Benchmarking LLMs for Translating Classical Chinese Poetry: Evaluating Adequacy, Fluency, and Elegance

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown remarkable performance in translation tasks. However, the increasing demand for high-quality translations that are not only adequate but also fluent and elegant. To evaluate the extent to which current LLMs can meet these demands, we introduce a suitable benchmark (**PoetMT**) for translating classical Chinese poetry into English. This task requires not only adequacy in translating culturally and historically significant content but also a strict adherence to linguistic fluency and poetic elegance. To overcome the limitations of traditional evaluation metrics, we propose an automatic evaluation metric based on GPT-4, which better evaluates translation quality in terms of adequacy, fluency, and elegance. Our evaluation study reveals that existing large language models fall short in this task. To evaluate these issues, we propose RAT, a **R**etrieval-**A**ugmented machine **T**ranslation method that enhances the translation process by incorporating knowledge related to classical poetry. Our dataset and code will be made available.

1 Introduction

> *The three difficulties in translation are: adequate, fluent, and elegant.*
>
> – Yan, 1898

The emergence of large language models (LLMs), especially ChatGPT, has demonstrated impressive performance in translation tasks (Tyen et al., 2023; Liang et al., 2023; Guerreiro et al., 2023; Ranaldi et al., 2023; Zhao et al., 2024; Zhang et al., 2024; Chen et al., 2024a). As the requirements for translation quality continues to rise, translated results not only be adequate but also fluent and elegant (Wang et al., 2024; Huang et al., 2024; Gao et al., 2024; Wu et al., 2024). This raises a question: can existing LLMs meet such translation requirements, and if so, to what extent can they achieve this performance?

To answer this question, we introduced a suitable benchmark (PoetMT): translating classical Chinese poetry into English. Firstly, these poems carry culture and history, so the translated results need to adequately convey these meanings. Secondly, classical Chinese poetry has strict rules on rhyme, tone, and structure, making fluent translation a significant challenge. Lastly, classical Chinese poetry has aesthetic value, with the concise expressions of the classical Chinese language showing unique elegance, which needs to be preserved in translated results.

Based on the proposed PoetMT benchmark, previous automatic evaluation metrics for machine translation only analyze entire sentences without evaluating classical poetry translation quality explicitly (Papineni et al., 2002; Rei et al., 2022; Sellam et al., 2020a). To overcome the limitations of traditional evaluation metrics, we propose an automatic evaluation metric based on GPT-4 (OpenAI 2024), which better evaluates translation quality from the perspectives of adequacy, fluency, and elegance. Additionally, evaluating current LLM-based MT methods reveals that these translated results often lack historical and cultural knowledge, strict rhyme and structure rules, and concise expressions. To address these issues, we introduce RAT, a retrieval-augmented machine translation method. This method enhances the translation process by retrieving knowledge related to classical poetry, ensuring the translated output remains adequate to the original text while also achieving fluency and elegance.

To our knowledge, this is the first study evaluating the translation performance of LLMs based on the task of translating classical Chinese poetry. Through this effort, we aim not only to test the capabilities of LLMs in translating classical

Chinese poetry but also to inspire community discussion on the potential and future development of LLMs in translated texts that are adequate, fluent, and elegant.

Our contributions are summarized as follows:

- We have introduced the first classical poetry translation benchmark (PoetMT), which allows for a better evaluation of LLMs in terms of adequacy, fluency, and elegance.
- We have designed a new evaluation metric based on GPT-4 to evaluate classical poetry translation. This metric aligns more closely with human annotations and is better suited for the PoetMT benchmark.
- Based on the limitations of current LLM-based translation methods on the PoetMT benchmark, we have proposed a retrieval-augmented translation method to further enhance the performance of large models in this task.

2 Related Work

Poetry Machine Translation The initial attempts at poetry machine translation utilized phrase-based machine translation systems to translate French poetry into metrical English poetry (Genzel et al., 2010). This study demonstrated how statistical machine translation systems can produce translations while adhering to the rhythmic and rhyming rules of poetry. Chakrabarty et al. (2021) conducted an empirical study that highlighted a critical but often overlooked issue: even though they can maintain meaning and fluency, advanced machine translation systems trained on large amounts of non-poetry data struggle to preserve the original style of poetry. To address the issue of style preservation, some studies have managed to generate diverse styles by inputting sentences with specific styles into the encoder and embedding the target style into the decoder (Zhang et al., 2018; Liu and Wang, 2012). Considering that poetry often contains richly stylized semantic information and deep historical and cultural connotations, such as in classical Chinese poetry, Rajesh Kumar Chakrawarti and Bansal (2022) proposed a Hybrid Machine Translation model to improve the semantic and syntactic accuracy of poetry translations. Recently, Wang et al. (2024) have successfully implemented translations of modern poems from English to Chinese using the knowledge and multilingual capabilities of ChatGPT.

Chinese Classical Poetry Dataset Chinese classical poetry has several open datasets. Chen et al. (2019) published the first fine-grained dataset, annotating a manually fine-grained emotional poetry corpus containing 5,000 Chinese quatrains. Yutong et al. (2020) released a dataset containing 3,940 quatrains with automatically annotated themes and 1,917 quatrains with automatically annotated emotions, using a template-based method. Additionally, Liu et al. (2020) collected a parallel bilingual dataset of ancient and modern Chinese and aligned the lines using a string-matching algorithm. Based on these bilingual pairs, Li et al. (2021) constructed a matching dataset to evaluate models' semantic understanding capabilities. To our knowledge, this is the first benchmark for English translation of Chinese classical poetry, aimed at evaluating the translation performance of current large models in terms of "adequate, fluent and elegant".

