10		
Total penalty points					

Note. Judgment of terminology accuracy is based on ISNTCM and ISTTM standards. For example, the standard translation for "面色无华" in this evaluation is "lusterless complexion." If translated as "pale" or "sallow," it is considered an "inaccurate term meaning."

2.4 Scoring Procedure

Two experts with TCM backgrounds and extensive translation experience (hereinafter referred to as "score-raters") independently used the scorecard above to evaluate all translations. Before scoring, the two raters underwent unified training on the understanding and operational standards of the scorecard. After scoring, the average of the two raters' total penalty points for each translation was calculated as the final quality score for that translation. A lower total penalty score indicates higher translation quality.

3. Results

The evaluation results show significant differences in translation quality among different categories of translation tools. The distribution of total penalty points for each translation exhibits certain characteristics, as shown in Figure 1.

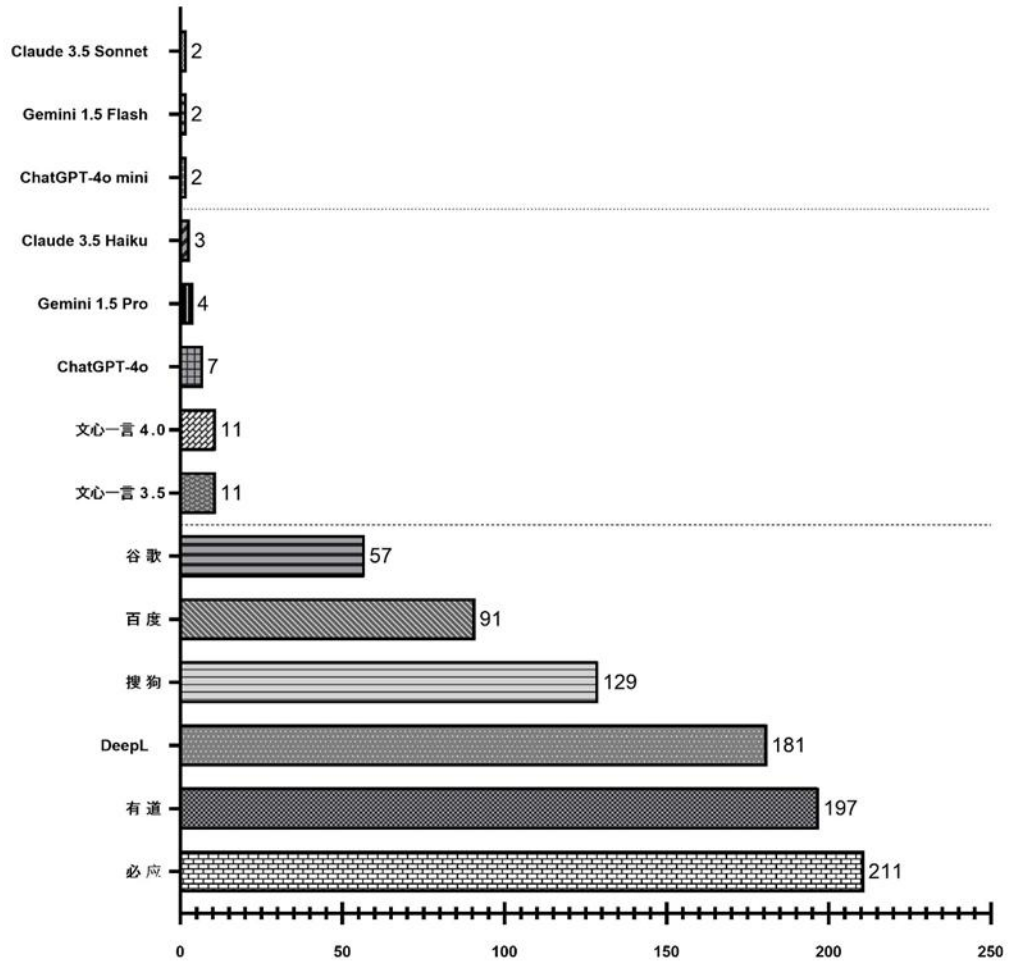

Figure 1. Comparison of Total Penalty Points for Translation Quality of Various Online Translation Tools

