Academic Education Publishing House

Higher Education and Practice Vol. 3 No. 1, 2026

A Study on the Accuracy and Comparison of Metaphor Translation of the Character "Liu" (Willow) in Ancient Chinese Poetry by Three Large Language Models—An Empirical Analysis based on ChatGPT, DeepSeek and Doubao

Zheng Mengli

Hubei University of Technology, Wuhan, Hubei, China

Abstract: Objective: To evaluate the accuracy and differences in metaphor translation of the character "Liu" (willow) in ancient Chinese poetry by three large language models: ChatGPT, DeepSeek, and Doubao.Methods: A comparative method was adopted, in which the three models were used to translate 61 ancient poems containing "Liu" in batches. The accuracy of metaphor translation (translated/ untranslated) was judged manually, the accuracy rate was quantified, and visual analysis was conducted via Python.Results: The accuracy rates of GPT and DeepSeek were both 86.89%, while that of Doubao was slightly lower at 85.25%. GPT performed well in metaphors unrelated to lovesickness and loneliness between men and women. Doubao exhibited stable translation of high-frequency metaphors with the smallest temporal fluctuation. However, all three models performed poorly in low-frequency metaphors such as frustration in official career and gestures of beauties, especially Doubao.Conclusion: Mainstream large language models possess basic capabilities in metaphor translation of ancient Chinese poetry. Model selection should be based on specific scenarios, and their ability to process low-frequency and uncommon metaphors needs targeted improvement.

Keywords: Large Language Models; Translation of Ancient Chinese Poetry; Metaphor of "Liu" (Willow); Translation Accuracy; Multidimensional Evaluation

1. Introduction

As a core carrier of Chinese culture, the imagery metaphors in ancient Chinese poetry (such as the homophonic association between liu (willow) and liu (stay), symbolizing farewell) combine literary merit and cultural uniqueness, thus presenting major challenges in machine translation. In recent years, large language models including ChatGPT, DeepSeek and Doubao have achieved remarkable performance. Nevertheless, systematic evaluation remains lacking regarding the translation of culturally specific metaphors in ancient Chinese poetry. Existing studies mostly focus on general texts, and few conduct multi-dimensional comparisons of the transmission effect of imagery metaphors such as liu, which can hardly support model selection and optimization. The academic community has established a foundation in the research of large-model translation, metaphor translation of ancient Chinese poetry, and the willow imagery. Conceptual Metaphor Theory and Perceptual Simulation Theory provide theoretical support for the translation of willow metaphors. The mainstream translation strategies consist of literal translation with annotations and metaphor reconstruction. The practice of scholars such as Xu Yuanchong and related studies offer methodological references for this research. Overall, gaps still exist in current research: first, the absence of empirical comparisons among the three models in translating metaphors in ancient Chinese poetry; second, the inconsistency of evaluation criteria for willow metaphor translation; third, the insufficient development of prompt engineering research targeting the willow imagery. This review systematically summarizes the research status of the three domains, identifies achievements and deficiencies, lays a foundation for the subsequent empirical comparison of the three models, fills the research gap, and provides a reference for AI-enabled communication of traditional Chinese culture.

2. Research Methods

Manual Annotation Criteria The English translation results of the three large language

models were evaluated by the author using a two-level annotation scheme:

1. Metaphorical meaning successfully rendered: The metaphorical semantics are completely and accurately conveyed (e.g., the metaphor of willows representing parting and sorrow is translated as “willows shed tear”).

2. Metaphorical meaning failed to render: The metaphor is lost or mistranslated (e.g., the metaphor of willows symbolizing longing between lovers is translated as “In gold-thread dress, with moth or willow ornaments,”). Data Visualization Python libraries including matplotlib and seaborn were employed to generate multi-dimensional visualizations, including:Bar charts of metaphor frequency;Heatmaps of single-poem translation quality;Radar charts or line charts of metaphor types;Moving average charts for temporal stability;Bar charts of overall accuracy.

