

Machine translation of modern Chinese poetry into Japanese: A comparative evaluation of LLMs, neural machine translation systems, and prompt strategies

Yuhuan Wang

wyh@neuq.edu.cn

Northeastern University at Qinhuangdao

Qilong Yu

Northeastern University at Qinhuangdao

Lianghong Hui

Northeastern University

Guanglu Peng

Northeastern University

Article

Keywords: Machine Translation, Large Language Models, Poetry Translation, Chinese-To-Japanese Translation, Prompt Strategy, Human Evaluation

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10122799/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

The rapid development of large language models (LLMs) has renewed interest in the role of machine translation (MT) in literary contexts, yet its applicability to poetry translation remains insufficiently understood. This study examines the Chinese-to-Japanese MT of modern Chinese poetry by comparing two LLM-based systems, ChatGPT and DeepSeek, with two widely used online NMT-based systems, Google Translate and DeepL. It further investigates how prompt strategy affects LLM-based poetry translation. Twelve poems were selected from Yu Xiuhua’s 2024 poetry collection *Houshan Huakai (Flowers Blooming on the Back Hill)* through stratified random sampling. The resulting translations were evaluated by eight expert raters using a multidimensional human evaluation framework covering fidelity, fluency, language style, machine-translation style, and overall quality. Linear mixed-effects models were used to account for variation across raters and source poems. The results show that, under the minimal prompt condition, the LLM-based systems received higher ratings than the online NMT-based systems in fidelity, fluency, language style, and overall quality, and showed fewer machine-translation traces. The effect of prompt strategy was model-dependent: the explanatory prompt improved DeepSeek’s performance in several dimensions, but did not improve ChatGPT’s performance. Under the explanatory prompt condition, no statistically significant differences were found between ChatGPT and DeepSeek, suggesting a degree of prompt-induced convergence in evaluated translation quality. These findings indicate that LLM-based poetry translation should be understood as an interaction between model characteristics and prompt design, and that machine-translated poetry should be evaluated beyond fluency and semantic accuracy, with closer attention to poetic style and machine-translation traces.

Keywords: Machine Translation; Large Language Models; Poetry Translation; Chinese-To-Japanese Translation; Prompt Strategy; Human Evaluation

1 Introduction

Machine translation (MT)[1] has changed substantially over the past decade. The development of neural machine translation (NMT)[2], especially after the introduction of the Transformer architecture, has improved translation quality across a wide range of practical domains, including news, technical documentation, and institutional communication[3, 4].

More recently, large language models (LLMs) [5] such as ChatGPT and DeepSeek have reshaped expectations about MT. Unlike conventional NMT systems, LLMs generate translations through instruction-based interaction and broader contextual modelling, allowing them to respond flexibly to task descriptions, stylistic requirements, and prompt variations [6]. However, whether these potential advantages extend to literary translation, especially poetry translation, remains insufficiently understood.

Poetry translation presents a particularly demanding test case for MT. In contrast to informative or transactional texts, poetry is characterized by semantic density, stylistic foregrounding, rhythmic organization, and a high degree of interpretive openness. Translating poetry therefore requires more than the transfer of propositional meaning; it also involves the recreation of tone, imagery, textual rhythm, and aesthetic effect in the target language[7]. These features make poetry a difficult object for both MT and translation evaluation. A translation that is lexically accurate may still fail to function as a poem, while a version that departs from the source at the surface level may better preserve its poetic effect. For this reason, poetry translation cannot be adequately assessed through automatic metrics alone, especially when the evaluation concerns literary quality rather than semantic adequacy in a narrow sense.

These difficulties become especially relevant in the context of modern Chinese poetry translated into Japanese. Chinese and Japanese share certain cultural and script-based resources, particularly through the historical role of Chinese characters. At the same time, they differ considerably in grammar, poetic convention, rhythm, and modes of literary expression [8, 9]. This combination of proximity and difference makes Chinese-to-Japanese poetry translation a revealing site for evaluating whether current MT systems can handle not only lexical meaning but also stylistic and poetic reconstruction [10]. Despite the growing body of research on MT and LLM-based translation, empirical studies on this language pair and text type remain limited.

Existing research on literary MT has shown that MT systems may perform reasonably well in terms of fluency and local accuracy, but they often struggle with style, figurative language, ambiguity, and literary voice [11, 12]. Recent studies have also begun to examine the potential of LLMs in literary and poetry translation, yet many focus on a limited set of models, specific genres, or single evaluation methods [13-15]. Less attention has been paid to systematic comparison between LLMs and widely used online NMT-based systems under the same experimental conditions [16]. Moreover, although prompting has been shown to influence LLM performance in many natural language processing tasks[6, 17, 18], its role in poetry translation has not been examined sufficiently. It remains unclear whether more explicit translation prompts improve poetic translation quality, and whether such effects are consistent across different LLMs.

To address these gaps, the present study evaluates MT quality in the Chinese-to-Japanese translation of modern Chinese poetry. It compares two LLM-based systems, ChatGPT and DeepSeek, with two widely used online NMT-based systems, Google Translate and DeepL. The study further examines how different prompt strategies affect the performance of LLMs in this task. Rather than relying solely on automatic evaluation, this study adopts a multidimensional human evaluation framework to assess translation quality in terms of

fidelity, fluency, language style, MT Style (machine-translation style), and overall quality. This design allows the study to capture both linguistic accuracy and poetic effectiveness, which are central to literary translation but often underrepresented in conventional MT evaluation.

This study makes three main contributions. First, it provides empirical evidence on MT quality in the underexplored context of modern Chinese poetry translated into Japanese. Second, it offers a comparative analysis of LLMs and online NMT-based systems under a unified evaluation framework, thereby clarifying their respective strengths and limitations in literary translation. Third, by incorporating prompt-based experimental conditions and multi-rater human assessment, it examines how prompt strategy interacts with model-specific performance in LLM-mediated poetry translation. In doing so, the study contributes to ongoing discussions on the applicability, limitations, and evaluation of MT in literary contexts.

2 Literature Review

This section reviews the theoretical and empirical work relevant to the present study. It first discusses literary translation and the specific features of poetic language, with particular attention to why poetry requires a multidimensional approach to translation evaluation. It then considers previous research on MT of poetry, the role of LLMs in literary translation, and the remaining gaps concerning Chinese-to-Japanese poetry translation and prompt-based evaluation.

2.1 Literary Translation and Poetic Language

Literary translation is generally understood as a process of reconstructing meaning, style, and aesthetic function across languages rather than transferring linguistic information in a straightforward manner [19, 20]. This complexity becomes more pronounced in poetry, where meaning is produced through the interaction of imagery, rhythm, syntax, ambiguity, and textual form. Unlike informative texts, poetry does not rely solely on linear semantic transmission. Its meaning is often layered, condensed, and open to interpretation[21].

In this sense, poetry translation involves the reconfiguration of both form and meaning. Lefevere [22] argues that translating poetry is not merely a matter of semantic equivalence, since poetic effects are embedded in formal and stylistic choices. His later concept of translation as “rewriting” further suggests that literary translation is shaped by the poetic norms and cultural expectations of the target context[23]. From this perspective, translators must negotiate between clarification and ambiguity, readability and poetic openness, and fidelity to the source and acceptability in the target language. This echoes Berman’s [24] argument that strategies of clarification and naturalization may reduce the foreignness, density, or aesthetic tension of the source text. Poetry translation evaluation should therefore move beyond semantic equivalence and consider how poetic function and aesthetic effect are reconstructed in the target language.

These concerns are particularly relevant to modern Chinese poetry. Modern Chinese poems often construct meaning through the interplay of linguistic form, imagery, rhythm, and fragmented expression. Their poetic force may lie not only in what is stated, but also in what

remains implicit, condensed, or structurally unresolved [25]. Such features require translation to preserve not only lexical meaning but also rhythm, imagery, and overall poetic effect. In the context of Chinese-to-Japanese translation, this task is further complicated by the need to negotiate between script-based proximity and differences in grammar, rhythm, poetic convention, and modes of expression.

For this reason, the evaluation of poetry translation should not be limited to accuracy or fluency. A translation may be grammatically smooth but stylistically flattened, or semantically close but unable to function as a poem in the target language. Poetry translation evaluation should therefore consider multiple dimensions, including fidelity, fluency, language style, MT Style, and overall poetic quality. This multidimensional perspective underpins the evaluation framework adopted in the present study.

2.2 Machine Translation of Poetry

Poetry has long posed difficulties for MT. Earlier rule-based MT (RBMT) systems depended heavily on manually constructed linguistic rules and dictionaries, which made them poorly suited to non-standard syntax, figurative expression, and creative language use. Statistical MT (SMT) later improved translation through data-driven modelling, but its performance depended strongly on the availability and quality of parallel corpora [26].

In literary domains, and especially in poetry, high-quality parallel corpora remain limited and unevenly distributed across language pairs[27-29]. Taken together, RBMT and SMT were both limited in handling the creativity, openness, and stylistic complexity of poetic language.