The study found that advanced international LLMs achieved the best translation quality. Among them, Claude 3.5 Sonnet, Gemini 1.5 Flash, and ChatGPT-4o mini produced the highest quality translations, with a total penalty score of only 2, significantly lower than other tools. The reason for the penalty points was their handling of the specific symptom term "面色无华," which did not fully align with the standard, reflecting that LLMs still have subtle deviations when processing some high-frequency but abstract terms (the standard translation is "lusterless complexion"; the three translations were "sallow complexion," "sallow complexion," and "pale complexion," respectively). Closely following were Gemini 1.5 Pro and ChatGPT-4o, whose translations demonstrated high fluency and accuracy, with only minor flaws in individual word choices or style. The domestic LLM representative, ERNIE Bot 4.0, also performed remarkably well, with penalty points significantly lower than all traditional online translation tools, proving its strong capability in processing Chinese academic texts. In contrast, the translation quality of all traditional online translation platforms was unsatisfactory, commonly suffering from serious issues such as term mistranslation (e.g., translating "生化" as "biochemistry", omission of key information, and chaotic sentence structures, resulting in total penalty points far exceeding those of LLMs.

4. Discussion

This study is the first to systematically compare the performance of LLMs and traditional online translation tools in translating TCM texts, revealing the profound impact of AI development on specialized domain translation.

4.1 Significant Advantages of LLMs and Technical Explanation

The research results clearly indicate that advanced LLMs represented by Claude 3.5 Sonnet, the GPT-4 series, and the Gemini 1.5 series have achieved an overwhelming advantage over traditional online translation tools in translating modern TCM texts. This advantage is mainly reflected in two aspects:

Contextualized term understanding: LLMs, based on the Transformer architecture and vast pre-training data, can better understand the specific meaning of words within their context. For example, traditional tools often mechanically translate "生化" as "biochemistry," while LLMs accurately render it as "the source of qi and blood production" according to the context of "气血生化之源." This indicates that LLMs not only learn the literal meaning of words but also capture their semantic networks within specific domains. Recent evaluation studies of LLMs in Chinese medical examinations also confirm their strong contextual adaptability when processing professional terms (Zhang et al., 2024).

Fluency and pragmatic appropriateness of generated translations: The sentence structures and expressions in LLM-generated translations are closer to human translation. For instance, the translation by Gemini 1.5 Pro even exhibits a certain stylistic awareness. This stems from the nature of generative models, which aim to create coherent, natural language rather than perform simple word-for-word substitution. Some researchers have already begun exploring intelligent TCM translation algorithms based on deep search and data node retrieval to further enhance translation fluency (He, 2022).

4.2 Current Limitations of LLMs: Take"面色无华"for Instance

Despite the overall excellent performance of LLMs, the translation deviations for the term "面色无华" warrant consideration. GPT-4o mini and Claude 3.5 Sonnet/Gemini 1.5 Flash provided translations like "pale" and "sallow," respectively. In the TCM context, "无华" specifically refers to a complexion losing its moist luster. Although related to the biomedical descriptions of "pale" or "sallow," the connotations are not entirely equivalent. "Lusterless complexion" is the professionally recommended translation according to industry standards, better reflecting the essence of TCM inspection diagnosis. This phenomenon suggests that the "knowledge" of LLMs primarily derives from statistical patterns in their training corpora, which may contain numerous non-standard or variably quality translations. Therefore, LLMs have not yet fully mastered the terminology standards established by authoritative bodies; there remains a gap between their internal "terminology knowledge base" and professional standards. Recent comparative studies on the English translation standards for TCM otorhinolaryngology, rhinology, and acupuncture terms also indicate that terminology standardization itself still faces numerous challenges, which to some extent increases the difficulty for LLMs to learn accurately (Zeng & Jiang, 2024; Zhu & Zhou, 2023).

4.3 Implications for TCM Translation Practice and Research

This study offers multiple insights for future TCM translation practice and research.

First, constructing a new model of human-computer collaboration: LLMs can serve as efficient tools for initial translation and assistance, significantly enhancing translation efficiency. The focus of translators' work should shift from basic word and sentence conversion to pre-editing (optimizing the source text) and post-editing (especially ensuring terminology consistency and handling culturally loaded terms). Through carefully designed prompts, LLMs can be guided to assume different roles such as translation expert, terminology consultant, or style polisher.

Second, new challenges for terminology standardization: The widespread application of LLMs poses new requirements for the standardization of TCM terminology translation. How to effectively "embed" authoritative terminology standards into LLMs, or develop LLM-based terminology agents, is key to ensuring the quality of AI translation. Concurrently, researchers need to pay attention to the impact of LLMs on the actual dissemination and use of terminology standards. The establishment of the Comprehensive Medical Benchmark in Chinese (CMB) provides an important tool for systematically evaluating the performance of LLMs in the medical field (Wang et al., 2023).

