Journal of Linguistics & Cultural Studies

Article

A Study on Translation Quality Assessment of English Versions of the Shang Han Lun Based on Corpus Linguistics and Large Language Models

Xintong Zhang[1,*]

[1] Beijing University of Chinese Medicine, Beijing, China
* Correspondence: Xintong Zhang, Beijing University of Chinese Medicine, Beijing, China

Abstract: Driven by the dual national strategies of TCM internationalization and cultural digitization, high-quality English translation of TCM classics has become a core component of TCM cross-cultural communication. However, the quality of existing English translations of TCM classics varies greatly, and there is a lack of quantifiable and reusable evaluation standards. Traditional evaluation models often rely excessively on experts' subjective judgments, which hinders the scientific selection of translations and the improvement of communication effectiveness. Taking the core text of Shang Han Lun as the research object, this study selects five translations (Luo Xiwen, Wiseman, Cheng Zhaozhi and Chen Jiaxu, Greta Young Jie De, Huang Hai) to construct a Chinese-English parallel corpus, and uses the GPT-4o large language model to generate reference translations. By integrating automated evaluation metrics with in-depth case analysis, this study constructs a composite evaluation system covering three dimensions: accuracy, fluency and domain adaptability. Empirical results show that there are significant differences among the translations in terminology consistency and the handling of culture-loaded words. Among them, Greta Young's translation has the highest comprehensive score, showing the best fit with the reference translation in terminology standardization and semantic transmission, and demonstrating strong potential for standardized dissemination. At the methodological level, this study couples the quantitative advantages of corpus linguistics with the semantic understanding capabilities of large language models, providing a reusable intelligent research paradigm for the translation quality assessment of TCM classics.

Keywords: Large language models; Corpus linguistics; TCM classics translation; Shang Han Lun; Translation quality assessment

Received: 23 June 2026
Revised: 06 August 2026
Accepted: 20 August 2026
Published: 24 August 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Traditional Chinese Medicine (TCM) is a scientific and technological resource with indigenous innovative advantages in China. The Strategic Plan for the Development of Traditional Chinese Medicine (2016--2030) clearly proposes "promoting the overseas dissemination of TCM" [1]. In 2023, the Implementation Plan for the Major Project of Revitalizing and Developing Traditional Chinese Medicine further listed "strengthening the development of standardization for TCM classics translation" as a core task of the TCM opening-up and development initiative [2]. As a foundational classic of TCM clinical theory, Shang Han Lun has an English translation quality that directly bears on the overseas interpretation of core TCM theories and the construction of TCM academic discourse system. Due to the complex language of the classical Chinese text and its profound cultural connotations, translations have long suffered from challenges such as inconsistent terminology rendering and outdated evaluation methods. Although existing translation evaluation research has accumulated many results in areas such as translation strategy studies, it still mainly relies on experts' subjective evaluations and case studies of single translations. Systematic quantitative cross-translation comparisons are still

insufficient, making it difficult to meet the needs of large-scale and international translation of classics.

In recent years, corpus and large language model (LLM) technologies have promoted changes in the field of translation quality assessment. Generative models represented by ChatGPT, with their language conversion and semantic understanding capabilities, can already handle the deep meaning and cultural connotations of texts [3]. Accordingly, this study integrates three automated evaluation metrics-BLEU, TER, and METEOR-uses GPT-4o to generate English translations of the Shang Han Lun, and conducts quantitative comparisons with five existing mainstream translations, providing objective data support for the translation quality assessment of TCM classics and verifying the applicability of automated metrics in high-cultural-density texts. The research results can intuitively present the relative advantages and disadvantages of each translation in terminology accuracy, syntactic fidelity, and semantic fluency. The highly computable characteristics of Greta Young's translation can also provide reference corpus for subsequent AI translation model training, offering standardized support for the international dissemination of TCM.

