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

Andong Chen[1], **Lianzhang Lou**[3], **Kehai Chen**[2], **Xuefeng Bai**[*2], **Yang Xiang**[3], **Muyun Yang**[1], **Tiejun Zhao**[1], **Min Zhang**[2]
[1] Harbin Institute of Technology, Harbin, China
[2] Harbin Institute of Technology, Shenzhen, China
[3] Pengcheng Laboratory, Shenzhen, China
ands691119@gmail.com, {loulzh, xiangy}@pcl.ac.cn
{chenkehai, baixuefeng, yangmuyun, tjzhao, zhangmin2021}@hit.edu.cn

Abstract

Large language models (LLMs) have shown remarkable performance in general translation tasks. However, the increasing demand for high-quality translations that are not only adequate but also fluent and elegant. To assess 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. Our study reveals that existing LLMs fall short of this task. To address 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. Additionally, we propose an automatic evaluation metric based on GPT-4, which better assesses translation quality in terms of adequacy, fluency, and elegance, overcoming the limitations of traditional metrics. Our dataset and code will be made available [1].

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 need to be not only 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 introduce 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 linguistic poetic elegance, which needs to be preserved in translated results.

Compared with 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., 2020; Post, 2018a). To overcome the limitations of traditional evaluation metrics, we propose an automatic evaluation metric based on GPT-4 (Achiam et al., 2023), 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 powered by LLMs. This method enhances translation by retrieving classical poetry knowledge, ensuring adequacy, 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

* Corresponding author.
[1] https://github.com/andongBlue/PoetMT

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33019–33036
November 4-9, 2025

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 enhance the performance of LLMs in this task.

2 Related Work

2.1 Literary Text Translation

Poetry machine translation is a specific subfield within literary text translation (Wang et al., 2023b), which itself encompasses the challenges of translating artistic forms such as poetry. Early research by Genzel et al. (2010) utilized phrase-based systems to translate French poetry into metrical English, demonstrating that statistical MT can respect poetic rhythm and rhyme. Chakrabarty et al. (2021) highlighted that advanced systems, while fluent, often miss poetic style when trained on non-poetic data. To address this, studies embedded stylistic features into the translation process, such as encoding stylistically varied sentences in the encoder and incorporating target style in the decoder (Zhang et al., 2018; Liu and Wang, 2012). Given the cultural and historical significance of poetry, particularly in classical Chinese works, Rajesh Kumar Chakrawarti and Bansal (2022) proposed a Hybrid Machine Translation model to enhance both semantic and syntactic accuracy. More recently, Wang et al. (2024) leveraged ChatGPT's multilingual and knowledge-enhancing capabilities to translate modern English poems into Chinese, highlighting LLMs' potential in literary translation.

2.2 Ancient Text Datasets

The translation of ancient texts, particularly Chinese classical text, presents its own set of challenges due to the complexity and depth of these texts (alt, 2023; Wang et al., 2023a; McManus et al., 2023). Several datasets have been developed to address these challenges. Chen et al. (2019) introduced the first fine-grained emotional poetry dataset with 5,000 annotated Chinese quatrains. Yutong et al. (2020) expanded on this by releasing a dataset of 3,940 quatrains with automated theme annotations and 1,917 emotional annotations using a template-based method. Liu et al. (2020) compiled a bilingual parallel dataset of ancient and modern Chinese, aligning lines via a string-matching algorithm. This served as the foundation for Li et al. (2021), who developed a matching dataset to evaluate models' semantic understanding. Our proposed dataset is the first benchmark for evaluating the translation of Chinese classical poetry into English, focusing on "adequacy, fluency, and elegance."

2.3 LLM-as-a-Judge

LLM-as-a-Judge has emerged as an innovative evaluation paradigm, particularly in translation quality assessment. Leveraging the intrinsic capabilities of LLMs, it enables fine-grained evaluations and has shown high consistency with human evaluators Dong et al. (2023); Zheng et al. (2023); Gu et al. (2024). Kocmi and Federmann, 2023a introduced the GEMBA technique, using GPT-4 (Achiam et al., 2023) for DA score prediction, demonstrating that LLMs can match the performance of state-of-the-art multilingual models. Building on this, Fernandes et al., 2023 proposed fine-tuning LLMs for DA score prediction and error categorization, enabling more detailed evaluation. While these studies focus on general translation, this work examines multiple dimensions of translation in the context of Chinese classical poetry, offering a new perspective on evaluation in this field.

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

We collect a batch of classical Chinese poetry data and corresponding human English translations

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.

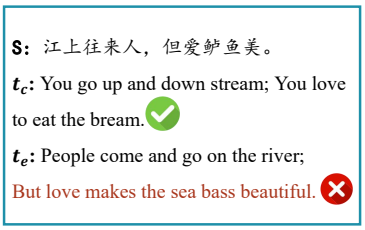


Figure 2: An example block in sentence-level poetry translation adequacy.

from online resources [2]. We manually screen 608 classical Chinese poems [3] and their corresponding translations from Tang Poems, Song Poems, and Yuan Opera [4]. An example of a single data is shown in Figure 1. Chinese Tang poetry from the Tang Dynasty (AD 618–907) is renowned for its strict forms and precise rhyming, highlighting mastery of structure and technique. Chinese Song poetry from the Song Dynasty (AD 960–1279) emphasizes individual emotion with a refined, restrained style that popularized diverse lyrical forms. Chinese Yuan opera from the Yuan Dynasty (AD 1271–1368) adopts a freer form, using colloquial language and dramatic elements to capture everyday life.

