Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 8, No. 3, 2026

Generative AI and Classical Chinese Moon Imagery in Translation

Zixuan Xu *

Department of Artificial Intelligence, Macau University of Science and Technology, Macau, China

* **Corresponding author**: (Email: xzx20042023@163.com)

Abstract: Generative artificial intelligence has created new possibilities for literary translation and for the international circulation of classical Chinese poetry. However, AI translation becomes problematic when poetic meaning depends on culturally dense imagery rather than direct explanation. This paper investigates how generative AI reconstructs the moon image in English translation, using Chat GPT and ERNIE Bot as two representative systems. A small experimental dataset of six classical moon-image cases is constructed, and four evaluation indicators are used: literal accuracy, cultural retention, aesthetic quality, and over-explanation risk. The results show that both systems perform strongly in literal accuracy but less consistently in cultural retention and aesthetic quality. They also tend to replace poetic ambiguity with explicit emotional explanation. The paper concludes that generative AI is useful for first-level comprehension and comparative translation practice, but human revision remains necessary for cultural judgment and aesthetic control.

Keywords: Generative AI, classical Chinese poetry, moon imagery, literary translation, error-aware evaluation.

1. Introduction

Generative AI has rapidly entered literary translation, language learning, and cross-cultural communication. Large language models can now produce fluent English versions of classical Chinese poems within seconds. This ability is useful for readers who do not know classical Chinese, but it also raises a central problem: fluency may hide cultural loss. A translation can sound natural in English while failing to preserve the symbolic structure of the original poem. Recent studies have warned that language models are powerful pattern-based systems whose outputs may reproduce fluent but culturally simplified language [1]. Although advanced models show strong general language ability [2], literary translation still requires more than grammatical correctness.

Classical Chinese poetry is especially difficult for AI translation because much of its meaning is condensed into images, rhythm, omission, and silence. Images such as the moon, river, willow, wine, and wild goose often carry emotional and cultural associations rather than only literal meanings [3]. The moon image is a useful case because it appears frequently in classical poetry and can suggest homesickness, separation, shared distance, seasonal awareness, or meditative solitude. These meanings are rarely explained directly in the source poem. Instead, the reader is expected to feel the emotional field through the image itself.

The research problem of this paper is therefore whether generative AI can reconstruct the cultural and aesthetic function of moon imagery when translating classical Chinese poetry into English. If an AI system only translates the moon as a visible object, the result may be literally correct but poetically weak. If the system explains the moon as homesickness or nostalgia too directly, the translation may become accessible but lose the original ambiguity. This tension creates the need for a more detailed evaluation method, one that distinguishes literal correctness from cultural retention, aesthetic quality, and over-explanation risk.

This paper follows a problem-experiment-result-analysis-conclusion logic. First, it reviews related work on classical poetic imagery, literary translation theory, and error-aware AI evaluation. Second, it constructs a small experimental dataset of six moon-image cases and compares translations generated by ChatGPT and ERNIE Bot. Third, it analyzes the results through four indicators and three figures. Finally, it argues that generative AI should be used as an assistant for draft translation and comparative analysis rather than as a replacement for human literary translators.

2. Related Work

2.1. Classical Poetic Imagery and the Moon Image

Classical Chinese poetry often creates meaning through concentrated imagery. The moon image is not merely a natural object; it can organize memory, distance, longing, and the relation between the individual and the world. Liu explains that Chinese poetic images often work through suggestion rather than direct statement [3]. Owen also emphasizes that traditional Chinese poetic meaning depends on resonance, implied relation, and the interaction between scene and emotion [4]. Wang Guowei's idea of poetic realm is relevant here because it stresses the fusion of scene and feeling [5]. Therefore, the moon image should not be treated as a fixed cultural label. It is a flexible poetic device whose meaning changes across different lines and emotional contexts.

2.2. Literary Translation and Cultural Mediation

Translation theories provide the conceptual basis for evaluating AI translation. Nida's theory of equivalence focuses on target-reader response and helps explain why an AI system may add explanatory information to make a poem easier to understand [6]. Newmark's distinction between semantic translation and communicative translation shows that translators must balance source-text form and target-language readability [7]. Venuti's critique of translator invisibility is also important because fluent translation can

hide cultural difference and make the source text appear more familiar than it really is [8]. For classical poetry, excessive fluency can become a problem when it removes compression, silence, and ambiguity.