Table 1. Model Versions

Model Name	Version
ChatGPT	GPT-3.5 Turbo
DeepSeek	DeepSeek-NLP v3
Doubao	8.9.0

3. Main Title

A Study on the Accuracy and Comparison of Metaphor Translation of the Character "Liu" (Willow) in Ancient Chinese Poetry by Three Large Language Models.

4. Multi-Dimensional Evaluation of the Translation Performance of Three Large Models

4.1 Micro Single Poem Dimension: Heat Map Analysis of Translation Accuracy

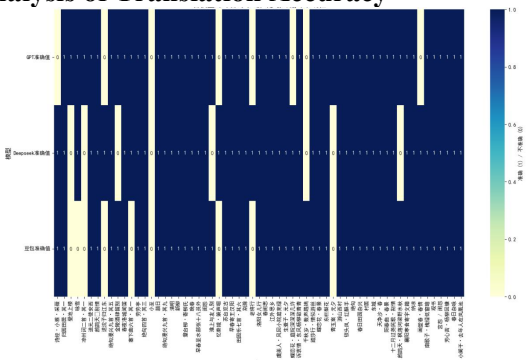

Figure 1. Heat Map Analysis of Translation Accuracy

With poem serial numbers as the horizontal axis and models as the vertical axis, the translation results are displayed in dark blue (1, metaphor translated) and light yellow (0, metaphor not translated). The key findings are as follows: ChatGPT translation accuracy rate: 86.9%, with 8 incorrect translations; DeepSeek translation accuracy rate: 86.9%, with 8 incorrect translations, which is the same as ChatGPT; Doubao translation accuracy rate: 85.2%, with 9 incorrect translations, and the accuracy rate is slightly lower.

4.2 Comparative Analysis of Radar Charts and Line Graphs

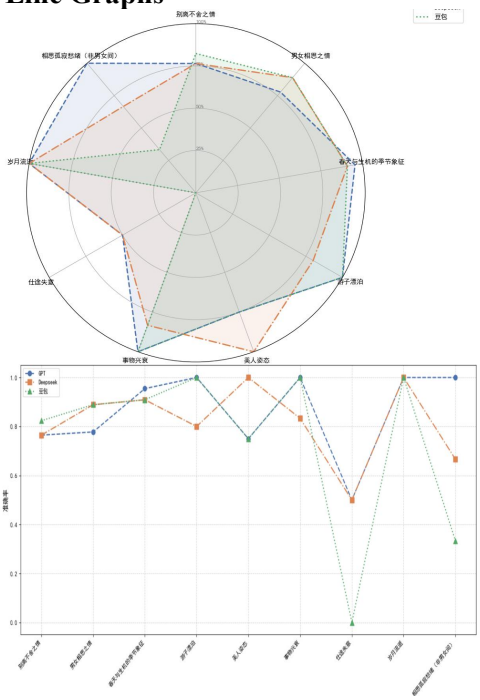

Figure 2. Comparative Analysis of Radar Charts and Line Graphs

Combining (Metaphorical Type Radar Chart) with (Metaphorical Type Translation Accuracy Line Graph), this study evaluates the model's processing capabilities for different metaphor types through dual dimensions of "expressive form" and "quantitative values":High-Frequency Metaphors: Demonstrates overall excellent and stable performance. Spring and Vitality: All three models achieve accuracy rates $\geqslant$ 90%, with radar charts clustered near outer circles indicating easy semantic learning. Separation and Reluctance: Rates $\geqslant$75% (DouBao highest), where contextual expressions enhance separation metaphors. Medium-Frequency Metaphors: Model differences begin to emerge. Love Between Men and Women: DeepSeek and DouBao $\geqslant$ 80% with parity, ChatGPT lowest. Rise and Fall of Things, Wandering Son: Both $\geqslant$ 80%, DeepSeek slightly lower. Low-

Frequency Metaphors: Significant model disparities. Loneliness in Love Between Men and Women: ChatGPT (100%) optimal, DouBao lowest. Beauty Posture: DeepSeek highest, ChatGPT and DouBao identical. Career Setbacks: ChatGPT and DeepSeek similar, DouBao zero showing extreme disparity. 4.2.4 ChatGPT's radar chart exhibits the most "rounded" shape, indicating balanced adaptability to various metaphors. DouBao's radar chart shows pronounced inward convergence in "Career Setbacks" and "Beauty Posture" dimensions, reflecting limitations in niche metaphor handling. DeepSeek falls between these extremes, demonstrating overall preference for "high-frequency metaphor optimization."

