Journal of Contemporary Educational Research, 2025, Volume 9, Issue 4
http://ojs.bbwpublisher.com/index.php/JCER
ISSN Online: 2208-8474
ISSN Print: 2208-8466

A Study on the Performance of Large Language Models in Translating the Texts of Red Culture into English

Yiming Li, Yuanpeng Huang*

Shanxi Normal University, Taiyuan 030031, Shanxi, China

***Corresponding author:** Yuanpeng Huang, 570291955@qq.com

Abstract: This paper takes Chinese red culture resources as its research subject and focuses on evaluating the Chinese-English translation quality of three major AI platforms: ChatGPT-4.0, ERNIE Bot, and DeepSeek. Through automatic quantitative evaluation, it systematically analyzes their performance in translating red culture texts. The study selects a diverse range of corpora, including historical documents, red classic texts, and culturally loaded terms. Three automatic evaluation metrics—GLEU, METEOR, and COMET—are employed for a comprehensive assessment.

Keywords: Red culture; Translation; AI translation evaluation; ChatGPT; ERNIE Bot; DeepSeek

Online publication: April 29, 2025

1. Introduction

Language barriers present a persistent challenge in our increasingly globalized world, often impeding the free flow of information and cultural exchange [1]. As individuals and organizations seek to expand their reach beyond national borders, the necessity for accurate and efficient translation becomes ever more pressing. In recent years, the rapid evolution of artificial intelligence (AI) and natural language processing (NLP) technologies has given rise to powerful AI-driven machine translation platforms. These tools are transforming the way people access, communicate, and disseminate knowledge across linguistic and cultural divides. Among the various applications of AI in language services, the translation of culturally specific material is especially pivotal in ensuring that audiences can fully grasp the societal, historical, and ideological nuances embedded within the source texts.

This paper concentrates on one such domain of culturally rich content: Chinese red culture resources. "Red culture" in the Chinese context refers to heritage and resources that symbolize revolutionary spirit, socialist values, and patriotic traditions. These resources typically include narratives of historical figures, accounts of revolutionary movements, and literary works that celebrate national pride and reflect socio-political ideologies. They hold not

only historical significance but also embody cultural and ideological values that resonate strongly with Chinese identity. For this reason, translating red culture texts into other languages—particularly English—is a delicate task. It requires maintaining the linguistic fidelity, conveying nuanced cultural and historical references, and preserving the emotional resonance that red culture texts often evoke among native Chinese speakers.

In response to this challenging task, three AI platforms have emerged as frontrunners in providing machine translation services: ChatGPT-4.0, ERNIE Bot, and DeepSeek. Developed by leading research entities, these platforms leverage state-of-the-art AI models that have undergone extensive training on diverse linguistic corpora. Their architectures include sophisticated encoder-decoder mechanisms and attention-based frameworks, designed to capture syntactic structures and semantic meanings. By doing so, they each strive to deliver translations that are simultaneously fluent, accurate, and culturally sensitive. Nevertheless, despite these advancements, questions remain as to how well these platforms perform when tasked with highly specialized, context-laden content such as Chinese red culture texts.

To address this gap in the literature, this paper employs an automatic and quantitative evaluation of translation quality, focusing on key metrics that are widely accepted in the NLP community. Three metrics—GLEU, METEOR, and COMET—are used to measure the performance of ChatGPT-4.0, ERNIE Bot, and DeepSeek on a diverse corpus representative of Chinese red culture resources. GLEU and METEOR both offer insight into aspects of lexical choice, syntactic structure, and semantic congruence, while COMET incorporates aspects of contextual and cross-lingual alignment to capture subtle discrepancies that might otherwise go unnoticed. By combining these three well-regarded metrics, the research provides a multilayered evaluation that encompasses fluency, adequacy, and semantic alignment—three crucial dimensions in any assessment of translation performance.

In constructing the dataset for this study, particular care was taken to ensure a representative and varied selection of texts. The corpora encompass revolutionary historical documents, red classic literature, and culturally loaded expressions—each holding distinctive challenges for machine translation. Historical documents often contain archaic language and references to pivotal events, whereas red classic literature demands a nuanced portrayal of sentiment and ideology. Meanwhile, culturally loaded expressions, such as idiomatic phrases and symbolic references, test the platforms' capacity to capture meaning that extends beyond direct lexical correspondence. Together, these texts form a rigorous test bed for evaluating the capacity of AI translation tools to handle culturally and historically significant material.