The statistics of the PoetMT benchmark are shown in Table 1. We present the number of classical Chinese poems, 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 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. The PoetMT benchmark includes a Classical Chinese Poetry Knowledge Base collected from 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 D.6.

3.3 Adequacy in Sentence-Level 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.

Following related works (He et al., 2020; Yao et al., 2024), we select sentences containing historical knowledge and commonsense from the collected 608 data of classical Chinese poetry. For historical knowledge and commonsense, the criteria are primarily based on the knowledge base we built. More specifically, the knowledge base corresponding to the poem includes historical knowledge, and if the words in the poem express clear commonsense, the poem is selected. We avoid selecting semantically similar words to ensure diversity in the test set. Additionally, we prefer to select words that have different English translations depending on the context. The final test set comprises 758 sentences, each representing as a triplet (s, t_c, t_e), where s is the source with ambiguous words, t_c is the correct translation, and t_e is the incorrect one (Figure 2).

4 LLM-based Evaluation Method

4.1 Evaluation Criteria

The translation of classical poetry requires not only artistic expression but also an understanding of the cultural background, yet the premise of correctness does not imply a singular or unique

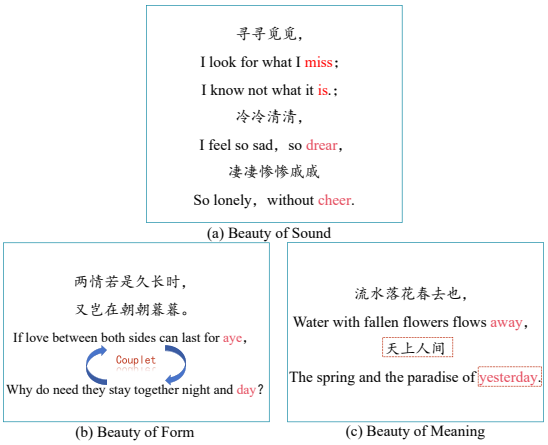
[2]We select professional translations by Xu Yuanchong, a renowned scholar of Chinese ancient poetry (Wikipedia: https://en.wikipedia.org/wiki/Xu_Yuanchong), ensuring high-quality results from experienced translators.

[3]In Appendix A, we discuss the details of the copyright of Chinese classical poetry.

[4]In the data we manually screened, we collect a total of 19 tang poems with 2 translation results. We have released these 19 poems as a subset in our open-source project

Poem Type	Number of Poems	Unique Tokens	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	11.7/13.6	40856/42524

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



(a) Beauty of Sound

(b) Beauty of Form

(c) Beauty of Meaning

Figure 3: Examples of evaluation metrics: (a) final word rhyme, (b) matching word count and couplet structure, (c) accurate translation of implied time passage.

expression. Following this line of thought, we evaluate classical poetry translations based on adequacy, fluency, and elegance.

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).

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).

4.1.3 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).

4.2 LLM-based Classical Poetry Metric

We propose a method for evaluating classical Chinese poetry translation using LLMs, inspired by QE research (Li et al., 2023; Kocmi and Federmann, 2023b). Our approach employs a 1-5 scoring prompt to assess translation quality across Beauty of Sound (LLM-BS), Beauty of Form (LLM-BF), and Beauty of Meaning (LLM-BM). A score of 1 indicates poor quality, 3 represents a basic but flawed translation, and 5 denotes excellence. The LLM generates scores, and we compute the LLM-Avg for overall evaluation. Prompt details are in Appendix C.6–C.8.

5 Proposed Method: RAT

The RAT method enhances translation by leveraging contextual information from the Classical Chinese Poetry Knowledge Base. Unlike traditional retrieval-based methods (Hoang et al., 2023), our approach uses retrieved content directly for translation with LLMs, employing natural language rather than representations. The workflow first retrieves poetry-related knowledge via text-matching, then integrates multi-view knowledge for translation.

5.1 The First Workflow

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

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 string-matching methods (Glück and Yokoyama, 2022)