2. Literature Review

2.1. Current Status of Translation Quality Assessment Research on Ancient TCM Classics

Over the past decade, the evaluation of English translations of TCM classics has gradually shifted from translation strategy analysis to quality assessment exploration, but it is still in its infancy. Existing evaluation studies mostly focus on basic theoretical classics such as the Huangdi Neijing (Yellow Emperor's Inner Canon). Luo Zhuzhu conducted a case analysis of the Huangdi Neijing (Suwen, referring to Plain Questions) under the House Translation Quality Assessment model, pointing out that the model has cumbersome parameters and cannot be applied to long texts [4]. Ning Jing compared Google Neural Machine Translation with human translations, noting that machine translation still has obvious shortcomings in conveying cultural metaphors [5]. There is still a lack of specific translation quality assessment research on the Shang Han Lun. A visual analysis by Dong Fangxu and Xu Wenjie showed that "translation quality assessment" did not enter the keyword burst sequence of Shang Han Za Bing Lun translation research in the past 30 years [6]. Han Rui constructed a machine translation model for ancient Chinese and carried out quality assessment, which is one of the few attempts to address the automatic assessment of ancient TCM texts [7].

In summary, although existing research on English translations of ancient TCM texts has accumulated many results in case analysis and translation strategy classification, there is still considerable space for expansion in systematic quantitative cross-translation assessment.

2.2. Research Progress in Translation Quality Assessment Methods

The methodology of translation quality assessment (TQA) as a whole shows a trend of development from purely manual judgment to human-machine collaboration. Mainstream evaluation methods include human evaluation, reference-based automatic evaluation, and reference-free automatic evaluation [8].

As a traditional evaluation paradigm, human evaluation can cover complex dimensions such as stylistic features and deep semantics, but it is easily affected by the subjective cognition of the evaluator, and large-scale evaluation incurs high costs and low reproducibility. Reference-based automatic evaluation calculates the matching degree between the translation under test and the reference translation at the lexical, syntactic, and semantic levels. It has high efficiency and strong reproducibility, making it the mainstream tool for current large-scale translation assessment. Reference-free automatic evaluation directly uses a model to judge the quality of the translation under test. Kocmi et al. verified the feasibility of LLMs as evaluation subjects through prompt engineering [9]. The BISU-AiTQA evaluation results show that mainstream LLMs perform fairly well in assessing the consistency of TCM classic terminology, but their accuracy in judging the

transmission of cultural metaphors is low [10]. Corpus linguistics provides a quantitative perspective for this research field. Since Baker proposed the corpus-based translation studies paradigm, the academic community has realized the quantitative description of translator styles and translation strategies through methods such as parallel corpus alignment and word frequency statistics [11].

In summary, existing research has laid a foundation in the case analysis and strategy review of TCM classic translations, and has also established the feasibility of corpus and automatic metrics in the field of general translation quality assessment. However, for classic texts such as Shang Han Lun that have three characteristics - low resource availability, high cultural density and classical Chinese syntax - existing research paths are mostly confined to subjective qualitative criticism under a single theoretical framework, or limited to formal matching based on pure automatic metrics. There is still no systematic evaluation path that organically couples corpus alignment, automatic metric quantification, and LLM semantic verification. Based on this, this study constructs a parallel corpus of the Shang Han Lun, introduces GPT-4o to generate standardized reference translations, builds a composite evaluation system of automatic metric quantification and manual case verification, and provides a reusable path for the translation quality assessment of ancient TCM texts.

3. Research Design

3.1. Overall Framework and Technical Route

This study takes 276 core clauses of the Shang Han Lun as the object to construct a structured parallel corpus. It guides GPT-4o to generate reference translations that meet academic norms via prompt engineering and controls randomness. Subsequently, it relies on the Shiyibao platform to calculate three metrics in batches: BLEU, TER, and METEOR. On this basis, the Friedman test is comprehensively used for verification, and typical cases are combined for qualitative analysis to cross-validate the results of the human and machine paths, thereby ensuring the validity of the conclusions.

3.2. Corpus Construction and Preprocessing

This study took 398 core clauses as the object and selected five translations belonging to different translation schools (Luo Xiwen, Wiseman, Cheng Zhaozhi & Chen Jiaxu, Greta Young Jie De, Huang Hai), covering different eras and strategic orientations to reduce single-translation bias. In the corpus preprocessing stage, the study first took the clauses as basic units, removed annotations and appendices, and retained 276 valid corpus entries. Secondly, it used the SY-XXX coding system for structured numbering. Finally, after double independent proofreading and arbitration by a third person in case of disagreement, the data were entered. A structured dataset aligned with the original clauses, human translations, and GPT-4o translations was finally formed.