3 Classical Chinese Poetry Dataset Construction

In this section, we discuss the design and construction of the PoetMT benchmark, including the rules and steps for building this benchmark.

3.1 Discourse-Level Poetry Translation

千山鸟飞绝，
From hill to hill no bird in flight;
万径人踪灭。
From path to path no man in sight.
孤舟蓑笠翁，
A straw-cloak'd man in a boat,
独钓寒江雪。
Fishing on river clad in snow.

Figure 1: An example block in the fluency and elegance in discourse-level poetry translation. The red parts indicate rhymes in both English and Chinese.

We collected a batch of classical Chinese poetry data and corresponding human English translations from online resources. The classical Chinese poetry database contains 1200 poems from Tang Poems, Song Poems, and Yuan Opera. Then, We manually screened 608 classical Chinese poems and their corresponding translations to serve as

Poem Type	Number of Poem	Unique Token	Average Tokens Per Sentence	Total Token Numbers
Tang	197	1980/3839	11.7/13.4	11727/13115
Song	189	2214/4899	10.9/14.1	16984/18212
Yuan	222	2006/3650	12.8/13.2	12145/1197
Total	608	3059/9223	11.69/13.6	40856/42524

Table 1: Statistics on the benchmark. Numbers a/b denote the corresponding number in source sentences/target sentences.

gold standards for translation source and target alignment. As shown in Figure 1.

The statistics of the PoetMT benchmark are shown in Table 1. We present the number of classical Chinese poetry, the number of unique tokens, the average number of tokens per sentence, and the total number of tokens in different poetry types. The source sentences in this benchmark have a moderate length, and the selected target translation sentences are well-aligned with the source in terms of length, indirectly reflecting the high quality of the reference sentences.

3.2 Adequacy in Sentence-Level Translation

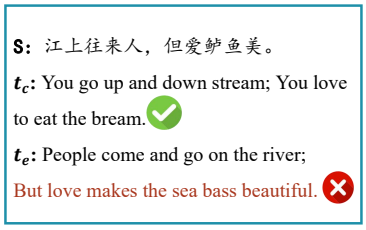

Figure 2: An example block in the adequacy in sentence-level poetry translation.

Due to the inclusion of historical background and common knowledge in classical Chinese poetry, achieving adequacy in translation poses a significant challenge. Therefore, to conduct a more detailed evaluation of adequacy, we have constructed a sentence-level test set.

We selected sentences containing historical knowledge and commonsense from the collected data of classical Chinese poetry. We avoid selecting semantically similar words to ensure diversity in the test set. Additionally, we preferred to select words that have different English translations depending on the context. Understanding the correct meaning requires considering factors like the specific context, historical background, writing style, and commonsense knowledge. This test set consists of 758 sentences, with each translated sentence represented as a triplet (s, t_c, t_e). As shown in Figure 2, s contains the source sentence with ambiguous words, t_c and t_e are the contrast translations, with the former being correct and the latter incorrect.

3.3 Classical Chinese Poetry Knowledge Base

Classical Chinese poetry holds rich historical and cultural nuances, but due to the limited resources for Classical Chinese, modern Chinese knowledge can greatly mitigate this issue. We collected a Classical Chinese Poetry Knowledge Base using open-source projects and internet resources. This Knowledge Base consists of 30,000 entries, including 30,000 Classical Chinese poems along with knowledge such as their corresponding historical background, dynasty name, modern Chinese translation, author introduction, modern Chinese analysis, and poetry type. The case is displayed in Appendix B.8.

4 LLM-based Evaluation Method

4.1 Evaluation Criterias

The translation of classical poetry requires not only artistic expression but also an understanding of the cultural background, making it nearly impossible to produce a single, universally correct translation. Following this line of thought, we evaluate the quality of classical poetry translations through three aspects: “Adequate," “Fluent," and “Elegant."

4.1.1 Adequate Criteria

Accuracy (Acc)↑: Focus on the precision of each element in the translation, accurately translating historical, cultural, and factual aspects, including words and phrases, to maintain the correct semantic and logical relationships of the poem.

4.1.2 Fluent Criteria

Beauty of Sound (BS)↑: The beauty of sound in Chinese classical poetry is primarily reflected in its rhyme. This standard examines whether the translation achieves harmonious sound, adherence to strict metrical rules, and a rhythm that is both smooth and dynamic. As shown in Figure 3(a).

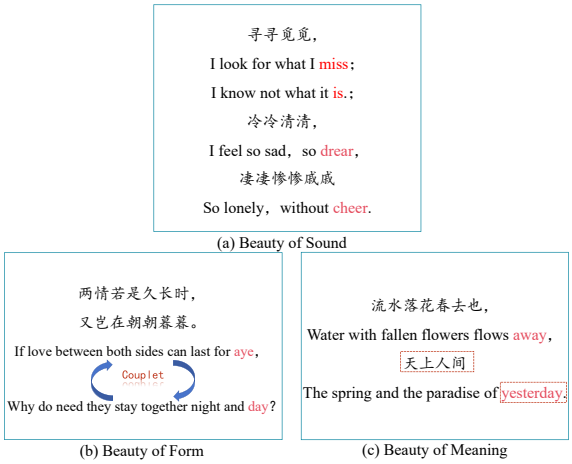

Figure 3: Examples of different evaluation metrics. Figure (a) represents the rhyme of the final words; Figure (b) shows that the two translated sentences have the same word count and couplet structure; Figure (c) indicates the accurate translation capturing the implied meaning of time passing.

Beauty of Form (BF)↑: Chinese classical poetry emphasizes symmetrical structures, with common forms including the "Five-character eight-line regulated verse (wulü)", "Seven-character eight-line regulated verse (qilü)", and "Extended forms (pailü)" among others. Each form showcases the structural characteristics of Chinese poetry. This standard evaluates whether the translation maintains consistency with the source poem's structure, including the alignment of line numbers and balanced phrasing. As shown in Figure 3(b).

