What is the Best Way for ChatGPT to Translate Poetry?

Shanshan Wang[1] **Derek F. Wong**[1*] **Jingming Yao**[2] **Lidia S. Chao**[1]
[1]NLP^{2}CT Lab, Department of Computer and Information Science, University of Macau
nlp2ct.shanshan@gmail.com, {derekfw,lidiasc}@um.edu.mo
[2]Department of Portuguese, Faculty of Arts and Humanities, University of Macau
jmyao@um.edu.mo

Abstract

Machine translation (MT) has historically faced significant challenges when applied to literary works, particularly in the domain of poetry translation. The advent of Large Language Models such as ChatGPT holds potential for innovation in this field. This study examines ChatGPT's capabilities in English-Chinese poetry translation tasks, utilizing targeted prompts and small sample scenarios to ascertain optimal performance. Despite promising outcomes, our analysis reveals persistent issues in the translations generated by ChatGPT that warrant attention. To address these shortcomings, we propose an Explanation-Assisted Poetry Machine Translation (EAPMT) method, which leverages monolingual poetry explanation as a guiding information for the translation process. Furthermore, we refine existing evaluation criteria to better suit the nuances of modern poetry translation. We engaged a panel of professional poets for assessments, complemented evaluations by using GPT-4. The results from both human and machine evaluations demonstrate that our EAPMT method outperforms traditional translation methods of ChatGPT and the existing online systems. This paper validates the efficacy of our method and contributes a novel perspective to machine-assisted literary translation.[1]

1 Introduction

Foreign poems translated into Chinese still have to be like poems.

—— Guo (1957)

Poetry translation is widely regarded as one of the most challenging tasks in the field of translation. When translating a foreign poem into Chinese, the resulting text should still be recognizable as a poem (Guo, 1957; Bian, 1989). In discussing poetry, we often refer to specific genres or styles (Stevenson, 1957). Unlike classical poetry, the term "modern" in modern poetry refers to the poetic styles of the 20th and 21st centuries, which represent a significant departure from traditional forms. The primary characteristic of modern poetry is its embrace of freedom and lack of restrictions (Awan and Khalida, 2015). Modern poetry is specifically characterized by open forms, diverse genres, a break from conventional narratives, and innovative language combinations (Guo, 1957; Wang, 2006; Skerratt, 2013). Unlike classical poetry, rhythm is no longer an essential feature of modern poetry. Consequently, when translating modern poetry, it is not necessary to adhere to the original poem's rhythm (Duan, 2008). However, the poeticity must not be overlooked; the poetic essence of the source poem must be preserved throughout the translation process (Guo, 1957; Skerratt, 2013).

For Chinese poetry, there are significant differences between various genres. Classical Chinese poetry is characterized by strict constraints on format, meter, sentence length, and rhyme. In contrast, modern poetry is free from these constraints and breaks away from the rigid structures of classical poetry (Skerratt, 2013; Awan and Khalida, 2015).

Previous work has successfully applied machine translation to poetry, but these poems typically have clear format or rhyme restrictions (Genzel et al., 2010; Ghazvininejad et al., 2018; Chakrabarty et al., 2021; Chakrawarti et al., 2022; Song et al., 2023), which differ significantly from modern Chinese poetry. Recent research has taken an innovative approach by first obtaining an initial translation of the input prose using traditional Neural Machine Translation (NMT) methods. This initial translation is then mapped to a set of masked sequences via a designed heuristic method. Finally, these sequences are used to generate poetry translations through a pre-trained Masked Language Modeling (MLM) technique (Khanmohammadi et al., 2023).

---

[*] Corresponding Author

[1] https://github.com/NLP2CT/Poetry-Translation

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14025–14043
August 11-16, 2024 ©2024 Association for Computational Linguistics

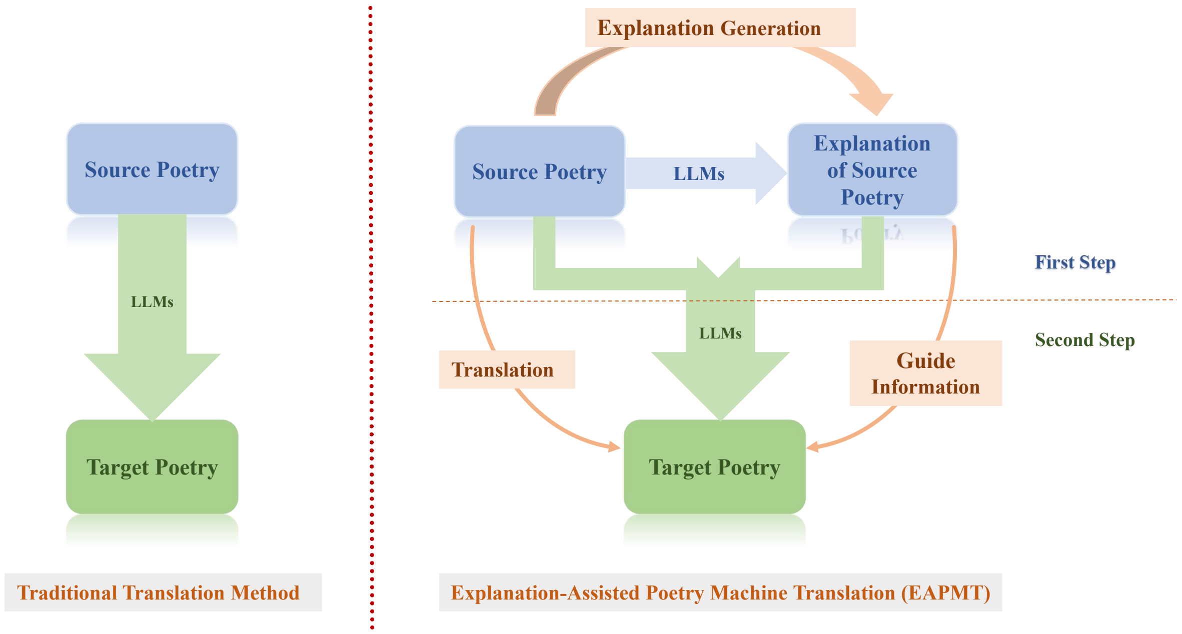

Figure 1: Comparison between the framework of the traditional translation method and the proposed Explanation-Assisted Poetry Machine Translation (EAPMT).

Another study compared the differences between machine translation and human translation of Arabic poetry into English (Alowedi and Al-Ahdal, 2023). The authors concluded that machine translation is not suitable for translating Arabic poetry into English as it fails to comprehend the socio-cultural background of poetry creation and the contextual nuances, particularly the genre-specific elements.