3.3. Large Language Model Reference Translation Generation

This study selected GPT-4o as the reference translation generation model. Existing research has verified the adaptability of this type of model in ancient text translation scenarios: Ren Yuxin et al. confirmed that GPT-4o is superior to Google Translate in retaining the cultural connotations of TCM terminology and in quantitative metrics [12]; Wen Lanfang and Liu Yan verified in the Analects task that the ChatGPT series performs better than open-source models such as DeepSeek in BLEU/TER in high-cultural-density scenarios of ancient texts [13]. It can be seen that GPT-4o has a stable hit rate for the TCM terminology system, and its translation style fits the norms of academic translations, making it suitable as a reference.

This study adopted a unified instructional prompt template, explicitly requiring that the logical relationships between diagnosis and treatment in the original text be fully preserved. The generation parameter Temperature was set to 0.7. Three translations were independently generated for each original text, and the metrics were calculated separately and then averaged to reduce sampling randomness.

3.4. Construction of a Multi-Dimensional Translation Quality Assessment System

3.4.1. Assessment Dimensions and Metric Selection

The English translation of the Shang Han Lun must not only faithfully express the source culture but also fully convey traditional medical thinking. For this purpose, the study selected three types of reference-based automatic metrics, corresponding to three dimensions: 1 BLEU: calculates similarity based on n-gram matching [14]. The higher the score, the higher the matching degree of vocabulary and sentence patterns.2 TER: calculates the translation modification cost based on edit distance [15]. The lower the score, the higher the translation fluency, which can reflect problems such as omissions and errors. 3 METEOR: calculates based on unigram recall and weighted harmonic mean, adding a synonym matching mechanism [16]. The higher the score, the higher the semantic matching degree. The combination of the three can comprehensively evaluate the translation from multiple levels of form, fluency, and semantics, reducing the evaluation deviation of a single metric.

3.4.2. Automated Evaluation Tools and Operation Process

The main tool of the study was the Shiyibao batch evaluation platform, which automatically outputs three core metrics. The Hugging Face BLEU evaluation space was introduced for cross-validation to avoid calculation deviations from a single tool. During evaluation, the 5 human translations were set as reference texts, and the corresponding GPT-4o translations were set as texts to be evaluated. The results were saved by"translator-metric-clause."

3.5. Reliability and Validity Control

This study set quality control from the following three levels. First, at the corpus reliability level, the study used the Shang Han Lun published by the People's Medical Publishing House as the base text. The corpus underwent triple verification, effectively eliminating factors such as misaligned lines and missing translations. Second, at the metric reliability level, the study compared the BLEU values output by Shiyibao with the recalculation results of the Hugging Face evaluate-metric/bleu space, controlling the difference to ≤0.01. For TER and METEOR, the output values from Shiyibao were adopted as the standard. At the same time, each reference translation was independently generated 3 times to reduce sampling noise. Third, at the result validity level, the three-dimensional metric distribution of this study basically matched the translator style characteristics revealed by existing corpus research [17-19], indicating that the discriminative power of the automatic metrics in this scenario was basically established. To further verify the significance of the differences in the three-dimensional metrics of the five translations, the study used the Friedman two-way rank variance analysis on the metric matrix of 157 fully paired corpora, and the results were shown in Table 1. The differences in the three metrics among the five translations reached extremely significant levels, indicating that the leading advantage of Greta Young's translation had statistical stability and was not caused by random fluctuations.

Table 1. Friedman Test Results of Translation Automatic Evaluation Metrics

Metric	Chi-square($\chi 2$)	Degrees of Freedom(df)	Asymptotic Significance(p)
BLEU	124.776	4	<.001
TER	104.435	4	<.001
METEOR	144.474	4	<.001

4. Empirical Results and Analysis

4.1. Quantitative Evaluation Results of Three-Dimensional Metrics

The study summarized the metric results of 276 corpus entries and obtained the average values of core metrics for the 5 translations, as shown in Table 2.