By systematically investigating how these platforms fare in translating red culture content, this paper offers insights into their strengths and limitations. It aims to guide users—ranging from researchers and educators to policy-makers and cultural organizations—in selecting and deploying AI translation tools that best suit their specific needs. Equally important, the research sheds light on areas that require further improvements in both technology and methodology. Concluding with practical recommendations for users and strategic suggestions for future development, the study endeavors to support the digital transformation and global dissemination of Chinese red culture. Through this analysis, it contributes to the broader discourse on cross-cultural communication and underscores the immense potential that AI translation holds for bridging linguistic and cultural boundaries in an ever-more interconnected world.

2. Literature review

Translation quality assessment (TQA) has been one of the central topics in Translation Studies (TS), spawning

heated debate among scholars, educators, and practitioners over the past decades [2]. It is essential to measure the quality of a machine translation system to improve how it performs. There is, however, a great lack of consensus and standardization in relation to translation quality assessment—both human and machine—given the complicated cognitive, linguistic, social, cultural, and technical process this supposes. According to Lauscher, translation quality assessment will mean a constant to and from a macro-analytic approach, wherein questions of ideology, function, gender, or register are considered, to a micro-analytical one in which the value of collocations and individual linguistic units are considered. Nevertheless, it should be taken into account that these approaches can differ enormously according to the individuals, groups, or contexts in which quality is assessed [3]. Thus, quality assessment in the industry is normally focused on the final product or customer, whereas in the field of research, the purpose can be to demonstrate significant improvements over prior studies or different translation processes.

In recent years, neural machine translation (NMT) based on deep learning has gradually become mainstream, achieving breakthroughs in semantic understanding and contextual awareness through training on massive corpora. For instance, Baidu's Chinese-English translation system significantly improved its handling of Chinese-specific vocabulary (such as idioms and colloquialisms) from 2017 to 2019 by optimizing its models. Similarly, domestic large models like DeepSeek have enhanced their ability to recognize technical terms and disciplinary associations through the construction of knowledge graphs, and have already been applied in the automatic summarization of academic literature. These technologies provide a technical foundation for translating complex historical terms in red culture texts. Such texts often contain unique expressions related to revolutionary historical events, political slogans, and material cultural terms (e.g., "Hongchuan Jingshen," "Jinggangshan Huishi"), which pose challenges due to semantic gaps and difficulties in conveying intended meaning in the target language. Research shows that while ChatGPT can produce literal translations of material culture terms (e.g., "Tu Su Jiu," "He Huan Tang"), it often fails to convey the deeper cultural connotations, necessitating human annotation or interpretive translation strategies to compensate. Moreover, red classic texts frequently involve ideological expressions, and machine translation may misrepresent politically sensitive content due to a lack of contextual understanding, highlighting the need for dynamic translation adjustment mechanisms.

Many scholars have already begun to explore the opportunities and challenges that ChatGPT brings to language teaching and academic writing. Salvagno and Taccone [4] argued that ChatGPT can assist writers in organizing materials, generating drafts, and proofreading, while also acknowledging risks such as plagiarism and inaccuracy. Thorp [5] had explicitly expressed serious concerns about the use of ChatGPT in academic writing, asserting that ChatGPT cannot replace the role of the author. Scholars also provided a detailed explanation of ChatGPT's technical principles, engineering framework, and related experiments, offering a comprehensive analysis of its potential and limitations in natural language processing, including in the field of translation.

The machine translation service provided by ChatGPT goes beyond that of conventional translation tools, making its performance in translation quality, proofreading, and sentence optimization a new focal point in the fields of machine translation and translation studies. However, existing literature focusing on ChatGPT and translation remains limited and mainly centers on two aspects.

The first aspect involves translation quality assessment using automatic evaluation metrics such as BLEU, TER, and CHRF. Jiao *et al.* [6] evaluated ChatGPT's performance in translation prompting, multilingual translation, and translation robustness. They found that among three prompt styles, the instruction

"Please provide the [TGT] translation for these sentences" yielded the best machine translation results. Ghassemiazghandi evaluated ChatGPT-4's Persian-English translation accuracy using BLEU and accuracy scores, comparing it to MateCat. Results show ChatGPT-4 outperforms traditional tools. The findings highlight the effectiveness of large language models in enhancing translation quality and AI-driven translation technologies [7]. Nemergut compared post-editing and from-scratch translation quality of English-to-Slovak NMT output using Translation Edit Rate. Findings show weak time-TER ties and call for explicit acceptability limits in training and pricing [8]. Wang *et al.* described the first "Ancient Chinese Translation Bakeoff," offering new parallel corpora and two tasks—ancient-to-modern Chinese and ancient-to-English using automatic evaluation metrics BLEU. Results underline low-resource obstacles, benefits of data augmentation, and the promise of fine-tuned large models [9].