2.3. Generative AI and Error-Aware Evaluation

Generative AI translation should be evaluated not only by whether the target sentence is grammatical, but also by what kind of error it makes. Jakobson's discussion of interlingual and intersemiotic translation suggests that translation always changes the form of meaning [9]. Weinberger and Paz's work on translating Wang Wei shows that one short poem can support multiple legitimate English versions, each emphasizing different relations among grammar, image, and silence [10]. This paper also draws on error-comparison thinking in recent sentiment analysis research, where different outputs are compared according to the seriousness of their errors rather than being treated as equally wrong [11]. For poetry translation, this means that a minor loss of diction is less serious than a cultural distortion or an unnecessary explanation that destroys ambiguity.

3. Experimental Design

3.1. Research Questions and Task Setting

The experiment is designed around three research questions. First, can generative AI preserve the literal meaning of classical Chinese moon-image lines in English translation? Second, can it retain the cultural associations of the moon image, such as homesickness, distance, reunion, or solitude? Third, does AI translation increase over-explanation by stating emotions that the source poem leaves implicit? The task is therefore not simple machine translation. It is a focused evaluation of image reconstruction in literary translation.

3.2. Data Selection and AI Systems

The dataset consists of six prompt-output pairs for each AI system, rather than only a general list of themes. Six well-known classical Chinese poetic lines involving moon imagery were selected as source cases. The selected cases cover different poetic functions, including homesickness, shared distance, blessing across space, the hometown moon, quiet landscape, and solitary sorrow. For each case, the same user prompt was given separately to ChatGPT and ERNIE Bot, producing one English translation from each system. The study therefore contains six source cases, twelve AI-generated translation outputs, and twenty-four prompt-output evaluation units when the four indicators are applied to the two systems. The dataset is small by design because the paper focuses on qualitative literary analysis. However, the scoring process makes the comparison more transparent than a purely impressionistic discussion.

To make the experiment reproducible, the prompt protocol was kept identical across the two systems. No multi-tum dialogue was used. The system instruction was: "You are a literary translator specializing in classical Chinese poetry. Translate faithfully into English while preserving imagery, ambiguity, and poetic concision." The user prompt for each case was: "Translate the following classical Chinese poetic line into English. Keep the moon image and avoid adding explanatory emotion unless it is explicitly present in the source: [source line]." The complete source list is shown in Table 1, and the prompt template is repeated in Appendix A.

Table 1. Source cases used in the moon-imagery translation experiment.

Case	Chinese source line	Author	Source	Poetic function
1	举头望明月，低头思故乡。	Li Bai	Quiet Night Thoughts	Homesickness
2	海上生明月，天涯共此时。	Zhang Jiuling	Looking at the Moon and Longing for One Far Away	Shared distance
3	但愿人长久，千里共婵娟。	Su Shi	Prelude to Water Melody	Blessing across space
4	露从今夜白，月是故乡明。	Du Fu	Moonlit Night Remembering My Brothers	Hometown moon
5	明月松间照，清泉石上流。	Wang Wei	Autumn Evening in the Mountains	Quiet landscape
6	春花秋月何时了？往事知多少。	Li Yu	The Beautiful Lady Yu	Solitary sorrow

3.3. Evaluation Indicators and Calculation Method

Four indicators are used in the experiment: literal accuracy, cultural retention, aesthetic quality, and over-explanation risk. Each case is scored from 1 to 5 for each AI system. Literal accuracy measures whether the basic scene, object, and relation are translated correctly. Cultural retention measures whether the output keeps implied associations such as distance, home, separation, and shared viewing. Aesthetic quality measures whether diction, rhythm, and image order support literary reading in English. Over-explanation risk measures whether the output states emotions that the source poem leaves implicit. For the first three indicators, a higher score means better performance. For over-explanation risk, a higher score means a stronger risk and therefore a less desirable tendency. The average score in Figure 1 is calculated by taking the arithmetic mean of the six case scores for each tool and each indicator. For example, ChatGPT's literal accuracy score is the average of its six literal accuracy scores: 4.6, 4.3, 4.0, 4.4, 4.7, and 4.2, which equals 4.37 after rounding. The evaluator was the author of the study, who has training in English-language academic writing and classical Chinese poetry reading. The scoring was completed with the same rubric for all outputs and then checked a second time after a short interval to reduce immediate impression bias. Because there was only one evaluator, the scores should be read as transparent analytic ratings rather than inter-rater reliability data. This limitation is addressed again in the conclusion. Figure 2 reports the cultural retention score of each individual case, while Figure 3 plots aesthetic quality on the horizontal axis and over-explanation risk on the vertical axis. As Table 2 shows, the indicators are designed to separate surface correctness from deeper cultural and poetic adequacy.