Table 2. Average Values of Three-dimensional Evaluation Metrics for Mainstream English Translations of the Shang Han Lun

Evaluation Dimension	Metric	Luo Xiwen	Wiseman	Cheng Zhaozhi & Chen Jiaxu	Greta Young Jie De	Huang Hai
Accuracy	BLEU	0.32	0.31	0.39	0.48	0.40
Fluency	TER	0.87	0.81	0.82	0.65	0.84
Domain Adaptability	METEOR	0.14	0.14	0.16	0.23	0.17
Comprehensive Score	-	21.02	21.57	/	35.2	25.9

The BLEU value reflects the matching degree between the GPT-4o reference translation and each human translation in vocabulary and sentence patterns. Greta Young's translation has the highest BLEU value (0.48), indicating the highest fit with the standardized reference, with better terminology uniformity and formal standardization. The Huang and Cheng & Chen translations are in the middle, with translation strategies focusing on literal translation and moderate domestication. The Luo and Wiseman translations are the lowest; the former uses Latin plant names for domestication, while the latter focuses on source terminology fidelity and thus weakens sentence pattern correspondence, resulting in lower formal matching with the reference.

The TER value reflects the editing cost required to adjust the human translation to the reference translation. Greta Young's translation has the lowest TER value, which is more in line with the normative needs of contemporary clinical translations in logical coherence. The TER values of Huang and Luo are high, indicating high terminology density and sentence patterns that retain the elliptical features of classical Chinese, leading to higher editing costs for alignment with the reference.

The METEOR value reflects the semantic matching degree of each translation with the reference in culture-loaded words. Greta Young's translation has a significantly higher METEOR value than other translations, balancing TCM category specificity and English academic expression habits. Luo and Wiseman are the lowest, adopting complete domestication and complete transliteration strategies respectively, making it difficult to capture the unique TCM connotations of classic concepts.

Overall, Greta Young's translation performs best in all three dimensions, with a comprehensive score significantly higher than other translations. This indicates that the GPT-4o reference translation has the highest match with this translation. Its clear language style, unified terminology, and standardized sentence patterns are easy for AI to learn and reproduce, making it suitable as a reference for machine translation of TCM classics.

4.2. In-Depth Analysis of Typical Cases

4.2.1. Cases of TCM Terminology Translation

Case 1:" TaiYangBing, ChuFuGuiZhiTang, FanFanBuJieZhe, XianCiFengChi, FengFu, QueYuGuiZhiTangZeYu "(SY-007)

In the translation of "Cinnamon Twig Decoction" (GuiZhiTang), Luo Xiwen adopted a complete domestication strategy, using the Latin plant name "Decoction of Ramulus Cinnamomi." Huang Hai used pure free translation "Cinnamon twig decoction." Wiseman used free translation plus Pinyin annotation "Cinnamon Twig Decoction (gui zhi tang)." Cheng & Chen and Greta Young both used pure Pinyin transliteration "GuiZhiTang," which is completely consistent with the GPT-4o reference translation. Quantitative results show that the BLEU value of Greta Young's translation with GPT-4o reaches 0.51, significantly higher than other translations. It can be seen that BLEU prefers concise transliteration. Greta Young's translation has significant advantages in word order, capitalization, and spelling, confirming that the "Pinyin transliteration plus class name"

form has both terminology specificity and AI adaptability, and is more suitable for building reference translation for machine translation of TCM classics.

Case 2: " TaiYangBing, HuoYiFaRe, HuoWeiFaRe, BiEHan, TiTong, OuNi, MaiYinYangJuJinZhe, MingWeiShangHan ."(SY-020)

In the translation of "Cold Damage" (ShangHan), Luo Xiwen adopted a complete domestication strategy, translating it as "febrile disease caused by Cold (Shanghan)." Wiseman used the free translation strategy "cold damage." Cheng & Chen translated it as "cold invasion syndrome." Huang Hai used the "transliteration plus class name" form "Shanghan syndrome." Greta Young directly translated it as "shang han," which completely corresponds to the transliteration annotation part of the reference translation "Cold Damage (Shang Han)." Quantitative results show that Greta Young's translation has the highest BLEU value. Its translation is consistent with the reference annotation in word order and spelling, and its sentence pattern is highly isomorphic with the reference in the conditional framework and term sequence arrangement logic.