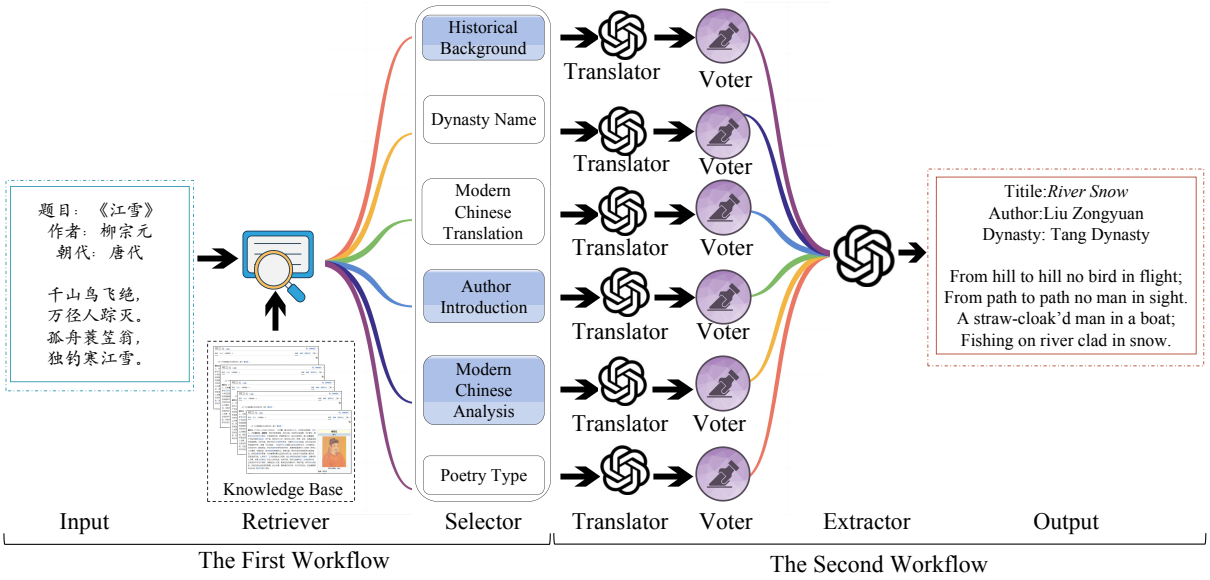


Figure 4: The proposed RAT framework. The "Historical Background," "Author Introduction," and "Modern Chinese Analysis" parts are at the discourse level, so the Selector needs to make selections based on the content.

to retrieve uniquely relevant knowledge from multiple perspectives [5]. These perspectives include historical background, dynasty name, modern Chinese translation, author introduction, modern Chinese analysis, and type.

Selector. The selector filters out irrelevant content from retriever results to enhance translation quality. As Table 16 shows, document-level knowledge often includes noise. Acting as an LLM agent, the selector understands the historical background, author, and modern Chinese analysis of the source poem, and outputs content more relevant to the input. Prompt details are in Appendix C.1.

5.2 The Second Workflow

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

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

Voter. The Voter integrates translations from different retrieval results to enhance quality. Acting as an LLM agent, it selects the highest-quality translations for each sentence based on the source input and concatenates them into the final result. Specific prompts are in Appendix C.3.

Extractor. The Extractor refines the Voter's output by filtering noise based on the source input, producing the final translation. Specific prompts are in Appendix C.4.

6 Experiment Setup

6.1 Comparing Systems

RAT is compared with various translation methods, including Zero-shot (Wei et al., 2022), 5-shot (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 test generalizability, we use closed-source models ChatGPT (Ouyang et al., 2022) and GPT-4 (Achiam et al., 2023) [6], and open-source models Llama3-8B (Dubey et al., 2024) [7], and Vicuna-7B (Chiang et al., 2023) [8]. For Chinese classical poetry translation, we also used the Chinese LLM Qwen-72B (Bai et al., 2023) [9]. Details on methods are in Appendix C.5.

6.2 Evaluation Metrics

LLM-based Automatic Evaluation. We propose an automatic evaluation method for translation

[5]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.

[6]via gpt-3.5-turbo and gpt-4-0613 APIs

[7]https://huggingface.co/meta-llama/Meta-Llama-3-8B

[8]https://huggingface.co/lmsys/vicuna-7b-v1.5

[9]https://huggingface.co/Qwen/Qwen-72B

Metric	Pearson's $r \uparrow$	Spearman's $\rho \uparrow$	Kendall's $\tau \uparrow$
	Traditional Automatic Evaluation		
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
	Qwen-72B-based Automatic Evaluation		
LLM-BM	0.63	0.59	0.61
LLM-BF	0.53	0.55	0.50
LLM-BS	0.54	0.53	0.55
LLM-AVG	0.57	0.53	0.54
	GPT-4-based Automatic Evaluation		
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.

based on LLMs as described in Section 4. The model used is GPT-4 (Achiam et al., 2023) [10].

Traditional Automatic Evaluation. We follow LLM-based translation standards (He et al., 2023; Huang et al., 2024), using COMET (Rei et al., 2022) and BLEURT (Sellam et al., 2020) as automatic metrics, and BLEU (Post, 2018a) for traditional evaluation.

7 Experimental Results

7.1 Can LLM evaluate Classical Poetry ?

We first translate randomly selected 100 discourse-level translation results from the PoetMT benchmark by the RAT method. Then, we calculate the translation scores using traditional automatic evaluation and LLM-based automatic evaluation methods. Furthermore, we score the translation results according to the criteria outlined in Figures 7, 8, and 9 through human evaluation (details in Appendix B.1). Finally, we compare the different evaluation results of the automatic methods with the human-evaluated 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. These correlation methods are indeed widely used in MT evaluation (Fomicheva and Specia, 2019; Isozaki et al., 2010; Gupta et al., 2015; Isozaki et al., 2010).

Table 2 shows that large language models effectively evaluate classical Chinese poetry translation, while BLEU, COMET, and BLEURT lack correlation with human judgment, underscoring our method's advantages (the multiple reference experiment in Appendix D.3). To assess potential bias from using ChatGPT in both RAT and evaluation (Panickssery et al., 2024), we test Qwen-72B, a Chinese-corpus-based model. Qwen-72B aligned better with human evaluation than traditional metrics but remained inferior to GPT-4, supporting the validity of our evaluation setup.

