Proceedings of the 5th International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2026.38879

Comparative Study on the Translation Quality of "Three-Character Classic" by Large Language Models and Multimodal Large Models

Wenxi Zhao

Guangzhou Huali College, Guangzhou, China
3118614170@qq.com

Abstract. With the widespread application of large language models (LLM) and multimodal large models (MLLM) in the field of machine translation, how to evaluate the translation quality of different models for ancient Chinese classics has become an urgent issue to be explored. This study takes the children's educational classic "Three Character Classic" as the research object, selects eight representative original sentences, and translates them into English using both LLM and MLLM. The translated texts are then quantitatively compared using the BLEU automatic evaluation tool. The aim of this study is to determine which type of model performs better in the English translation of classical Chinese texts and which model is more suitable for the translation tasks of Chinese classical literature. The study found that, here, fill in the actual experimental data: the BLEU and METEOR scores show which type of model performs better. This study provides empirical reference for AI-assisted English translation of Chinese ancient classics and also offers an empirical basis for the applicability of automatic translation quality assessment tools in the context of classical Chinese translation.

Keywords: Large language model, Multimodal large model, "Sanzi Jing", Machine translation quality assessment, BLEU

1. Introduction

1.1. Research background

In recent years, large language models (LLMs) such as ChatGPT and GPT-4o have made significant breakthroughs in the field of natural language processing, with machine translation being one of their important application scenarios. At the same time, multimodal large language models (MLLMs) like GPT-4o and KIMI, with their ability to process both text and images simultaneously, have demonstrated unique advantages in translating culturally loaded words and handling context-sensitive texts. This technological evolution has brought a paradigmatic change to traditional machine translation - translation is no longer limited to the input and output of pure text, but can integrate multimodal information such as images to enhance the understanding of the cultural context of the source language.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.38879

The translation of ancient Chinese classics is an important means of the dissemination of Chinese culture to the outside world. The "Three-Character Classic" is a traditional Chinese vernacular learning classic that condenses Confucianism, historical knowledge, and moral teachings in a rhyming form of three characters per line, and has been listed by UNESCO as part of the "World Children's Education Series". Its English translation faces multiple challenges such as concise writing, strict rhythm, and profound cultural connotations. How to use artificial intelligence technology to improve the quality of the English translation of such classics, which is a traditional Chinese classic, is a research topic with both theoretical value and practical significance. Recent research has explored multi-agent collaborative frameworks for classical Chinese translation based on large language models, demonstrating promising potential for AI-assisted translation of ancient texts [1].

1.2. Current problems

Current academic research mainly relies on automatic evaluation indicators such as BLEU (Bilingual Evaluation Understudy) for the assessment of machine translation quality [2]. However, most existing studies focus on general text or specific domain translation evaluations [3], and there is a lack of comparative studies on the translation quality of ancient Chinese classics, especially rhyming vernacular learning texts, under the two model architectures of LLM and MLLM. In addition, the actual effect of multimodal large models in the translation of classics, such as whether the introduction of visual information can significantly improve the quality of the translation, still lacks sufficient empirical verification.

1.3. Research objectives

This paper aims to answer the following core questions:

(1) Under the two automatic evaluation indicators of BLEU and METEOR, do large language models and multimodal large language models have significant differences in the quality of the English translation of the "Three-Character Classic"?

(2) Which type of model's translation is more advantageous in terms of word overlap and semantic accuracy?

(3) Based on the above findings, which type of large model is more suitable for the English translation task of classical Chinese?

1.4. Research significance

Theoretical significance: This study expands the application scenarios of automatic evaluation tools for translation quality to the translation evaluation of ancient Chinese rhyming classics, tests the applicability and effectiveness of BLEU in the special language pair of Chinese classical - English, and provides new empirical materials for the development of translation quality assessment theory.

Practical significance: The research conclusions can provide reference for the selection of models for AI-assisted English translation of ancient Chinese classics, and contribute to the digital dissemination and international expression of excellent Chinese cultural traditions.

DOI: 10.54254/2753-7064/2026.38879

2. Literature review

2.1. Core concept definitions

Large Language Model (LLM): A language model based on large-scale corpora pre-training, with strong text understanding and generation capabilities, and capable of completing various tasks such as translation, summarization, and question answering through natural language instructions.