Promisingly, the artificial intelligence chatbot ChatGPT, released by OpenAI, has demonstrated excellent performance across various tasks and domains, including translation tasks (Jiao et al., 2023; Zhan et al., 2024; Pang et al., 2024; Lan et al., 2024). Although previous work has studied the performance of ChatGPT on translation tasks (Hendy et al., 2023; Jiao et al., 2023; Yang et al., 2023), and recent studies have explored the application of ChatGPT to poetry-related tasks, these investigations primarily focused on poetry generation. For instance, recent research examined the effectiveness of ChatGPT-4 in generating Arabic poetry and found the results to be unsatisfactory (Antar, 2023). The study highlighted several issues with the text generated by ChatGPT-4, including poor language quality, superficial content, lack of emotion, inconsistent speech, inappropriate word choices, and an ease of recognition by human evaluators.

Unlike previous work, this paper focuses on the capabilities of ChatGPT in translating English poetry into modern Chinese poetry. We explored optimal strategies for utilizing ChatGPT to translate poetry and evaluated its maximum performance in this specific task. Inspired by Jiao et al. (2023) and Gao et al. (2023), we investigated ChatGPT's performance on modern poetry translation tasks by designing appropriate prompts and providing example shots for the model.

Experimental results demonstrate the effectiveness of our designed prompts. Despite these promising outcomes, our analysis reveals persistent issues in the translations generated by ChatGPT that warrant further attention. Consequently, we propose a new poetry translation method called Explanation-Assisted Poetry Machine Translation (EAPMT). Our method leverages the explanation of monolingual poetry as guidance information to achieve high-quality translations from English poetry to modern Chinese poetry. Furthermore, existing evaluation criteria are typically designed for ordinary texts or poems with specific restrictions and are not fully applicable to modern poetry. Therefore, we refined these criteria to better capture the nuances of contemporary poetry translation. We engaged a panel of professional poets for assessments and complemented their evaluations with those conducted using GPT-4. The results from both human and machine evaluations demonstrate that our EAPMT method outperforms traditional translation techniques of ChatGPT and existing online systems.

The contributions of our work are as follows:

- We are the first to examine ChatGPT's capabilities in English-Chinese modern poetry translation tasks.
- We construct and release a high-quality bilingual poetry dataset.
- We identify the optimal prompts and examples (shots) for ChatGPT to effectively translate poetry.
- We propose a novel method for poetry translation that uses monolingual poetry explanations as guiding information. This method significantly enhances ChatGPT's performance in translating modern poetry and can be extended to other language pairs and models of language understanding and generation.
- We design a new framework for human evaluation criteria specifically applicable to modern poetry translation and engage several professional poets to evaluate the translation results.

2 Related Work

The earliest work on poetry machine translation employed phrase-based machine translation systems to translate French poetry into metrical English poetry (Genzel et al., 2010). This research explored how statistical machine translation systems could produce translations that adhere to constraints such as length, rhythm, and rhyme. Subsequently, Greene et al. (2010) used statistical methods to translate rhymed poetry, achieving the translation of Italian poetry into English. Through unsupervised learning, they identified word stress patterns in an original poetry corpus, which were then utilized in generating English poetry. Ghazvininejad et al. (2018) introduced the first neural poetry translation system, capable of translating French poetry into English while adhering to user-specified target rhythm and rhyme patterns. This system demonstrated greater flexibility than phrase-based machine translation (PBMT) systems, consistently generating translations that fit any specified scheme. Chakrabarty et al. (2021) conducted an empirical study on poetry translation, highlighting a crucial yet often overlooked issue: advanced machine translation systems, trained on large amounts of non-poetry data, struggle to preserve poetic style even if they can maintain meaning and fluency. In 2022, a Hybrid Machine Translation (HBMT) model was proposed (Chakrawarti et al., 2022), which improves the semantic and syntactic accuracy of the translation framework. This model effectively translates Hindi poetry into English, demonstrating notable advancements in maintaining the integrity of the original poetic content.

	Poems	Lines	Tokens
Datasets	400	7K	40K

Table 1: The statistics of ModePoem dataset.

3 Construction of ModePoem Dataset

Translators need to have a strong foundation in both the source and target languages (Guo, 1983; Wang, 2006). Additionally, poetry translators must understand poetry and ideally write poems themselves. In this paper, we constructed a high-quality bilingual poetry dataset, ModePoem, consisting of 400 modern poems (7,000 lines) in English and their Chinese translations. Our dataset is selected from the high-quality online platform Poetrysky.[2] We were meticulous in our selection process, ensuring that only high-quality data was included. This rigor is in part due to the expertise of one of our authors, a professional poet who has rigorously filtered out any data of inferior quality. It is important to note that the judges involved in our human evaluations did not overlap with the translators of the dataset, ensuring unbiased assessments. The source poems in this dataset were written by multiple professional English poets, with translations from various bilingual poetry translators who are native Chinese speakers. These translators are also accomplished poets, proficient in English, and have extensive experience translating poetry from English to Chinese. Therefore, the translations in this dataset can be considered the gold standard. The statistics of our dataset are shown in Table 1.

Data Validation Experiments To address the concerns about the potential exposure of poems in ChatGPT's training data (Shi et al., 2023), we used all the poems (10 poems) in the test set of the EAPMT experiment to conduct data verification experiments on GPT-4[3] to verify the originality of ChatGPT responses. Our data verification experiments consisted of two stages:

---

[2] http://poetrysky.com/

[3] We used the version gpt-4-1106 of GPT-4 in this study.

	50%	70%	90%
Source Poem	0.4	0.6	0.8
Translation	0.3	0.5	2.5

Table 2: SacreBLEU Scores from data validation experiments

1. *Poem Continuation*: We prompted the models with the initial 50%, 70%, and 90% of a poem, asking the models to generate the remaining portions. This aimed to verify if the poem had been seen by ChatGPT.
2. *Translation Continuation*: We applied the same approach to the Chinese translations, providing varying degrees of content (50%, 70%, and 90%) and instructing the models to generate the complete translations. We then evaluate the similarity against the human translations using SacreBLEU (Post, 2018).

As shown in Table 2, the outcomes indicated that the model's continuations and translations received low scores in the evaluation, suggesting that the model might not have seen these specific texts. Although we cannot categorically confirm the model's unfamiliarity with the poems, the results highlight areas where model-generated translations could be improved. We refer readers to Appendix A.1 for the validation details.

4 Task Setups

4.1 Appropriate Prompts

The output of ChatGPT depends on the input prompts (Antar, 2023), so it is essential to design effective prompts. Based on the work of Jiao et al. (2023) and Gao et al. (2023), and considering the characteristics of modern poetry, we asked GPT-4 for advice and obtained five prompts that can guide it to translate English poetry into modern Chinese poetry at the highest quality. The five prompts advised by GPT-4 are recorded as $\mathcal{P}_{i\in\{1\ldots5\}}$, respectively. More details about GPT-4 providing translation prompts are provided in Appendix A.2.