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 task of translating Classical Chinese Poetry is challenging. Experiments show that translating classical Chinese poetry is highly challenging. Traditional metrics like COMET, BLEURT, and BLEU yield low scores, with BLEU particularly unsuited for poetry. GPT-4-based evaluation also highlights significant gaps in BS, BM, and BF aspects.

The effectiveness of RAT method. The proposed RAT method outperforms all baselines across metrics, proving its effectiveness.

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 PoetMT task is more challenging.

The effectiveness of retrieved knowledge. The RAT method, leveraging retrieval-based knowledge, provides more accurate information than LLMs' self-generated approaches (e.g., EAPMT), leading to better translation quality and enhancing the PoetMT task.

7.3 Evaluation of Adequacy

To evaluate the translation performance of LLMs in terms of Adequacy, we employ 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 three main dimensions: manual evaluation of translation adequacy (see Appendix B.2 for details), the LLM-BM score based on GPT-4, and the BM score given by human (details in Appendix B.1). Results (Table 4 and 13) show that RAT achieves

[10]This work uses GPT-4 via the gpt-4-0613 API.

Methods	Discourse-Level Poetry Translation									
	COMET ↑	BLEURT ↑	LLM-BM ↑	LLM-BS ↑	LLM-BF ↑	LLM-Avg ↑	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑
GPT-4	60.3	43.0	4.0	3.7	3.6	3.8	22.1	7.8	3.3	1.7
ChatGPT	61.1	42.4	3.3	3.2	2.9	3.1	23.4	8.7	3.1	1.8
+5shot	61.0	42.5	3.5	3.3	3.3	3.4	22.0	7.7	3.2	1.6
+Rerank	61.0	42.5	3.7	3.7	3.9	3.8	22.5	8.0	3.4	1.7
+MAD	59.9	42.3	3.7	3.6	3.8	3.7	23.2	8.8	3.7	1.8
+Dual-Reflect	58.2	40.9	3.8	3.8	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	21.6	7.5	3.1	1.5
+RAT	**62.7**	**43.9**	**4.1**	**3.9**	**3.9**	**4.0**	**23.9**	**9.8**	**3.9**	**2.2**
Vicuna-7B	52.2	26.4	2.4	2.4	1.8	2.2	16.5	4.7	3.4	1.0
+5shot	52.4	26.1	2.5	2.6	2.3	2.4	17.1	4.3	3.6	1.3
+Rerank	52.8	26.3	3.0	2.6	3.3	2.8	17.5	5.0	3.7	1.6
+RAT	**60.1**	**26.9**	**3.0**	**2.5**	**3.3**	**2.9**	**17.6**	**5.3**	**3.9**	**1.9**
Llama3-8B	54.3	37.4	2.7	2.6	2.4	2.5	17.4	6.1	3.5	1.3
+5shot	54.5	37.6	2.9	2.8	2.6	2.7	17.4	6.2	3.4	1.3
+Rerank	54.8	38.1	3.0	3.3	3.5	3.2	17.9	6.6	3.6	1.5
+RAT	**55.6**	**38.4**	**3.4**	**3.3**	**3.6**	**3.4**	**18.2**	**7.0**	**3.9**	1.8
Qwen-72B	60.9	43.5	3.4	3.4	3.3	3.3	22.1	7.1	**3.0**	2.0
+5shot	60.4	43.8	3.6	**3.7**	3.4	3.5	21.5	7.2	2.9	1.5
+Rerank	59.8	43.2	3.0	3.3	3.5	3.2	20.6	6.7	2.7	1.3
+RAT	**61.7**	**43.5**	**3.7**	3.6	**3.6**	**3.6**	**22.9**	**8.0**	2.9	**2.0**

Table 3: The main results from the PoetMT benchmark are presented. The bold indicates the highest scores. The bolded results indicate the highest statistically significant scores (p-value < 0.05 in the paired t-test against all compared methods).

Methods	LLM-BM ↑	Human-BM ↑	ACC ↑
GPT-4	3.9	3.6	69.1
ChatGPT			
+Zero-Shot	3.2	3.2	60.5
+Rerank	3.2	3.3	64.4
+Dual-Reflect	3.7	3.6	66.4
+MAD	3.7	3.8	67.3
+RAT	**3.9**	**3.9**	**69.9**
Qwen-72B			
+Zero-Shot	3.1	2.2	43.9
+Rerank	3.3	2.3	42.7
+Dual-Reflect	3.0	2.0	46.3
+MAD	3.1	2.4	47.5
+RAT	**3.3**	**2.8**	**55.4**

Table 4: LLM-BM and human-annotated results for Adequacy in Sentence-Level PoetMT. Llama3-8B and Vicuna-7B results are in Appendix D.4.

the best adequacy scores. This suggests that retrieving accurate information improves adequacy. RAT achieves the highest LLM-BM score, best capturing the themes, emotions, and messages of the original poems.