The second aspect concerns the advantages and challenges ChatGPT brings to language and translation. Ali surveyed 58 Saudi EFL faculty and found that ChatGPT enhances English learning by improving motivation, writing, and vocabulary skills, offering quick feedback, and saving time. However, concerns include plagiarism, reduced originality, and overreliance on AI, which may hinder critical thinking and ethical learning practices in language education [10]. Sahari *et al.* surveyed 34 Saudi translation teachers and students, finding that ChatGPT can enhance translation by improving speed, accuracy, and editing efficiency. It supports learning and idea generation. On the other hand, challenges include reduced creativity, limited cultural nuance handling, and ethical concerns like plagiarism [11].

While these studies have analyzed the significant impact of ChatGPT on the field of translation, few have paid attention to its translation quality when applied to Chinese-specific discourse. In the context of China's entry into a new era, the theoretical framework and macro perspective for exploring and reflecting on new technologies still require further development.

Current studies evaluating ChatGPT's translation quality either rely on the author's own assessment or use only one or two machine evaluation methods, lacking human evaluation. Traditional metrics like BLEU and TER rely on lexical matching and are insufficient for measuring cultural adaptability. For example, directly translating "Qunzhong Luxian" as "mass line" may be lexically accurate, but cultural differences may lead target readers to misunderstand its political connotation. Research shows that automatic evaluation can serve only as a preliminary screening tool in the translation of red culture texts and must be supplemented by human evaluation. Wang and Wen [12] argued that BLEU should not be used alone to assess translation quality and must be complemented with other similar metrics. Moreover, since machine evaluation tends to be rigid and reflects only partial aspects of a translation, the scores can vary significantly depending on text complexity, and semantic nuances are often missed. Therefore, human evaluation remains essential.

More critically, most current academic studies rely on ChatGPT-3.5 to evaluate translation capabilities. Akhtarshenas *et al.* [13] noted that ChatGPT-3.5 is based on a model with 175 billion parameters, while the newly released GPT-4 is expected to exceed 100 trillion parameters. Once GPT-4 reaches the 100 trillion-parameter scale, it would approach the scale of neural connections in the human brain. Given the substantial capability gap between ChatGPT-3.5 and ChatGPT-4, evaluations based on the former may lack accuracy and reliability in light of the latest developments.

In short, current research reveals both the potential and limitations of AI translation in the domain of red culture. On the technical level, NMT models can handle basic terminology, but cultural adaptability still needs improvement. On the methodological level, there is a need to construct an integrated evaluation system combining automated metrics and multidimensional human assessment. Therefore, this study takes Chinese

red culture resources as its research subject and focuses on evaluating the Chinese-English translation quality of three major AI platforms—ChatGPT-4.0, ERNIE Bot, and DeepSeek—by systematically analyzing their performance in translating red culture texts.

3. Research design

This study adopts an automatic evaluation approach to compare and analyze the GLEU, METEOR, and COMET scores of translations produced by ChatGPT-4.0 and two other large language models on the same corpus, in order to assess the translation quality of ChatGPT-4.0. It further compares the strengths and weaknesses of these large language models in translating red culture resources and provides corresponding recommendations for their effective use.

3.1. Research questions

This study aims to address the following two questions:

(1) How does the translation quality of ChatGPT-4.0 compare with that of other large language models?

(2) What are the respective strengths and weaknesses of these large language models in translating red culture texts?

3.2. Research methodology

This study employs automatic evaluation of machine translation quality and comparative analysis. Given that large language models can generate different outputs for the same prompt each time—avoiding uniformity and repetition—three translation outputs are collected under the same prompt to account for this variability. The average score of these translations is reported to ensure the validity and reliability of the evaluation results.

3.2.1. Translation texts

The translation test texts in this study are selected from the Chinese and English introductions of various red tourism sites in Shanxi Province. A total of 7,896 Chinese characters and 5,443 corresponding English reference characters are collected. The dataset includes 78 terms for word-level translation, 151 sentences for sentence-level translation, and 10 paragraphs for paragraph-level translation.

3.2.2. Machine translation tools

This study compares the translation performance of ChatGPT-4.0 with two other large language models: ERNIE Bot and DeepSeek. These three machine translation tools support real-time translation across 133, 109, and 29 languages, respectively. All data used in this study for ChatGPT are based on the GPT-4.0 version, with data collection completed by February 28, 2025.