4.2.2. Cases of Culture-Loaded Word Translation

Case 3:" TaiYangZhongFeng, XiaLiOuNi, BiaoJieZhe, NaiKeGongZhi. QiRenZhiZhiHanChu, FaZuoYouShi , TouTong , XinXiaPiYingMan , YinXieXiaTong , GanOuDuanQi , HanChuBuEHanZhe , CiBiaoJieLiWeiHeYe , ShiZaoTangZhuZhi ."(SY-105)

In this clause, Luo Xiwen still used domestication expressions, translating "attack" (GongZhi) as "purgative," weakening it to an indication for ordinary laxatives, dissolving the specific treatment connotation of "attacking and expelling water rheum." He translated "the exterior is resolved but the interior is not harmonized" as "the Interior is not in good harmony...," simplifying the logic of "exterior-interior transmission and struggle between pathogenic qi and healthy qi" to ordinary "internal-external disharmony." In contrast, Greta Young's translation directly translates "attack" as "one may still purge," and uses "exterior is resolved but the interior is not harmonized" to retain the dynamic treatment connotation of "attacking and expelling" and the unique pathological logic of "exterior-interior disharmony." Although Wiseman used foreignization, he translated "attack" as "attack," failing to reflect the specific direction of "attacking and expelling water rheum." Quantitative results show that Greta Young's translation has a higher METEOR value, indicating that his literal translation of the treatment expression retains the original cultural logic and has a higher semantic recall with the reference.

Case 4:" FaHanGuoDuo, QiRenChaShouZiMaoXin, XinXiaJi, YuDeAnZhe , GuiZhiGanCaoTangZhuZhi ."(SY-059)

At the semantic transmission level of physical sign description, Luo Xiwen translated " ChaShouZiMaoXin " as "tends to press his chest with interlaced fingers," replacing the dynamic sign of "clasping hands over the heart" with a concrete expression, weakening it to an ordinary pressing action. Huang Hai's "press his chest with interlacing hands" is highly similar, also failing to convey the deep connotation of "covering and guarding" in " Mao." Wiseman translated it as "hands are crossed over the heart." Greta Young translated it as "clasps their hands to their heart," where "clasp" accurately conveys the dynamic sign of "clasping hands to protect and press." Moreover, "palpitations below the heart" is translated as "palpitations below the heart," and "desires pressure" as "a desire for pressure," retaining the dynamic logic of the patient actively seeking pressure to relieve palpitations. All these align with the reference in the symptom semantic domain. The METEOR value reaches 0.22, significantly higher than other translators.

4.2.3. Cases of Classical Chinese Sentence Pattern Conversion

Case 5:" JiaLingChiZhongChiZhe, BuKeFaHan, HeYiZhiRan? YiYingQiBuZu, XueShaoGuYe"(SY-025)

The core sentence pattern feature of this clause is a series of causal inference chains. At the level of logic manifestation, Luo Xiwen used a single causal clause structure, merging the hypothetical "if the cubit is slow" and the rhetorical question "how do we know this" into one clause, but the syntactic level differs greatly from the reference's

clause-by-clause progressive structure, thus the TER value is high. Wiseman used a bracketed annotation structure, which is academically normative but increases the edit distance due to text noise. Huang Hai mixed the text logic into a misplaced prepositional structure, resulting in a large edit distance. Greta Young's translation is highly isomorphic with the reference's progressive logic, both starting with a conditional clause, taking "reason" to carry the rhetorical question, and ending with "It is because" to close the causal inference. The TER value is 0.71, the lowest among the participating translations.

Case 6:" TaiYangBing, XiangBeiQiangJiJi, FanHanChuEFengZhe, GuiZhiJiaGeGenTangZhuZhi."(SY-013)

In this sentence, Luo Xiwen employed a single conditional clause structure to merge the symptoms and directly introduced the prescription using an imperative sentence, resulting in a flat sentence hierarchy and the largest edit distance. Huang Hai utilized a relative clause as a post-modifier; however, he mistranslated " FanHanChu " (unexpected sweating) as "without perspiration." Wiseman adopted a "When... there is..." conditional framework, whose syntactic logic is close to that of the reference, yet a formal deviation exists between his term structure with preposed free translation and the reference Pinyin transliteration. In contrast, Greta Young's translation exhibits a high degree of consistency with the reference in syntactic logic: both commence with the disease category, employ parallel phrases to describe the symptoms, present the prescription conclusion via Pinyin transliteration, and conclude with the verb "governs." Greta Young's TER value is the lowest among all participating translations, further substantiating that his sentence conversion strategy aligns with the AI reference generation paradigm and compensates for the loss of implicit logical connections in classical Chinese during English translation through the explicit manifestation of logical connectives.