7.4 Data Validation Experiments

Type of Poetry	Tang		Song		Yuan	
Language	Chinese	English	Chinese	English	Chinese	English
ChatGPT	6.6	0.4	4.4	0.6	1.7	0.4
GPT4	8.1	0.8	7.3	0.9	4.2	0.6
Qwen-72B	8.2	0.6	6.7	0.7	5.0	0.2

Table 5: BLEU Scores from data validation experiments

To explore whether PoetMT poems are included in the training data of closed-source LLMs like GPT-4,ChatGPT and Qwen-72B (§7.2), we conduct an experiment using 150 poems (50 each from Tang poetry, Song lyrics, and Yuan opera). Following concerns raised by Shi et al. (2024), we prompt GPT-4/ChatGPT/Qwen with the title and author to generate poems, then evaluate the similarity to human reference using SacreBLEU (Post, 2018b). As shown in Table 5, the results indicate low BLEU scores for both Chinese and English, suggesting limited task-specific data in the LLM training corpus.

7.5 Impact of Different Knowledge on Translation Performance

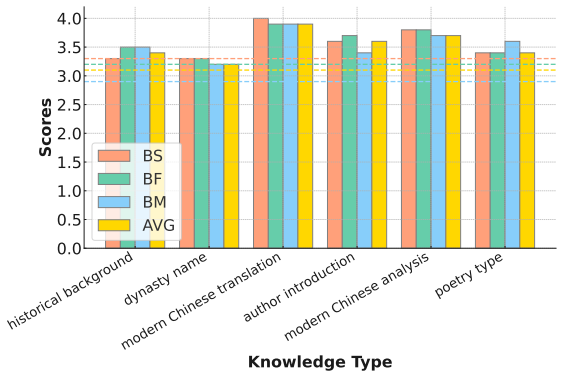


Figure 5: Experiment on the impact of different knowledge of classical Chinese poetry on translation. The dashed line indicates not using knowledge, but directly translating the result through ChatGPT.

The RAT method utilizes the Classical Chinese Poetry Knowledge Base for translation. To identify

the most helpful knowledge type, we modify RAT to use only one knowledge type at a time, removing the Voter module (Figure 4). Results (Figure 5) confirm that retrieval-based methods enhance performance, highlighting the importance of knowledge in poetry translation. Among them, modern Chinese translation knowledge contributes the most, suggesting its potential as an intermediary to mitigate PoetMT task.

7.6 Ablation Study on Modern Chinese Translations in RAT Framework

In Section 7.5, Modern Chinese translations in the RAT framework significantly impact output quality. To assess whether this improvement stems solely from these translations, we conduct an ablation experiment and case study.

	COMET ↑	BLEURT ↑	LLM-BM ↑	BS ↑	BF ↑	Avg ↑
ChatGPT-RAT	61.1	42.4	3.3	3.2	2.9	3.1
↪ only MC	57.2	38.1	3.1	2.6	2.7	2.8
Vicuna-RAT	60.1	26.9	3.0	2.5	3.3	2.9
↪ only MC	53.1	26.9	2.7	2.4	2.5	2.5

Table 6: Ablation study comparing RAT with and without Modern Chinese (**MC**) translations in the Knowledge Base.

Table 6 shows that while Modern Chinese aids translation, the multi-knowledge RAT method performs better. Case studies in Table 7 (with more in Appendix D.5) further highlight its limitations, as Modern Chinese-based translations resemble general-domain text and lack BF, BM, and BS.

7.7 Ablation Study on Components of RAT Framework

Since the RAT method we proposed requires retrieval, translation, selection of the best result, and extraction of translated text, we perform ablation experiments on each component to explore the effectiveness of each step in the current setup.

The experimental results, as shown in Table 8, indicate that the current settings of the RAT method are reasonable and yield the best translation results. Additionally, it is found that the w/o selector setup, which omits the knowledge selection step, significantly impacts the final translation performance due to the excessively long context.

7.8 Translation Challenges Across Different Types of Classical Chinese Poetry

To examine translation difficulty across Classical Chinese poetry (Tang, Song, Yuan) from 608 poems, we apply the RAT method and evaluate results using LLM-BF, LLM-BM, LLM-BS, and LLM-AVG (Figure 6). Findings reveal consistent trends: Tang poetry is easier to translate due to its stricter structure and brevity. Lower LLM-BF and LLM-BS scores highlight challenges in preserving poetic structure and rhythm, while higher LLM-BM scores suggest that retrieval-based methods enhance translation elegance.

Source: 红豆生南国，春来发几枝？愿君多采撷，此物最相思

RAT: Red beans grow in the south, sprouting many branches in spring. Pick them often, as they hold deep feelings of longing.

RAT-only Modern Chinese: Red beans grow in the sunny south, sprouting countless new branches every spring. I hope those who are missed will pick more of them, as they best express longing and love.

Reference: Red beans grow in the southern land, In spring, how many branches sprout? I wish you would gather them often, For they most evoke longing thoughts.

Table 7: Comparison of RAT, RAT-only Modern Chinese, and Reference Translations.

Methods	COMET ↑	BLEURT ↑	BS ↑	BM ↑	BF ↑
RAT	62.7	43.9	4.1	3.9	3.9
↪ w/o selector	61.0	42.5	3.6	3.3	3.4
↪ w/o voter	61.4	43.2	3.9	3.5	3.7
↪ w/o extractor	62.5	43.7	4.0	3.8	3.9

Table 8: Ablation results for RAT components.