The emergence of NMT substantially improved translation quality, particularly in terms of fluency and readability. Compared with statistical approaches, NMT systems are generally better at producing grammatically coherent and natural-sounding translations[28, 30]. However, fluency does not necessarily imply literary adequacy. Existing studies suggest that NMT systems still struggle with figurative language, stylistic consistency, ambiguity, and literary voice [11]. In poetry translation, this limitation is especially visible because the target text must not only convey meaning but also sustain rhythm, imagery, and aesthetic effect. Even when formal constraints or line-level alignment are introduced, machine-translated poems may still fail to maintain poetic tension and structural coherence across the text [31].

Current online translation systems such as Google Translate and DeepL are widely used in practical translation contexts and often serve as baseline systems in translation-related studies. In this study, they are treated as representative general-purpose online MT systems commonly associated with NMT-based architectures. Their inclusion provides a practical benchmark for evaluating whether newer LLM-based systems offer advantages over widely used online NMT-based systems in the translation of modern Chinese poetry into Japanese.

2.3 Large Language Models in Literary Translation

LLMs have introduced an instruction-based approach to MT. Unlike conventional NMT systems, which are usually optimized for translation through task-specific training, LLMs can perform translation through natural-language instructions and prompt-mediated interaction. Their ability to follow instructions, model broader contexts, and generate flexible outputs has led to growing interest in their use for translation tasks [6, 32].

In literary translation, LLMs may offer potential advantages. Because they are exposed to large-scale multilingual and heterogeneous text data, they can generate translations that appear natural, coherent, and contextually sensitive. Recent research suggests that LLMs can produce readable literary translations and may perform well in maintaining discourse-level coherence[13]. Nevertheless, their performance in poetry translation remains unstable. Poetry requires careful handling of ambiguity, imagery, rhythm, and stylistic density. In such contexts, LLMs may produce fluent but interpretively loose translations, introduce unsupported meanings, weaken imagery, or normalize stylistic features that are central to the source poem[33].

A further issue concerns prompting. The output of LLMs is shaped not only by the model itself but also by how the task is formulated. Different prompts may guide the model toward literal translation, stylistic recreation, explanation-oriented rendering, or greater target-language naturalness. Previous research has shown that prompting can affect LLM performance across a range of natural language processing tasks[34]. However, its role in poetry translation remains insufficiently examined. In particular, it is still unclear whether explanation-before-translation prompts improve poetic translation quality, whether such effects occur across evaluation dimensions, and whether different LLMs respond to the same prompt conditions in similar ways.

2.4 Research Gaps and Research Questions

The existing literature points to three major gaps. First, although MT has been widely studied, empirical research on poetry translation remains limited, especially in the context of modern Chinese poetry translated into Japanese [16, 35]. This language pair is theoretically significant because it combines cultural and script-based proximity with substantial grammatical and poetic differences.

Second, direct comparison between general-purpose online NMT-based systems and LLM-based systems under the same experimental conditions remains insufficient. Many existing studies examine a single system, a limited set of texts, or different evaluation procedures [16, 28, 36], making it difficult to determine whether and in what respects LLMs differ from established online MT systems in literary translation.

Third, while prompt strategy is widely recognized as an important variable in LLM performance, its effect on poetry translation has not been systematically tested [17, 37]. It remains unclear whether prompt design significantly influences LLM translation quality, and whether such effects differ between models.

To address these gaps, this study examines the Chinese-to-Japanese MT of modern Chinese poetry. It compares two widely used online NMT-based systems, Google Translate and DeepL, with two LLM-based systems, ChatGPT and DeepSeek, under a unified experimental and evaluation framework. It also investigates the effect of prompt strategy on LLM translation performance. Translation quality is assessed through a multidimensional human evaluation framework covering fidelity, fluency, literary style, machine-translation style, and overall quality.

The study is guided by the following research questions:

RQ1. When LLMs are used under the minimal prompt condition (P1), are there

significant differences in translation quality between LLM-based systems, represented by ChatGPT and DeepSeek, and online NMT-based systems, represented by Google Translate and DeepL, in the Chinese-to-Japanese translation of modern Chinese poetry?

RQ2. Does prompt strategy significantly affect the poetry translation performance of ChatGPT and DeepSeek? If so, do the effects of prompt strategy differ between the two models?

RQ3. Under the explanatory prompt condition (P2), are there significant differences between ChatGPT and DeepSeek across the evaluation dimensions?

3 Methodology

This study adopts an experimental design to evaluate the quality of machine-translated modern Chinese poetry into Japanese. The methodology consists of four main components: source text selection, translation system and prompt design, multidimensional human evaluation, and statistical analysis. As shown in Figure 1, the research framework proceeds from source texts to translation systems, prompt conditions, translation outputs, human evaluation, and statistical analysis. This design makes it possible to compare different translation systems under controlled conditions and to examine the effect of prompt strategy on LLM-based poetry translation.

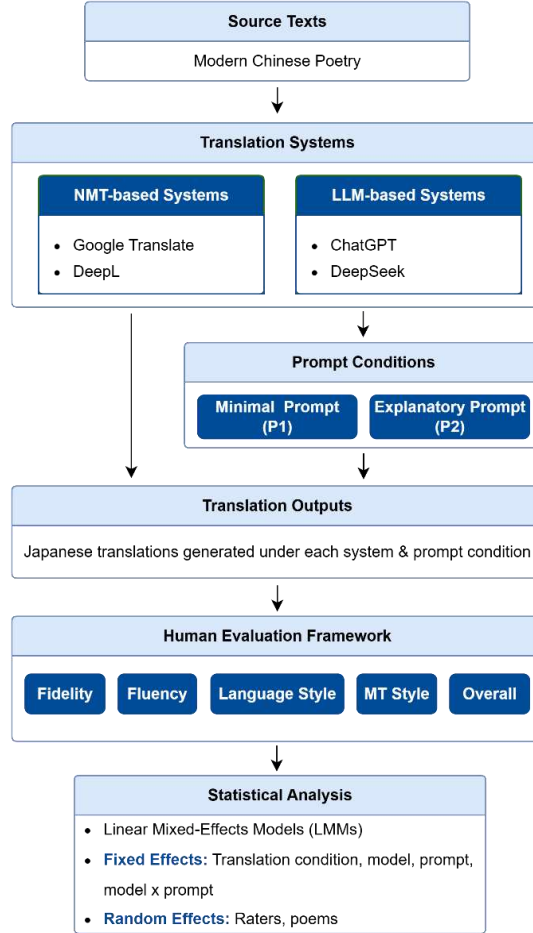


Figure 1. Overview of the research framework

3.1 Research Materials

The source texts used in this study were selected from *Houshan Kaihua* (*Flowers Blooming on the Back Hill*; 后山开花), a poetry collection by the contemporary Chinese poet Yu Xiuhua published in 2024 [38]. This collection was chosen for two reasons. First, as a recently published collection, it helped reduce the risk that the selected poems had been widely circulated in existing translated or machine-readable forms before the study. Second, at the time of this study, the collection had no published foreign-language translation, which helped avoid the possible influence of existing translated versions.

The collection consists of six sections and contains 152 poems covering themes such as hometown, family affection, love, everyday life, and reflections on existence. To ensure a balanced distribution of source texts, stratified random sampling was adopted. Two poems were randomly selected from each of the six sections, resulting in a final set of 12 poems. Given the expert-based nature of the evaluation, this sample size was considered appropriate for balancing evaluative depth and rater workload. In total, the 12 source poems generated 72 Japanese translation samples across six translation conditions.

3.2 Translation Systems and Prompt Design

3.2.1 Translation Systems

Four translation systems were selected for comparison. Google Translate and DeepL were included as representative general-purpose online MT systems commonly associated with NMT-based translation architectures. ChatGPT and DeepSeek were selected as representative instruction-driven LLM-based systems. This selection allowed the study to compare established online MT systems with newer LLM-based systems within the same experimental framework.

For the LLM-based systems, ChatGPT was accessed using GPT-5.2, and DeepSeek was accessed using DeepSeek-V3.2. All translations were generated in January 2026 within the same time window in order to reduce the potential influence of model updates. The same source texts were used across all systems. No post-editing was applied to any machine-generated translation.

3.2.2 Prompt Design

Prompt design was included as an experimental variable for the two LLM-based systems. Previous studies have shown that LLM outputs can be affected by the wording and structure of prompts, especially in tasks requiring stylistic or domain-specific generation[32]. To examine whether prompt strategy influences poetry translation quality, two prompt conditions were designed. To examine whether prompt strategy influences poetry translation quality, two prompt conditions were designed. Prompt 1 (P1) was designed as a minimal translation instruction, following the simple task-oriented prompt format commonly used in LLM-based translation studies [32]. Prompt 2 (P2) was designed as an explanatory prompt, drawing on research that highlights the role of task-specific and domain-sensitive prompting and on

poetry translation prompt designs that use semantic explanation to guide translation. Together, the two prompts allow the study to compare a minimal instruction with an interpretation-oriented instruction.