Multimodal Large Language Model (MLLM): An extension of LLM that expands the processing capacity of multi-modal information such as images, capable of accepting both text and image inputs and achieving cross-modal understanding and generation. BLEU (Bilingual Evaluation Understudy): An automatic evaluation metric for machine translation based on n-gram matching, which generates scores by calculating the number of overlapping words and phrases between the machine translation and the reference translation.

2.2. Relevant research abroad

In the field of automatic evaluation of machine translation quality, the effectiveness of BLEU has been widely verified. Studies have used BLEU to evaluate the quality of translations produced by various machine translation methods in the French-English translation task, and found that neural machine translation performed the best overall, followed by large language models [4].

Additionally, Al Amer et al. conducted a comparative evaluation of machine translation models using human-translated social media posts as references, providing valuable datasets for low-resource language translation research [5].Canga Alonso and Napoletano further examined the role of metrics in assessing the quality of machine translation across different language pairs, highlighting that appropriate metric selection significantly affects evaluation outcomes [6].

Moreover, Lei evaluated different translation methods through a case study of Chinese graduate students in an MTI program, offering empirical evidence for translation method effectiveness in educational contexts [7].

In the field of multimodal translation, some studies have compared the translation performance of GPT-4o, KIMI, and Google Translate on pure text and "text + image" conditions using cultural specific items from "A Qian Zhongshu's Story" [8]. It was found that visual cues significantly improved the translation quality scored by human raters, and GPT-4o (text + image) achieved the highest scores in all dimensions.

2.3. Relevant research in China

Chinese scholars have achieved rich results in the evaluation of the quality of English translations of classic works using large language models. Yu Qiangfu et al. used BLEU values, Julian House's evaluation model, and the MQM error hierarchy scale to conduct multi-dimensional evaluations of six large language models (Zhinan Qingyan, Xunfei Xinghuo, Wenxin Yiyan, Baichan Intelligent, ChatGPT, and DeepL) on the English translations of "Mencius - Li Huaiwang Chapter" [9]. They found that domestic large language models (especially Wenxin Yiyan) had higher English translation quality for ancient Chinese classics.

Another study evaluated the translation quality of DeepSeek-R1, Wenxin Yiyan-X1 Turbo, Xunfei Xinghuo-X1, Zhinan Qingyan-4.5, ChatGPT-5, and Google Translate using the "Confucian Classic - Mencius" as an example [10]. It was found that domestic large language models performed

DOI: 10.54254/2753-7064/2026.38879

overall better than foreign large language models in terms of semantic accuracy, cultural load word handling, and style restoration in the text.

Wen Lanfang et al. selected 20 sentence segments from "The Analects" and used BLEU, TER, and METEOR to evaluate the translation quality of GPT-4o, DeepL, Google Translate, and Youdao Translate [11]. The results showed that GPT-4o outperformed other tools in terms of language fluency, translation accuracy, and cultural understanding. In the field of traditional Chinese medicine terminology translation, the BLEU, TER, and METEOR automatic evaluation indicators of ChatGPT-4 were significantly better than those of Google Translate [12].

2.4. Research review

In summary, scholars at home and abroad have made significant progress in the following aspects: (1) The effectiveness of BLEU and METEOR as automatic evaluation tools for machine translation quality has been verified to a considerable extent; (2) The application of large language models in the English translation of ancient Chinese classics has attracted attention and generated several valuable empirical studies; (3) The potential of multimodal large models in cultural translation has been initially demonstrated.

However, the existing research still has the following gaps: First, no study has directly compared the performance differences between large language models and multimodal large models in the translation of the same classic text; second, the "three-character sentence" unique form of the rhymed classical education classic "The Three-Character Classic" poses special requirements for translation, but existing research rarely conducts systematic machine translation quality evaluations using this as an object; third, the applicability and sensitivity of BLEU in evaluating the translation of rhymed classical Chinese have not been fully tested. This study precisely addresses the aforementioned gap by using the "Three-Character Classic" as the corpus and comparing the performance of LLM and MLLM in the translation of classical Chinese texts.

3. Research methods

3.1. Research design

This study adopts a quantitative comparative design within the framework of mixed research methods. Specifically, eight representative original sentences from the "Three-Character Classic" were selected and input into the large language model and the multimodal large model to obtain English translations. Subsequently, BLEU was used to quantitatively score and compare the quality of the translations. This study does not involve an artificial evaluation process and focuses on objective comparisons within the framework of automatic evaluation indicators.

