

Poller: Are LLMs Suitable for Evaluating the Poetry Understanding Task?

Shanshan Wang¹ Derek F. Wong^{1*} Jingming Yao² Lidia S. Chao¹

¹NLP²CT Lab, Department of Computer and Information Science, University of Macau
nlp2ct.shanshan@gmail.com, {derekfw, lidiasc}@um.edu.mo

²Department of Portuguese, Faculty of Arts and Humanities, University of Macau
jmyao@um.edu.mo

Abstract

Traditional automatic evaluation methods have been shown to be unsuitable for modern Chinese poetry because of the distinct nature of this literary genre. Human evaluation remains reliable, but is expensive and not applicable to large-scale data. In this paper, we propose Poller (Poetry LLM Evaluator), a novel method leveraging large language models (LLMs) to evaluate the poetry understanding task. Specifically, our method requires LLMs to play the role of a poem’s author with detailed information, thereby emulating human evaluation and judgment by adopting the poet’s perspective. We conducted comprehensive experiments on multiple LLMs, evaluating the interpretations of poems across eight specialized dimensions. Experimental results demonstrate that our method effectively reduces the evaluation error between LLMs and humans. Especially for specific dimension evaluation, Poller-based LLMs achieve a 94.55% and 89.53% error reduction for rhetorical techniques and defamiliarization, respectively, compared to baseline methods. These performances are unattainable by conventional LLM evaluation methods. Experimental results from multiple LLMs across various dimensions validate the efficacy of our method. This work bridges the gap between automated efficiency and human expertise, establishing a foundation for automated evaluation in poetry-related tasks.

1 Introduction

Poetry has attracted the attention of researchers in natural language processing (NLP), with topics such as poetry understanding (Wang et al., 2026), generation (Hutson and Schnellmann, 2023; Antar, 2023; Deng et al., 2024), detection (Wang et al., 2025; Li et al., 2026), and translation (Virvou et al., 2023; Song et al., 2023; Wang et al., 2024).

Effective and reliable evaluation is indispensable for promoting the development and progress

*Corresponding Author



Human

From now on, you are a professional Chinese modern poet and need to play the role of a judge. Your details are as follows:

- Basic Information: You are a native Chinese speaker... You are a professional poet...
- Your Own Experience and Views on Poetry: The earliest motivation for writing...
- Others' Comments (Professional Poetry Critics) on You and Your Published Poems: One distinctive feature of his poetry lies in its capture of ephemeral insights...

The above are your details. Now, you need to play the role of a professional poet with the above characteristics. Your task is to play the role of a judge. Specifically, you need to score the following "reader's understanding of the poem" from the eight dimensions of content, language... and give detailed reasons for your score. Please note that you only need to score and give reasons for the "Reader's Understanding of this Poem", not the poem itself.

- Meaning of Each Scoring Dimension: (1) Content: Summarize...
- Scoring Rules: Each dimension range from 0 to 100...

Please remember to state your own role before giving your formal answer.

Poem:

The Matter of Being Drunk (Title)
In the kingdom of wine, turtles might
Depart the sea, and birds need not perch in trees...

Reader's Understanding of this Poem:

- (1) Content: This poem describes the state of a drunkard...
- (2) Language: This poem uses direct and concrete...



LLM-Poller

Role: Professional Poet, University Professor...

Evaluation and Reason:

- (1) Content (90 points): The reader's understanding of the poem's content is relatively accurate, capturing the drunk's...
- (2) Language (85 points): The reader correctly points out the poem's use of direct and figurative language...but...

Summary: The reader's understanding of the poem is generally accurate and in-depth, especially in terms of thought and...