5. Discussion

5.1. Analysis of the Causes of Quality Differences among Translations

The performance differences among translations are essentially the result of the combined effect of translator background and translation strategies. Greta Young has both a TCM clinical background and translation ability. While prioritizing terminology standardization, she balances clinical readability, thus performing best in all three metrics. Wiseman's translation, as a terminology standardized translation led by Western translators, focuses on terminology system uniformity but has rigid sentence expression, leading to insufficient fluency. Luo Xiwen's translation, as an early translation, takes the readability for target language readers as the core and adopts a highly domesticated strategy, thus having the lowest formal matching. Huang Hai and Cheng & Chen's translations attempt to balance fidelity and fluency, but limited by the depth of understanding of ancient medical texts, their terminology consistency and cultural transmission are insufficient.

5.2. Reference Value and Limitations of LLM Reference Translations

The reference translation generated by GPT-4o shows dual value in this study. First, the study uses its unified and stylistically standardized translation as a stable reference for quantitative evaluation, effectively avoiding terminology confusion among human translations. Second, the study uses its instant interactivity to quickly generate a large number of translations for batch evaluation, effectively changing the dilemma of low efficiency and difficulty in scaling up traditional manual evaluation. Its limitations are also obvious. Facing expressions carrying TCM pathogenesis logic such as "contention of wind and dampness", GPT-4o can only output surface semantics and cannot convey deep cultural connotations. When facing special classical Chinese sentence patterns such as elliptical sentences and inverted sentences, GPT-4o occasionally has logical misreading, leading to semantic deviation. Therefore, LLM translations can only be used as a standardized reference for evaluation, and their cultural value cannot replace human translations.

5.3. Applicability and Promotion Value of the Coupled Evaluation Paradigm

The composite translation quality assessment model constructed in this study has many advantages over traditional evaluation methods. On one hand, the study reduces manual subjective bias through quantitative indicators, ensuring the reproducibility of results. On the other hand, the model can simultaneously process multiple translations for batch comparison, greatly improving evaluation efficiency. More fundamentally, it features good extensibility. In addition, the evaluation framework can adjust metric weights according to the characteristics of different classics to adapt to the evaluation needs of other TCM classics. In terms of promotion value, this paradigm promotes the transformation of TCM classic translation evaluation from subjective criticism to objective measurement at the academic level. At the practical application level, the evaluation paradigm can be implemented in multi-dimensional scenarios: first, as a translation aid tool, it can detect the matching degree between the translation and the standard terminology system in real time and prompt the translator to correct deviations; second, as publishing quality control, it can batch pre-screen full translations, quickly locate problems such as terminology inconsistency and missing translations, and reduce manual proofreading workload; third, as AI translation training, it can use highly standardized translations as reference corpus to optimize the translation performance of large language models in the TCM field.

6. Conclusion

This study develops a composite translation quality assessment system that couples corpus linguistics and large language models, evaluating five mainstream English translations of Shang Han Lun from three dimensions: accuracy, fluency, and domain adaptability. Using GPT-4o-generated standardized reference translations and three automated metrics (BLEU, TER, METEOR), the study identifies significant cross-translation differences in terminology consistency and culture-loaded word rendering. Greta Young's translation achieves the highest comprehensive score, as its unified terminology, standardized syntactic structure and clinically oriented readability demonstrate the closest alignment with the AI reference within the scope of this study. The findings further confirm that automated metrics can effectively capture translator style differences documented in prior corpus studies, verifying the discriminative power and feasibility of the proposed coupled evaluation paradigm for low-resource, high-cultural-density classical Chinese medical texts. Despite the limitation that LLM-generated references cannot fully capture the pathological connotations embedded in TCM theories and complex logical relations in classical Chinese, the framework bears notable methodological and practical implications. It advances TCM classic translation assessment from subjective qualitative criticism to objective quantitative measurement, and can be adapted to other TCM classics via metric weight adjustment. For future research, the framework can be further optimized by integrating expert human evaluation and exploring reference-free evaluation approaches, so as to support translation aid, publishing quality control and domain-specific AI translation model training, and ultimately facilitate the standardized international communication of TCM.