3.2. Research objectives and sample selection basis

Translation Source Text: The full text of the "Three-Character Classic" is approximately 1,140 words, organized in the form of three characters per line and two lines per rhyme. In this study, eight sentences were selected as translation samples. The selection criteria are as follows:

(1) Cover different themes such as ethical education, historical knowledge, and natural common sense, ensuring semantic diversity of the samples;

(2) Include cultural-loaded words and core concepts (such as "nature", "goodness", "righteousness", "filial piety", etc.), to examine the model's ability to handle cultural specificities;

DOI: 10.54254/2753-7064/2026.38879

(3) Balance the typicality and variability of the syntactic structure.

The selected eight sentences are as follows:

1. People are born with nature's goodness. Nature is similar, but habits vary. (The opening summary, presenting the core proposition of the theory of human nature)

2. If not educated, nature will change. The way of education is to be dedicated. (Discussing the necessity of education)

3. Once upon a time, Mencius' mother chose a neighborhood. If the child does not learn, she will break the loom. (Historical anecdote, containing cultural specificities)

4. If one does not educate, the father is at fault. If education is not strict, the teacher is lazy. (Discussion of ethical responsibility)

5. Jade does not become a tool without polishing. People do not know the meaning without learning. (Comparative argument, containing metaphorical rhetoric)

6. As a child, one should be respectful to parents. Respect for teachers and friends, and learn etiquette. (Behavioral norms guidance)

7. Huang Xiang warmed the mat when he was young. Filial piety is what one should adhere to. (Historical anecdote, containing cultural specificity "Huang Xiang warming the mat")

8. First, be filial and respectful, then acquire knowledge. Know certain numbers and recognize certain texts. (Discussion of the order of learning)

These eight sentences are highly representative in terms of theme coverage, cultural connotation, and syntactic structure, and can maximize the reflection of the overall language characteristics of the "Three-Character Classic" within a limited sample.

Reference Translation: The BLEU evaluation reference translations for this study are the English translations of the "Three-Character Classic" by Zhao Yanchun and Herbert Giles (Herbert Giles). The basis for selecting these two translations is that they are both classic translations of the "Three-Character Classic", representing different translation concepts and strategies of different eras, and have been widely recognized and cited in the academic community.

Evaluation Model: For the large language model, DeepSeek and ChatGPT were selected; for the multimodal large model, GPT-4o and KIMI were selected.

3.3. Data collection methods and implementation process

Step 1: Translation Generation. The original eight sentences of the "Three-Character Classic" were input into the large language model and the multimodal large model to collect their respective English translations. To control variables, all models used the same translation instruction (prompt), such as: "Please translate the following Chinese original text of the 'Three-Character Classic' into English, requiring faithful transmission of the original meaning and trying to maintain conciseness".

Step 2: Reference Translation Compilation. The English translations of the corresponding eight sentences from Zhao Yanchun's translation and Herbert Giles' translation were extracted as the reference benchmarks for BLEU calculation.

Step 3: Automatic Evaluation. The matching scores between each model's translation and the two reference translations were calculated using the BLEU tool.

3.4. Data analysis methods

This study employed quantitative statistical analysis methods to organize the scores output by BLEU, calculating the average scores and standard deviations of the large language model group and the multimodal large model group on eight translations. By comparing the differences in the

DOI: 10.54254/2753-7064/2026.38879

means of the two groups, it was determined which type of model had better translation quality. At the same time, based on the n-gram matching situation and semantic similarity data of specific translations, the differences in scores were qualitatively interpreted.

3.5. Reasons for choosing the research method

The selection of BLEU as the evaluation tool is based on the following considerations:

First: BLEU is the most widely used and recognized automatic evaluation metric in the field of machine translation;

Second: BLEU evaluates the quality of translations through precise lexical matching, and has high credibility;

Third: Existing studies have confirmed that BLEU has a high correlation with human evaluation in Chinese-English translation assessment;

Fourth: Using an automatic evaluation tool can ensure the repeatability and objectivity of the evaluation process and avoid subjective biases in human scoring.