In addition, we designed three simple poetry translation prompts, recorded as $\mathcal{H}_{i\in\{1\ldots3\}}$, respectively:

- $\mathcal{H}_1$: *Please provide the Chinese translation for these sentences:*
- $\mathcal{H}_2$: *Please provide the Chinese translation for this poem:*
- $\mathcal{H}_3$: *Please translate this English poem into modern Chinese poetry:*

To choose the most suitable prompt for modern poetry translation from the above eight prompts, we randomly selected eight poems from the dataset as the test set for the prompt experiment and translated them using GPT-3.5[4] and GPT-4 with different prompts.

4.2 Evaluation

Automatic Evaluation Compared with human evaluation, automatic evaluation is widely used because of its lower cost and faster speed. Most natural language generation task evaluations rely on BLEU (Novikova et al., 2017). For our automatic evaluation, we used SacreBLEU (Post, 2018), BERTScore (Zhang et al., 2019), and COMET (Rei et al., 2022) to evaluate the performance of different candidate translations, as BLEU tends to give very low scores. The results are reported in Table 3. From the results of SacreBLEU, it can be seen that $\mathcal{H}_1$ achieved the highest scores on both GPT-3.5 and GPT-4, which are at least 3.2 points and 1.5 points higher than the prompts provided by GPT-4, respectively. According to BERTScore and COMET, our designed $\mathcal{H}_2$ achieves the highest score on GPT-3.5, while $\mathcal{P}_4$ achieves the highest score on GPT-4. However, the differences in scores among the eight prompts within the same model are relatively minor. Based on these observations, our subsequent experiments were carried out using the designed prompts rather than those generated by the GPT models.

Human Evaluation Automatic evaluation metrics are not always appropriate for natural language generation, especially for poetry-related tasks (Novikova et al., 2017; Refaee, 2023), where the translation of poetry cannot be directly evaluated due to its unique characteristics (Guo, 1957, 1983). The best poetry translators are often both poets and translators (Guo, 1983; Wang, 2006). Therefore, we invited six judges to evaluate the experimental results of eight modern English poems translated under eight different prompts. These judges are Chinese professional poets born between the 1960s and 1990s, who are members of the Chinese Writers Association.[5] They have a high level of poetry creation among poets of their age group

---

[4] We used the gpt-3.5-turbo version in this study.

[5] https://www.chinawriter.com.cn

Metrics	Models	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$
SacreBLEU	GPT-3.5	6.6	12.1	11.7	10.6	11.4	**15.3**	15.1	14.2
	GPT-4	7.8	9.1	9.9	13.9	11.1	**15.4**	14.2	13.1
BERTScore	GPT-3.5	0.5603	0.5648	0.5658	0.5601	0.5616	0.5766	**0.5777**	0.5728
	GPT-4	0.5455	0.5583	0.5533	**0.5721**	0.5589	0.5686	0.5693	0.5662
COMET	GPT-3.5	0.7750	0.7895	0.7865	0.7746	0.7816	0.7891	**0.7991**	0.7917
	GPT-4	0.7636	0.7677	0.7627	**0.7876**	0.7732	0.7812	0.7815	0.7774

Table 3: Results of English poetry translated into modern Chinese poetry under different prompts, evaluated with SacreBLEU, BERTScore (F1), and COMET metrics. $\mathcal{P}_{i\in\{1\ldots5\}}$ are the five prompts provided by GPT-4 itself, while $\mathcal{H}_{i\in\{1\ldots3\}}$ are the prompts we designed.

	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$
GPT-3.5	11	16	**21**
GPT-4	14	**19**	15

Table 4: Voting results of six professional judges on the output translation of eight poems by GPT-3.5 and GPT-4 under different prompts.

	0-shot	1-shot	3-shot	5-shot
GPT-3.5	**25**	13	12	10
GPT-4	**19**	16	13	12

Table 5: Voting results on the output translation of ten poems in different shot scenarios.

and possess their own representative poems or poetry collections. These judges represent different eras in Chinese poetry. Among them are university professors specializing in literature, PhDs in poetry, and professional translators of English-Chinese poetry. Importantly, they are native Chinese speakers who understand English, with some having engaged in poetry writing or English-Chinese poetry translation for many years. Specifically, we asked the judges to vote on the output translations of GPT-3.5 and GPT-4 under three human prompts through a questionnaire survey to select the most appropriate prompt. For three translations of the same source language poem, each judge had only one vote to choose the poem they believed to be of the best quality. Thus, each judge had a total of eight votes for GPT-3.5 and GPT-4, respectively. To obtain an objective and fair evaluation and avoid any bias against artificial intelligence or large models, we did not inform the judges that the poems were translated by ChatGPT when we invited them to complete the questionnaire. Table 4 shows the voting results of the six professional judges on the output translations of eight poems by GPT-3.5 and GPT-4 under different prompts. For GPT-3.5, $\mathcal{H}_3$ received 21 votes, at least 5 votes more than the other two prompts. For GPT-4, $\mathcal{H}_2$ received the most votes. Ultimately, we concluded that for GPT-3.5, the best prompt is $\mathcal{H}_3$, while for GPT-4, the best prompt is $\mathcal{H}_2$.

The results from both automatic and human evaluation demonstrate that the prompts we crafted achieve better results in poetry translation tasks. Compared to other prompts, $\mathcal{H}_3$ is the most appropriate for GPT-3.5, whereas $\mathcal{H}_2$ is the most appropriate for GPT-4.

4.3 Appropriate Demonstrations

By exploring the translation performance of GPT-3.5 and GPT-4 on modern poetry in different small sample scenarios, we aim to find the most appropriate method. Specifically, we randomly selected another ten poems (with no overlap with the poems in previous experiments) from the dataset as a new test set. GPT-3.5 and GPT-4 were tested under $\{0, 1, 3, 5\}$-shot scenarios in sequence using the most appropriate prompts identified from previous experiments. To ensure the fairness and reliability of the experiment, the demonstration examples used were excluded from the test samples (Gao et al., 2023). The prompts we designed for few-shot scenarios are provided in Appendix A.3.

Evaluation We invited an additional six professional poets to evaluate the model's outputs through a questionnaire, aiming to select the highest-quality translations under different shot scenarios. For translations of the same poem in different cases, each judge had only one vote to choose the poem they deemed to be of the best quality. Thus, each judge had a total of ten votes for GPT-3.5 and GPT-4, respectively. To ensure an unbiased evaluation, the judges were not informed of the origins of the different poems when casting their votes. Table 5 presents the voting results. Both GPT-3.5 and

GPT-4 received the most votes in the 0-shot scenario, which is contrary to the findings of Gao et al. (2023). Unexpectedly, as the number of samples increases, the performance of the models decreases. This suggests that the conclusions of previous work are not applicable to modern poetry translation. More samples did not promote the models to learn the characteristics of poetry. The superior performance in the zero-shot scenario is likely due to the inherent nature of modern poetry, which is characterized by its freedom and lack of constraints (Awan and Khalida, 2015). The essence of modern poetry lies in its open forms, free genres, departure from traditional narratives, and the absence of fixed language patterns (Guo, 1957; Wang, 2006; Skerratt, 2013). Providing demonstrations in the prompts appears to curtail the model's capacity for generating diverse translations, thereby constraining the inherently flexible nature of modern poetry. This is evidenced by the results presented in Table 5, where an increase in provided examples inversely correlates with human votes. Based on these observations, we can draw the following conclusions: The optimum performance of GPT-3.5 on the poetry translation task occurs in the 0-shot scenario using the third prompt ($\mathcal{H}_3$), and the best translation result is denoted as GPT3.5-Best. The optimum performance of GPT-4 occurs in the 0-shot scenario using the second prompt ($\mathcal{H}_2$), and the best translation result is recorded as GPT4-Best.