The original prompts were written in Chinese, as the source texts were Chinese poems. For reporting purposes, English translations of the prompts are provided below.

Prompt 1 (P1: minimal prompt; original Chinese: 请提供以下材料的日文翻译。): Please provide a Japanese translation of the following text.

Prompt 2 (P2: explanatory prompt; original Chinese: 以下是汉语现代诗歌, 请先解释其含义, 并将其翻译成日语现代诗歌。): The following text is a modern Chinese poem. Please first explain its meaning and then translate it into modern Japanese poetry.

The two prompts differ in whether they explicitly introduce an interpretation step before translation. P1 represents a minimal translation instruction, while P2 guides the model to process the source poem semantically before producing the Japanese translation. This design allows the study to test whether an explanation-before-translation prompt improves LLM performance in poetry translation. For P2, only the final Japanese translation was extracted for evaluation; the explanatory text generated before the translation was not included in the rating materials. For Google Translate and DeepL, no prompt condition was applied, as these systems do not operate through instruction-based prompting in the same way as LLMs.

3.3 Human Evaluation of Translation Quality

Poetry translation involves not only semantic transfer but also the reconstruction of imagery, rhythm, style, and aesthetic effect. For this reason, automatic metrics such as BLEU[39] are insufficient for evaluating literary translation quality, particularly when the evaluation concerns poetic function and stylistic adequacy rather than surface-level lexical similarity[11]. This study therefore adopted a multidimensional human evaluation framework.

3.3.1 Evaluation Procedure

Eight expert raters participated in the evaluation. All raters had bilingual competence in Chinese and Japanese and held doctoral degrees in fields related to Chinese-Japanese translation, Japanese language studies, or related areas. Seven raters were native speakers of Chinese, and one was a native speaker of Japanese. All had experience in translation, teaching, or academic research related to Chinese and Japanese.

Before the evaluation, the raters provided informed consent and received written instructions explaining the scoring procedure and evaluation dimensions. The evaluation materials consisted of Japanese translations of 12 Chinese poems produced under six translation conditions: Google Translate, DeepL, ChatGPT-P1, ChatGPT-P2, DeepSeek-P1, and DeepSeek-P2.

For each poem, the original Chinese text and six anonymized Japanese translations were presented on the same page. The raters were not informed of the systems that produced the translations. To reduce order effects and source bias, both the order of poems and the order of translation versions were randomized for each rater.

3.3.2 Evaluation Dimensions and Rating Scale

The raters evaluated each translation using a five-point Likert scale. Drawing on MT evaluation and literary translation studies, five dimensions were used, including fidelity, fluency, language style, MT Style, and overall quality. The evaluation framework is presented in Table 1.

Table 1. Human evaluation framework

Dimension	Definition
Fidelity	The extent to which the translation accurately conveys the core meaning of the source poem, without major omissions, additions, or semantic shifts.
Fluency	The extent to which the translation is grammatically correct, coherent, and natural in Japanese.
Language Style	The extent to which the translation preserves poetic features, including imagery, metaphor, tone, rhythm, and overall literary effect.
MT Style	The extent to which the translation shows obvious machine-translation traces, such as literal rendering, rigid sentence patterns, or unnatural expressions.
Overall	The overall quality of the translation, considering both linguistic adequacy and poetic effectiveness.

Each dimension was rated on a scale from 1 to 5. For fidelity, fluency, language style, and overall quality, higher scores indicate better translation quality. MT Style was treated as a reverse-oriented dimension: a higher score indicates stronger machine-translation traces and therefore lower translation quality in that specific dimension. This dimension was analysed separately to avoid interpretive confusion with the other positively oriented dimensions.

3.4 Statistical Analysis

The rating data had a hierarchical structure, as observations were nested within both raters and source poems. Conventional analysis of variance is limited in handling this non-independence. Therefore, linear mixed-effects models (LMMs) [40] were used to examine the effects of translation system and prompt condition while accounting for variation associated with individual raters and source texts.

Separate models were fitted for the five evaluation dimensions: fidelity, fluency, language style, MT style, and overall quality. In all models, rater and poem were included as random intercepts. The general model structure is given in equation (1):

$$Score_{ijk} = \beta_0 + X_i\beta + u_j + v_k + \varepsilon_{ijk} \quad (1)$$

where $Score_{ijk}$ represents the rating score for a translation condition, i indexes the translation condition or fixed-effect level, j indexes the rater, and k indexes the source poem. β_0 is the intercept, representing the baseline predicted score, $X_i\beta$ represents the fixed effects of interest, u_j represents the random intercept for rater, v_k represents the random intercept for poem, and ε_{ijk} represents the residual error.

The fixed-effect structure differed according to the research question. For RQ1, the analysis compared four conditions: Google Translate, DeepL, ChatGPT-P1, and DeepSeek-P1. Translation condition was treated as the fixed effect. For RQ2, only the LLM outputs were analysed, with model type, prompt type, and their interaction entered as fixed effects. For

RQ3, only the P2 outputs were analysed, comparing ChatGPT-P2 and DeepSeek-P2 across the five evaluation dimensions.

All analyses were conducted in R[41]. The statistical significance of fixed effects was tested using Satterthwaite’s approximation for degrees of freedom, implemented through the lmerTest package[42]. When significant effects were detected, post hoc comparisons were conducted using estimated marginal means (EMMs) with the emmeans package[43]. Pairwise contrasts were reported with 95% confidence intervals, and Tukey adjustment was applied for multiple comparisons. Model assumptions, including residual normality and homogeneity of variance, were assessed through diagnostic plots. The reported results include effect estimates, standard errors (SEs), test statistics, and p -values.

4 Results

This section reports the results for the three research questions. Following the experimental design and statistical procedures described in Section 3, the findings are presented in three parts: the comparison between LLM-based systems and online NMT-based systems under P1, the effect of prompt strategy on LLM translation performance, and the comparison between ChatGPT and DeepSeek under P2.

4.1 Results for RQ1

RQ1 examined whether LLM-based systems differed from online NMT-based systems in the Chinese-to-Japanese translation of modern Chinese poetry when the LLMs were used under P1. Four translation conditions were compared: ChatGPT-P1, DeepSeek-P1, DeepL, and Google Translate. The EMMs are shown in Table 2 and Figure 2. Since MT Style is a reverse-oriented dimension, lower scores indicate fewer machine-translation traces.

Table 2. EMMs and SEs across translation conditions

System	Fidelity	Fluency	Language Style	MT Style	Overall
ChatGPT-P1	4.56 (0.12)	4.66 (0.16)	4.48 (0.18)	1.52 (0.15)	4.53 (0.15)
DeepSeek-P1	3.99 (0.12)	3.92 (0.16)	3.72 (0.18)	2.35 (0.15)	3.75 (0.15)
DeepL	3.66 (0.12)	3.53 (0.16)	3.20 (0.18)	2.83 (0.15)	3.28 (0.15)
Google Translate	3.23 (0.12)	3.25 (0.16)	2.85 (0.18)	3.26 (0.15)	2.96 (0.15)

Note. Values are EMMs, with SEs in parentheses.

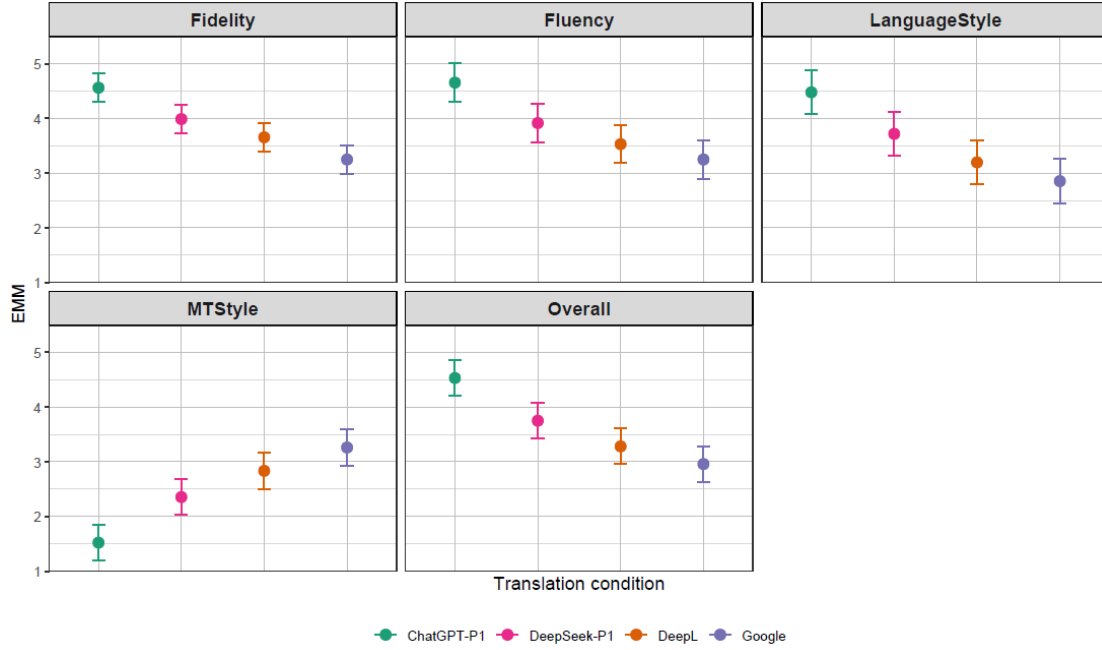


Figure 2. EMMs across the four translation systems

Overall, the results show a consistent pattern across the five evaluation dimensions. The two LLM-based systems received higher scores than DeepL and Google Translate in fidelity, fluency, language style, and overall quality, and lower scores in MT Style. Descriptively, ChatGPT-P1 obtained the highest scores in the positively oriented dimensions, followed by DeepSeek-P1.