4.2 Elegant Criteria

Beauty of Meaning (BM)↑: Chinese classical poetry uses concise and precise language to create vivid imagery and a rich atmosphere for readers. The criteria evaluate the depth and richness of the translation, focusing on the effectiveness of conveying themes, emotions, and messages. As shown in Figure 3(c).

4.3 LLM-based Classical Poetry Metric

In this section, we aim to establish a method for evaluating the translation of classical Chinese poetry based on LLMs. Drawing from current QE (Quality Estimation) research (Li et al., 2023; Kocmi and Federmann, 2023), we have developed an approach to utilize LLM in determining the quality of classical Chinese poetry translations. As shown in Figure 4, we designed a 1-5 scoring prompt to help LLM focus on the translation quality in terms of Beauty of Sound (LLM-BS). In the evaluation criteria, a score of 1 reflects a poor translation of classical Chinese poetry, a score of 3 indicates a basic but imperfect translation result and a score of 5 represents an excellent and accurate translation in this aspect. Subsequently, test data is provided to the LLM, and the LLM is asked to generate an evaluation with scores only. Additionally, we calculate the average scores (LLM-Avg) of BF, BF, and BM to evaluate the overall translation performance. The prompt details for Beauty of Form (LLM-BF) and Beauty of Meaning (LLM-BM) are provided in Appendix B.6 and B.7.

/* Task prompt */
Evaluate the beauty of sound in the given Chinese translation of classical poetry. Focus on whether the translation achieves harmonious sound, adherence to strict metrical rules, and a rhythm
1 point: Poor translation, lacks harmony and adherence to metrical rules, and fails to capture the beauty of sound.
2 point: Below average, some rhyme and meter present but with noticeable imperfections and awkwardness.
3 point: Basic translation, captures some aspects of sound beauty but with several imperfections in rhyme, meter, or rhythm.
4 point: Good translation, mostly harmonious with minor imperfections in sound quality or adherence to metrical rules.
5 point: Excellent translation, achieves harmonious sound, precise wording, strict adherence to metrical rules, and a smooth, dynamic rhythm.
/* Input Data */:

Original Chinese poem: {source}
English translation: {translation}
Evaluation (score only):

/*Output Text */:

{score}

Figure 4: Evaluation of the beauty of sound in Chinese translation of classical poetry

5 Proposed Method: RAT

The RAT (**R**etrieval **A**ugmented **T**ranslate) method was introduced to enhance translation performance by leveraging rich contextual information from the Classical Chinese Poetry Knowledge Base. We introduce the two workflows implemented in the RAT method in detail (in Figure 5). The first workflow uses poetry as input and retrieves relevant knowledge about the poetry from the Classical Chinese Poetry Knowledge Base through text-matching methods, obtaining different views of

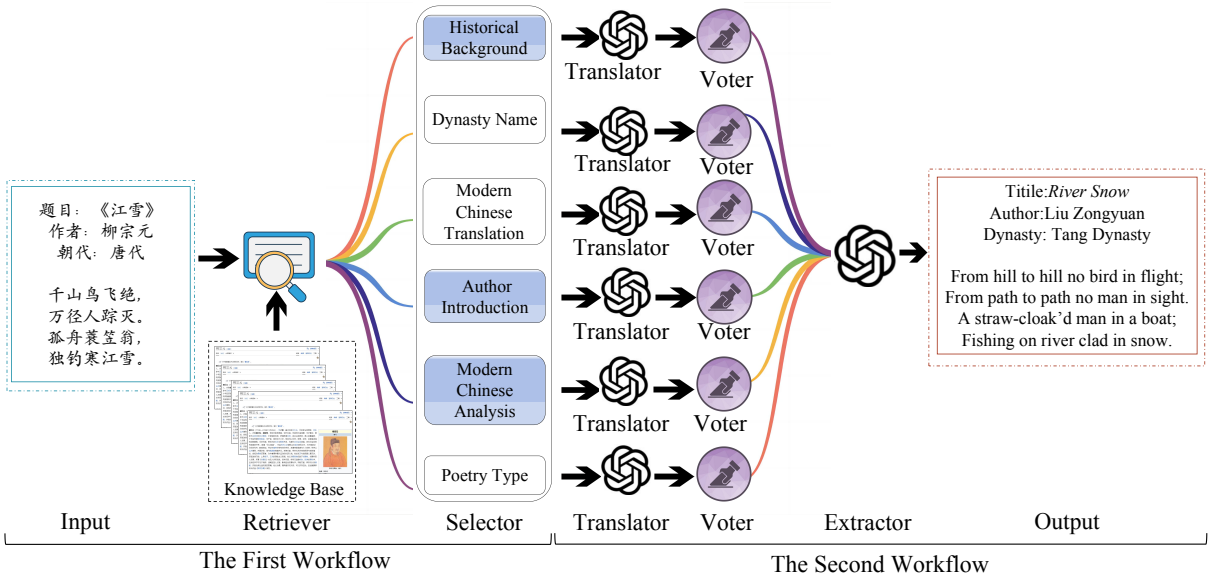

Figure 5: The proposed RAT framework. The "Historical Background," "Author Introduction," and "Modern Chinese Analysis" parts are at the discourse level, so the Selector needs to make selections based on the content.

the relevant knowledge. The second workflow translates based on the relevant knowledge from different views and integrates these to produce the final translation result.

5.1 The First Workflow

In the first workflow of RAT, there are two modules: Retriever and Selector.