The rapid development of large language models has revolutionized code intelligence in software development. However, the predominance of closed-source models has restricted extensive research and development. The DeepSeek-Coder series, a range of open-source code models with sizes from 1.3B to 33B, trained from scratch on 2 trillion tokens. These models are pre-trained on a high-quality project-level code corpus and employ a fill-in-the-blank task with a 16K window to enhance code generation and infilling [14]. DeepSeek-Coder not only achieves state-of-the-art performance among open-source code models across multiple benchmarks but also surpasses existing closed-source models like Codex and GPT-3.5. Furthermore, DeepSeek-Coder models are under a permissive license that allows for both research and unrestricted

commercial use. To ensure the preservation of LLM capabilities during pretraining, DeepSeek investigates an effective VL pretraining strategy by integrating LLM training from the beginning and carefully managing the competitive dynamics observed between vision and language modalities. Starting with a focus on text, it gradually adjusts the ratio to facilitate a balanced integration of both modalities. The DeepSeek-VL family (both 1.3B and 7B models) showcases superior user experiences as a vision-language chatbot in real-world applications, achieving state-of-the-art or competitive performance across a wide range of visual-language benchmarks at the same model size while maintaining robust performance on language-centric benchmarks [15]. It has made both 1.3B and 7B models publicly accessible to foster innovations based on this foundation model.

ERNIE Bot is an AI-powered large language model developed by Baidu, based on the ERNIE (Enhanced Representation through Knowledge Integration) architecture. Unlike traditional language models that rely solely on text data, ERNIE is trained by deeply integrating structured knowledge, such as facts from knowledge graphs, enabling it to better understand context, reason logically, and generate more accurate responses. This allows it to understand language with deeper semantic awareness and deliver more accurate, informed responses. ERNIE Bot demonstrates particularly strong performance in Chinese language tasks, including natural language understanding, summarization, translation, and content generation, often outperforming other LLMs in this domain. Through integration with Baidu's search engine and real-time data, ERNIE Bot can provide more current and context-aware information compared to some static models.

3.2.3. Assessment methods

This paper uses three methods to assess the quality of translation. GLEU is a variation of BLEU developed by Google to evaluate machine translation quality. Unlike traditional BLEU, which favors longer n-gram precision, GLEU balances both precision and recall by penalizing missing words more effectively. It is particularly suitable for comparing individual sentence-level translations against references. General Language Understanding Evaluation (GLUE): a collection of NLU tasks including question answering, sentiment analysis, and textual entailment, and an associated online platform for model evaluation, comparison, and analysis [16]. GLUE does not place any constraints on model architecture beyond the ability to process single-sentence and sentence-pair inputs and to make corresponding predictions. For some GLUE tasks, training data is plentiful, but for others, it is limited or fails to match the genre of the test set. GLUE, therefore, favors models that can learn to represent linguistic knowledge in a way that facilitates sample-efficient learning and effective knowledge transfer across tasks.

METEOR, an automatic metric for machine translation evaluation that is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations [17]. Unigrams can be matched based on their surface forms, stemmed forms, and meanings; furthermore, METEOR can be easily extended to include more advanced matching strategies. Once all generalized unigram matches between the two strings have been found, METEOR computes a score for this matching using a combination of unigram-precision, unigram-recall, and a measure of fragmentation that is designed to directly capture how well-ordered the matched words in the machine translation are in relation to the reference. We evaluate METEOR by measuring the correlation between the metric scores and human judgments of translation quality [17]. We compute the Pearson *R* correlation value between its scores and human quality assessments of the LDC TIDES 2003 Arabic-to-English and Chinese-to-English datasets. It improves upon BLEU by considering synonymy, stemming, and word order. It aligns hypotheses and reference translations at the word level using modules like exact match, stem match, synonym match (via WordNet), and paraphrase match. METEOR is known for correlating well with human judgment, especially in fine-grained evaluations.

COMET, a PyTorch-based framework for training highly multilingual and adaptable machine translation evaluation models that can function as metrics [18]. It takes advantage of recent breakthroughs in cross-lingual language modeling to generate prediction estimates of human judgments such as Direct Assessments (DA) [19], Human-mediated Translation Edit Rate (HTER) [20], and metrics compliant with the Multidimensional Quality Metric framework [21]. It is a newer, neural-based evaluation metric that leverages pre-trained language models and supervised learning. It considers both the source sentence and the reference when scoring the hypothesis, making it more context-aware. COMET has been shown to outperform traditional metrics in correlating with human evaluations.