For **fidelity**, the effect of translation condition was statistically significant, $F(3, 362) = 89.99, p < .001$. ChatGPT-P1 received the highest score and significantly outperformed DeepL (4.56 vs. 3.66, $\beta = 0.91, SE = 0.08, t = 10.95, p < .001$) and Google Translate (4.56 vs. 3.23, $\beta = 1.31, SE = 0.08, t = 15.87, p < .001$). DeepSeek-P1 also obtained higher scores than the two online NMT-based systems. These results indicate that the LLM-based systems were more effective in preserving the core meaning of the source poems.

For **fluency**, the differences among translation conditions were also significant, $F(3, 362) = 88.54, p < .001$. ChatGPT-P1 scored significantly higher than DeepL (4.66 vs. 3.53, $\beta = 1.13, SE = 0.09, t = 12.27, p < .001$) and Google Translate (4.66 vs. 3.25, $\beta = 1.41, SE = 0.09, t = 15.34, p < .001$). DeepSeek-P1 likewise outperformed DeepL and Google Translate. This suggests that the LLM-based systems produced more natural and coherent Japanese translations.

For **language style**, the effect of translation condition was significant, $F(3, 362) = 107.73, p < .001$. ChatGPT-P1 showed the highest descriptive mean and scored significantly higher than DeepL (4.48 vs. 3.20, $\beta = 1.28, SE = 0.10, t = 13.30, p < .001$) and Google Translate (4.48 vs. 2.85, $\beta = 1.63, SE = 0.10, t = 16.87, p < .001$). DeepSeek-P1 also obtained higher scores than both online NMT-based systems. This finding shows that the LLM-based systems were more successful in maintaining poetic language features and literary expressiveness.

For **MT Style**, where lower scores indicate fewer machine-translation traces, the effect of translation condition was significant, $F(3, 362) = 96.32, p < .001$. ChatGPT-P1 obtained the

lowest score and differed significantly from DeepL (1.52 vs. 2.83, $\beta = -1.31$, $SE = 0.12$, $t = -12.21$, $p < .001$) and Google Translate (1.52 vs. 3.26, $\beta = -1.74$, $SE = 0.12$, $t = -16.18$, $p < .001$). DeepSeek-P1 also showed fewer machine-translation traces than DeepL and Google Translate. By contrast, the two online NMT-based systems received higher MT Style scores, indicating more visible traces of literal or rigid MT.

For **overall quality**, the effect of translation condition was again significant, $F(3, 362) = 116.40$, $p < .001$. ChatGPT-P1 scored significantly higher than DeepL (4.53 vs. 3.28, $\beta = 1.25$, $SE = 0.09$, $t = 13.96$, $p < .001$) and Google Translate (4.53 vs. 2.96, $\beta = 1.57$, $SE = 0.09$, $t = 17.57$, $p < .001$). DeepSeek-P1 also obtained higher scores than the two online NMT-based systems. Taken together, the results for RQ1 show that, when the LLMs were used under P1, the LLM-based systems performed better than the online NMT-based systems in the poetry translation task.

4.2 Results for RQ2

RQ2 examined whether prompt strategy significantly affected the translation performance of the two LLM-based systems. The analysis compared ChatGPT and DeepSeek under two prompt conditions, P1 and P2. The EMMs are shown in Figure 3.

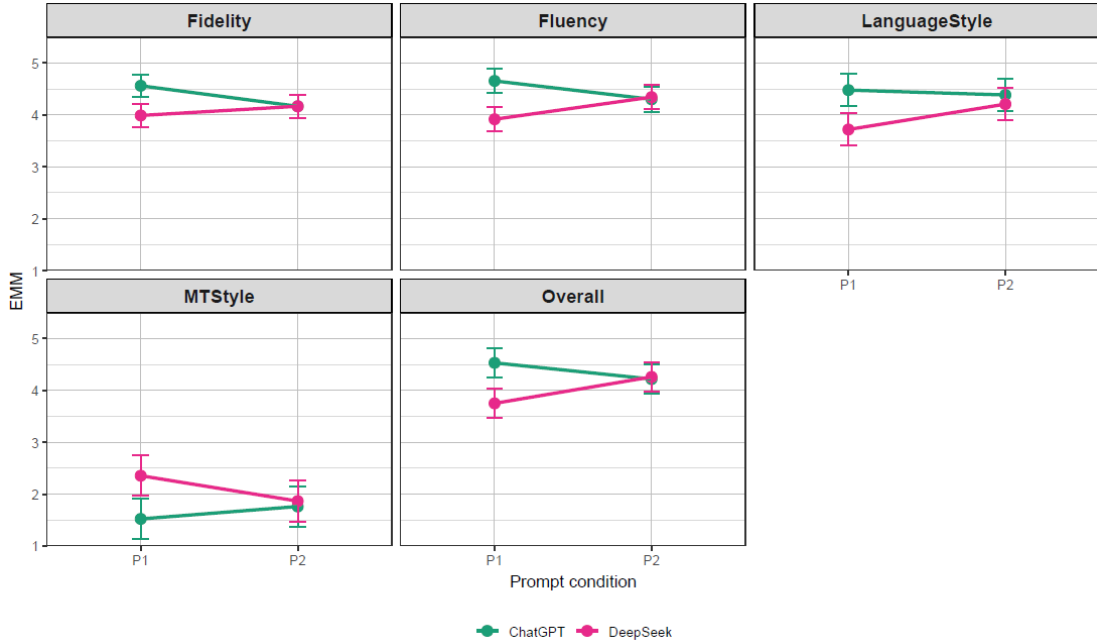


Figure 3. EMMs of ChatGPT and DeepSeek under P1 and P2

Significant model $\times$ prompt interactions were found across all five evaluation dimensions, with F values ranging from 19.47 to 44.39 and all p values below .001. This indicates that the effect of prompt strategy was model-dependent. In other words, P2 produced different, and in some cases opposite, effects for ChatGPT and DeepSeek.

For **ChatGPT**, P2 did not improve translation quality. In fidelity, ChatGPT-P2 scored significantly lower than ChatGPT-P1 (4.17 vs. 4.56, $\beta = -0.39$, $SE = 0.09$, $t = -4.49$, $p < .001$). A similar pattern was observed for fluency (4.30 vs. 4.66, $\beta = -0.35$, $SE = 0.09$, $t = -4.12$, $p < .001$). In MT Style, where lower scores indicate fewer machine-translation traces, ChatGPT-P2 scored significantly higher than ChatGPT-P1 (1.76 vs. 1.52, $\beta = 0.24$, $SE = 0.10$,

$t = 2.48, p = .01$), suggesting that the explanatory prompt increased visible machine-translation traces. In overall quality, ChatGPT-P2 also scored significantly lower than ChatGPT-P1 (4.22 vs. 4.53, $\beta = -0.31, SE = 0.09, t = -3.58, p < .001$). The only dimension in which no significant difference was found was language style (4.39 vs. 4.48, $\beta = -0.09, SE = 0.09, t = -1.00, p = .32$).

For **DeepSeek**, P2 had a generally positive effect. DeepSeek-P2 scored significantly higher than DeepSeek-P1 in fluency (4.34 vs. 3.92, $\beta = 0.43, SE = 0.09, t = 4.96, p < .001$), language style (4.21 vs. 3.72, $\beta = 0.49, SE = 0.09, t = 5.24, p < .001$), and overall quality (4.26 vs. 3.75, $\beta = 0.51, SE = 0.09, t = 5.84, p < .001$). In MT Style, DeepSeek-P2 scored significantly lower than DeepSeek-P1 (1.86 vs. 2.35, $\beta = -0.49, SE = 0.10, t = -5.07, p < .001$), indicating fewer machine-translation traces. For fidelity, DeepSeek-P2 showed a small improvement over DeepSeek-P1, but the evidence for this improvement was weaker ($p = .05$).

Overall, the results for RQ2 show that prompt strategy affected LLM-based poetry translation in a model-dependent manner. P2 was associated with lower or unchanged performance for ChatGPT, but with improved performance for DeepSeek in several evaluation dimensions.