1.Retriever. We propose a retrieval augmentation method to obtain knowledge relevant to translating classical Chinese poetry. Based on the Classical Chinese Poetry Knowledge Base, we use text-matching methods to retrieve uniquely relevant knowledge from multiple perspectives[1]. These perspectives include historical background, dynasty name, modern Chinese translation, author introduction, modern Chinese analysis, and poetry type.

2.Selector. The goal of the selector is to filter out irrelevant content to translated sentences from the results of the retriever. As an agent of the LLM, the selector first understands the historical background, author introduction, and modern Chinese analysis based on the source sentences, and then outputs content relevant to the translated source sentences. Specific prompts are displayed in Appendix B.1.

[1]The Classical Chinese Poetry Knowledge Base contains all 608 classical poems presented in the paper, ensuring a one-to-one correspondence between the poems and the knowledge.

5.2 The Second Workflow

In the second workflow of RAT, there are three modules: Translator, Voter, and Extractor.

1.Translator. The goal of the Translator is to translate classical Chinese poetry based on different types of retrieved knowledge. Six types of related knowledge were retrieved for classical Chinese poetry, resulting in six different translation outputs. Specific prompts are displayed in Appendix B.2.

2.Voter. The purpose of the Voter is to integrate translations based on different retrieval results to improve translation quality. As an agent of the LLM, the Voter votes on each sentence of all translation outputs based on the source input selects the highest-quality translations and concatenates these selected sentences to form the final translation result. Specific prompts are displayed in Appendix B.3.

3.Extractor. The goal of the Extractor is to extract the final translation results generated by the Voter. As an agent of the LLM, the Extractor filters out noise from the content generated by the Voter based on the source input and outputs the final translation results. Specific prompts are displayed in Appendix B.4.

6 Experiment Setup

6.1 Comparing Systems

In our evaluation, RAT is compared with a range of translation methods, including Zero-shot (Wei

et al., 2022), 5-shot (Brown et al., 2020; Hendy et al., 2023), Rerank (Moslem et al., 2023a), Refine (Chen et al., 2023), MAD (Liang et al., 2023), EAPMT (Wang et al., 2024), and Dual-Reflect (Chen et al., 2024a). To validate its generalizability, we utilize three LLMs, which include closed-source models such as ChatGPT(Ouyang et al., 2022) and GPT-4(Achiam et al., 2023) [2], as well as open-source models like Llama3-8B (Dubey et al., 2024) [3] and Vicuna-7B (Chiang et al., 2023) [4]. Details on comparative methods are in Appendix B.5.

6.2 Evaluation Metrics

LLM-based Automatic Evaluation. We propose an automatic evaluation method for translation based on LLMs in the section 4.

Traditional Automatic Evaluation. We initially employ COMET[5] (Rei et al., 2022) and BLEURT[6] (Sellam et al., 2020b) as automatic metrics, aligning with the established standards in LLM-based translation literature (He et al., 2023; Huang et al., 2024). For traditional translation evaluation, we use BLEU [7] (Papineni et al., 2002).

7 Experimental Results

7.1 Can LLM evaluate Chinese classical Poetry ?

We first translate randomly selected 100 discourse-level data from the PoetMT benchmark by the RAT method. We then compare the different evaluation results of the different automatic methods with human-annotated results to calculate the Pearson correlation coefficient (Pearson, 1920), Spearman correlation coefficient (Spearman, 1961), and Kendall correlation coefficient (Kendall, 1948) to determine the level of consistency.

The results in Table 2 indicate that large language models can serve as effective tools for evaluating the translation quality of classical Chinese poetry. In contrast, BLEU, COMET, and BLEURT were found to be less consistent with human evaluations, highlighting the advantages of our proposed evaluation method in the translation of classical poetry.

Metric	Pearson's r	Spearman's ρ	Kendall's τ
BLEU	-0.23	-0.18	-0.12
BLEU-1	0.05	0.08	0.05
BLEURT	0.14	0.16	0.11
COMET	0.13	0.18	0.11
LLM-BM	0.85	0.81	0.85
LLM-BF	0.71	0.75	0.70
LLM-BS	0.73	0.73	0.76
LLM-AVG	0.77	0.73	0.75

Table 2: correlation metrics between human and BLEU, BLEU-1, COMET, BLEURT, LLM-BM, LLM-BF, LLM-BS or LLM-AVG evaluation on our PoetMT benchmark.

7.2 Main Results

We compare various different LLM-based methods on the PoetMT benchmark with RAT. The results are shown in Table 3. The analysis of the experimental results is as follows:

- **The task of translating Classical Chinese Poetry is challenging.** The experiment shows that translating Classical Chinese Poetry is difficult, with low COMET/BLEURT/BLEU scores for current LLM-MT methods. This is because the translation is challenging and n-gram-based evaluation methods are not ideal for poetry. Additionally, GPT-4-based evaluation metrics, particularly in BS/BM/BF/AVG aspects, still need considerable improvement.

- **The effectiveness of RAT method.** The proposed RAT method consistently outperforms other comparative methods across all evaluation metrics, demonstrating the effectiveness of our approach.

- **Performance Variations Among Different Types of LLMs.** Among all comparative methods, closed-source models perform better on this task than open-source models, possibly implying that closed-source models benefit from richer pre-training data, thus enabling higher-quality translations. This also suggests that the Chinese Classical Poetry translation task is more challenging.