4. Research results and discussion

4.1. Automatic evaluation results

Unlike traditional machine translation software that directly outputs translations of the original input sentences, large language models (LLMs) require specific prompt templates to format the input for translation generation. In this study, the prompt "Please provide the [TGT] translation for these sentences:" is used as the direct translation cue for ChatGPT to obtain its translation outputs. These are then compared with the outputs of the other two emerging LLMs. The quality of each translation is evaluated using GLEU, METEOR, and COMET scores. The results are shown in **Figure 1**.

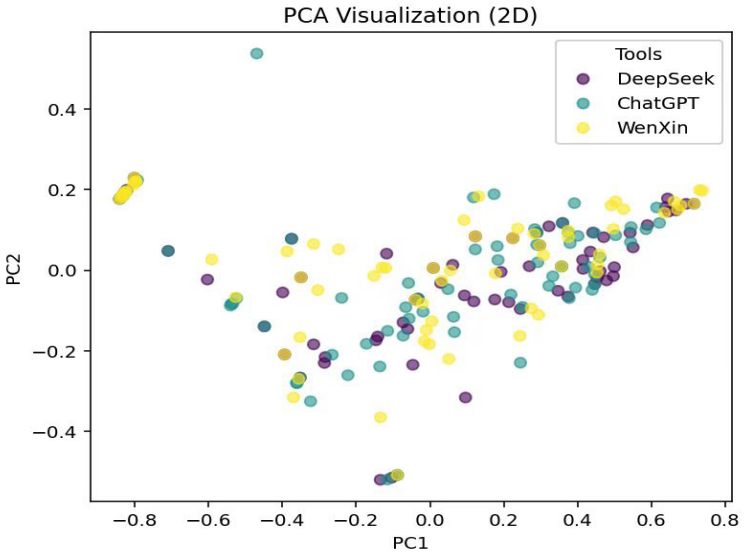

Figure 1. Translation quality metrics of ChatGPT and other machine translation tools

As shown in **Figure 1**, ERNIE Bot outperforms the other two large language models across all three evaluation metrics. This indicates that its translations demonstrate the highest performance in terms of lexical and grammatical accuracy, linguistic coherence, and cultural adaptability. ChatGPT ranks second, while DeepSeek's translations show slightly weaker performance in comparison.

4.2. Strengths and weaknesses of various large language models in translating red culture resources

4.2.1. Words

Based on the characteristics of red culture resources, the translation tasks are categorized into two types: word-

level and sentence-level translation. As shown in **Table 1**, ERNIE Bot achieves the highest scores across all three metrics in the translation of culturally loaded terms, indicating its superior performance in cultural adaptability.

Table 1. Performance of three large language models in translating red culture vocabulary

Tools	GLUE	METEOR	COMET
DeepSeek	0.519	0.616	0.779
ChatGPT	0.483	0.630	0.813
ERNIE Bot	0.587	0.682	0.800

The term "Xunfu" is officially translated as "governor." DeepSeek translates it as "provincial governor (in imperial China)," ERNIE Bot provides "governor," and ChatGPT outputs "Provincial Governor." ERNIE Bot's translation aligns exactly with the official version. While "Provincial Governor" with an explanatory note referencing "imperial China" is not identical to the official translation, it is still relatively accurate and aids comprehension. This is because "Xunfu" in the Ming and Qing dynasties was indeed responsible for provincial administration, akin to a modern-day governor. DeepSeek's translation, which includes additional context, enhances understanding and acceptability.

The official translation of "Wuxiang Gushu" is "Wuxiang Dagu Storytelling." ChatGPT translates it as "Wuxiang Drum Opera," DeepSeek provides "Wuxiang Drum Storytelling (a traditional folkart form)," and ERNIE Bot renders it as "Wuxiang Dagu—a traditional Chinese ballad accompanied by drum music." Although the translations by DeepSeek and ERNIE Bot do not exactly match the official version, they capture the core meaning of "Wuxiang Gushu" and provide helpful explanations of this culturally loaded term. In contrast, ChatGPT's version, "Wuxiang Drum Opera," reflects a weaker understanding of Chinese culture and context, which leads to a lower evaluation score compared to the other two models.