4.3 Results for RQ3

RQ3 examined whether ChatGPT and DeepSeek differed significantly under P2. The analysis directly compared ChatGPT-P2 and DeepSeek-P2 across the five evaluation dimensions. The results are shown in Table 3 and Figure 4.

Under P2, no significant differences were found between ChatGPT-P2 and DeepSeek-P2 in any of the five evaluation dimensions. The two models received highly similar scores in fidelity, fluency, and overall quality. In language style, ChatGPT-P2 obtained a slightly higher descriptive score than DeepSeek-P2 (4.38 vs. 4.21), but this difference was not statistically significant. In MT Style, ChatGPT-P2 received a slightly lower score than DeepSeek-P2 (1.76 vs. 1.86), indicating marginally fewer machine-translation traces, but this difference was also not significant.

These findings suggest that, under P2, the performance gap between ChatGPT and DeepSeek became less pronounced. While small descriptive differences remained across specific dimensions, they did not reach statistical significance. Thus, the results for RQ3 indicate that the two LLM-based systems performed comparably when guided by the same explanation-before-translation prompt.

Table 3. Comparison between ChatGPT and DeepSeek under P2

Dimension	EMM(SE)		Pairwise comparison			
	ChatGPT-P2	DeepSeek-P2	β	SE	t	p
Fidelity	4.17(0.13)	4.17(0.13)	0	0.09	0	1.00
Fluency	4.30(0.12)	4.34(0.12)	-0.04	0.09	-0.48	0.63
Language Style	4.38(0.15)	4.21(0.15)	0.17	0.09	1.93	0.06
MT Style	1.76(0.22)	1.86(0.22)	-0.10	0.09	-1.18	0.24
Overall	4.22(0.15)	4.26(0.15)	-0.04	0.09	-0.48	0.64

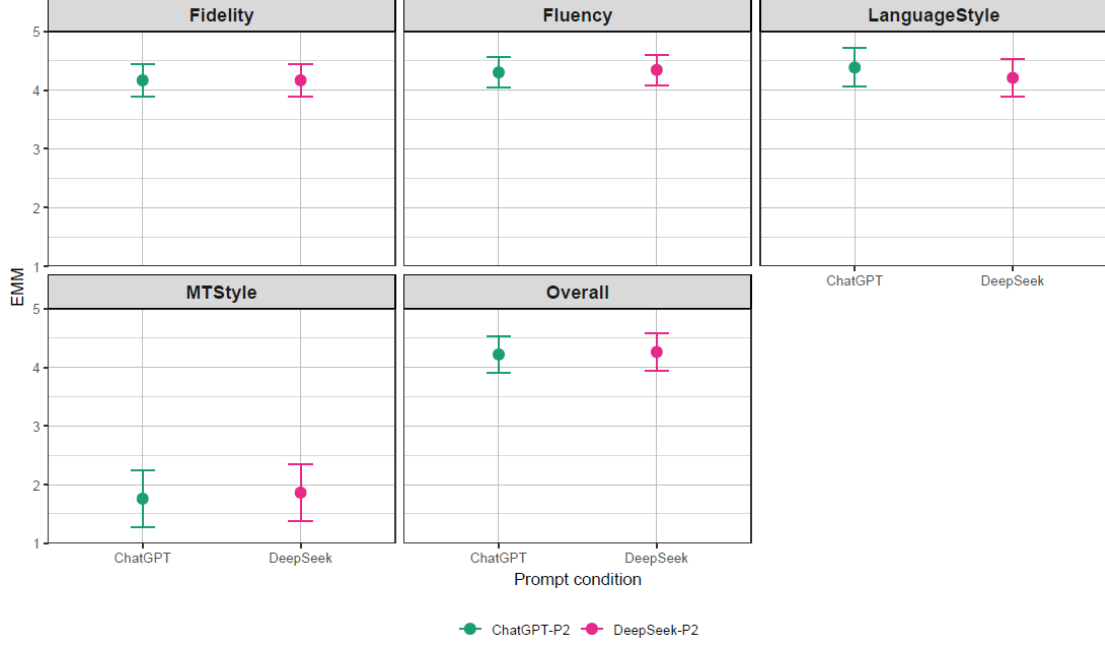


Figure 4. Comparison of ChatGPT and DeepSeek under P2

5 Discussion

This section discusses the findings in relation to the three research questions. Building on the statistical results reported in Section 4, the discussion focuses on how different translation systems and prompt conditions shaped the quality of Chinese-to-Japanese poetry translation. Particular attention is given to the ways in which semantic fidelity, poetic style, and machine-translation traces varied across systems, as these dimensions are central to evaluating the translation of highly stylized literary texts.

5.1 Response to RQ1

RQ1 examined whether LLM-based systems differed from online NMT-based systems in the Chinese-to-Japanese translation of modern Chinese poetry when the LLMs were used under P1. The results showed a clear overall pattern: ChatGPT-P1 and DeepSeek-P1 received higher ratings than Google Translate and DeepL across fidelity, fluency, language style, and overall quality, and lower ratings in MT Style. This suggests that, even under a minimal prompt, LLM-based systems were better able to produce translations that were not only semantically accurate but also more stylistically appropriate for poetic discourse.

This finding is consistent with the view that poetry translation requires more than lexical transfer. Modern Chinese poetry often constructs meaning through compressed imagery, implicit emotion, and non-linear semantic relations. As Boase-Beier [44] argues, the translation of poetic language depends on the interpretation of style and context rather than the reproduction of isolated linguistic units. In this respect, LLM-based systems appeared to benefit from their broader contextual modelling capacity. They were more likely to interpret the poem as an integrated textual unit, whereas the online NMT-based systems more often produced translations that were locally understandable but stylistically flattened or semantically narrowed.

In terms of **fidelity**, the LLM-based systems showed a stronger ability to preserve the semantic orientation and cultural implications of the source poems. Table 4 illustrates this pattern.

Table 4. Source text and example translations from poem 8 (“祭母辞”)

Text/Condition	Selected Excerpt 1	Selected Excerpt 2
Source text	祭母辞	我们在尘世里行走急急如律令
Google Translate	母への犠牲	私たちは法のように速くこの世を歩む
DeepL	母を偲ぶ辞	私たちはこの世を急いで歩む、命令のように
ChatGPT-P1	母を弔う詞	私たちはこの塵世を 律令のように急ぎ歩み
DeepSeek-P1	母を祭る辞	私たちはこの塵世を急ぎ歩む 律令のごとく

In the poem, the title “祭母辞” carries both ritual and elegiac meanings. The character “祭” points to an act of mourning or offering, while “辞” suggests a specific literary or rhetorical form. Google Translate rendered it as 「母への犠牲」, which misinterprets “祭母” as “sacrifice for the mother” and thus shifts the semantic focus. DeepL’s 「母を偲ぶ辞」 is closer in emotional orientation, but it weakens the ritual force of “祭” by turning it into a more general act of remembrance. By contrast, ChatGPT’s 「母を弔う詞」 and DeepSeek’s 「母を祭る辞」 retain the mourning context and ritual tone more effectively. This difference suggests that the LLM-based systems were better able to identify the cultural and functional meaning of the title rather than translating its components mechanically.

A similar contrast appears in the treatment of “急急如律令”. This expression does not simply mean “quickly”; it also carries a ritualized and imperative force. Google Translate’s 「法のように速く」 captures only the surface meaning of speed, while DeepL’s 「命令のように」 reflects command but loses the ritual register. ChatGPT and DeepSeek both retained the key term “律令” in expressions such as 「律令のように」 and 「律令のごとく」, thereby preserving the sense of command and ritualized compulsion. In this case, the LLM-based translations were more faithful not because they were more literal, but because they preserved a culturally loaded semantic function.

For **fluency**, all systems produced generally acceptable Japanese output, but the LLM-based systems received higher ratings. This result indicates that current MT systems have become relatively competent in generating grammatical and readable target texts. However, fluency alone is not sufficient for evaluating poetry translation. A fluent translation may still fail to preserve poetic density, imagery, or literary tone. Previous research has also noted that fluency and adequacy are often difficult to separate in human evaluation[45], and that naturalness at the sentence level does not necessarily indicate overall translation quality [46]. Therefore, the fluency results should be interpreted together with language style and MT Style, where the differences among systems became more revealing.

The advantage of LLM-based systems was especially clear in **language style**. Compared with Google Translate and DeepL, ChatGPT-P1 and DeepSeek-P1 were more capable of producing translations that retained poetic rhythm, imagery, and expressive tension. This difference reflects the tendency of LLMs to reorganize expression at the level of the whole line or poem, rather than translating each phrase as an isolated unit. In contrast, online NMT-

based systems often produced more explanatory or prose-like renderings[11, 47], which reduced the stylistic force of the source text. Table 5 illustrates this pattern.