Figure 1: An example of our proposed Poller (Poetry LLM Evaluator) method.

of NLP (Papineni et al., 2002; Sun et al., 2019; Wang et al., 2023b). Different tasks and domains require distinct evaluation criteria (Li et al., 2024). While automatic evaluation is prevalent in NLP, it is not well-suited for poetry-related tasks due to the unique characteristics of poetry, such as its use of figurative language, rhythm, and imagery (Refaee, 2023; Wang et al., 2024; Novikova et al., 2017). In addition, the reliability of static evaluation methods may decrease when the distribution shifts occur (Shen et al., 2026). Human evaluation, often regarded as the gold standard, serves as the reference for automatic evaluation methods (Resnik and Lin, 2010), particularly for creative tasks like poetry. Previous poetry-related works relied heavily on human evaluation (Manurung et al.,

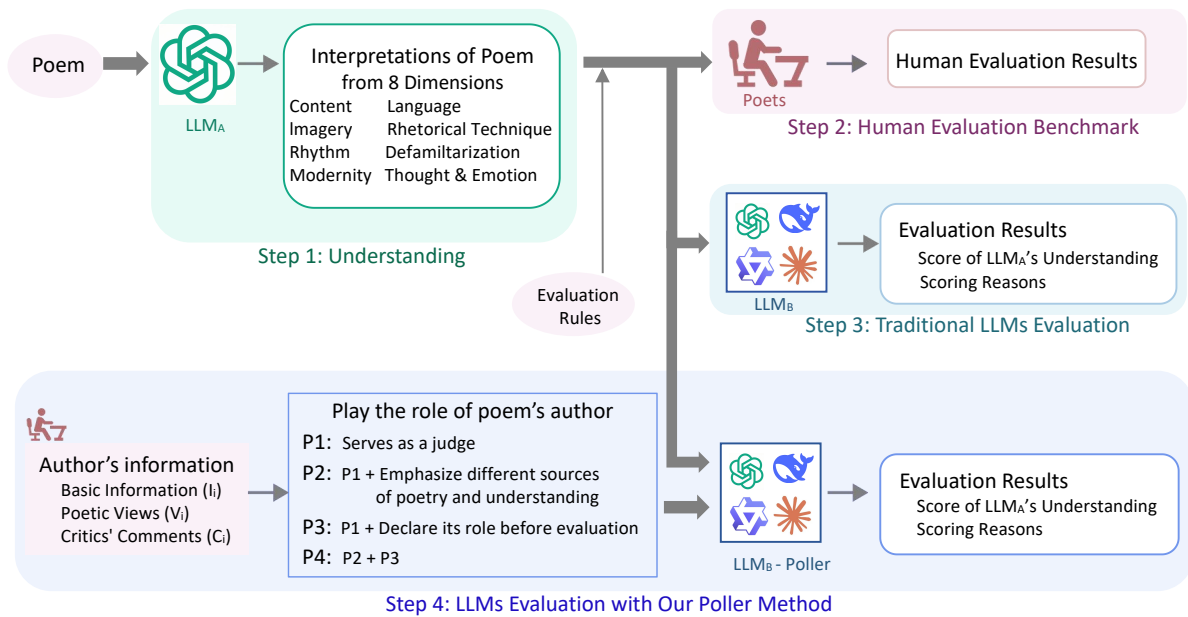


Figure 2: The overall workflow of our work, consisting of four components. Poller is the core method we propose.

2012; Zhang and Lapata, 2014; Yan, 2016; Yi et al., 2018). Although human evaluation is most suitable for poetry-related tasks, it is costly and not applicable to large-scale data (Papineni et al., 2002; Liu et al., 2024). The shortcomings of existing automatic and human evaluation methods have hindered the progress of poetry-related tasks. Therefore, it is urgent to propose new evaluation methods.

Fortunately, the emergence of large language models (LLMs) makes new evaluations possible, although LLMs have limitations on certain tasks (Lin et al., 2025). LLMs were used as evaluation tools due to their obvious potential and advantages (Chiang and Lee, 2023a; Wang et al., 2023a; Liu et al., 2023), and have made effective progress in the evaluation of multiple tasks such as open-domain conversations (Lin and Chen, 2023), text summarization (Chu et al., 2024), dialogue response generation, and open-ended question answering (Chan et al., 2023). However, evaluation based on LLMs has the potential for bias (Liu et al., 2024; Bitton et al., 2023; Koo et al., 2023; Lan et al., 2025a,b). For example, previous text summarization task (Panickssery et al., 2024) found that LLMs, including GPT-4 and Llama 2, tend to give higher scores to their own generation.