4.4 Case Study

Table 12 of Appendix A.4 presents the best translations of GPT-3.5 and GPT-4 on the poem, "Balance."[6]

Not Accurate Enough Every word in the poem cannot be replaced by another word, and it is necessary to find the most suitable translation (Owen, 1990). However, as we can see from the examples, GPT-3.5 and GPT-4 typically translate based on the surface meaning of the text, rather than choosing the most appropriate words based on context. For instance, "anchors" is translated literally as "锚定 (anchor)" by GPT-3.5 and "稳固 (stable)" by GPT-4. While "稳固 (stable)" is more suitable than "锚定 (anchor)," it still does not conform to the context of the original poem and is far from the meaning of "停泊 (anchor)" in the human translation. Additionally, "applauds" is translated by GPT-3.5 and GPT-4 as "拍手欢呼 (clapping and cheering)" and "鼓掌 (handclap)," respectively, losing the literary meaning.

Room for Improvement in Line-breaking Good poetry cannot exist without appropriate line-breaking. Line-breaking is more profound than the poem itself; it not only creates internal tension within the poem but also allows the reader to appreciate its beauty at a glance (Xue, 2016). However, there is still significant room for improvement in the line-breaking of translations generated by GPT-3.5 and GPT-4. For example, the sentence "She felt water on its ordained/journey to the sea" in the source poem was translated into a grammatically accurate sentence by GPT-4, resulting in the disappearance of tension between the lines of the poem. When translating English into Chinese, grammatical rules should not strictly be followed (Chen, 1993); the narrative can be incoherent, and the combination of words is free (Awan and Khalida, 2015).

Poeticity Is Not Well-perserved The poeticity of the source poem should be preserved in translation (Guo, 1957; Bian, 1989; Skerratt, 2013). However, we found that the most poetic sentence in the source poem, "pass beneath her rippling image," was lost when translated literally by GPT-3.5 and GPT-4. The original poem and the reference translation use anthropomorphic rhetoric to compare "water" to human beings, so it is translated as "经过." However, GPT-3.5 and GPT-4 translate "pass" as "流过," which clearly does not capture the meaning of the source poem, resulting in the omission of the poeticity during the translation process.

5 Explanation-Assisted Poetry Machine Translation

In light of the experimental results and analysis presented, it is evident that while ChatGPT can interpret and analyze poems (Chen and Chen, 2023), it still encounters numerous challenges. To enhance the performance of ChatGPT in poetry translation, we introduce a novel approach termed Explanation-Assisted Poetry Machine Translation (EAPMT). This method leverages monolingual poetry explanations as guidance information. Specifically, the process begins with ChatGPT generating an explanation for a poem presented in the source language. This explanation encompasses both the literal content and deeper meanings, contextualized within the cultural backdrop of the poem. Subsequently,

---

[6] http://poetrysky.com/quarterly/quarterly-4-carrigan.html

this explanation serves as guidance for ChatGPT to produce a translation into the target language. The two-step framework ensures that nuances and subtler aspects of the poem, often lost in conventional translation processes, are preserved. Figure 1 contrasts the novel Explanation-Assisted Poetry Machine Translation (EAPMT) framework with traditional translation methods, highlighting the advantages of the EAPMT approach. Table 6 presents the prompts designed for the translation process. A complete example of the EAPMT method is illustrated in Appendix A.5.

GPT-3.5:
Step 1
Please provide an explanation for this English poem:
English poem:{{X_i}}
Explanation:{{M_i}}

Step 2
Please translate this English poem into a modern Chinese poem based on its explanation:
Explanation:{{M_i}}
English poem:{{X_i}}
Modern Chinese poem:

GPT-4:
Step 1
Please provide an explanation for this poem:
Poem:{{X_i}}
Explanation:{{M_i}}

Step 2
Please provide the Chinese translation for this poem based on its explanation:
Explanation:{{M_i}}
Poem:{{X_i}}
Chinese translation:

Table 6: The prompts designed for Explanation-Assisted Poetry Machine Translation.

6 Experiments

Our experiments were performed using GPT-3.5 (gpt-3.5-turbo) and GPT-4 (gpt-4-1106), and involved comparisons with leading online translation systems. For Chinese literary texts, Baidu Translate[7] demonstrated superior performance over Google Translate (Razzaka et al., 2019). Consequently, Baidu Translate was selected as the online translation system for our study. Table 12 of Appendix A.4 presents the translations of our EAPMT method on the poem, "Balance."

7 Evaluation

7.1 Human Evaluation

Previous work on poetry generation has utilized diverse human evaluation criteria such as fluency, coherence, meaningfulness, poeticness, overall quality, and adequacy (Manurung et al., 2012; Zhang and Lapata, 2014; Yan, 2016; Yi et al., 2017, 2018; Seljan et al., 2020; Refaee, 2023). However, when translating English into Chinese, strict adherence to grammatical rules is not always necessary (Chen, 1993; Owen, 2003), and modern poetry often features discontinuous narration and a flexible combination of words (Awan and Khalida, 2015). Thus, our human evaluation framework does not prioritize fluency and coherence. Given the unique and irreplaceable nature of each word in a poem (Owen, 1990), accuracy remains an indispensable criterion in our evaluation of poetry translation. Additionally, the aspect of line-breaking, which distinguishes modern from classical poetry, plays a critical role in conveying the intrinsic rhythm and creating inner tension, thereby enhancing the visual and architectural beauty of the poem (Wang and Gao, 2018; Xue, 2016). For these reasons, appropriate line-breaking is also included in our evaluation criteria. Drawing from previous studies (Manurung et al., 2012; Zhang and Lapata, 2014; Yan, 2016; Yi et al., 2017, 2018; Seljan et al., 2020; Refaee, 2023) and reflecting the unique elements of modern poetry, we have designed a new human evaluation framework specifically tailored for the translation of modern poetry. We assess candidate translations comprehensively, focusing on eight key aspects, ranging from overall impact to specific details:

- **Overall Impression (OI)**: This criterion evaluates the general impact of the candidate translation as compared to the source poem or reference translation. It assesses whether the translation successfully captures the essence and tone of the original.
- **Similarity (Sim)**: Measures the degree of similarity between the candidate translation and the reference translation, focusing on stylistic and thematic alignment.
- **Fidelity (Fide)**: Assesses how faithfully the translation conveys the original poem's intent, emotions, and deeper meanings, thus evaluating whether the translation transcends mere linguistic equivalence to preserve the poem's core essence.