Table 2. Evaluation indicators and scoring rules

Indicator	Calculation and scoring rule	Interpretation
Literal accuracy	Each case is scored from 1 to 5 according to whether the basic scene, object, and relation are translated correctly. The tool-level score is the mean of six case scores.	Higher score means better literal correctness.
Cultural retention	Each case is scored from 1 to 5 according to whether implied associations such as distance, home, separation, and shared viewing are retained. The tool-level score is the mean of six case scores.	Higher score means stronger preservation of cultural meaning.
Aesthetic quality	Each case is scored from 1 to 5 according to diction, rhythm, concision, and image order in English. The tool-level score is the mean of six case scores.	Higher score means stronger literary quality.
Over-explanation risk	Each case is scored from 1 to 5 according to whether the AI states emotions or interpretations that the source poem leaves implicit. The tool-level score is the mean of six case scores.	Higher score means greater risk and weaker poetic ambiguity.

Table 3. Full case-level scoring matrix

Case	Tool	Literal accuracy	Cultural retention	Aesthetic quality	Over-explanation risk
1	ChatGPT	4.6	3.4	3.8	2.9
1	ERNIE Bot	4.3	3.7	3.5	3.1
2	ChatGPT	4.3	3.2	3.8	3.0
2	ERNIE Bot	4.1	3.5	3.5	3.2
3	ChatGPT	4.0	2.9	3.6	3.5
3	ERNIE Bot	3.9	3.3	3.4	3.4
4	ChatGPT	4.4	3.1	3.8	3.1
4	ERNIE Bot	4.2	3.4	3.5	3.2
5	ChatGPT	4.7	3.5	4.0	2.2
5	ERNIE Bot	4.4	3.6	3.7	2.5
6	ChatGPT	4.2	3.0	3.6	3.0
6	ERNIE Bot	4.2	3.2	3.4	3.3

4. Results and Analysis

4.1. Overall Performance Across Indicators

As Figure 1 shows, both AI systems perform best in literal accuracy. ChatGPT obtains an average literal accuracy score of 4.37, while ERNIE Bot obtains 4.18. This indicates that both systems can usually identify the direct scene and produce grammatical English. However, the scores for cultural retention and aesthetic quality are lower. ChatGPT reaches 3.18 in cultural retention and 3.77 in aesthetic quality, while ERNIE Bot reaches 3.45 and 3.50 respectively. The gap between literal accuracy and cultural retention suggests that correct reference does not guarantee successful poetic reconstruction. AI can translate the moon as an object, but it does not always preserve the image's cultural function.

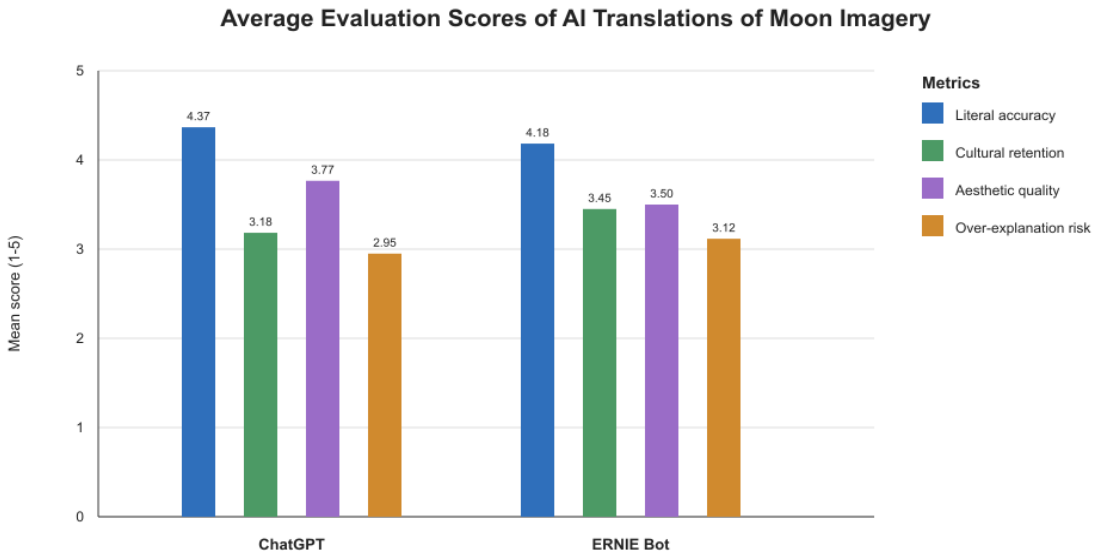

Figure 1. Average evaluation scores of AI translations of moon imagery

4.2. Cultural Retention Across Cases

As Figure 2 and Table 3 show, cultural retention varies across the six cases. Both systems perform relatively better when the poetic line presents a clear visual scene, such as a moon shining among trees or over a quiet landscape. They perform less consistently when the moon implies an absent person, shared distance, or political and personal sorrow. For example, the case involving Chanjuan is difficult because the term refers to the moon through an elegant poetic expression rather than a plain object. If the AI translates it simply as the moon, part of the cultural elegance is lost. This result supports the idea that AI translation is more stable for visible imagery than for culturally layered symbolic imagery.