The official translation of "Jiefang Zhanzheng" is "the War of Liberation." Both ERNIE Bot and DeepSeek translate it as "Chinese Civil War (or Liberation War, depending on context)," while ChatGPT renders it as "the Chinese Civil War." When prompted to explain its choice, ChatGPT provided the following reasoning: "The Chinese Civil War" is a more neutral and commonly used translation, widely adopted to describe the civil conflict in China between 1945 and 1949. It avoids aligning with any particular perspective and is typically used in historical research and international discourse. From a neutral or global historical standpoint, "the Chinese Civil War" is the most widely recognized and accepted term. In contrast, "the War of Liberation" reflects the Chinese Communist Party's perspective, emphasizing the notion of liberation, and is commonly found in official CCP narratives. In certain contexts—especially in international historical discussions—"the Chinese Civil War" is more appropriate.

ChatGPT is developed by OpenAI, a company primarily composed of non-Chinese researchers. As such, the ideological framing reflected in its translation is shaped by a Western perspective of China. This leads to translations that, while fluent and contextually neutral from an international standpoint, often lack alignment with Chinese-specific political and cultural contexts. In this case, ChatGPT's translation is not suitable for use in a red culture discourse setting due to its lack of contextual and ideological accuracy, despite maintaining grammatical fluency.

The official translation of "Geming Xianbei" is "revolutionary forefathers." ChatGPT translates it as "revolutionary pioneers," which is relatively close to the official version. DeepSeek offers "Revolutionary

Pioneers (or Revolutionary Forerunners)," while ERNIE Bot provides "revolutionary predecessors or revolutionary ancestors."

Literally, "Geming Xianbei" refers to those elder figures who contributed to the revolutionary cause, typically including individuals who struggled and sacrificed for the revolution. While some may suggest translating it as "martyrs of the revolution," this expression only conveys the idea of sacrifice and fails to capture the broader meaning of "Xianbei," which encompasses not only those who gave their lives but also those who made significant contributions without necessarily sacrificing themselves. Therefore, "martyrs" is not a comprehensive or accurate translation, as it narrows the scope and overlooks the broader connotation of "revolutionary forebears."

The official translation of "Beiyang Junfa" is "the Northern Warlords." Both ChatGPT-4.0 and DeepSeek translate it as "Beiyang Warlords," while ERNIE Bot's translation matches the official version exactly. ERNIE Bot demonstrates strong web-based search capabilities, enabling it to accurately handle culturally loaded terms and specialized terminology. In this case, its ability to align with the official translation reflects a better adaptation to established naming conventions and historical discourse.

The official translation of "Kangri Minzhu Zhengquan" is "Anti-Japanese Democratic Government." ChatGPT-4.0 translates it as "the Anti-Japanese Democratic Government," which is very close to the official version. DeepSeek provides a translation that exactly matches the official one, while ERNIE Bot renders it as "Anti-Japan Democratic political power."

The phrase "Kangri Minzhu Zhengquan" refers to the form of government established during the War of Resistance Against Japan, particularly in the anti-Japanese base areas in China. It emphasizes an organized political system and governmental structure. Although ERNIE Bot has strong web-searching capabilities, its weaker core translation ability sometimes leads to inaccuracies. In this case, its use of the phrase "political power" instead of "government" fails to accurately convey the structured, institutional nature of the term, resulting in a translation that is less precise and less aligned with the historical context.

"The New Yangge Movement" is the official translation of "Xinyangge Yundong." Both DeepSeek and ERNIE Bot use the official translation, while ChatGPT translates it as "New Yangge Movement," which is an inaccurate rendering of this culturally loaded term.

"The Taihang Taiyue Martyrs Cemetery" is the official translation of "Taihang Taiyue Lieshi Lingyuan." Both DeepSeek and ERNIE Bot follow the official version, whereas ChatGPT renders it as "Taihang and Taiyue Martyrs' Cemetery," showing an insufficient understanding of the Chinese structure.

The term "strategic fulcrum" is the official translation of "Zhanlue Zhidian." While ChatGPT and DeepSeek translate it as "strategic pivot," ERNIE Bot uses the official version. The official translation more accurately conveys the military connotation of the term.

For the 78th anniversary of China's victory in the Chinese People's War of Resistance Against Japanese Aggression (1931–1945) and the World Anti-Fascist War, both DeepSeek and ERNIE Bot use the official translation, while ChatGPT renders it as "78th Anniversary of the Victory of the Chinese People's Anti-Japanese War and the World Anti-Fascist War," which differs from the official version.

The official translation of "Jubu Kangzhan" is "Regional War of Resistance against Japanese Aggression." ChatGPT translates it as "local resistance," DeepSeek renders it as "Partial War of Resistance," while ERNIE Bot's translation matches the official version exactly. The official translation emphasizes both the regional scope and the confrontational nature of the resistance. In contrast, ChatGPT's and DeepSeek's translations simplify or distort the term's historical and political implications, making them less precise in this context.