---

[7] https://fanyi.baidu.com

- **Line-breaking (Line)**: Evaluates the appropriateness of line breaks in the translation in comparison to the source poem or reference translation, considering how these contribute to the poem's rhythm and tension.
- **Meaningfulness (Mean)**: Examines the extent to which the translation conveys the original poem's meanings, exploring both surface-level content and deeper interpretative layers.
- **Poeticity (Poet)**: Assesses how well the poetic qualities of the original text, such as imagery, metaphor, and overall poetic effect, are preserved in the translation.
- **Accuracy (Acc)**: Focuses on the precision of translated elements, including words and word combinations, crucial to maintaining the integrity of the poem.
- **Errors (Erro)**: Identifies and categorizes errors in the translation, with a detailed scoring system that ranges from minor, ignorable mistakes to significant errors that alter the poem's meaning.

Each of these criteria utilizes a 6-point rating scale, where a score of 5 indicates equivalence to the reference translation, and a score of 6 signifies a superior quality compared to the reference translation. To ensure an objective evaluation, we enlisted six professional judges who were provided with only the source poem and the reference translation. The identities of the translators (human or machine) behind the candidate translations were not disclosed to the judges. Table 7 presents the human evaluation results, comparing our proposed method with traditional translation methods and online translation systems, thereby elucidating the effectiveness of our approach in translating modern poetry.

8 Analysis

As shown in Table 7, both our proposed method and traditional methods using GPT-3.5 and GPT-4 outperform online systems across all evaluation criteria. Furthermore, under all evaluation criteria, our method scored higher than the best performances of various versions of ChatGPT. Notably, in terms of accuracy, line-breaking, and overall impression, our EAPMT method achieved significant improvements, with scores at least 0.20 points higher than those of traditional methods (GPT3.5-Best & GPT4-Best). Traditional machine translation faces two primary challenges: first, addressing the ambiguities arising from the different ways languages express the same concepts (Andrabi et al., 2021); second, handling poor translation results when encountering rare words (Arthur et al., 2016; Sennrich et al., 2016; Luong et al., 2015). Our method effectively mitigates these issues in traditional machine translation. Specifically, in terms of accuracy, our method surpasses traditional methods by 0.35 points on GPT-3.5 and 0.31 points on GPT-4. This improvement is attributed to GPT-4's ability to generate poetic explanations based directly on the source poem, capturing its exact meaning and appropriate emotions, and even replicating specific words from the source. By incorporating these explanations into the prompts as guidance, GPT models can more accurately grasp the meaning of the source poem, thus reducing issues related to ambiguity and polysemy in translation. In terms of line-breaking, our method shows a 0.20-point increase with GPT-3.5 and a 0.33-point increase with GPT-4 compared to traditional methods. This is because the source poem in the prompt includes explicit line-breaking information, enabling ChatGPT to better learn and apply this information while comprehending the specific meaning of the poem. Regarding overall impression, our method achieved scores of 0.24 and 0.23 points higher than traditional methods on GPT-3.5 and GPT-4, respectively. This aligns with our survey results, where we invited 10 poets to evaluate what factors influenced their judgment of overall impression. All poets indicated that poeticity was the most critical factor, followed by line-breaking.

Surprisingly, we observed a score of 6 in the human evaluation results. To further analyze this, we counted the instances where each evaluation criterion received a score of 6 across six human judges evaluating the candidate translations of ten poems. As illustrated in Table 8, our proposed method achieved a score of 6 for both line-breaking and accuracy when using GPT-3.5 and GPT-4, indicating superior performance compared to the reference translations in these criteria. Notably, in terms of accuracy, our method was awarded a score of 6 on seven occasions with GPT-3.5 and five occasions with GPT-4.

8.1 Evaluation Using GPT-4

Previous studies, such as those by Chen and Chen (2023), have demonstrated that models like ChatGPT can effectively analyze and interpret the meaning of poetry. Building on this foundation, we

	OI	Sim	Fide	Line	Mean	Poet	Acc	Erro
Online System	3.00	3.02	3.15	3.85	3.27	3.08	3.12	3.00
GPT3.5-Best	3.58	3.50	3.60	4.50	4.02	3.82	3.75	3.70
EAPMT-3.5	**3.82**	**3.58**	**3.67**	**4.70**	**4.08**	3.93	**4.10**	**3.80**
GPT4-Best	3.77	3.53	3.65	4.25	3.97	3.82	3.82	3.70
EAPMT-4.0	**4.00**	**3.60**	**3.80**	**4.58**	**4.15**	**4.05**	**4.13**	**3.87**

Table 7: Results from human evaluations comparing our proposed method with traditional approaches and online translation systems.

	Line	Acc
EAPMT-3.5	3	7
EAPMT-4.0	3	5

Table 8: Count of each criterion receiving a score of 6 in the evaluation.

	English	Chinese
GPT-3.5	**5**	2
GPT-4	2	**3**

Table 9: Consistency of GPT-3.5 and GPT-4 in identifying the most poetic sentences from ten poems compared to human judgments.

employed GPT-4 to evaluate various translations based on the human evaluation criteria we established. As detailed in Appendix A.6, the evaluation results for GPT-4 indicate that our method surpasses traditional approaches in terms of overall impression, fidelity, meaningfulness, and poeticity. These findings reaffirm that our method significantly enhances the performance of GPT-4 in poetry translation tasks, as shown in Table 17 in Appendix A.6.

Determining Poeticity in Translated Texts Poeticity is inherently subjective; however, within the lines of the reference translation, one distinctly stands out as the most poetic. To explore whether GPT-3.5 and GPT-4 can accurately assess poetic qualities, we engaged six professional poets to select the most poetic sentence from each of ten poems. The sentence receiving the majority of votes was considered the most poetic. Concurrently, we tasked GPT-3.5 and GPT-4 with identifying the most poetic sentence from each of the ten source poems in English and their corresponding translations into Chinese, using language-specific prompts. We then evaluated the consistency of the selections made by GPT-3.5 and GPT-4 against those chosen by the human poets. As illustrated in Table 9, GPT-3.5 correctly identified the most poetic sentences in five English poems and two Chinese poems, indicating significant potential in recognizing poetic elements within English poetry. In contrast, GPT-4 accurately identified the most poetic lines in two English poems and three Chinese poems, demonstrating superior comprehension of Chinese poetry compared to GPT-3.5. These findings underscore the effectiveness of our designed prompts and the reliability of the human evaluation results. Specifically, in HP2, which exclusively involved Chinese poetry, human evaluation results indicated that $\mathcal{H}_2$ optimally harnesses GPT-4's potential in poetry translation. Similarly, $\mathcal{H}_3$, which involved translating English poetry into Chinese, showed that HP3 maximizes GPT-3.5's capabilities in poetry translation. This study not only demonstrates the varying capabilities of GPT models in identifying poetic qualities across languages but also highlights the importance of tailored prompt design to optimize model performance in specific linguistic contexts.