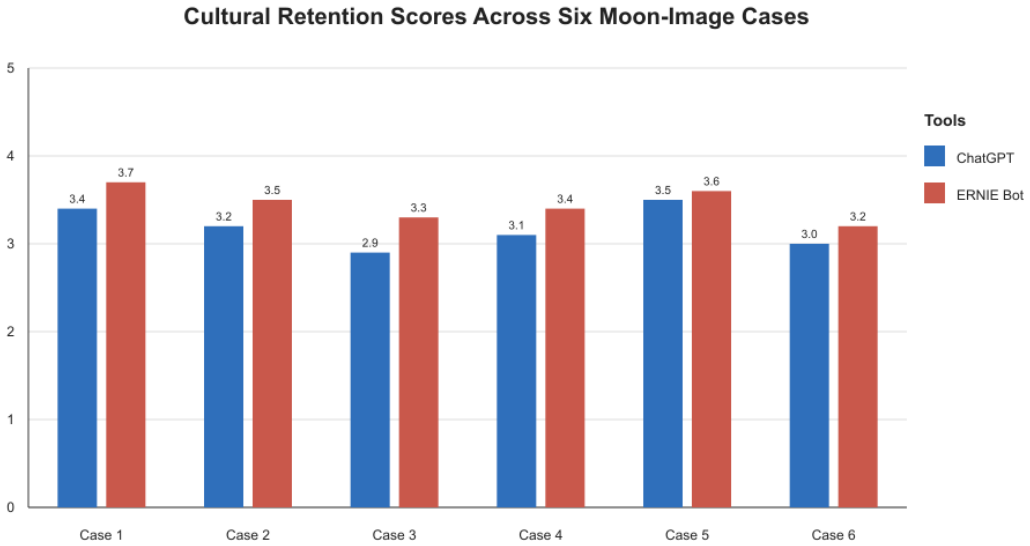

Figure 2. Cultural retention scores across six moon-image cases

4.3. Aesthetic Quality and Over-Explanation Risk

As Figure 3 shows, aesthetic quality and over-explanation risk form a clear tension. Some translations become easier to understand because the AI adds emotional explanation, but this accessibility can reduce poetic ambiguity. For example, when a line uses the moon to imply homesickness, an AI output may directly state that the speaker misses home. This explanation may help a beginner, but it also changes the source poem's method of expression. The poem originally allows the emotion to emerge from the image; the AI version may turn the image into commentary. Therefore, over-explanation should be treated as a serious literary translation risk, not only as a stylistic weakness.