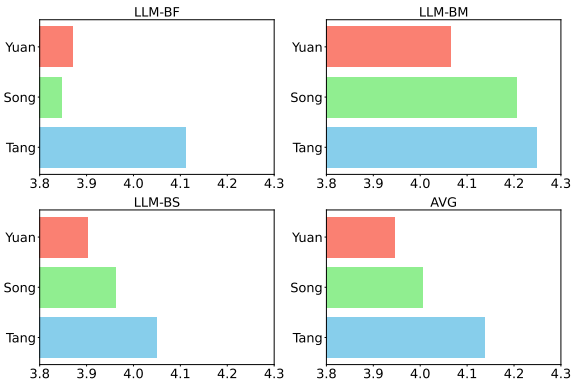


Figure 6: Experiment on the Impact of Different Types of Classical Chinese Poetry on Translation

7.9 Human-centered Error Analysis

To evaluate the RAT method's effectiveness and limitations, we manually assess 50 randomly selected poems from the 608 test samples. Using both direct translation and the RAT method based on ChatGPT, translations receive an average rating on a 1-5 scale for semantic adequacy, fluency, and elegance (see Figures 7, 8, 9). Table 9 shows that while RAT outperform the baseline, it still had a low proportion of Excellent (5-4) translations and a high proportion of Failed (2-1)

ones, underscoring PoetMT's challenges and the need for further improvement.

Categories	Number of Sentences	Rate
RAT		
Excellent	5	10%
Decent	23	46%
Failed	22	44%
ChatGPT		
Excellent	1	2%
Decent	12	24%
Failed	37	74%

Table 9: Manual evaluation results of 50 RAT and without RAT translations, categorized by performance.

Based on the results in Table 9, we manually categorize the failed outcomes from RAT and provide case examples for clearer illustration in Table 10. We further provide three cases in the paper with different ACC, BS, BF, and BM scores rated by both human annotators and LLM (GPT-4) in Table 15.

Categories	Rate	Source/Error Result/Reference
Errors in handling polysemous words	2/22	*Source:* 万壑树参天 *Error:* The trees in your valley scrape the sky *Right:* In myriad gorges, trees touch the sky
Lack of cultural context	7/22	*Source:* 秦时明月汉时关 *Error:* The moon still shines on mountain passes as of yore *Right:* Under the Qin moon, by the Han frontier
Confusion in long sentence structures	6/22	*Source:* 子弟每是个茅草岗沙土窝初生的兔羔儿乍向围场上走 *Error:* The young gallants are new-born bucks in chase of bunny *Right:* Young ones are like rabbits, new to the hunt, Born in a thatch of grass, on sandy ground
Incorrect translation of low-frequency vocabulary	7/22	*Source:* 缚虎手 *Error:* Binding a tiger with bare hands *Right:* Barehanded tiger fighting

Table 10: Translation Error Types with Examples.

8 Conclusion

Our research highlights the challenges LLMs face in translating classical Chinese poetry, particularly in cultural knowledge, fluency, and elegance. We introduce a GPT-4-based evaluation metric, demonstrating current models' limitations, and propose the RAT method to improve translation quality. This study is the first to evaluate LLM limitations in classical poetry translation, aiming to inspire future discussions in the MT community.

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.

Acknowledgements

We thank all the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (62276077, 62376075, 62406091, U23B2055), Guangdong Basic and Applied Basic Research Foundation (2024A1515011205), Shenzhen Science and Technology Program (ZDSYS20230626091203008, KQTD20240729102154066), and the Shenzhen College Stability Support Plan (GXWD20220817123150002, GXWD20220811170358002).

References

2023. *Proceedings of ALT2023: Ancient Language Translation Workshop*. Asia-Pacific Association for Machine Translation, Macau SAR, China.

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*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and T. Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *ArXiv*, abs/2304.06767.

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*.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, J. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Conference on Machine Translation*.

Marina Fomicheva and Lucia Specia. 2019. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.

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.

Robert Glück and Tetsuo Yokoyama. 2022. Reversible programming: A case study of two string-matching algorithms. In *Proceedings 9th Workshop on Horn Clauses for Verification and Synthesis and 10th International Workshop on Verification and Program Transformation, HCVS/VPT@ETAPS 2022, and 10th International Workshop on Verification and Program TransformationMunich, Germany, 3rd April 2022*, volume 373 of *EPTCS*, pages 1–13.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

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.

Rohit Gupta, Constantin Orăsan, and Josef van Genabith. 2015. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal. Association for Computational Linguistics.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics.

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*.

Cuong Hoang, Devendra Sachan, Prashant Mathur, Brian Thompson, and Marcello Federico. 2023. Improving retrieval augmented neural machine translation by controlling source and fuzzy-match interactions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 289–295, Dubrovnik, Croatia. Association for Computational Linguistics.

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*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Maurice George Kendall. 1948. Rank correlation methods.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. 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.

Stuart Michael McManus, Roslin Liu, Yuji Li, Leo Tam, Stephanie Qiu, and Letian Yu. 2023. The ups and downs of training RoBERTa-based models on smaller datasets for translation tasks from classical Chinese into modern standard Mandarin and Modern English. In *Proceedings of ALT2023: Ancient Language Translation Workshop*, pages 15–22, Macau SAR, China. Asia-Pacific Association for Machine Translation.

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.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

Matt Post. 2018a. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

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

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 P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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*.