9 Conclusion

In this work, we examine ChatGPT's capabilities in English-Chinese poetry translation tasks by utilizing targeted prompts and small sample scenarios to ascertain optimal performance. To address ChatGPT's shortcomings, we propose an Explanation-Assisted Poetry Machine Translation (EAPMT) method, which leverages monolingual poetry explanation as a guiding information for the translation process. Furthermore, we refine existing evaluation criteria to better suit the nuances of contemporary poetry translation. The results from both human and machine evaluations demonstrate the effectiveness of our EAPMT method. In future work, we will try to correct the explanation generated in the first step with human experts or ChatGPT, thereby further improving the accuracy of the translation.

Limitations

Our research focuses on modern poetry, which has a very free format (Skerratt, 2013; Awan and Khalida, 2015; Duan, 2008). Our method may not be applicable to classical poetry and rhyming modern poetry. Therefore, the results of our study cannot represent or cover all types of poetry.

Ethics Statement

- **Datasets** The dataset we built consists of high-quality poems and translations. Poetry may contain negative emotions, but there is no information harmful to society.

- **Prejudice** In order to ensure the fairness and reliability of the results, we did not tell the human judges that the content they evaluated was generated by artificial intelligence during the process of human evaluation, but we told the truth to the judges after the evaluation and won the support of the judges.

- **The impact of this work on human translators** After the human evaluation was completed, we held discussions with professional poetry translators who work across various language pairs including English-Chinese, Russian-Chinese, Portuguese-English-Chinese, and Japanese-English-Chinese. They all concur that artificial intelligence, while a significant advancement, poses no existential threat to the domain of poetry translation. This consensus stems from the belief that AI technologies, including our own contributions, have not yet achieved the nuanced comprehension of modern poetry that human expertise offers. Conversely, these translators view our work as a promising support tool for novices in poetry translation, facilitating their initial forays into this intricate field.

Acknowledgements

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant No. FDCT/060/2022/AFJ, the mainland China collaboration project, National Natural Science Foundation of China Grant No. 62261160648), the Science and Technology Development Fund, Macau SAR (Grant No. FDCT/0070/2022/AMJ, the mainland China collaboration project, China Strategic Scientific and Technological Innovation Cooperation Project Grant No. 2022YFE0204900), the Multi-year Research Grant from the University of Macau (Grant No. MYRG-GRG2023-00006-FST-UMDF), and the Tencent AI Lab Rhino-Bird Gift Fund (Grant No. EF2023-00151-FST).

References

Noha Abdullah Alowedi and Arif Ahmed Mohammed Hassan Al-Ahdal. 2023. Artificial intelligence based arabic-to-english machine versus human translation of poetry: An analytical study of outcomes. *Journal of Namibian Studies: History Politics Culture*, 33:1523–1538.

Syed Abdul Basit Andrabi et al. 2021. A review of machine translation for south asian low resource languages. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(5):1134–1147.

D. Antar. 2023. The effectiveness of using chatgpt4 in creative writing in arabic: Poetry and short story as a model. In *Information Sciences Letters*.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567.

Abdul Ghafoor Awan and Yasmin Khalida. 2015. New trends in modern poetry. *Journal of Literature, Languages and Linguistics*, 13:63–72.

Zhilin Bian. 1989. The merits and demerits of translation in chinese new poetry since the may fourth movement. *Yilin*, pages 182–188.

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. *arXiv preprint arXiv:2109.02972*.

Rajesh Kumar Chakrawarti, Jayshri Bansal, and Pratosh Bansal. 2022. Machine translation model for effective translation of hindi poetries into english. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(1):95–109.

Baoya Chen. 1993. *Cultural theory of language*. Language and Culture.

Hung-Cheng Chen and Zhongwen Chen. 2023. Using chatgpt and midjourney to generate chinese landscape painting of tang poem 'the difficult road to shu'. *International Journal of Social Sciences*, 3(2).

Jinyan Duan. 2008. *Research on Zheng Zhenduo's Literary Translation*. Ph.D. thesis, Xiamen University.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *arXiv e-prints*, pages arXiv–2304.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. “poetic” statistical machine translation: Rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166, Cambridge, MA. Association for Computational Linguistics.

Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2018. 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)*, pages 67–71.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 524–533.

Moruo Guo. 1957. Talking about literary translation work. In *Complete Works of Guo Moruo*.

Moruo Guo. 1983. Guo moruo on creation. In *Shanghai Literature & Art Publishing House*. Guo Moruo on Creation.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Reza Khanmohammadi, Mitra Sadat Mirshafiee, Yazdan Rezaee Jouryabi, and Seyed Abolghasem Mirroshandel. 2023. Prose2poem: The blessing of transformers in translating prose to persian poetry. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–18.

Kaixin Lan, Tao Fang, Derek F. Wong, Yabo Xu, Lidia S. Chao, and Cecilia G. Zhao. 2024. Focus: Forging originality through contrastive use in self-plagiarism for language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19.

Ruli Manurung, Graeme D. Ritchie, and Henry S. Thompson. 2012. Using genetic algorithms to create meaningful poetic text. *Journal of Experimental & Theoretical Artificial Intelligence*, 24:43 – 64.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Stephen Owen. 1990. What is world poetry? *The New Republic*, 19:28–32.

Stephen Owen. 2003. *Stepping Forward and Back: Issues and Possibilities for "World" Poetry*. Modern Philology.

Jianhui Pang, Fanghua Ye, Derek F. Wong, and Longyue Wang. 2024. Anchor-based large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Md Abdur Razzaka, Md Serajul Islamb, Sharowar Hossainc, and Jinzhang Jiang. 2019. Exploring and learning english: An analysis of baidu and google translation. In *International Journal of Linguistics, Literature and Translation*, volume 2, pages 275–284.

Eshrag Ali Refaee. 2023. Okaz: A deep-learning-based system for automatic arabic poem generation. In *2023 3rd International Conference on Computing and Information Technology (ICCIT)*, pages 396–403. IEEE.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Sanja Seljan, Ivan Dunder, and Marko Pavlovski. 2020. Human quality evaluation of machine-translated poetry. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1040–1045.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. In *NeurIPS 2023 Workshop on Regulatable ML*.

Brian Phillips Skerratt. 2013. *Form and Transformation in Modern Chinese Poetry and Poetics*. Ph.D. thesis, Harvard University.