References

1. State Council, *Strategic Plan for the Development of Traditional Chinese Medicine (2016–2030)*, 2016.
2. General Office of the State Council, *Notice of the General Office of the State Council on Issuing the Implementation Plan for Major Projects of Revitalization and Development of Traditional Chinese Medicine*, State Council General Office Doc. No. 3 [2023], 2023.
3. OpenAI, *GPT-4o Technical Report*, 2024. [Online]. Available: https://openai.com/research/gpt-4o. [Accessed: Aug. 7, 2026].
4. Z. Luo, "A case study on English translations of the *Huangdi Neijing* under the translation quality assessment model," M.S. thesis, Jiangxi University of Chinese Medicine, 2019. doi: 10.27180/d.cnki.gjxzc.2019.000087.
5. J. Ning, "Comparison between machine translation and human translation—Taking the terminology of *Huangdi Neijing* as an example," *Masterpieces Review*, no. 36, pp. 146–147, 2021.
6. F. Dong and W. Xu, "Current status and trends of translation research on *Shang Han Za Bing Lun* based on CiteSpace," *English Square*, no. 10, pp. 23–28, 2026. doi: 10.16723/j.cnki.yygc.2026.10.032.

7. R. Han, "Research on machine translation models for ancient Chinese and their quality assessment methods," M.S. thesis, North China University of Science and Technology, Tangshan, 2025.
8. C. Zhang, F. Chen, L. Hu, et al., "Research on English translation issues of *Shang Han Lun*," *Western Journal of Traditional Chinese Medicine*, vol. 37, no. 6, pp. 103–107, 2024.
9. T. Kocmi and C. Federmann, "Large language models are state-of-the-art evaluators of translation quality," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Dubrovnik, Croatia: ACL, 2023, pp. 193–203.
10. Beijing International Studies University, *BISU-AiTQA: Large Language Model Translation Quality Evaluation Report (v1.0)*, 2025.
11. M. Baker, "Corpus linguistics and translation studies: Implications and applications," in *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis, and E. Tognini-Bonelli, Eds. Amsterdam, The Netherlands: John Benjamins, 1993, pp. 233–250.
12. Y. Ren, Z. Chen, Y. Lin, et al., "Research on machine translation quality assessment of TCM terminology in the era of artificial intelligence—Taking ChatGPT-4 and Google Translate as examples," *Guiding Journal of Traditional Chinese Medicine and Pharmacy*, vol. 31, no. 12, pp. 284–293, 2025.
13. L. Wen and Y. Liu, "Quality evaluation of GPT-4o translations based on NLP metrics—Taking *The Analects of Confucius* as an example," *Journal of Yichun University*, vol. 47, no. 7, pp. 92–98, 2025.
14. K. Papineni, S. Roukos, T. Ward, et al., "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA: ACL, 2002, pp. 311–318.
15. M. Snover, B. Dorr, R. Schwartz, et al., "A study of translation edit rate with targeted human annotation," *Machine Translation*, vol. 20, no. 2, pp. 85–112, 2006.
16. A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Machine Translation*, Prague, Czech Republic: ACL, 2007, pp. 228–231.
17. J. Jiang, "A corpus-based comparative study on translation styles of English versions of *Shang Han Lun*," *Journal of Basic Chinese Medicine*, no. 9, pp. 1531–1534, 2023.
18. B. Gao and Y. Yang, "A comparative study of English translations of *Shang Han Lun* from the perspective of Skopos theory," *Journal of Yangling Vocational and Technical College*, no. 6, pp. 110–118, 2025.
19. J. Huang and X. Li, "A corpus-based study on translator styles of three English versions of *Shang Han Lun*," *Medicine & Philosophy*, no. 13, pp. 73–77+81, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.