For “Dashengchan Yundong,” the official translation is “the Great Production Campaign.” Both ChatGPT and DeepSeek translate it as “the Great Production Movement,” while ERNIE Bot aligns exactly with the official version. ERNIE Bot explains that this translation directly corresponds to the historical name of the movement and accurately reflects the historical context, making it the most appropriate rendering in this case.

4.2.2. Sentences

As shown in **Table 2**, ERNIE Bot achieves the highest scores across all three metrics in sentence translation, indicating stronger performance in grammatical accuracy, logical structure, and sentence coherence. Unlike in word-level translation, the differences among the three large language models are relatively small in sentence-level translation. This suggests that the overall performance of the models is quite comparable in this aspect, and the most significant variation still lies in the translation of vocabulary and culturally loaded terms.

Table 2. Performance of three large language models in translating red culture sentences

Tools	GLUE	METEOR	COMET
DeepSeek	0.277	0.556	0.752
ChatGPT	0.278	0.557	0.809
ERNIE Bot	0.292	0.577	0.811

The official translation is “pay tribute to the revolutionary martyrs and passing on the traditions of revolution.” DeepSeek translates it as “Remember the revolutionary martyrs and carry forward their revolutionary spirit.” ERNIE Bot provides two versions: “Cherish the memory of our revolutionary martyrs and carry forward their spirit” and “Commemorate the martyrs and inherit their revolutionary spirit.”

ChatGPT’s translation—“Commemorating the Revolutionary Martyrs and Inheriting the Revolutionary Spirit”—received the highest score among the models. Its version maintains both a formal tone and clear parallel structure, accurately conveying the original meaning while aligning well with typical expressions found in official or commemorative English discourse. The use of gerunds adds rhetorical strength and fluency, contributing to its high evaluation.

Let’s first examine DeepSeek’s translation. The original term “Mianhuai” conveys a sense of reverence and remembrance for revolutionary martyrs. The translation “remembering” slightly lacks the depth of solemn respect embedded in the original. If the tone is intended to emphasize the solemnity and formality of remembrance, alternatives like “honoring” or “commemorating” would be more appropriate. Nevertheless, “remembering” still communicates the basic meaning of the source text and is therefore acceptable. The translation of “Geming Xianlie” as “Revolutionary Martyrs” is accurate and consistent with standard historical terminology.

Turning to ERNIE Bot’s translation, the phrase “cherish the memory” is a particularly suitable rendering. It effectively captures both the respect and emotional attachment implied by “Mianhuai.” Compared to “remembering,” it emphasizes a deeper emotional connection and reverence toward historical figures, making it well-suited for conveying the intended sentiment in “Mianhuai Geming Xianlie.” The translation clearly reflects the core message of the original, especially in the way “cherish the memory” conveys a sense of heartfelt remembrance.

Finally, ChatGPT’s translation, using “commemorating,” is highly appropriate. It accurately conveys the reverent and memorial tone of “Mianhuai.” Compared with “remembering” or even “cherish the memory,”

the word “commemorating” evokes a more solemn, formal tone, making it particularly fitting for contexts like “Mianhuai Geming Xianlie.” It captures both the honor and ceremony intended in the original, which is why this version received the highest evaluation.

5. Conclusion

As machine translation tools, ChatGPT and the two other AI translation platforms demonstrate comparable performance in the Chinese-to-English translation of red culture texts, suggesting that ChatGPT has attained a certain level of competitiveness among existing MT systems. In vocabulary translation, ERNIE Bot achieves the highest overall scores, indicating superior cultural adaptability. In sentence translation, although ERNIE Bot again leads slightly, the gap among the three large language models is minimal, showing that all three models perform well in grammatical accuracy, logical consistency, and coherence, with ERNIE Bot having a slight edge.

The translation of red culture texts requires careful consideration of specialized terminology and rich historical and cultural context. Traditional translation in this field is time-consuming and mentally demanding, posing high requirements and significant challenges for translators. Thanks to their massive databases and deep learning capabilities, large language models can efficiently support pre-translation research and post-translation terminology revision for such complex texts.

This study also proposes a prompt optimization strategy—“manual prompt design → optimized AI-generated prompt → obtain translation”—to improve translation quality. Post-editing, a widely adopted practice in the translation industry, benefits from starting with higher-quality machine outputs, thereby enhancing both translation quality and efficiency while maximizing the advantages of human-machine collaboration. This has important implications for institutions such as museums and translation research centers, helping them achieve higher translation quality at lower costs. At the same time, it demands that translators in this field adapt quickly, continuously improve their domain knowledge and translation competence, and develop the ability to effectively utilize AI tools through prompt engineering, all in service of better disseminating red culture.