- **The effectiveness of retrieved knowledge.** For methods that rely on LLMs' self-generated knowledge (such as the EAPMT method), the RAT approach, which utilizes

[2]The ChatGPT and GPT-4 models used in this work are accessed through the gpt-3.5-turbo and gpt-4 APIs, respectively.
[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B
[4]https://huggingface.co/lmsys/vicuna-7b-v1.5
[5]https://huggingface.co/Unbabel/wmt22-comet-da
[6]https://github.com/lucadiliello/bleurt-pytorch
[7]https://github.com/mjpost/sacrebleu

Methods	Discourse-Level Poetry Transaltion										
	COMET	BLEURT	LLM-BM	LLM-BS	LLM-BF	LLM-Avg	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
GPT-4	60.3	43.0	4.0	3.7	3.6	3.8	4.0	22.1	7.8	3.3	1.7
ChatGPT	61.1	42.4	3.3	3.2	2.9	3.1	4.3	23.4	8.7	3.1	1.8
+5shot	61.0	42.5	3.5	3.3	3.3	3.4	3.7	22.0	7.7	3.2	1.6
+Rerank	61.0	42.5	3.7	3.7	3.9	3.8	3.7	22.5	8.0	3.4	1.7
+MAD	59.9	42.3	3.7	3.6	3.8	3.8	4.3	23.2	8.8	3.7	1.8
+Dual-Reflect	58.2	40.9	3.8	3.8	3.9	3.9	3.8	20.5	7.5	3.2	1.6
+EAPMT	61.1	42.9	3.8	3.7	3.8	3.7	3.7	21.6	7.5	3.1	1.5
+RAT	**62.7**	**43.9**	**4.1**	**3.9**	**3.9**	**4.0**	**4.6**	**23.9**	**9.8**	**3.9**	**2.2**
Llama3-8B	54.3	37.4	2.7	2.6	2.4	2.5	1.7	17.4	6.1	3.5	1.3
+5shot	54.5	37.6	2.9	2.8	2.6	2.7	1.8	17.4	6.2	3.4	1.3
+Rerank	54.8	38.1	3.0	3.3	3.5	3.2	2.0	17.9	6.6	3.6	1.5
+RAT	**55.6**	**38.4**	**3.4**	**3.3**	**3.6**	**3.4**	**2.2**	**18.2**	**7.0**	**3.9**	1.8
Vicuna-7B	52.2	26.4	2.4	2.4	1.8	2.2	0.3	16.5	4.7	3.4	1.0
+5shot	52.4	26.1	2.5	2.6	2.3	3.6	1.0	17.1	4.3	3.6	1.3
+Rerank	52.8	26.3	3.0	2.6	3.3	2.8	1.2	17.5	5.0	3.7	1.6
+RAT	**60.1**	**26.9**	**3.0**	**2.5**	**3.3**	**2.9**	**1.8**	**17.6**	**5.3**	**3.9**	**1.9**

Table 3: The main results from the PoetMT benchmark are presented. The bold indicates the highest scores.

retrieval-based knowledge, provides more accurate information and thus enhances translation quality. This suggests that more accurate relevant knowledge can better assist in the task of translating Chinese classical poetry.

7.3 Evaluation of Adequacy

To evaluate the translation performance of LLMs in terms of Adequacy, we employed a constructed dataset of 758 Classical Chinese Sentence-Level Translations to evaluate various translation methods. This experiment follows the method of Liang et al., 2023 and Chen et al., 2024b, evaluating translation results from two main dimensions: manual evaluation of translation adequacy (see Appendix A for details) and direct evaluation of translation quality using the LLM-BM.

Methods	LLM-BM	ACC
GPT-4	3.9	**69.1**
ChatGPT		
+Zero-Shot	3.2	60.5
+Rerank	3.2	64.4
+Dual-Reflect	3.7	66.4
+MAD	3.7	67.3
+RAT	**3.9**	**69.9**

Table 4: The LLM-BM and human-annotated results of the Adequacy in Sentence-Level PoetMT benchmark Translation.

The experimental results are shown in Table 4. In terms of translation adequacy, RAT performed the best, indicating that retrieving accurate information helps to better address issues of translation adequacy. In the LLM-BM evaluation, RAT received the highest score, further indicating that translation results based on RAT more accurately convey the themes, emotions, and messages of the original poems.

8 Analysis

8.1 LLM-based Metric Consistency

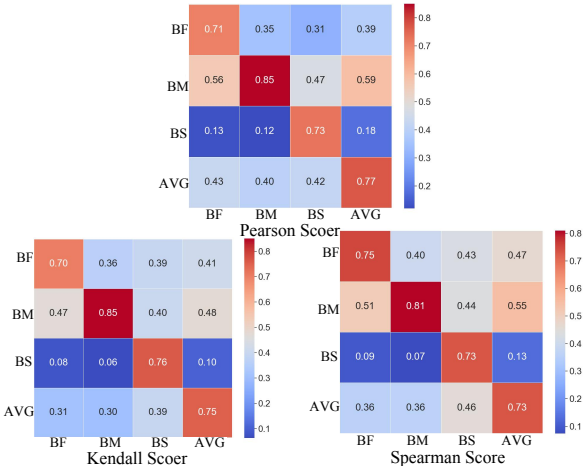

Figure 6: **LLM-based Metric Consistency Experiment.** In the heatmap, the horizontal axis represents the human evaluation results, and the vertical axis represents the LLM evaluation results.

This experiment evaluated whether the proposed LLM-based metrics (LLM-BS, LLM-BF, LLM-BM and LLM-AVG) accurately reflect Beauty of Sound, Beauty of Form, Beauty of Meaning, and overall translation quality. We conducted pairwise correlation tests between human and LLM-based evaluations using Pearson, Spearman, and Kendall correlation coefficients. The results are shown in Figure 6.

The experimental results indicate that, among all correlation coefficients, the consistency results based on the same annotations are significantly higher than the other results. This demonstrates the rationality of the evaluation settings for LLM-BS, LLM-B, LLM-BM, and LLM-AVG in the experiment.