Third, limitations and future research directions: Although this study yielded some enlightening conclusions, it has several limitations that need to be addressed and expanded upon in subsequent research. Firstly, this study only selected one passage from the textbook *Fundamentals of Chinese Medicine* as evaluation material. The sample size is small, and the generalizability of the conclusions needs to be tested with larger-scale corpora. Secondly, although the MQM evaluation strives for objectivity, it inevitably carries some degree of subjective judgment from the score-raters. Future research should expand the corpus scope, involve more evaluators, and potentially incorporate automatic evaluation metrics as auxiliary references. Finally, this study primarily evaluated the final translations, lacking in-depth exploration of the internal mechanisms of the translation process. This prevents us from precisely explaining why certain LLMs could accurately translate "the source of qi and blood production" but deviated when handling "lusterless complexion." This will be an important direction for future research.

5. Conclusion

Through an empirical comparison of 14 mainstream online translation tools, this study found that large language models represented by Claude 3.5 Sonnet, ChatGPT-4o mini, and Gemini 1.5 Flash demonstrate strong capability in translating modern TCM texts into English. Their translation quality significantly surpasses that of traditional online translation tools, approaching a professional level in contextualized term understanding and linguistic fluency. This signifies that AI is reshaping the technological landscape of TCM translation, providing a powerful new tool for the international dissemination of TCM. However, LLMs still fail to fully comply with international standards when handling terms like "lusterless complexion," which embody unique TCM diagnostic culture, exposing their limitations in precisely

mastering and applying domain-specific terminology standards. Therefore, future TCM translation should move towards a deeply integrated "human-computer collaboration" model, fully leveraging the efficiency advantages of LLMs while strengthening the terminological gatekeeping and cultural interpretation roles of translation experts. Subsequent research should focus on exploring the application of LLMs in translating ancient TCM classics and developing strategies to effectively integrate authoritative terminological knowledge into AI translation workflows.

Fund Project

This paper is a phased achievement of Shandong Provincial Social Science Planning Research Project entitled "Construction and Application of an Online Corpus Platform for Traditional Chinese Medicine English".

References

He, L. (2022). Design and robustness test of TCM English smart translation algorithm based on data node database retrieval and deep search. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 881-884). IEEE.

Sun, G. R., & Zheng, H. X. (2012). *Fundamentals of Chinese medicine*. China Traditional Chinese Medicine Press.

Tian, P. (2020). Multidimensional metrics for translation quality: Review and implications. *East Journal of Translation*, (3), 23-30. (In Chinese)

Wang, X., Chen, G. H., Song, D., Zhang, L., Li, M., & Liu, Y. (2023). *CMB: A comprehensive medical benchmark in Chinese*. arXiv. https://arxiv.org/abs/2308.08833

Wang, Y., & Zhang, Z. (2024). The hidden concerns and solutions of ChatGPT AI translation. *Chinese Translators Journal*, *45*(2), 95-102. (In Chinese)

WHO Regional Office for the Western Pacific. (2007). *WHO international standard terminologies on traditional medicine in the Western Pacific Region*.

World Federation of Chinese Medicine Societies. (2007). *International standard nomenclature of traditional Chinese medicine*. People's Medical Publishing House.

Wu, X. Z. (2021). Computer aided translation and traditional Chinese medicine translation. In V. Sugumaran, Z. Xu, & H. Zhou (Eds.), *Application of intelligent systems in multi-modal information analytics* (Vol. 1385). Springer. https://doi.org/10.1007/978-3-030-74814-2_112.

Zeng, Z., & Jiang, J. (2024). Comparative analysis of English translation standards for TCM otorhinolaryngology and rhinology disease terms. *Journal of Jiangxi University of Chinese Medicine*, *36*(1), 115-119. (In Chinese)

Zhang, S. W., Chu, Q., Li, Y. J., Liu, J. L., Wang, J. Y., Yan, C., ... Chen, Y. W. (2024). Evaluation of large language models under different training background in Chinese medical examination: A

comparative study. *Frontiers in Artificial Intelligence*, *7*, Article 1442975. https://doi.org/10.3389/frai.2024.1442975

Zhu, Q., & Zhou, E. (2023). A corpus-based comparative study on English translation standards of TCM acupuncture terms. *Chinese Journal of Basic Medicine in Traditional Chinese Medicine*, *29*(10), 1725-1729. (In Chinese)