4.3 Temporal Stability Dimension

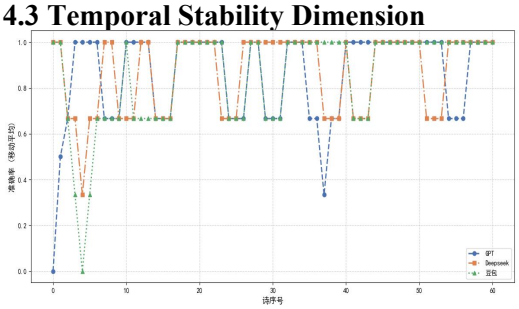

Figure 3. Temporal Stability Dimension

Analysis of Moving Average Fluctuations for Three Poems Using poem sequence numbers as the horizontal axis and moving average accuracy rates as the vertical axis to evaluate model translation stability:Model Fluctuation Characteristics: ChatGPT: Exhibits extreme volatility with steep fluctuations, dropping to around 0.3, demonstrating poor stability. DeepSeek: Shows relatively smooth fluctuations, quickly recovering after declines, indicating stable performance. DouBao: Demonstrates minimal fluctuations and overall stability, exhibiting the strongest robustness. Causes of Fluctuations: Impact of High-Frequency Metaphor Concentration Zones: Between poem sequence numbers 10-30, all three models maintain high and stable accuracy rates approaching 1. High-frequency metaphors possess typical semantic characteristics and are frequently encountered during training, allowing models to comprehend and translate them relatively easily, thus sustaining high and stable accuracy rates. Impact of Low-Frequency Metaphor Shock Zones: Within poem sequence numbers 0-5, all models experience significant accuracy drops, with ChatGPT showing the most pronounced decline. Poems in these zones likely contain predominantly obscure "willow" metaphors, for which models lack sufficient training exposure and weak understanding capabilities. Continuous exposure to low-frequency metaphor-containing poems leads to inaccurate interpretation and translation, resulting in substantial accuracy fluctuations. Particularly for ChatGPT — a general cross-language model — its inadequate adaptation to Chinese-specific low-frequency and obscure metaphor logic makes it most susceptible to concentrated low-frequency metaphor occurrences, causing the most severe accuracy volatility.

4.4 Macro-Level Dimension

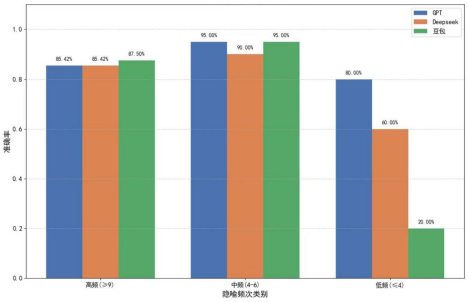

Figure 4. Macro-Level Dimension

Translation Accuracy Comparison of Different Frequency Metaphors The three models demonstrate the following translation accuracy rates for metaphors at varying frequencies: High-frequency metaphors: Both GPT and DeepSeek achieved 85.42%, while Doubao scored slightly higher at 87.50%, showing comparable and satisfactory performance. Medium-frequency metaphors: Both GPT and Doubao reached 95.00%, with DeepSeek achieving 90.00%, indicating overall outstanding performance. Low-frequency metaphors: GPT showed 80.00%, DeepSeek 60.00%, and Doubao only 20.00%, showing a significant decline in accuracy rates, with Doubao exhibiting the most pronounced drop. Overall, the three models achieved the highest accuracy rates for medium-frequency metaphors, while low-frequency metaphors generally showed declining performance, with Doubao being the most affected.