Wai Lei Song, Haoyun Xu, Derek F. Wong, Runzhe Zhan, Lidia S. Chao, and Shanshan Wang. 2023. Towards zero-shot multilingual poetry translation. *MT Summit 2023*, page 324.

Charles L Stevenson. 1957. On" what is a poem?". *The philosophical review*, 66(3):329–362.

Yuchun Wang. 2006. Moruo guo's view on literary translation. *Moruo Guo Journal*, page 5.

Zelong Wang and Zhouquan Gao. 2018. Review and reflection on the research of chinese modern poetry branch. *Journal of Central China Normal University: Humanities and Social Sciences Edition*, 57(4):8.

Shichang Xue. 2016. Something deeper than poetry itself – on the artistic intention of the line arrangement of poetry. *Write*, page 6.

Rui Yan. 2016. I, poet:automatic poetry composition through recurrent neural networks with iterative polishing schema. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, page 2238–2244.

Xinyi Yang, Runzhe Zhan, Derek F. Wong, Junchao Wu, and Lidia S. Chao. 2023. Human-in-the-loop machine translation with large language model. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 88–98.

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. 2017. Generating chinese classical poems with rnn encoder-decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings 16*, pages 211–223. Springer.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018. Automatic poetry generation with mutual reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3143–3153. Association for Computational Linguistics.

Runzhe Zhan, Xinyi Yang, Derek F. Wong, Lidia S. Chao, and Yue Zhang. 2024. Prefix text as a yarn: Eliciting non-english alignment in foundation language model. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.

A Appendix

A.1 Validation Experiments

To validate if the poems have been seen by the models, we prompted the models by providing various portions of the poem (or target translation) and asked the models to generate the rest of it. Table 10 shows an example of Poem Continuation using GPT-4 to complete an English poem, “Balance,” by providing the first 50% of the poem’s context.

A.2 GPT-4 Generated Prompts

To identify the best candidate prompts, we asked GPT-4 to generate five prompts for translating English poetry into modern Chinese poetry at the highest quality by using the following prompt: *“Provide five concise prompts or templates that can make you translate English poetry into modern Chinese poetry at the highest quality.”*

The prompts advised by GPT-4 are recorded as $\mathcal{P}_{i\in\{1\ldots5\}}$, respectively:

- $\mathcal{P}_1$: *Please translate the following modern English poem into modern Chinese poetry, considering its cultural and historical context in which it was written. Maintain the tone, style, and emotional impact of the original poem.*
- $\mathcal{P}_2$: *Translate this modern English poem into modern Chinese poetry, focusing on preserving the vivid imagery and metaphoric language. Ensure the translation conveys the same visual and sensory experiences as the original.*
- $\mathcal{P}_3$: *Translate the following modern English poem into modern Chinese poetry, making sure to maintain its rhyme scheme and rhythm to the best extent possible. The translation should attempt to replicate the musicality and flow of the original text.*
- $\mathcal{P}_4$: *Translate this modern English poem into modern Chinese poetry, ensuring the literal meaning of each line is accurately conveyed. The focus here is on the direct translation of the words and phrases, rather than on preserving the poetic devices used in the original.*
- $\mathcal{P}_5$: *Translate the following modern English poem into modern Chinese poetry, taking into account the unique style of the poet. Try to capture the author’s voice, style, and idiosyncrasies in the translated version.*

A.3 Few-shot Prompt

Table 11 provides the prompt template utilized for conducting few-shot translation experiments, illustrating the structured input provided to the model.

A.4 Examples of Poem Translation

Table 12 showcases an instance of the poem “Balance,” featuring the original English poem, its target Chinese translation, the translations generated by the GPT models based on their corresponding optimum prompts, as well as the translations generated by our Explanation-Assisted Poetry Machine Translation (EAPMT) method.

A.5 Examples of the EAPMT method

Taking the English poem “Balance” as an example, the prompts, explanations, and results of the Explanation-Assisted Poetry Machine Translation (EAPMT) through GPT-4 are illustrated in Table 13, Table 14, Table 15, and Table 16 respectively.

A.6 GPT-4 Evaluation Details

In addition to human evaluation, we conducted parallel evaluation experiments using GPT-4 to determine the consistency of the results with those obtained from human assessors. The experimental setup, including the specifically designed prompts, is detailed in Table 17. Table 18 presents a comparative analysis of the evaluation results for our proposed method against traditional methods and an online translation system, demonstrating the superiority of our approach.

Source Poem: Balance

The white crane
anchors to her liquid mirror
on one improbable leg.
She feels water on its ordained
journey to the sea
pass beneath her rippling image.
It is, she knows,
the way of things.
On the shore, bamboo
applauds
such grace
with swaying hands
clapping in the wind.

Input Prompt

Please continue writing the next 6 lines of the modern poem entitled “Balance”, which requires a total of 13 lines:
The white crane
anchors to her liquid mirror
on one improbable leg.
She feels water on its ordained
journey to the sea
pass beneath her rippling image.
anchors to her liquid mirror
It is, she knows,

Model Output

a delicate dance of give and take,
the art of holding still while letting go.
Around her, life in ceaseless flux,
yet she centers herself in the flow,
a serene pivot in the world’s relentless churn,
balancing the simple with the sublime.

Table 10: This illustrates the continuation of the English poem “Balance” using GPT-4, which was provided with the first 50% of the poem as context. The lower part displays the six sentences generated by the model. Notably, there is a significant divergence between GPT-4’s continuation and the original poem’s latter half.

GPT-3.5

Please translate this English poem into modern Chinese poetry:
Example(s):
English Poem:{{X_1}}
Modern Chinese Poem:{{Y_1}}
...
English Poem:{{X_k}}
Modern Chinese poem:{{Y_k}}
English Poem:{{English Poem}}
Modern Chinese Poem:

GPT-4

Please provide the Chinese translation for this poem:
Example(s):
Poem:{{X_1}}
Chinese Translation:{{Y_1}}
...
Poem:{{X_k}}
Chinese Translation:{{Y_k}}
Poem:{{English Poem}}
Chinese Translation:

Table 11: This table presents the prompts specifically crafted for our few-shot learning experiments, showcasing the tailored inputs used to optimize model performance in minimal data scenarios.