Dongbo Wang, Litao Lin, Zhixiao Zhao, Wenhao Ye, Kai Meng, Wenlong Sun, Lianzhen Zhao, Xue Zhao, Si Shen, Wei Zhang, and Bin Li. 2023a. EvaHan2023: Overview of the first international Ancient Chinese translation bakeoff. In *Proceedings of ALT2023: Ancient Language Translation Workshop*, pages 1–14, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

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.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

A Copyright and Open-Source Licensing of Chinese Classical Poetry Resources

Regarding the copyright licensing of online resources, under current Chinese law, the copyright protection period is 50 years after the creator's death. Therefore, Tang poetry, Song lyrics, and Yuan drama have all exceeded the protection period and are in the public domain. Specifically, Tang poetry originates from the Tang Dynasty (618-907 AD), Song lyrics from the Song Dynasty (960-1279 AD), and Yuan drama from the Yuan Dynasty (1271-1368 AD), so collecting these works does not involve any copyright issues. In addition, several open-source projects related to Chinese classical poetry on GitHub have adopted the MIT license, which further demonstrates the feasibility of using an open-source license. Our work will open source the test data under the MIT license to ensure the legality and openness of the resources.

B Human Evaluations

B.1 Human Evaluation for BM/BF/BS score

Human evaluation is the core part of this study, providing a benchmark for automatic evaluation metrics. Each translation hypothesis is scored by 5 annotators using the "beauty of sound (BS), beauty of form (BF), and beauty of meaning (BM)" framework (see Figures 7/8/9). To ensure a high standard of evaluation, all annotators have a solid background in translation studies and at least one year of experience in poetry translation. Before the evaluation begins, they participate in calibration sessions where they review the scoring criteria and discuss examples to align their understanding of each dimension. This process helps to minimize subjective biases and ensures consistency across evaluations. After individual evaluations, the final annotation for each hypothesis is determined based on majority agreement. In instances where a clear majority is not reached, the median score is adopted to reduce the impact of any outlier ratings.

B.2 Human Evaluation for ACC

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.

C Detail Prompt

C.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.

C.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 Poetry Type

Output Text:

Translated English Poem

C.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

C.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

C.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.

C.6 Detailed prompt for Beauty of Sound

For evaluation of the beauty of form, the detailed prompt is displayed in Figure 7

C.7 Detailed prompt for Beauty of Form

For evaluation of the beauty of form, the detailed prompt is displayed in Figure 8

/* 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 7: Evaluation of the beauty of sound in Chinese translation of classical poetry

C.8 Detailed prompt for Beauty of Meaning

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

D Supplementary Experiment

D.1 LLM-based Metric Consistency

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 10.

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.

/* 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 8: Evaluation of the beauty of form in Chinese translation of classical poetry

/* 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 9: Evaluation of the beauty of meaning in Chinese translation of classical poetry

D.2 Impact of Smaller LLM Ensembles on RAT Performance

Further, although we discussed in Table 3, 4, and 13 that smaller LLMs do not yield better results for this task, we would like to further explore whether combining smaller LLMs with different characteristics can eliminate the bias introduced by a single smaller LLM. Here, we replace the Selector in RAT with the Chinese-based Qwen-72B, and the Voter with Vicuna-7B. The experimental results are as follows:

Method	COMET	BLEURT	BS	BM	BF
RAT-ChatGPT	62.7	43.9	4.1	3.9	3.9
RAT-QWen-Vicuna	60.4	42.1	3.7	3.0	2.6

Table 11: Performance comparison between RAT-ChatPT and RAT-Qwen-Vicuna.

Experimental results in Table 11 demonstrate that, despite using a model ensemble approach, the performance of methods based on smaller LLMs remains inferior to the current settings based on ChatGPT. This further attests to the effectiveness of our proposed method design.

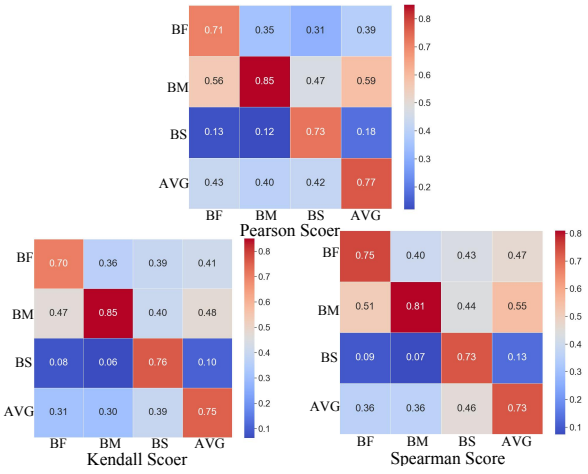


Figure 10: **LLM-based Metric Consistency Experiment.** In the heatmap, the horizontal axis represents the human evaluation results, and the vertical axis represents the LLM evaluation results.

D.3 Impact of Multiple References on BLEU Evaluation

In the MT community, BLEU can evaluate results with multiple references. Therefore, to explore the impact of multiple references on translation evaluation, we conducted experiments on 19 Tang poems with two translation outputs. The

translations were first generated using RAT and then manually evaluated following the settings in Section 7.1. Subsequently, the results were scored using BLEU, COMET, BLEURT, and LLM-BM/BF/BS. Finally, we determined the level of consistency through Pearson correlation coefficient (Pearson, 1920), Spearman correlation coefficient (Spearman, 1961), and Kendall correlation coefficient (Kendall, 1948).