Table 5. Example translations of poem 1(“啜泣”)

Text/Condition	Selected Excerpt 1	Selected Excerpt 2
Source text	像一个异乡人, 缀满对这个村庄的敌意	千里迢迢背回了一尊泥菩萨
Google Translate	まるでよそ者のように、この村に敵意を抱く	粘土製の仏像を遠くから運んできた
DeepL	異国のよそ者のように、この村への敵意で飾られている	私は粘土の菩薩を何千マイルも持ち帰った
ChatGPT-P1	まるで異郷の者のように この村への敵意を縫い合わせながら歩く	はるばる背負って帰ったのは 泥でできた菩薩像
DeepSeek-P1	異郷人のように この村への敵意をまとめて	千里の道を引き返し 泥の菩薩を背負って帰った

In the line “像一个异乡人, 缀满对这个村庄的敌意”, Google Translate and DeepL tended to present the image through complete explanatory sentence structures, making the translation read more like a prose statement. ChatGPT and DeepSeek, by contrast, used more flexible syntactic segmentation and line arrangement, which helped preserve part of the rhythm and tension of the original poem.

A similar contrast can be found in the image “千里迢迢背回了一尊泥菩萨”. The online NMT-based systems tended to make the image explicit by rendering “泥菩萨” as expressions such as 「粘土製の仏像」 and by clarifying the action through ordinary descriptive phrasing. Although such renderings are understandable, they reduce the compactness and symbolic force of the original image. The LLM-based translations, especially through expressions such as 「背負って帰った」, better maintained the bodily action embedded in “背回” and allowed the image to remain more poetic in Japanese. This suggests that the LLM-based systems were more successful in preserving the relation between image, movement, and poetic form.

The results for **MT Style** further support this interpretation. Since MT Style is a reverse-oriented dimension, lower scores indicate fewer machine-translation traces. The LLM-based systems received lower MT Style scores than Google Translate and DeepL, suggesting that their translations were less rigid, less literal, and less visibly machine-translated. Table 6 illustrates this pattern.

Table 6. Example translations of poem 3(“如果我也在桃花源”)

Text/Condition	Selected Excerpt
Source text	一定是桃花谢后。一定是人群离散 一定是又一段无着落的情缘 一定是源已无水
Google Translate	桃の花が散り、人々が散り散りになった後のことだろう またしても行き先のない恋に違いない 源泉の水が尽きたのかもしれない

DeepL	必ず桃の花が散った後だろう。必ず人々が散り散りになった後 だろう
	必ずまた一つ、行き場のない恋の縁だろう
	必ず源流に水がないだろう
ChatGPT-P1	きっと、それは桃の花が散ったあとのこと
	きっと、それは人々が去ったあとのこと
	きっと、それはまた行き場のない縁（えにし）
DeepSeek-P1	きっと、それはもはや水の尽きた源（みなもと）
	きっと桃の花が散った後だろう。きっと人々は散り散りに
	きっとまた一つの行き場のない情縁だろう
	きっと水源はもう枯れ果てているだろう

In this poem, the repeated structure “一定是……” creates a rhythm of emotional insistence and gradual accumulation. Google Translate’s output lacked consistency in sentence rhythm and appeared as a set of separately generated lines, which weakened the cumulative force of the repetition. DeepL maintained formal repetition through 「必ず」, but 「必ず」 in Japanese often suggests logical certainty rather than emotionally charged supposition. As a result, the translation became more explanatory than lyrical.

By contrast, the LLM-based translations showed fewer machine-translation traces in handling this parallel structure. ChatGPT used repeated 「きっと……」 structures together with line segmentation, allowing the repetition to support emotional progression rather than merely reproduce a surface pattern. DeepSeek also produced a more natural poetic flow, although some repeated forms, such as 「だろう」, still showed traces of mechanical patterning. Overall, the example indicates that the LLM-based systems were better able to reduce template-like rendering and pragmatic mismatch at the level of poetic lines.

Taken together, the findings for RQ1 show that the advantage of LLM-based systems was not limited to one evaluation dimension. Their higher overall ratings were supported by stronger performance in semantic fidelity, poetic style, and reduced machine-translation traces. For modern Chinese poetry translated into Japanese, this suggests that LLMs may be better suited than general-purpose online NMT-based systems to tasks requiring contextual interpretation and stylistic reconstruction. At the same time, this advantage should not be understood as a complete solution to poetry translation. Rather, it indicates that LLM-based systems offer a stronger starting point for literary MT, especially when the evaluation concerns both meaning and poetic effect.

5.2 Response to RQ2

RQ2 examined whether prompt strategy affected the performance of LLM-based systems in the Chinese-to-Japanese translation of modern Chinese poetry. Unlike RQ1, which compared different translation systems, RQ2 focused on the role of prompt design within LLM-mediated translation. The results showed that prompt strategy did not produce a uniform effect. Instead, its influence was clearly model-dependent: P2 improved DeepSeek’s performance in several dimensions, but did not improve ChatGPT’s performance and, in some dimensions, was associated with lower ratings.

This finding suggests that prompt strategy should not be understood as a general quality-

enhancement technique. In poetry translation, a more explicit prompt may guide one model toward better semantic organization and stylistic control, while leading another model toward over-interpretation or stylistic normalization. The effectiveness of a prompt therefore depends not only on its wording, but also on how a specific model responds to the task structure introduced by that prompt.

5.2.1 ChatGPT: Decline under P2

For ChatGPT, P2 did not lead to higher translation quality. Compared with P1, P2 produced lower scores in fidelity, fluency, and overall quality, and higher scores in MT Style, indicating more visible machine-translation traces. This result suggests that, for ChatGPT, the explanation-before-translation structure may have functioned as an over-constraining condition in the poetry translation task.

This interpretation is consistent with previous research showing that more explicit prompts do not necessarily improve translation quality and that prompt effects may vary substantially across task settings, prompt designs, and model responses [48, 49]. One plausible explanation is that P2 encouraged ChatGPT to make the poem’s meaning explicit before translating it. In highly stylized poetic texts, especially modern Chinese poetry, such explicit interpretation may interfere with semantic openness, image tension, and non-declarative expression. When the model is asked to explain the poem first, it may become more inclined to translate the interpreted meaning rather than reconstruct the poetic form. As a result, the final translation may become semantically clearer but stylistically flatter, more explanatory, or less condensed. This pattern can be observed in Table 7.

Table 7. Source text and example translations from poem 5(“江水至此亦不回”)

Text/Condition	Selected Excerpt 1	Selected Excerpt 2
Source text	江水至此亦不回	我真想长久地活下去
ChatGPT-P1	江水ここに至りて 亦た還らず	——私は、ただ長く生きたいと願った
ChatGPT-P2	川の水はここまで来て、二度と帰らない	僕はただ—— 永く生きたいと心の底で呟く

In the title “江水至此亦不回”, ChatGPT-P1 rendered the line as 「江水ここに至りて亦た還らず」. This version retains a compact and literary register, preserving the solemn and condensed quality of the original title. Under P2, however, the title was translated as 「川の水はここまで来て、二度と帰らない」. Although this version is semantically understandable, it shifts the expression toward a modern explanatory sentence and weakens the concise poetic force of the source text.

A similar shift appears in the line “我真想长久地活下去”. Under P2, ChatGPT added interpretive wording such as 「心の底で呟く」, turning a restrained poetic statement into a more explicit form of inner monologue. This addition may make the emotion easier to understand, but it also changes the tone of the original line. The source poem presents the desire to live in a direct yet controlled manner, whereas the P2 version makes the emotion more narratively framed.

These examples indicate that P2 was not neutral. For ChatGPT, it appears to have shifted

the translation process from poetic reconstruction toward semantic interpretation. In doing so, it reduced some of the ambiguity, compression, and symbolic tension that are central to the source poem. This may help explain why P2 did not improve ChatGPT’s ratings, even though it provided more explicit task guidance.

5.2.2 DeepSeek: Improvement under P2

For DeepSeek, P2 had a different effect. Compared with P1, P2 led to higher ratings in fluency, language style, and overall quality, and to lower MT Style scores, indicating fewer machine-translation traces. This suggests that, for DeepSeek, the explanation-before-translation structure did not over-constrain the translation process. Instead, it provided useful guidance for organizing the poem’s meaning and producing a more coherent poetic rendering.

One plausible explanation is that DeepSeek was more responsive to the semantic structuring introduced by P2. By introducing an explanation-before-translation procedure, P2 provided a more explicit process for organizing the poem’s imagery, emotional direction, and central expressive focus before translation. This may have encouraged DeepSeek to generate translations based on the poem’s broader context rather than on local textual cues alone. Without such semantic guidance, the model may have produced translations that were understandable but sometimes more prose-like or insufficiently condensed. This positive effect of prompt strategy is illustrated in Table 8.