8.2 Impact of Different Knowledge on Translation Performance

The RAT method utilized knowledge from the Classical Chinese Poetry Knowledge Base for translation. To investigate which specific knowledge is more effective in translating Classical Chinese Poems within knowledge-based translation methods, this experiment employed a modified RAT method that used only one type of retrieved knowledge at a time and removed the Voter module (The process of the RAT method is shown in Figure 5). The experimental results are presented in Figure 7.

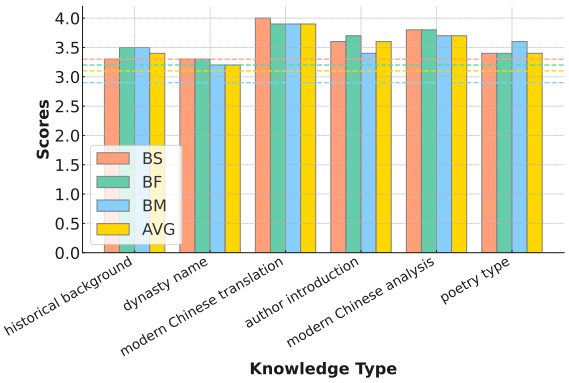

Figure 7: Experiment on the Impact of Different Knowledge of Classical Chinese Poetry on Translation. The dashed line indicates not using knowledge, but directly translating the result through ChatGPT.

On the one hand, retrieval-based translation methods showed overall performance improvement, indicating that knowledge plays a significant role in the task of translating Classical Chinese Poetry. On the other hand, the results revealed that knowledge of modern Chinese translation contributed the most to performance enhancement. This suggests that there are differences in form, semantics, and rhythm between Classical Chinese Poems and modern Chinese, further demonstrating the considerable challenges of directly translating Classical Chinese Poetry.

8.3 Translation Challenges Across Different Types of Classical Chinese Poetry

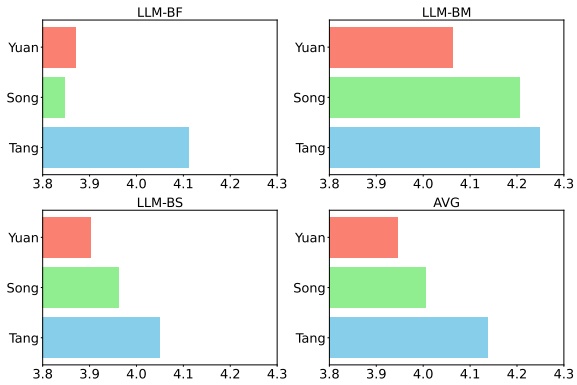

Figure 8: Experiment on the Impact of Different Types of Classical Chinese Poetry on Translation

To explore the translation difficulty of different types of poems (Tang, Song, and Yuan) in the collected 608 Chinese Classical Poems, we employed the RAT method to generate translation results for each poetry type. These results were then evaluated using LLM-BF, LLM-BM, LLM-BS, and LLM-AVG. The experimental results are shown in Figure 8.

From the experimental results, we can observe that the translation trends are consistent across different types. Firstly, the translation of Tang poetry is relatively easier compared to other types, likely due to its stricter format and fewer words. Secondly, the lower scores of LLM-BF and LLM-BS indicate that the translation of poetic structure and rhythm remains a significant challenge for LLMs. Lastly, the generally higher results of LLM-BM suggest that retrieval-based methods can produce more elegant translations.

9 Conclusion

Our research reveals the challenges LLMs face in translating classical Chinese poetry, especially in cultural knowledge, fluency, and elegance. We introduced a GPT-4-based evaluation metric, demonstrating that existing large language models fall short in this task. Additionally, we introduced the RAT method to improve translation quality. This study is the first to evaluate the limitations of large language models in the translation of classical poetry It aims to inspire discussion within the translation community about the potential and future development of LLMs.

10 Limitations

The inherent challenges of translating classical poetry, such as the preservation of rhyme, tone, and aesthetic qualities, remain complex and subjective. Although the proposed GPT-4-based automatic evaluation metric has demonstrated consistency with human evaluation, these subjective dimensions still pose a significant challenge.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. 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, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024a. DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 693–704, Bangkok, Thailand. Association for Computational Linguistics.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024b. Dual-reflect: Enhancing large language models for reflective translation through dual learning feedback mechanisms. arXiv preprint arXiv:2406.07232.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pages 4925–4931. ijcai.org.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models. arXiv preprint arXiv:2306.03856.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Ruiyao Gao, Yumeng Lin, Nan Zhao, and Zhenguang G Cai. 2024. Machine translation of chinese classical poetry: a comparison among chatgpt, google translate, and deepl translator. Humanities and Social Sciences Communications, 11(1):1–10.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Josef 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.

Nuno Miguel Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. CoRR, abs/2303.16104.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. ArXiv, abs/2305.04118.

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.

Yichong Huang, Xiaocheng Feng, Baohang Li, Chengpeng Fu, Wenshuai Huo, Ting Liu, and Bing Qin. 2024. Aligning translation-specific understanding to general understanding in large language models. arXiv preprint arXiv:2401.05072.

Maurice George Kendall. 1948. Rank correlation methods.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. ArXiv, abs/2308.13961.

Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. CCPM: A chinese classical poetry matching dataset. CoRR, abs/2106.01979.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2020. Ancient-modern chinese translation with a new large training dataset. ACM Trans. Asian Low Resour. Lang. Inf. Process., 19(1):6:1–6:13.

Ying Liu and Nan Wang. 2012. Sentence alignment for ancient and modern chinese parallel corpus. In International Conference on Artificial Intelligence and Computational Intelligence, pages 408–415. Springer.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive machine translation with large language models. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023, pages 227–237. European Association for Machine Translation.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023b. Adaptive machine translation with large language models. arXiv preprint arXiv:2301.13294.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Karl Pearson. 1920. Notes on the history of correlation. Biometrika, 13(1):25–45.