Moreover, in the face of the growing impact of generative AI and the language models on which they are trained, linguists and translation scholars must take responsibility for ensuring the accuracy and appropriateness of linguistic expression. They should preserve the authenticity of the Chinese language and ensure the faithful transmission of China’s fine traditional culture in translation.

In summary, large language models can serve as effective assistants in the translation of red culture texts, supporting the entire process from preparation to post-editing, and significantly improving both quality and efficiency. As highly intelligent systems still under rapid development, ERNIE Bot, ChatGPT, and DeepSeek continue to evolve, and it is foreseeable that the translation quality of LLMs will only improve. Optimized prompt-generation techniques are also expected to play an increasingly important role in the future.

Funding

Shanxi Normal University Graduate Innovation Project (2024XSY31)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Rivera-Trigueros I, 2022, Machine Translation Systems and Quality Assessment: A Systematic Review. Language Resources and Evaluation, 56(2): 593–619.

[2] Han C, 2020, Translation Quality Assessment: A Critical Methodological Review. The Translator, 26(3): 257–273.

[3] Lauscher S, 2014, Translation Quality Assessment: Where Can Theory and Practice Meet? Evaluation and Translation, Routledge, New York, 149–168.

[4] Salvagno M, Taccone FS, Gerli AG, 2023, Can Artificial Intelligence Help for Scientific Writing? Critical Care, 27(1): 75.

[5] Thorp H, 2023, ChatGPT Is Fun, But Not An Author. Science, 379(6630): 313.

[6] Jiao W, Wang W, Huang JT, et al., 2023, Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine. arXiv. https://doi.org/10.48550/arXiv.2301.08745

[7] Ghassemiazghandi M, 2024, An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score. Theory and Practice in Language Studies, 14(4): 985–994.

[8] Nemergut M, 2024, Machine Translation Quality Based on TER Analysis from English into Slovak. L10N Journal, 3(2): 60–86.

[9] Wang D, Lin L, Zhao Z, et al., 2023, EvaHan2023: Overview of the First International Ancient Chinese Translation Bakeoff, Proceedings of ALT2023: Ancient Language Translation Workshop, 1–14.

[10] Ali JKM, 2023, Benefits and Challenges of Using ChatGPT: An Exploratory Study on English Language Program. University of Bisha Journal for Humanities, 2(2): 629–641.

[11] Sahari Y, Al-Kadi AMT, Ali JKM, 2023, A Cross-Sectional Study of ChatGPT in Translation: Magnitude of Use, Attitudes, and Uncertainties. Journal of Psycholinguistic Research, 52(6): 2937–2954.

[12] Wang J, Wen Q, 2010, A Review of Automatic Scoring Systems at Home and Abroad and the Enlightenment for Chinese Students. Foreign Languages, 2010(1): 75–81.

[13] Akhtarshenas A, Dini A, Ayoobi N, 2025, ChatGPT or A Silent Everywhere Helper: A Survey of Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2503.17403

[14] Guo D, Zhu Q, Yang D, et al., 2024, DeepSeek-Coder: When the Large Language Model Meets Programming—The Rise of Code Intelligence. ArXiv. https://doi.org/10.48550/arXiv.2401.14196

[15] Lu H, Liu W, Zhang B, et al., 2024, Deepseek-VL: Towards Real-World Vision-Language Understanding. arXiv. https://doi.org/10.48550/arXiv.2403.05525

[16] Wang A, Singh A, Michael J, et al., 2018, GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv. https://doi.org/10.48550/arXiv.1804.07461

[17] Banerjee S, Lavie A, 2005, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65–72.

[18] Rei R, Stewart C, Farinha AC, et al., 2020, COMET: A Neural Framework for MT Evaluation. arXiv. https://doi.org/10.48550/arXiv.2009.09025

[19] Graham Y, Baldwin T, Moffat A, et al., 2013, Continuous Measurement Scales in Human Evaluation of Machine Translation, Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, 33–41.

[20] Snover M, Dorr B, Schwartz R, et al., 2006, A Study of Translation Edit Rate with Targeted Human Annotation, Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (ACL), Cambridge, 223–231.

[21] Lommel A, Uszkoreit H, Burchardt A, 2014, Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. Revista Tradumatica: Tecnologies de la Traduccio, 12: 455–463.