Metric	Pearson's $r \uparrow$	Spearman's $\rho \uparrow$	Kendall's $\tau \uparrow$
	Traditional Automatic Evaluation		
BLEU	-0.27	-0.25	-0.15
BLEURT	0.06	0.10	0.08
COMET	0.07	0.12	0.07
	GPT-4-based Automatic Evaluation		
LLM-BM	0.79	0.79	0.80
LLM-BF	0.68	0.67	0.65
LLM-BS	0.70	0.69	0.72
LLM-AVG	0.72	0.69	0.71

Table 12: Correlation metrics between human evaluation and BLEU, COMET, BLEURT, LLM-BM, LLM-BF, LLM-BS, or LLM-AVG evaluation on our PoetMT dataset.

As shown in Table 12, although multiple references were considered in BLEU evaluation, the experimental results remain consistent with those in Section N. The findings suggest that the evaluation method of BLEU with multiple references does not lead to a significant improvement.

D.4 Additional Evaluation of Adequacy of Open-source LLMs

Methods	LLM-BM	Human-BM	ACC
Llama3-8B			
+Zero-Shot	2.1	1.1	32.5
+Rerank	2.1	1.2	32.4
+Dual-Reflect	2.5	1.7	34.4
+MAD	2.6	1.9	37.3
+RAT	**2.9**	**2.4**	**59.9**
Vicuna-7B			
+Zero-Shot	2.1	0.8	26.9
+Rerank	2.3	1.2	31.7
+Dual-Reflect	2.0	1.0	33.0
+MAD	2.2	1.3	-
+RAT	**2.5**	**2.1**	**43.4**

Table 13: The LLM-BM and human-annotated results of the Adequacy in Sentence-Level PoetMT benchmark Translation.

D.5 Additional Examples of Modern Chinese-Based Translation Limitations

To provide a more comprehensive understanding of the limitations inherent in Modern Chinese-based translation, Table 14 presents two additional examples. These examples further illustrate scenarios where the RAT method demonstrates superior performance due to its multi-knowledge integration, highlighting the gaps left by relying solely on Modern Chinese interpretations.

D.6 Knowledge Base Case

A case regarding the Classical Chinese Poetry Knowledge Base is shown in Table 16, demonstrating relevant knowledge of classical Chinese poetry.

Source: 水纹珍簟思悠悠，千里佳期一夕休。从此无心爱良夜，任他明月下西楼。

RAT: Ripples stir the jade mat as thoughts drift endlessly afar, A thousand miles of vows dissolve beneath the evening star. No longer shall fine nights find a place within my heart, Let the moonlight wane and vanish, where its path departs.

RAT-only Modern Chinese: Lying on the precious bamboo mat, my thoughts drift away with the flowing water. Suddenly, the winds and clouds have changed, and the cherished promise spanning a thousand miles has been shattered in an instant. From now on, I no longer have the heart to appreciate the beautiful moments, whether the moon rises above the eastern tower or sets beneath the western tower.

Reference: Ripples on the jade mat, thoughts drift far and wide, A thousand miles of promise, in one night, all denied. From now on, I'll have no heart for such fine nights, Let the bright moon set where it will, beyond my sight.

Source: 林暗草惊风，将军夜引弓。平明寻白羽，没在石棱中。

RAT: Through shadowed woods and rustling grass, The general's bow is drawn at last. By dawn he seeks the arrow's flight, Its white fletch lodged in stone, out of sight.

RAT-only Modern Chinese: In the dim woods, the grass suddenly rustles in the wind, prompting the general to quickly draw his bow and shoot in the cover of night. At dawn, he searches for the white-fletched arrow fired the previous evening, finding its tip deeply embedded in a massive stone.

Reference: Amid the dark woods and startled grasses, The general draws his bow at night. At dawn he seeks the white fletching, Embedded in the stone's edge, out of sight.

Table 14: Comparison of RAT, RAT-only Modern Chinese, and Reference Translations.

Source	Model	Translation	ACC/BS/BF/BM
江上往来人，但爱鲈鱼美。	ChatGPT	People come and go on the river; but love makes the sea bass beautiful.	0/3/2/2
	Qwen-72B	People pass by the river, loving the delicious bass.	1/2/3/3
	RAT(ChatGPT)	You go up and down the stream; you love to eat the bream.	1/4/4/3
白日依山尽，黄河入海流。	ChatGPT	The white sun stays on the mountain; the Yellow River goes into the sea.	1/2/3/2
	Qwen-72B	The setting sun leans on the mountain; the Yellow River flows to the sea.	1/3/3/2
	RAT(ChatGPT)	The sun beyond the mountains glows; the Yellow River seawards flows.	1/4/3/3
昨夜雨疏风骤，浓睡不消残酒。	ChatGPT	Last night the rain was sparse, the wind blew hard. Even deep sleep could not dispel the wine still lingering.	1/2/3/3
	Qwen-72B	Last night, the rain was sparse and the wind was strong. A deep sleep did not eliminate the leftover wine.	0/2/2/2
	RAT(ChatGPT)	Rain thinned, wind roared through the night; Drunken sleep could not wear off the wine.	1/3/4/4

Table 15: Comparison of different models' translations with evaluation scores.

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's 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 16: A case about Classical Chinese Poetry Knowledge Base.