Source Poem	**Reference Translation**
Balance	平衡
The white crane	白鹤
anchors to her liquid mirror	停泊在她流体的镜子里
on one improbable leg.	以一只不太可能的腿。
She feels water on its ordained	她感到水在它注定的
journey to the sea	流向海洋的旅程上
pass beneath her rippling image.	从她起皱的影子下经过。
It is, she knows,	事情，她知道，
the way of things.	就是这样。
On the shore, bamboo	岸上，竹子们
applauds	为如斯的优美
such grace	喝彩
with swaying hands	用摆动的手
clapping in the wind.	在风中频频鼓掌。
GPT3.5-Best Translation	**GPT4-Best Translation**
平衡	平衡
白鹤	白鹤
用一条匪夷所思的腿	以一只难以置信的脚
锚定在她的液态镜面上。	稳固于她的水面镜像。
她感受着水	她感受到水正按其注定
在它注定的旅程中	的旅程向大海流去，
在她涟漪的映像下流过。	流过她波动的倩影下。
她知道，	这是，她知道，
这就是万物的方式。	万物的运行方式。
在岸边，竹子	岸上，竹子们
拍手欢呼	为这般优雅
这样的优雅	鼓掌，
随着风摇摆的双手	随风摇摆的手
鼓掌。	在风中相互拍击。
EAPMT-3.5	**EAPMT-4**
平衡	平衡
白鹤	白鹤
用一条难以置信的腿	以一脚不可思议地
根植于她的水镜中。	稳稳扎根于她的液态镜面。
她感受着水	她感觉到水在它注定的
在它注定的	通往大海的旅程中
流向大海的旅途中	流过她波动的倒影。
在自己泛起的波纹下穿行。	这，她知道，
她明白，	是万物的运行之道。
这就是事物的运行方式。	岸上，竹子
在岸边，竹子	以摇曳的手掌
用巨大的鼓掌	为这般优雅
向如此的优雅	鼓掌
鼓掌致敬，	在风中拍打。
随风摇曳的双手。	

Table 12: This table showcases an instance of the poem "Balance," featuring the original English text, its target Chinese translation, translations generated by the GPT models based on their corresponding optimum prompts, and the results produced by our proposed Explanation-Assisted Poetry Machine Translation (EAPMT) method.

Please provide an explanation for this poem:

Balance

The white crane
anchors to her liquid mirror
on one improbable leg.
She feels water on its ordained
journey to the sea
pass beneath her rippling image.
It is, she knows,
the way of things.
On the shore, bamboo
applauds
such grace
with swaying hands
clapping in the wind.

Explanation:

Table 13: Step (1) Prompt for Generating Poem Explanations: This table presents the prompts used to guide models in generating explanations for poems, laying the foundation for subsequent translation tasks.

"Balance" is a poem that draws upon the natural world to convey themes of stability, harmony, and the acceptance of life's flow. Here's a breakdown of the poem's imagery and meaning:

"The white crane
anchors to her liquid mirror
on one improbable leg."

The opening lines introduce a white crane standing on one leg, a common position for cranes, which they maintain for various reasons, including conserving body heat and resting. The crane's reflection on the water's surface is described as a "liquid mirror," emphasizing the stillness and clarity of the moment. The word "improbable" suggests a sense of wonder or marvel at the bird's ability to maintain such a poised stance.

"She feels water on its ordained
journey to the sea
pass beneath her rippling image."

The crane is acutely aware of the water moving around and under her, on its "ordained journey to the sea." This line introduces the concept of destiny or natural order—the water is following a path it is meant to take. The crane's "rippling image" suggests that while she is a part of this environment, she is also distinct from it, her presence momentarily altering the water's surface.

"It is, she knows,
the way of things."

These lines affirm the crane's understanding of the natural process, accepting the inevitability and the flow of life. This acceptance is reflective of a broader philosophical or spiritual understanding of life's transient nature.

"On the shore, bamboo
applauds
such grace
with swaying hands
clapping in the wind."

The final lines shift focus from the crane to the bamboo on the shore, personifying it as an audience that "applauds" the crane's grace. The bamboo's "swaying hands" clapping in the wind create an auditory image that complements the visual imagery of the poem. Bamboo is often associated with resilience, flexibility, and strength in Asian cultures, further enhancing the themes of balance and harmony.

Overall, the poem "Balance" uses the imagery of a crane balancing on one leg and the responsive natural environment to explore themes of equilibrium and acceptance. The crane's stillness amidst the flowing water mirrors the idea of finding stability within change, and the bamboo's applause symbolizes a recognition of the beauty and grace in living in harmony with the natural world. The poem suggests that there is an intrinsic order and beauty to life's processes, and there is grace in understanding and embodying this balance.

Table 14: GPT-4 Generated Explanations: This table displays the explanations generated by GPT-4, illustrating the model's interpretative output for a poem.

Please provide the Chinese translation for this poem based on its explanation:

Explanation: {Explanation generated by GPT-4 (Same as step 1)}
Poem: {English Poem "Balance" (Same as step 1)}

Chinese translation:

Table 15: Step (2) Integrating Explanations into the Translation Prompt: This prompt is specifically designed to incorporate the explanations as additional information, guiding the model in accurately translating the poem by providing deeper contextual insights.

平衡

白鹤
以一脚不可思议地
稳稳扎根于她的液态镜面。
她感觉到水在它注定的
通往大海的旅程中
流过她波动的倒影。
这，她知道，
是万物的运行之道。
岸上，竹子
以摇曳的手掌
为这般优雅
鼓掌
在风中拍打。

Table 16: Final Translation Result Produced by GPT-4: This entry provides the final translation outcome as generated by the GPT-4 model, showcasing its ability to interpret and render the original poem into the target language.

GPT-4 Evaluation Prompt Template

Please evaluate the following five candidate translations based on eight criteria: Overall Impression, Similarity, Fidelity, Line-breaking, Meaningfulness, Poeticity, Accuracy, and Errors. In this context, English poetry serves as the source language, and the reference translation is considered the gold standard. The score for each criterion should range from 1 to 6, with higher scores indicating superior translation quality. A score of 5 signifies that the translation is of comparable quality to the reference translation, while a score of 6 indicates that the translation surpasses the quality of the reference translation.

Each criterion is defined as follows: (Same as human evaluation criteria)

Source Language Poem: $\{\{X_i\}\}$

Reference translation: $\{\{Y_i\}\}$

Candidate translation 1: $\{\{C_{i1}\}\}$

Candidate translation 2: $\{\{C_{i2}\}\}$

Candidate translation 3: $\{\{C_{i3}\}\}$

Candidate translation 4: $\{\{C_{i4}\}\}$

Candidate translation 5: $\{\{C_{i5}\}\}$

The scores of different candidate translations under various criteria:

Table 17: The prompts we used for GPT-4 Evaluation.

	OI	Sim	Fide	Line	Mean	Poet	Acc	Erro
Online System	3.10	3.20	3.30	4.00	3.20	3.10	2.70	3.20
GPT3.5-Best	4.20	**4.70**	4.60	5.00	4.60	4.40	**4.50**	**4.70**
EAPMT-3.5	**4.30**	4.60	**4.60**	**5.00**	**4.60**	**4.40**	4.30	4.60
GPT4-Best	4.40	**4.80**	4.70	5.00	4.60	4.50	4.60	**4.90**
EAPMT-4.0	**4.70**	4.70	**4.80**	**5.00**	**4.80**	**4.80**	**4.70**	4.80

Table 18: Comparison of GPT-4 evaluation results for our proposed method, traditional methods, and the online translation system, Baidu Translate.