Table 8. Source text and example translations from poem 4(“悲悯”)

Text/Condition	Selected Excerpt 1	Selected Excerpt 2
Source text	站在风口就听到无数窟窿呼啸的身体	仿佛肯定我所有的灿烂是为了苍老
DeepSeek(P1)	風の吹き抜ける場所に立つと無数の穴が唸るような体	あたかも私のすべての輝きが老いるためにあったかのように肯定してくれること まるで 僕のすべての輝きも
DeepSeek(P2)	風穴に立てば無数の穴が唸る体	老いるためにあったと 肯定してくれること

In the line “站在风口就听到无数窟窿呼啸的身体”, DeepSeek under P1 produced a more explanatory rendering, including expressions such as “風の吹き抜ける場所”. This translation conveys the basic meaning but weakens the compressed physical force of the original image. Under P2, however, DeepSeek rendered “风口” as 「風穴」, producing a shorter and more forceful expression. This choice gives the Japanese line a more concentrated rhythm and better preserves the bodily and spatial intensity of the source image.

A similar improvement can be seen in the translation of “仿佛肯定我所有的灿烂是为了苍老”. Under P1, the translation followed a more continuous, prose-like structure. Under P2, the use of line division and rhythmic segmentation created a clearer poetic movement in Japanese. The translation became less explanatory and more attentive to the formal qualities of the source poem.

These examples suggest that P2 helped DeepSeek move from semantic transfer toward poetic organization. Rather than simply making the meaning explicit, P2 appeared to provide

a structure through which DeepSeek could better integrate imagery, rhythm, and emotional progression. This may help explain why DeepSeek showed improved performance under P2 in several evaluation dimensions.

Overall, the results for RQ2 show that the effect of prompt strategy in poetry translation is strongly model-dependent. The same prompt may weaken the performance of a model that already generates effective poetic translations under P1, while improving the performance of another model that benefits from additional semantic guidance. This finding has important implications for LLM-based literary translation. It suggests that prompt design should not be treated as a fixed solution, but as a variable that interacts with model-specific generation patterns. In poetry translation, the optimal prompt may depend on whether a model needs stronger semantic scaffolding or greater freedom for stylistic reconstruction.

5.3 Response to RQ3

RQ3 examined whether ChatGPT and DeepSeek differed significantly under P2. The results showed that no statistically significant differences were found between the two models across the five evaluation dimensions. Although small descriptive differences remained in language style and MT Style, these differences did not form a stable performance advantage. This finding suggests that the performance gap between the two LLM-based systems became less pronounced when they were guided by the same prompt condition.

Viewed together with the findings for RQ2, this result points to a form of prompt-induced convergence. P2 affected the two models differently, but the combined effect was that their evaluated translation quality became more comparable. By asking the models to explain the poem before translating it, P2 introduced a shared interpretation-before-translation procedure. This may have directed the models toward more similar task priorities, such as semantic clarity, thematic coherence, and explicit reconstruction of imagery. This interpretation is consistent with studies suggesting that explicit prompt structures can reduce model-specific variation by guiding models toward more similar task orientations [37, 49].

However, convergence in quantitative ratings does not imply complete stylistic equivalence. The two models still showed different tendencies in how they handled poetic imagery, rhythm, and emotional expression. Table 9 illustrates this pattern.

Table 9. Source text and example translations from poem 2(“冬至”)

Text/Condition	Selected Excerpt 1	Selected Excerpt 2
Source text	桂花树里的夜晚，茈萆草的夜晚 一杯茶里的夜晚	我将在这些事物里扶起自己的 倒影
ChatGPT-P2	キンモクセイの夜、 カヤツリグサの夜、 一杯のお茶の中の夜	私はこれらのものの中で 自分の倒影をそっと起こす
DeepSeek-P2	キンモクセイに沈む夜、アシカキ草の夜 茶碗に揺れる夜	私はこれらのものに 映った己れを引き上げる

In the lines “桂花树里的夜晚，茈萆草的夜晚 / 一杯茶里的夜晚”, ChatGPT maintained the parallel structure and image order of the source text, presenting “夜晚” as a perceptual space carried by different objects. Its rendering was relatively restrained and structurally close to the original. DeepSeek, by contrast, introduced more dynamic

expressions such as “沈む” and “揺れる”, which strengthened the interaction among the images and gave the lines a stronger sense of movement. Both approaches are defensible, but they reflect different stylistic tendencies: ChatGPT favored structural preservation and emotional restraint, whereas DeepSeek tended toward more dynamic image activation.

A similar difference appears in the line “我将在这些事物里扶起自己的倒影”. ChatGPT used a lower-intensity expression such as “そっと起こす”, producing a more inward and restrained emotional tone. DeepSeek’s use of “引き上げる” made the action more explicit and physically forceful. Again, the difference lies less in basic semantic accuracy than in stylistic orientation. ChatGPT’s translation was more subdued, while DeepSeek’s was more dynamic and externally articulated.

These examples show that the absence of significant differences in the quantitative evaluation does not mean that the two models produced stylistically equivalent translations. Rather, under P2, their differences are better understood as tendencies in poetic rendering than as statistically stable differences in translation quality.

Overall, the findings for RQ3 suggest that prompt strategy may not only affect the output of an individual LLM, but also reshape the relative differences between models. In poetry translation, model selection and prompt design should therefore be considered together. A model that performs strongly under P1 may not necessarily benefit from P2, while another model may require P2 to reach a comparable level of performance. This supports the view that LLM-based literary translation depends on the interaction between model characteristics, prompt structure, and the specific demands of the literary text.

5.4 Limitations and Remaining Challenges

Although this study revealed clear patterns across the three research questions, several limitations should be acknowledged. First, the study focused on modern Chinese poetry translated into Japanese. Since modern Chinese poetry is characterized by condensed imagery, implicit emotion, and open-ended meaning, the findings should be generalized with caution. The results may not apply directly to other literary genres, classical poetry, or non-literary texts. Future research could extend the comparison to other genres and language pairs to examine whether the model-dependent effects of prompt strategy observed in this study remain stable across different translation contexts.

Second, although the study used expert human evaluation and LMMs to account for variation across raters and source poems, human judgment inevitably involves a degree of subjectivity. This is especially relevant for dimensions such as language style, poetic naturalness, and overall literary effect. Future studies could combine expert evaluation with reader-response data, qualitative annotation, or more fine-grained error analysis. Such approaches would provide a more detailed understanding of how different readers perceive poetic quality in machine-translated texts.

Beyond these methodological limitations, the findings also point to a broader challenge in LLM-based poetry translation: the tension between semantic explicitness and poetic openness. P2 may help some models organize semantic relations more clearly, but it may also reduce ambiguity, compression, and symbolic space, which are often central to poetic language. This tension is not merely a technical issue. It also concerns a fundamental question

in poetry translation: how to balance interpretation and poetic preservation when translating across languages.

Finally, this study examined only two prompt conditions. It did not systematically explore the wider prompt design space or more complex combinations of instruction, explanation, stylistic guidance, and examples. Future research could investigate how prompt structure, explanation depth, and stylistic constraints interact with model characteristics. Such work would help clarify the controllability and boundaries of LLM-based literary translation and provide more practical guidance for using LLMs in poetry translation.

6 Conclusion

This study examined the MT of modern Chinese poetry into Japanese by comparing LLM-based systems with general-purpose online NMT-based systems and by investigating the role of prompt strategy in LLM-based poetry translation. Using a unified experimental design, multidimensional expert evaluation, and linear mixed-effects modelling, the study addressed three research questions concerning system differences, the effects of prompt strategy, and model comparison under the same prompt condition.

The findings can be summarized in three main points. First, under the minimal prompt condition (P1), the LLM-based systems performed better than Google Translate and DeepL across the major evaluation dimensions. Their advantages were particularly evident in fidelity, language style, and overall quality, suggesting that LLMs may be better able to handle semantic density, poetic imagery, and stylistic reconstruction in modern Chinese poetry. Second, the effect of prompt strategy was model-dependent: the explanatory prompt condition (P2) improved DeepSeek’s performance in several dimensions, but did not improve ChatGPT’s performance and was associated with lower scores in some aspects. Third, under P2, no statistically significant differences were found between ChatGPT and DeepSeek across the five evaluation dimensions, indicating that prompt strategy may reduce model-specific differences and lead to a certain degree of convergence in evaluated translation quality.

These findings suggest that LLM-based poetry translation should be understood as an interaction between model characteristics and prompt design. Prompts do not simply enhance translation quality in a uniform way; rather, they may guide, constrain, or reshape model output depending on how a specific model responds to the task. In this sense, model selection and prompt design should be treated as a joint optimization process in literary MT. Different combinations of models and prompts may be suitable for different translation goals, and no single model or prompt strategy can be assumed to be universally optimal.

Overall, this study contributes empirical evidence to the evaluation of LLMs in literary translation, particularly in the underexplored context of modern Chinese poetry translated into Japanese. It also highlights the need to evaluate LLM-based translation beyond fluency and semantic accuracy, with closer attention to poetic style, machine-translation traces, and the tension between semantic explicitness and poetic openness. Future research should further examine how these findings extend across other genres, language pairs, prompt designs, and human-machine collaborative translation settings.