Are LLMs suitable for evaluating poetry-related tasks? Wang et al. (2024) applied GPT-4 to evaluate the poem’s translations using the same rules as human evaluation, but the evaluation results of GPT-4 were far from those of humans, and there

were many cases where full marks or close to full marks appeared, proving that GPT-4 is not suitable for the evaluation of poetry translation.

Understanding of poetry serves as an essential foundation for all research involving poetry (Pierce, 2003; Tate, 1940; McGrath, 2018). This paper aims to study and enhance LLMs’ ability as evaluators to evaluate the poetry understanding task. Specifically, we first explored LLMs’ performance of understanding poetry from multiple dimensions. To that end, we collected 40 high-quality unpublished modern Chinese poems by 5 professional modern Chinese poets, of which 8 poems are written by the same poet. Human evaluation results show that GPT-4 can understand poetry with an accuracy rate of over 76%. Then, we explored the evaluation capabilities of several mainstream LLMs on “understanding of poetry” from multiple dimensions, but the results were not satisfactory because the scores from LLMs were far from those of humans. These initial results evaluated by multiple LLMs are consistent with the results of Wang et al. (2024) evaluating poems’ translation with GPT-4, which once again proves that current LLMs are not yet viable as evaluation tools for poetry-related tasks.

To enhance the evaluation ability of LLMs, we propose a novel method called Poller (Poetry LLM Evaluator). Poetry is a special literary genre, and the author of poetry (i.e., poet) usually writes with assumed characters. To accurately understand poetry, readers should first determine the specific role

assumed by the author of the poem before reading it (Crowe Ransom, 1938; Culler, 2017). Therefore, our method requires LLMs to play the role of the authors of poems and then evaluate the understanding of poems in the same way as human (real author) evaluation. Specifically, we provide multiple LLMs with detailed information about these authors through prompts, including the basic information of the authors, the authors' own poetry experience and views, and the comments of others (professional poetry critics) on the authors and their published poems. Subsequently, these models evaluate the interpretations of poems from eight dimensions under different prompts and provide reasons for the score. Experimental results demonstrate that our Poller method effectively reduces the evaluation error between LLMs and humans, achieving a 57.37% reduction in average error for GPT-4-Turbo and 46.95% for DeepSeek-R1 compared to baseline methods. Especially for specific dimension evaluation, Poller-based LLMs achieve a 94.55% reduction in error for rhetorical techniques and an 89.53% reduction in error for defamiliarization. These performances are unattainable by conventional LLM evaluation methods. Experimental results of multiple LLMs from multiple dimensions validate the efficacy of our method.

The main contributions of our work include:

- We are the first to investigate large language models (LLMs) as evaluation tools for the poetry understanding task.
- We propose a novel LLMs-based evaluation method, which can be extended to other poetry-related tasks.
- Experimental results using multiple LLMs across multiple dimensions validate the efficacy of our method.

2 Related Work

Previous work evaluated poetry tasks in different languages and genres. Walsh et al. (2024) evaluated LLMs on the recognition of poetic forms of English poetry, focusing on sonnets, sestinas, and pantoums. Their experimental results demonstrated that when the form of the poems is fixed, LLMs can accurately recognize these poems regardless of whether they are common or uncommon poems. However, LLMs still have difficulties in recognizing poetic forms that are not fixed, such as those based on topic or visual features. Possi et al. (2023)

proposed a specific metric to automatically evaluate morphological features such as rhyme and meter in poetry, filling the gap in automatic poetry evaluation. Their experiments involved ballad and sonnet and found that the metric is not limited to form, which can distinguish poetry from prose texts (such as news articles). Agirrezabal et al. (2023) proposed an automatic evaluation framework that can be used for both automatically generated poems and human-written poems. Their evaluation focuses on four aspects of poetry, including poetic features, novelty features, lexico/semantics, and poetic fluency. They conducted experiments on English and Spanish poems and found that, in addition to evaluation, the framework can also be used for the analysis of syllable and stanza structure of sonnets. Another work by Agirrezabal and Oliveira (2024) evaluated whether Llama2 can generate poetry with predefined metric constraints in zero-shot scenarios. Specifically, they measured how well the generated poetry matched simple constraints, including the number of lines, stanzas, and syllables. Quantitative analysis showed that in zero-shot scenarios, Llama2 had difficulty generating text that followed specific constraints, especially in relation to the number of syllables.