Jayshri Bansal Rajesh Kumar Chakrawarti 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.

Leonardo Ranaldi, Giulia Pucci, and André Freitas. 2023. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. CoRR, abs/2308.14186.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT 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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7881–7892, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020b. Bleurt: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696.

Charles Spearman. 1961. The proof and measurement of association between two things.

Gladys Tyen, Hassan Mansoor, Peter Chen, Tony Mak, and Victor Cărbune. 2023. Llms cannot find reasoning errors, but can correct them! arXiv preprint arXiv:2311.08516.

Shanshan Wang, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2024. What is the best way for chatgpt to translate poetry?

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. arXiv preprint arXiv:2405.11804.

Fu Yan. 1898. Evolution and ethics.

Liu Yutong, Wu Bin, and Bai Ting. 2020. The construction and analysis of classical chinese poetry knowledge graph. Journal of Computer Research and Development, 57(6):1252–1268.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024. Paying more attention to source context: Mitigating unfaithful translations from large language model. In Findings of the Association for Computational Linguistics ACL 2024, pages 13816–13836, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zhiyuan Zhang, Wei Li, and Xu Sun. 2018. Automatic transferring between ancient chinese and contemporary chinese. CoRR, abs/1803.01557.

Tiejun Zhao, Muven Xu, and Antony Chen. 2024. A review of natural language processing research. Journal of Xinjiang Normal University (Philosophy and Social Sciences), pages 1–23.

A Human Evaluations

In this section, we conduct a human evaluation to measure translation quality. We evaluated the adequacy of the translation. Four native English speakers were invited to participate. In the sentence-level adequacy task, the four experts scored each sentence for adequacy against the reference, awarding 1 point for fully adequate and 0 points for inadequate.

B Detail Prompt

B.1 Detailed prompt for Selector

Part-1: Selector: Please identify the knowledge related to the content to translating this classical Chinese poem {text} from the {rag context} knowledge base.

Input Text:

Source Poem, Sentence Length and Retrieved knowledge

Output Text:

Refined knowledge.

B.2 Detailed prompt for Translator

Part-2: Translator: Please translate this classial a Chinese poem {translate type} into a English poem {translate type}: Explanation:{rag context} Poem:{text}

Input Text:

Source Poem, Retrieved knowledge and Potery Type

Output Text:

Translated English Poem

B.3 Detailed prompt for Voter

Part-3: Iterative Refinement: Using the classical Chinese poem {src_text} as a source, compare six translation candidates to determine the highest quality result. Avoid including unrelated content. Here are the candidates: First, {s1}; second, {s2}; third, {s3}; fourth, {s4}; fifth, {s5}; sixth, {s6}.

Input Text:

Source Sentence, Translated Resluts based on six knowledge

Output Text:

Translated Result

B.4 Detailed prompt for Extractor

Part-4:Understanding-Based Translation: Extract only translation-relevant content from {target text} based on {text}. **Input Text**:

The final translation result.

Output Text:

Target Sentence t

B.5 Comparative Methods

The following content will provide detailed descriptions of these comparative methods:

- **Baseline**, standard zero-shot translation is performed in ChatGPT (Ouyang et al., 2022) and GPT-4 (Achiam et al., 2023). The temperature parameter set to 0, which is the default value for our experiments.
- **5-Shot** (Hendy et al., 2023), involves prepending five high-quality labelled examples from the training data to the test input.
- **Rerank** (Moslem et al., 2023a) was conducted with the identical prompt as the baseline, employing a temperature of 0.3 (Moslem et al., 2023b). Three random samples were generated and combined with the baseline to yield four candidates. The optimal candidate was chosen through GPT4.
- **Refine (Chen et al., 2023)** first requests a translation from ChatGPT, then provides the source text and translation results, and obtains a refined translation through multiple rounds of modifications by mimicking the human correction process.
- **MAD (Liang et al., 2023)** enhance the capabilities of LLMs by encouraging divergent thinking. In this method, multiple agents engage in a debate, while an agent oversees the process to derive a final solution.
- **EAPMT** (Wang et al., 2024) leverages the explanation of monolingual poetry as guidance information to achieve high-quality translations from Chinese poetry to English poetry.
- **Dual-Reflect(Chen et al., 2024a)** provide supervisory signals for large models to reflect on translation results through dual learning,

thereby iteratively improving translation performance (the maximum number of iterations is set to 5).

- **RAT** is the proposed method in this work.

B.6 Detailed prompt for Beauty of Form

For evaluation of the beauty of form, the detailed prompt is displayed in Figure 9

/* Task prompt */
Evaluate the translation of the given Chinese classical poem into English. Focus on whether the translation maintains consistency with the source poem's structure, including the alignment of line numbers and balanced phrasing.
1 point: Poor translation, disregards the poem's structure, and fails to convey its aesthetic qualities.
2 point: Some attempt to maintain structure but lack alignment and aesthetic consistency.
3 point: Basic structural elements are maintained but with noticeable imperfections in alignment and phrasing.
4 point: Good translation, with most structural elements preserved and minor issues in phrasing and alignment.
5 point: Excellent translation, accurately preserving the structure, alignment, and aesthetic qualities of the original poem.
/* Input Data */:

Original Chinese poem: {source}
English translation: {translation}
Evaluation (score only):

/*Output Text */:

{score}

Figure 9: Evaluation of the beauty of form in Chinese translation of classical poetry

B.7 Detailed prompt for Beauty of Meaning

For evaluation of the beauty of meaning, the detailed prompt is displayed in Figure 10