References

1. Freitag, M., et al., *Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation*. Transactions of the Association for Computational Linguistics, 2021. **9**: p. 1460-1474.
2. Liu, Y., et al., *Multilingual Denoising Pre-training for Neural Machine Translation*. Transactions of the Association for Computational Linguistics, 2020. **8**: p. 726-742.
3. Vaswani, A., et al., *Attention is all you need*. Advances in neural information processing systems, 2017. **30**.
4. Kocmi, T., et al. *Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet*. in *Proceedings of the Ninth Conference on Machine Translation*. 2024.
5. Zhao, W.X., et al., *A survey of large language models*. Frontiers of Computer Science, 2026. **20**(12): p. 2012627.
6. Li, J., et al., *Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions*. Transactions of the Association for Computational Linguistics, 2024. **12**: p. 576-592.
7. Boase-Beier, J., *Stylistics and Translation*, in *The Routledge Handbook of Stylistics*, M. Burke, Editor. 2023, Routledge: London. p. 420-435.
8. Kornicki, P.F., *Languages, scripts, and Chinese texts in East Asia*. 2018: Oxford University Press.
9. Chu, C., T. Nakazawa, and S. Kurohashi. *Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese*. 2012. Istanbul, Turkey: European Language Resources Association (ELRA).
10. Chakrabarty, T., A. Saakyan, and S. Muresan. *Don't Go Far Off: An Empirical Study on Neural Poetry Translation*. 2021. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
11. Toral, A. and A. Way, *What Level of Quality Can Neural Machine Translation Attain on Literary Text?*, in *Translation Quality Assessment*. 2018. p. 263-287.
12. Cespedosa Vázquez, A.I. and R. Mitkov. *Machine Translation of literary texts: genres, times and systems*. 2023. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria.
13. Karpinska, M. and M. Iyyer. *Large language models effectively leverage document-level context for literary translation, but critical errors persist*. in *Proceedings of the Eighth Conference on Machine Translation*. 2023.
14. Wang, S., et al. *What is the best way for ChatGPT to translate poetry?* in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024.
15. Zhang, R., W. Zhao, and S. Eger. *How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs*. in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2025.
16. Gao, R., et al., *Machine translation of Chinese classical poetry: a comparison among ChatGPT, Google Translate, and DeepL Translator*. Humanities and Social Sciences Communications, 2024. **11**(1): p. 835.
17. Wei, J., et al., *Chain-of-thought prompting elicits reasoning in large language models*. Advances in neural information processing systems, 2022. **35**: p. 24824-24837.
18. Chen, B., et al., *Unleashing the potential of prompt engineering for large language models*. Patterns (N Y), 2025. **6**(6): p. 101260.
19. Jakobson, R., *On linguistic aspects of translation*, in *The translation studies reader*. 2021, Routledge. p. 156-161.
20. Bassnett, S., *Translation studies*. 3rd ed. 2002, London: Routledge. 82-86.
21. Jones, F.R., *Poetry Translating as Expert Action: Processes, Priorities and Networks*. Benjamins Translation

- Library. Vol. 93. 2011, Amsterdam; Philadelphia: John Benjamins Publishing Company. 1-227.
22. Lefevere, A., *Translating Poetry: Seven Strategies and a Blueprint*. Approaches to Translation Studies. Vol. 3. 1975, Assen: Van Gorcum. 1-127.
 23. Lefevere, A., *Translation, Rewriting, and the Manipulation of Literary Fame*. Translation Studies. 1992, London; New York: Routledge.
 24. Berman, A., *Translation and the Trials of the Foreign*. The Translation Studies Reader, ed. L. Venuti. 2021, London; New York: Routledge.
 25. van Crevel, M., *Chinese Poetry in Times of Mind, Mayhem and Money*. Sinica Leidensia. Vol. 86. 2008, Leiden; Boston: Brill. 1-518.
 26. Costa-Jussà, M.R., et al., *Study and comparison of rule-based and statistical catalan-spanish machine translation systems*. Computing and informatics, 2012. **31**(2): p. 245-270.
 27. Tiedemann, J. *Parallel data, tools and interfaces in OPUS*. in *Lrec*. 2012.
 28. Chakrabarty, T., A. Saakyan, and S. Muresan. *Don't go far off: An empirical study on neural poetry translation*. in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.
 29. Song, W.L., et al. *Towards Zero-Shot Multilingual Poetry Translation*. 2023. Macau SAR, China: Asia-Pacific Association for Machine Translation.
 30. Kuzman, T., Š. Vintar, and M. Arcan. *Neural machine translation of literary texts from English to Slovene*. in *Proceedings of the qualities of literary machine translation*. 2019.
 31. Ghazvininejad, M., Y. Choi, and K. Knight. *Neural poetry translation*. in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 2018.
 32. Jiao, W., et al., *Is ChatGPT a good translator? A preliminary study*. arXiv preprint arXiv:2301.08745, 2023. **1**(10).
 33. Chen, A., et al., *Large language models for classical Chinese poetry translation: Benchmarking, evaluating, and improving*. arXiv preprint arXiv:2408.09945, 2024.
 34. Jiao, W., et al. *ParroT: Translating during chat using large language models tuned with human translation and feedback*. in *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023.
 35. Chen, A., et al. *Benchmarking LLMs for Translating Classical Chinese Poetry: Evaluating Adequacy, Fluency, and Elegance*. 2025. Suzhou, China: Association for Computational Linguistics.
 36. Yang, W. and M. Yang, *Evaluating literary translation by large language models: a multidimensional quality assessment of Shen Congwen's Border Town*. Humanities and Social Sciences Communications, 2026. **13**(1).
 37. Liu, P., et al., *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. ACM Computing Surveys, 2023. **55**(9): p. 1-35.
 38. Yu, X., *Houshan Kaihua [后山开花]*. 2024, Guilin: Guangxi Normal University Press.
 39. Papineni, K., et al., *BLEU: a method for automatic evaluation of machine translation*, in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 2002, Association for Computational Linguistics: Philadelphia, Pennsylvania. p. 311–318.
 40. Bates, D., et al., *Fitting Linear Mixed-Effects Models Using lme4*. Journal of Statistical Software, 2015. **67**(1): p. 1 - 48.
 41. Team, T.R.C. R: *A language and environment for statistical computing*. 2026 2026-4-26 [cited 2026 4-30]; Version 4.6.0]. Available from: <https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf>.
 42. Kuznetsova, A., P.B. Brockhoff, and R.H.B. Christensen, *lmerTest Package: Tests in Linear Mixed Effects*

- Models*. Journal of Statistical Software, 2017. **82**(13).
43. Lenth, R.V., et al., *emmeans: Estimated Marginal Means, aka Least-Squares Means*. CRAN Package Reference Manual, 2026.
 44. Boase-Beier, J., *A Critical Introduction to Translation Studies*. Bloomsbury Critical Introductions to Linguistics. 2011, London: Continuum. 1-200.
 45. Callison-Burch, C., et al. *(Meta-) evaluation of machine translation*. in *Proceedings of the Second Workshop on Statistical Machine Translation*. 2007.
 46. House, J., *Translation Quality Assessment: Past and Present*. 1st ed. 2015, London; New York: Routledge.
 47. Vanmassenhove, E., D. Shterionov, and A. Way. *Lost in translation: Loss and decay of linguistic richness in machine translation*. in *Proceedings of Machine Translation Summit XVII: Research Track*. 2019.
 48. Cao, B., et al., *On the worst prompt performance of large language models*. Advances in Neural Information Processing Systems, 2024. **37**: p. 69022-69042.
 49. Schmidtová, P., et al. *How Important is 'Perfect' English for Machine Translation Prompts?* in *Findings of the Association for Computational Linguistics: EACL 2026*. 2026.

Figure legends

Figure 1. Overview of the research framework. The figure shows the workflow of source text selection, translation system and prompt design, translation output generation, expert evaluation, and statistical analysis. This is an original figure created by the authors.

Figure 2. EMMs across the four translation systems. The figure shows the estimated marginal means (EMMs) under the minimal prompt condition (P1). MT Style is reverse-oriented; lower scores indicate fewer machine-translation traces. This is an original figure created by the authors.

Figure 3. EMMs of ChatGPT and DeepSeek under P1 and P2. The figure shows prompt effects across the five evaluation dimensions. P1 indicates the minimal prompt condition, and P2 indicates the explanatory prompt condition. This is an original figure created by the authors.

Figure 4. Comparison of ChatGPT and DeepSeek under P2. The figure shows the estimated marginal means (EMMs) of ChatGPT and DeepSeek under the explanatory prompt condition (P2). MT Style is reverse-oriented; lower scores indicate fewer machine-translation traces. This is an original figure created by the authors.

Data availability

The anonymized rating data that support the findings of this study are available from the corresponding author upon reasonable request. The full source poems cannot be publicly shared due to copyright restrictions. Only selected short excerpts are reproduced in the manuscript for scholarly analysis.

Acknowledgements

Acknowledgements have been removed for peer review.

Author contributions

Author contributions have been anonymized for peer review.

Competing interests

The authors declare no competing interests.

Ethics approval

Author contributions have been anonymized for peer review.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixCorpusOverview.docx](#)