Obviously, these automatic evaluation works involve poems with features such as rhyme, line number, rhythm, etc., and none of them involve Chinese poetry. Different from these studies, modern Chinese poetry in our work is characterized by freedom and is not constrained by these characteristics (Awan and Khalida, 2015; Wang et al., 2024). Due to the open form, diverse genres, and innovative structure of modern Chinese poetry (Wang, 2006b; Skerratt, 2013), evaluating modern Chinese poetry is challenging. Therefore, to fill the gap in automatic evaluation of modern Chinese poetry, our work focuses on multiple mainstream LLMs and aims to enhance their evaluation performance.

3 Task Definition

The core objective of this work is to enhance LLMs' capability in evaluating poetry understanding by playing the role of the poem's author, which simulates a poet's interpretive perspectives. It is worth emphasizing that our task evaluates the model's understanding of poetry, not the poetry itself. Figure 2 presents the overall workflow of our work. Our methodology comprises three key phases:

1. Multidimensional Understanding Genera-

tion: Prompting LLMs to generate interpretations U of poem P across eight defined dimensions k .

2. **Human Evaluation Benchmark:** Establishing reference evaluation $H^{(k)}(P, U)$ through human assessment.
3. **LLMs Evaluation:** Evaluating the same interpretations using mainstream LLMs under diverse prompting strategies θ .

Evaluation criteria and rules remain identical between human and LLM evaluators. The evaluation deviation is formalized as:

$$\varepsilon_{\theta}^{(k)}(P, U) = f_{\theta}^{(k)}(P, U | \mathcal{I}) - H^{(k)}(P, U) \quad (1)$$

where $k \in \{1, \dots, 8\}$ denotes the evaluation dimension index (covering dimensions like content, language, imagery, etc.). θ represents the prompt strategies (\mathcal{P}_0 to \mathcal{P}_4). $\mathcal{I} = \{I_i, V_i, C_i\}$ encapsulates the detailed information of the poem’s author, including basic information I_i , poetic views V_i , and critics’ comments C_i . $\varepsilon_{\theta}^{(k)}$ signifies the evaluation deviation of strategy θ for dimension k . And the optimization objective is to minimize the expected absolute deviation:

$$\min_{\theta} \mathbb{E} \left[\left| \varepsilon_{\theta}^{(k)} \right| \right] \quad (2)$$

This framework optimizes the evaluation error in each dimension and identifies the optimal evaluation strategy θ_k^* for specific dimensions.

4 Evaluation Capabilities of LLMs

4.1 LLM Generate Understanding of Poetry

Data Referring to the scale of poetry datasets used in previous studies (Wang et al., 2024; Virvou et al., 2023; Hutson and Schnellmann, 2023; Antar, 2023), we collected 40 high-quality **unpublished** modern Chinese poems by 5 professional modern Chinese poets, of which 8 poems are written by the same poet. These poems total 110 stanzas and 600 lines. It is worth mentioning that the poets participating in this study are all senior experts in the field of modern Chinese poetry. They have extensive experience in poetry writing and are familiar with poetry theory.

Dimensions Based on the recommendations of eight professional modern Chinese poets, and referring to previous paper on modern poetry (Chang, 2012), we summarized eight key dimensions (Yu,

2001; Wang, 1998; Xi, 2019; Wang, 2006a, 2019; Cheng, 2005; Wang et al., 2024) for understanding modern Chinese poetry: (1) **Content:** Summarize what the poem is about. (2) **Language:** What are the characteristics of the language of this poem? Is it innovative? (3) **Imagery:** What imageries are used in this poem, and what are their functions? (4) **Rhetorical Techniques:** What rhetorical techniques are used in this poem? What is the function? (5) **Rhythm:** How is the rhythm of this poem? Such as pauses, lines, and stanzas. (6) **Defamiliarization:** Does this poem contain any expression of defamiliarization? If there is any unfamiliar expression, please point it out; if not, you can skip it. (7) **Thought & Emotion:** What thought and emotion does this poem express? (8) **Modernity:** Is the poem modern or modern-minded?