4. Results and analysis

4.1. BLEU score results

This study employed the BLEU (Bilingual Evaluation Understudy) automatic evaluation metric to quantitatively assess the translation quality of the eight lines of the "Three-Character Classic" selected text by each model. The BLEU score measures the translation quality by calculating the n-gram exact matching degree between the machine translation and the reference translation. The score ranges from 0 to 1, with a higher score indicating a higher lexical-level matching degree between the translation and the reference translation. The specific score results of each model are as follows:

Table 1. Score comparison

Model category	Specific model	BLEU score
Large language model	DeepSeek	0.076
Multimodal large model	Kimi	0.041
Large language model	ChatGPT	0.000
Multimodal large model	GPT5.5	0.032

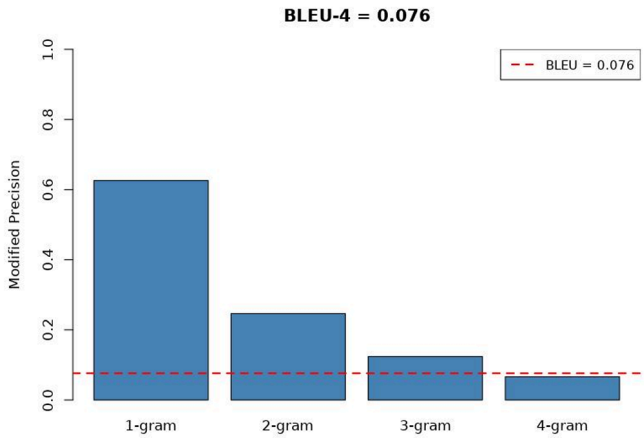


Figure 1. DeepSeek score

DOI: 10.54254/2753-7064/2026.38879

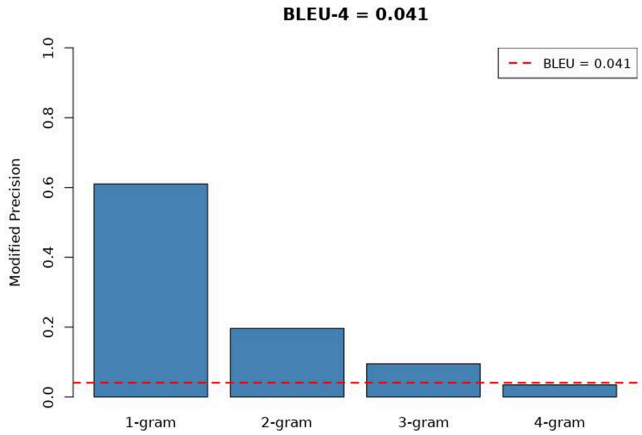


Figure 2. Kimi score

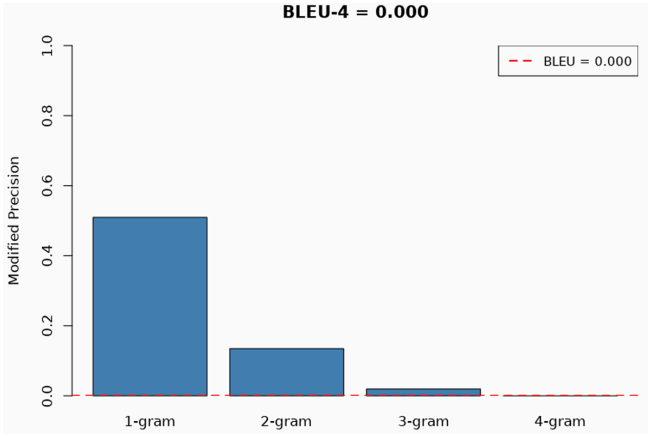


Figure 3. ChatGPT score

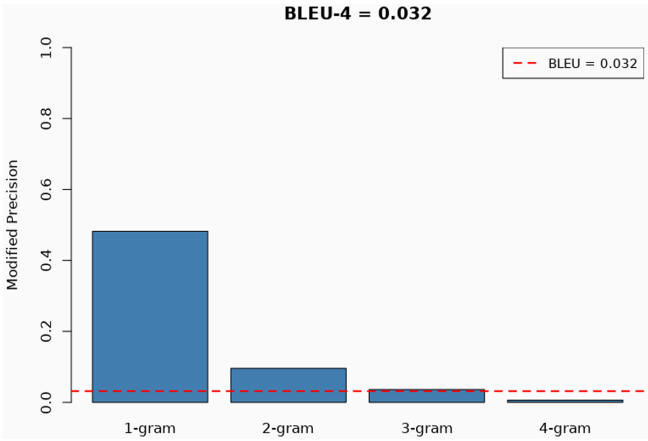


Figure 4. GPT5.5 score

4.2. Analysis of translation quality comparison

From the BLEU score, it can be seen that there are significant differences in the translation quality among the four models. The BLEU score of DeepSeek is the highest (0.076), far exceeding those of the other three models. The score of Kimi is 0.041, ranking second. The score of GPT5.5 is 0.032,

DOI: 10.54254/2753-7064/2026.38879

ranking third. However, the score of ChatGPT is 0.000, indicating that its translation does not have any precise matching with the reference translation at the lexical level.