/* Task prompt */
Evaluate the translation of Chinese classical poetry for the beauty of meaning, focusing on whether the translation effectively conveys the themes, emotions, and messages of the original. This includes the use of concise and precise language to create vivid imagery and a rich atmosphere.
1 point: Poor translation, fails to convey the depth and richness of the original poetry.
2 point: Basic translation with significant shortcomings in capturing themes, emotions, and messages.
3 point: Satisfactory translation, conveys basic themes and emotions but lacks refinement or depth.
4 point: Good translation, effectively captures most themes, emotions, and messages with minor imperfections.
5 point: Excellent translation, accurately conveys the depth, richness, and atmosphere of the original poetry with full thematic and emotional resonance.
/* Input Data */:

Original Chinese poem: {source}
English translation: {translation}
Evaluation (score only):

/*Output Text */:

{score}

Figure 10: Evaluation of the beauty of meaning in Chinese translation of classical poetry

B.8 Knowledge Base Case

A case regarding the Classical Chinese Poetry Knowledge Base is shown in Table 5, demonstrating relevant knowledge of classical Chinese poetry.

Classical Chinese Poem: 慈母手中线，游子身上衣。临行密密缝，意恐迟迟归。谁言寸草心，报得三春晖。 (From the threads a mother's hand weaves, A gown for parting son is made.Sown stitch by stitch before he leaves, For fear his return be delayed. Such kindness as young grass receives from the warm sun can't be repaid.)

Historical Background: 《游子吟》写在溧阳。孟郊早年漂泊无依，一生贫困潦倒，直到五十岁时才得到了一个溧阳县尉的卑微之职，结束了长年的漂泊流离生活，便将母亲接来住。诗人仕途失意，饱尝了世态炎凉，此时愈觉亲情之可贵，于是写出这首发于肺腑，感人至深的颂母之诗。 ("Song of the Parting Son" was written in Liyang. In his early years, Meng Jiao lived a wandering and destitute life, experiencing poverty throughout his existence. It was not until he was fifty that he obtained a modest position as a county official in Liyang, which finally ended his years of wandering. He then brought his mother to live with him. Having faced the disappointments of his career and the coldness of society, he grew increasingly aware of the preciousness of familial bonds. Thus, he composed this deeply heartfelt poem in honour of his mother.)

Dynasty Name: 唐代 (Tang Dynasty)

Morden Chinese Translation: 慈母用手中的针线，为远行的儿子赶制身上的衣衫。临行前一针针密密地缝缀，怕的是儿子回来得晚衣服破损。有谁敢说，子女像小草那样微弱的孝心，能够报答得了像春晖普泽的慈母恩情呢？ (A loving mother uses her needle and thread to make clothes for her son, who is about to embark on a journey. She stitches each seam tightly, fearing that her son may return late and the clothes will be worn out. Who can dare say that a child's feeble filial piety, like a small blade of grass, can repay the boundless kindness of a mother, akin to the nurturing warmth of spring sunlight?)

Author Introduction: 孟郊，(751-814)，唐代诗人。字东野。汉族，湖州武康（今浙江德清）人，祖籍平昌（今山东临邑东北），先世居洛阳（今属河南）。唐代著名诗人。现存诗歌500多首，以短篇的五言古诗最多，代表作有《游子吟》。有“诗囚”之称，又与贾岛齐名，人称“郊寒岛瘦”。元和九年，在阌乡(今河南灵宝)因病去世。张籍私谥为贞曜先生。 (Meng Jiao (751-814) was a poet of the Tang Dynasty. His courtesy name was Dongye. He was of Han ethnicity and hailed from Wukang, Huzhou (present-day Deqing, Zhejiang), with ancestral roots in Pingchang (northeast of present-day Linyi, Shandong). His family originally resided in Luoyang (now in Henan). A renowned poet of the Tang era, he has over 500 surviving poems, most of which are short five-character ancient verses. His notable works include "Song of the Parting Son." He was known as the "Poet Prisoner" and was contemporaneous with Jia Dao, with the phrase "Jiao Han, Dao Shou" used to describe them together. He passed away in the ninth year of the Yuanhe era, in Wanquan (present-day Lingbao, Henan), due to illness. Zhang Ji posthumously honoured him with the title of "Mr Zhenyao.")

Modern Chinese Analysis: 开头两句用“线”与“衣”两件极常见的东西将“慈母”与“游子”紧紧联系在一起，写出母子相依为命的骨肉感情。三、四句通过慈母为游子赶制出门衣服的动作和心理的刻画，深化这种骨肉之情。母亲千针万线“密密缝”是因为怕儿子“迟迟”难归。前面四句采用白描手法，不作任何修饰，但慈母的形象真切感人。最后两句是作者直抒胸臆，对母爱作尽情的讴歌。这两句采用传统的比兴手法：儿女像区区小草，母爱如春天阳光。 (The opening two lines connect "the loving mother" and "the wandering son" through the commonplace items of "thread" and "clothes," highlighting the deep bond of flesh and blood between them. In the third and fourth lines, the mother's actions and thoughts as she makes clothes for her son further deepen this familial affection. The mother's meticulous stitching is driven by her fear that her son will return late. The first four lines employ a straightforward style, without embellishment, yet the image of the loving mother is vivid and touching. The final two lines express the author's heartfelt emotions, celebrating maternal love. These lines use traditional metaphorical techniques: children are like fragile blades of grass, while maternal love resembles the warm sunlight of spring.)

Poetry Type: 唐诗三百首,乐府,赞颂,母爱 (Three Hundred Tang Poems, Yuefu, Panegyric, Maternal Love.)

Table 5: A case about Classical Chinese Poetry Knowledge Base.