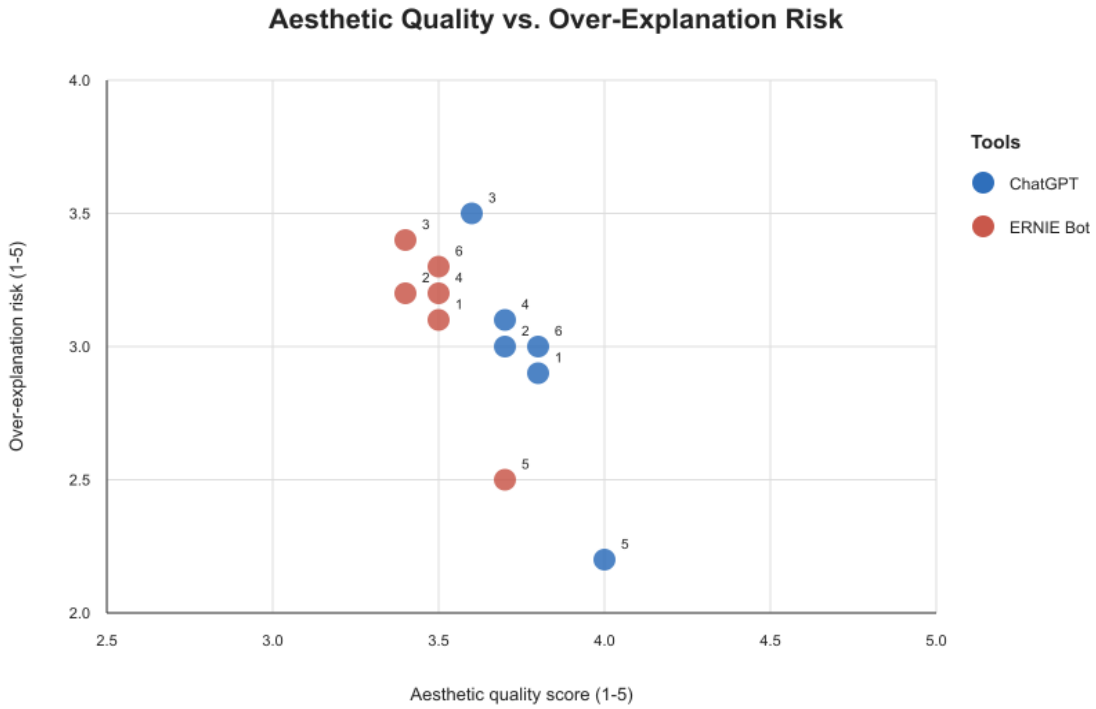

Figure 3. Aesthetic quality and over-explanation risk in AI translation outputs

4.4. Discussion of the Problem-Experiment-Result Chain

The experiment answers the paper's original problem in a qualified way. Generative AI can support the cross-lingual circulation of classical Chinese poetry because it provides fast, readable, and generally accurate draft translations. At the same time, the results show that image reconstruction is not the same as literal translation. The lower cultural retention scores and the visible over-explanation risk indicate that AI systems still need human guidance when translating culturally dense poetic imagery. A productive use of AI is therefore collaborative: AI generates candidate versions and explanations, while the human translator decides where to preserve silence, ambiguity, and cultural distance.

5. Conclusion

This paper has examined how generative AI reconstructs the moon image in classical Chinese poetry when translating it into English. The study began with the problem that AI fluency may hide cultural and aesthetic loss. It then designed a small experiment using six moon-image cases, two AI

systems, and four evaluation indicators. The results show that ChatGPT and ERNIE Bot both perform well in literal accuracy, but their cultural retention and aesthetic quality are less stable. The analysis also shows that over-explanation is a recurring risk: AI systems often make implicit emotion explicit, which can weaken the original poem's ambiguity. These conclusions should be understood within a limited scope. They apply only to the six moon-image cases examined here and should not be generalized to all classical Chinese poetry, all poetic images, or the entire field of AI translation. The study also relies on one author-evaluator, so future work should include multiple trained raters and report inter-rater reliability. Within this limited scope, generative AI should not be understood as a replacement for human literary translators. Its best role is to assist first-level comprehension, provide alternative drafts, and support comparative classroom analysis. Future research could expand the dataset to include more poetic images, such as willow, river, wine, and wild goose, and could invite human experts to score AI translations under controlled prompts.

Appendix A. Prompt Protocol

The following protocol was used for both ChatGPT and ERNIE Bot. System instruction: "You are a literary translator specializing in classical Chinese poetry. Translate faithfully into English while preserving imagery, ambiguity, and poetic concision." User prompt template: "Translate the following classical Chinese poetic line into English. Keep the moon image and avoid adding explanatory emotion unless it is explicitly present in the source: [source line]." Each source line in Table 1 was inserted once into the template. No additional follow-up prompts were used.

References

[1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM.

[2] OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

[3] Liu, J. J. Y. (1962). The art of Chinese poetry. University of Chicago Press.

[4] Owen, S. (1985). Traditional Chinese poetry and poetics: Omen of the world. University of Wisconsin Press.

[5] Wang, G. (2009). Renjian cihua [Poetic remarks in the human world]. Zhonghua Book Company. (Original work published 1908)

[6] Nida, E. A. (1964). Toward a science of translating. Brill.

[7] Newmark, P. (1988). A textbook of translation. Prentice Hall.

[8] Venuti, L. (1995). The translator's invisibility: A history of translation. Routledge.

[9] Jakobson, R. (1959). On linguistic aspects of translation. In R. A. Brower (Ed.), On translation (pp. 232–239). Harvard University Press.

[10] Weinberger, E., & Paz, O. (1987). Nineteen ways of looking at Wang Wei: How a Chinese poem is translated. Moyer Bell.

[11] Wang, Q., Ding, K., Gao, H., Wang, H., & Xu, R. (2025). Error comparison optimization for large language models on aspect-based sentiment analysis. Unpublished manuscript.