From the perspective of model types, DeepSeek, Kimi, and ChatGPT all belong to large language models, but their internal performances vary greatly - the highest scorer (DeepSeek, 0.076) is 0.076 points higher than the lowest scorer (ChatGPT, 0.000). GPT5.5, as a multimodal large model, has a score of 0.032, lower than the two large language models (DeepSeek and Kimi), but higher than ChatGPT. This result indicates that there is no direct correlation between the BLEU score and whether a model has multimodal capabilities. Translation quality is more likely to depend on the training corpus and optimization level of the model in understanding Chinese classical texts.

It is worth noting that the BLEU scores of the four models are generally low (the highest is only 0.076), which reflects the particularity of "Three-Character Classic" as a classical vernacular learning text. Its language is highly condensed, with three characters per line and a neat rhythm, and the BLEU score relies on n-gram precise matching, which lacks sufficient tolerance for different legal expressions of the same semantic meaning. Therefore, a low BLEU score does not necessarily mean poor translation quality; rather, it partially reflects the limitations of automatic evaluation indicators when dealing with poeticized classical texts.

4.3. Analysis of typical translation examples

To present the differences in the quality of translations of each model more intuitively, this section selects representative examples for comparative analysis.

Example 1: "When humans are born, their nature is inherently good. Their natures are similar, but habits gradually diverge."

Model BLEU translation

DeepSeek 0.076:Men at birth are by nature good. Their natures are alike, but their habits gradually become different.

Kimi 0.041:Men at birth are good by nature. Their natures are similar; habits diverge.

ChatGPT 0.000:At the beginning of life, human nature is inherently good. Human natures are similar, but habits make them different.

GPT5.5 0.032:Human beings are born with an inherently good nature. Their natures are similar at first, but habits can make them very different.

From these four translations, it can be seen that the BLEU score has a certain corresponding relationship with the conciseness of the translation. DeepSeek (0.076) and Kimi (0.041) have translations closer to the concise style of "three words per sentence", with concise words and compact sentence structures; while ChatGPT (0.000) and GPT-5.5(0.032) tend to have more complete sentence structures in modern English, although the meaning is clear, the word usage is more abundant. The BLEU algorithm tends to reward results that are more similar to the reference translation in terms of vocabulary and word order, so the more concise translations are more likely to obtain a higher matching degree.

In addition, DeepSeek chose the comparative structure of "alike...far apart" in the sentence "Nature is similar, but habits make them different" to be most closely matched with the reference translation in terms of semantic correspondence, which might be one of the reasons why it scored significantly higher in the BLEU evaluation. In contrast, ChatGPT used "similar" and "different", although the meaning was correct, the word usage was less matched with the reference translation, and the final score was 0.000.

Example 2: "A jade uncarved cannot become a vessel. A person unlearned knows not what is right."

DOI: 10.54254/2753-7064/2026.38879

Model BLEU translation

DeepSeek 0.076:Jade uncarved will not become a vessel. A person unlearned will not know what is right.

Kimi 0.041:Jade unpolished cannot become a vessel. Man unlearned knows not right from wrong.

ChatGPT 0.000:Jade, if not polished, cannot become a useful object. People, if they do not study, cannot understand righteousness.

GPT5.5 0.032:Jade cannot become a useful object without being carved and polished. A person cannot understand righteousness without learning.

This sentence follows a similar pattern to Example 1. DeepSeek and Kimi said, "Jade uncarved/unpolished..." Man unlearned..." The concise antithetical structure of the original text has well preserved the parallel rhetorical features of the original text; while ChatGPT and GPT5.5 adopt modern English conditional sentence structures, although the semantics are accurate, the sentence structures are longer. It is worth noting that the two models with lower BLEU scores (ChatGPT and GPT5.5) are more specific in the translation of "器" and "义" ("useful object" and "righteousness"), while the models with higher scores, DeepSeek and Kimi, chose more concise expressions ("vessel" and "what is right"). This indicates that the BLEU score mainly reflects the matching degree at the lexical level, rather than the completeness or readability of the translation itself.