These dimensions are clearly delineated to ensure clarity and precision in the evaluation process. Each dimension has its own unique focus:

- **Cont:** Focuses on the subject matter or theme of the poem—essentially, what the poem describes.
- **Lang:** Refers to the linguistic features of the poem, such as word choice, syntax, and diction.
- **Imag:** Examines the use of visual and sensory images in the poem to evoke specific feelings or ideas.
- **Rhet:** Focuses on the rhetorical strategies employed, such as metaphor, personification, and other literary techniques.
- **Rhyt:** Assesses the internal rhythm of the poem, often expressed in external forms like pauses, line breaks, and stanza structure.
- **Defa:** Highlights a common technique in modern poetry that challenges conventional perceptions, emphasizing innovation and originality.
- **Thou:** Captures the ideas and emotions conveyed by the poem, focusing on its depth and resonance.
- **Mode:** Reflects the contemporary consciousness or modern sensibility embedded in the poem, which distinguishes it as a product of modern literature.

Evaluator ↓		Cont	Lang	Imag	Rhet	Rhyt	Defa	Thou	Mode	Avg
Human		80.33	79.05	81.18	77.83	76.15	79.40	78.80	79.88	79.08
GPT-4	\mathcal{P}_0	95.25	92.25	97.00	93.13	88.75	90.38	96.50	94.38	93.46
GPT-4-Poller	\mathcal{P}_1	90.88	89.00	93.88	89.38	86.38	83.00	92.63	91.63	89.60
DeepSeek	\mathcal{P}_0	91.45	89.93	92.58	89.20	86.85	96.38	92.75	94.35	91.69
DeepSeek-Poller	\mathcal{P}_1	89.00	88.93	90.50	86.80	82.90	87.83	90.15	92.65	88.60
Qwen-Turbo	\mathcal{P}_0	90.00	88.25	90.00	90.00	80.00	83.00	90.00	90.00	87.66
Qwen-Turbo-Poller	\mathcal{P}_1	97.25	90.88	93.75	91.38	90.88	89.00	94.38	93.88	92.66
Claude3.5	\mathcal{P}_0	91.45	88.55	91.55	84.15	88.45	89.25	92.83	94.50	90.09
Claude3.5-Poller	\mathcal{P}_1	91.90	88.00	93.73	84.95	89.35	91.88	92.68	95.18	90.96

Table 1: Evaluation results of humans and different models using \mathcal{P}_0 and \mathcal{P}_1 . The models in the first column are GPT-4-Turbo, DeepSeek-R1, Qwen2.5-Turbo, and Claude3.5-sonnet. The models that apply our method are named with the suffix Poller. Avg represents the average score of 8 dimensions. The bolded scores are relatively close to human scores.

Subsequently, GPT-4 (gpt-4-turbo) (Achiam et al., 2023) generates an understanding of all poems from eight dimensions. The prompt and experimental details are shown in Appendix A.2.

In this work, we use Chinese prompts directly in the source language without translating them into English. This is because translating Chinese poetry into English for use as prompts introduces several challenges. For instance, this process often results in the loss of key poetic elements such as poeticity, style, cultural nuances, and rhythm (Chakrabarty et al., 2021; Song et al., 2023; Wang et al., 2024). These elements are crucial for accurately evaluating a model’s understanding of poetry. To ensure fairness and preserve the integrity of our analysis, we avoided translating Chinese poems into English. Instead, we used the original Chinese poetry directly as input prompts for ChatGPT. This approach ensures that the evaluation reflects the full richness and complexity of the original poems’ text without the distortions introduced by translation.