5. Conclusion

This study conducts a multi-dimensional comparison of three models' performance in translating ancient poetry imagery related to "willow." Results demonstrate that all models

achieve overall accuracy rates exceeding 85%, with minimal performance disparities, indicating foundational capabilities in translating traditional cultural imagery. Particularly noteworthy is their precise handling of high-frequency metaphors such as spring and parting, meeting basic cultural communication needs. The three models exhibit distinct characteristics suited to different scenarios: ChatGPT excels in processing low-frequency metaphors but demonstrates poor stability, making it ideal for academic contexts requiring in-depth interpretation and manual verification. DeepSeek delivers stable translation consistency for high-frequency metaphors, suitable for educational materials and slogans featuring common imagery. DouBao exhibits the most robust output with strong robustness, particularly adept at translating frequent parting metaphors for mass cultural dissemination. A shared limitation across models lies in low-frequency and obscure metaphors, primarily due to insufficient cultural allusion corpora and lack of specialized cultural reasoning modules.

In summary, the existing model can meet basic translation requirements and should be selected according to specific scenarios. Future improvements should include supplementing low-frequency corpora and enhancing the cultural inference module to enhance the accuracy of traditional cultural imagery translation.

References

[1]Ed-Dali R. Assessing DeepSeek R1 and ChatGPT 4.5 in Arabic-English literary translation[J].Cogent Arts & Humanities,2025,12(1).

[2]Li, RC; Nawi, AM and Kang, MS. Human-machine Translation Model Evaluation Based on Artificial Intelligence Translation[J].EMITTER-INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY,2023,11(2):145-159.

[3]Wael Etaiwi, Bushra Alhijawi.Comparative Evaluation of ChatGPT and DeepSeek Across Key NLP Tasks: Strengths, Weaknesses, and Domain-Specific Performance[J].ARRAY，2025，27.

[4]Hui Jiao,Bei Peng,Lu Zong,Xiaojun Zhang,Xinwei Li.Gradable ChatGPT Translation Evaluation [J].PROCESAMIENTO DEL LENGUAJE NATURAL,2024,72:73-85.

[5]L. David Ritchie & Xuede Zhao.To “Face the Powder” or “Powder the Face”? Contemporary Metaphor Theory and the Art of Chinese to English Translation [J].METAPHOR AND SYMBOL,2020,35(2):122-135.

[6]Li Hongyu, Che Mingming. Comparative Analysis of Linguistic Features of Scientific Texts Based on ChatGPT and DeepSeek Translation Models [J]. Journal of North China University of Science and Technology (Social Science Edition), 2025, 25(4): 88-96

[7]Du Lijuan. My View on the Translation of "Willow" in Farewell Poems [J]. Journal of Tongren University, 2014, 16(3): 67-69

[8]. Sun Xiuhua. An Overview of Plant Imagery and Cultural Implication Interpretation in "Nineteen Ancient Poems"[J]. Ningxia Social Sciences, 2010, 162(5): 168-172

[9]Feng Quangong, Li Lin. A Study on Emotional Metaphors in Chinese Classical Poetry and Their English Translation [J]. Language and Translation, 2017, 3: 57-64

[10]An Yong, Xie Xiaoli. Fragrant Grass and Elegant Trees Enhance Implication, Carefully Deliberate and Translate Elegantly and Truthfully - A Probe into the Translation of Plant Names and Related Expressions in Tang and Song Dynasty Poetry [J]. Journal of Shanghai Customs College, 2010, 2: 99-102

[11]Xu Yuanchong. Three Hundred Tang Poems in Chinese-English Parallel Text [M]. Beijing: Higher Education Press, 2000

[12]Xu Yuanchong, Lu Peixian, Wu Juntao. A New Translation of Three Hundred Tang Poems [M]. Beijing: China Translation Publishing House, 1997

[13]Yang Xianyi, Gladys Yang, et al. Ancient Poetry Garden: A Collection of Chinese-English Translations [M]. Beijing: Foreign Languages Press, 2001