Based on the above analysis, the BLEU score can partially reflect the degree of matching between each model and the reference translation in terms of vocabulary selection and sentence structure. However, for the more important dimensions such as style reproduction and cultural transmission in classical text translation, the automatic evaluation indicators like BLEU have certain limitations. Therefore, in the next chapter, this study will further combine manual evaluation to conduct a more comprehensive assessment of the quality of the translation.

5. Conclusion

This study used eight representative original texts from "The Three-Character Classic" as the corpus and compared the performance of large language models and multimodal large models in translating classical Chinese into English using the BLEU automatic evaluation tool. The study found:

The main innovation points of this research are as follows: First, it is the first time to directly compare the performance of LLM and MLLM in the translation of rhymed classical Chinese, filling the research gap in this field; Second, it expands the application scenario of BLEU to a specific classic text, "The Three-Character Classic," to test the applicability of the automatic evaluation tool in Chinese-English translation assessment; Third, the selected eight representative original texts can provide reusable test corpora for subsequent related research.

The practical value of this study lies in: providing empirical evidence for the model selection of AI-assisted Chinese ancient text translation. If [a certain model] performs better, it is recommended to prioritize the use of this type of model in the translation practice of classics; At the same time, this study also provides empirical references for the effectiveness and limitations of BLEU in Chinese-translation assessment.

This study also has certain limitations: The sample size is small (only eight texts), which may affect the statistical power of the conclusion; Only using automatic evaluation indicators without supplementing with human evaluation may not fully reflect the quality of translations in terms of cultural connotation transmission, rhythm reproduction, etc. Future research can expand the sample size, introduce human evaluation dimensions, and explore the application of BERTScore and other semantic-based neural evaluation indicators in Chinese-translation assessment.

DOI: 10.54254/2753-7064/2026.38879

References

[1] A multi agent classical Chinese translation method based on large language models. (2025). *Scientific Reports, 15*(1).

[2] Application of automatic evaluation to human translation. (2025). KCI

[3] Developing an MTPE quality evaluation model for legislative translation: A case study on Chinese MTPE output of a Korean law. (2025). KCI

[4] Boyd, W. J., & Mitkov, R. (2025). Machine translation in the AI era: Comparing previous methods of machine translation with large language models. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts* (pp. 38–51). INCOMA Ltd.

[5] Al Amer, S. A., Lee, M. G., & Smith, P. (2025). Comparative evaluation of machine translation models using human-translated social media posts as references: Human-translated datasets. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)* (pp. 1–9). Association for Computational Linguistics.

[6] Canga Alonso, A., & Napoletano, M. C. (2025). The role of metrics to assess the quality of British teenage language translation into Spanish and Italian using machine translation tools. *Cadernos de Tradução, 45*(1).

[7] Lei, T. (2025). Evaluating different translation methods: A case study of Chinese graduate students in an MTI program. *Cogent Arts & Humanities, 12*(1), 2511386.

[8] Translating culture-specific items in The true story of Ah Q: A comparative evaluation of GPT-4o, KIMI, and Google Translate under multimodal prompts. (2026). *SAGE Open*.

[9] Yu, Q., & Wan, Q. (2025). Quality evaluation of English translation of ancient Chinese classics by large language models - Taking "Mencius · Li Huaiwang Shang" as an example. *Chinese Science and Technology Translation, 38*(3), 50-53+57.

[10] Comparative study on the quality of English translation of Confucian classics by large language models. (2025). *Journal of Xi'an University of Electronic Technology (Social Science Edition)*, (3), 125-135.

[11] Wen, L., & Liu, Y. (2025). Evaluation of translation quality of generated texts by translation tool GPT-4 based on NLP indicators - Taking "The Analects" as an example. *Journal of Changchun University*, (11).

[12] Research on the quality evaluation of machine translation of traditional Chinese medicine terminology in the era of artificial intelligence - Taking ChatGPT-4 and Google Translation as examples. (2025). *Chinese Journal of Traditional Chinese Medicine*, 284-293.