4.2 Human Evaluation

Human evaluation solves the disadvantage that automatic evaluation does not consider the intrinsic characteristics of poetry. Reading poetry is inherently subjective, as interpretations often vary significantly among readers (Brodsky, 1999; Jiang, 2012; Zhang, 2022). However, the author’s original intent provides a unique benchmark for evaluation. To accurately confirm the model’s ability to understand poetry, we invited the five authors of poems to evaluate the experimental results from all dimensions.

Evaluation Rules Each dimension ranges from 0 to 100. A score of 0 means that authors think the

model’s understanding of a certain dimension is completely wrong, and 100 indicates that authors think the model’s understanding of a certain dimension is completely correct. The more correct the model’s understanding of a certain dimension is, the higher the score authors will give.

Evaluation results are reported in the first row of Table 1. Obviously, the model’s understanding of the poems scored over 76 points in eight dimensions, indicating that the model is able to grasp most of the meaning of the poem.

4.3 LLMs Initial Evaluation

To explore the evaluation capabilities of LLMs, we prompted LLMs to evaluate the interpretations of poems generated by GPT-4-Turbo in Section 4.1. Specifically, we experimented with several mainstream LLMs, including: GPT-4-Turbo, DeepSeek-R1, Qwen2.5-Turbo, Claude3.5-sonnet (Achiam et al., 2023; Guo et al., 2025; Yang et al., 2024; Anthropic, 2024). The text being evaluated and the evaluation rules used are exactly the same as those for human evaluation.

LLMs can generate interpretable evaluations and feedback (Li et al., 2024), and providing explanations can effectively improve the accuracy of LLMs’ evaluation (Chiang and Lee, 2023b). So in addition to scoring, these models were also tasked to provide reasons for the score so that we can fully grasp the models’ evaluation performance. Decomposing complex evaluation into multiple dimensions is more conducive to the LLMs’ evaluation (Song et al., 2024), so the models were required to evaluate “understanding of poems” from multiple specific dimensions similar to human evaluation. The prompts are shown in Table 6, denoted as \mathcal{P}_0 .

From now on, you are a professional Chinese modern poet and need to play the role of a judge. Your details are as follows:

Basic Information: $\{\{I_i\}\}$

Your Own Experience and Views on Poetry: $\{\{V_i\}\}$

Others' Comments (Professional Poetry Critics) on You and Your Published Poems: $\{\{C_i\}\}$

The above are your details. Now, you need to play the role of a professional poet with the above characteristics. Your task is to play the role of a judge. Specifically, you need to score the following "readers' understanding of the poem" from the eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thought and emotion, and modernity, and give detailed reasons for your score. Please note that you only need to score and give reasons for the "Readers' Understanding of this poem, not the poem itself.

Meaning of Dimension: Same as human evaluation

Scoring Rules: Each dimension ranges from 0 to 100. A score of 0 means that you think the reader's understanding of a certain dimension is completely wrong, and 100 indicates that you think the reader's understanding of a certain dimension is completely correct; the more correct the reader's understanding of a certain dimension is, the higher the score you will give.

Poem: $\{\{P_i\}\}$

Readers' Understanding of this Poem: $\{\{U_i\}\}$

Please remember to state your own role before giving your formal answer.

Table 2: The Prompt 3 (\mathcal{P}_3) we designed for evaluating models' understanding of poetry.

\mathcal{P}_0 is used as the baseline.

\mathcal{P}_0 in Table 1 present the evaluation results of LLMs from different dimensions. The results are unexpected, as we are disappointed to see that LLMs are not reliable for all dimensions. Specifically, Qwen2.5-Turbo performed best relative to the other eight models, as its average score (87.66) is closest to the human average score (79.08). In terms of specific dimensions, Qwen2.5-Turbo's score for defamiliarization (83.00) is closest to the human score (79.40). However, for other dimensions, the difference between Qwen2.5-Turbo's score and the human score is more than 8.82 points.

The average score of the other models except Qwen2.5-Turbo for poetry understanding is at least 11.01 points different from the human score (Claude3.5-sonnet), and the score for specific dimensions is at least 6.32 points different from the human score (Claude3.5-sonnet). The model with the worst evaluation performance is GPT-4-Turbo, whose average score (93.46) for poetry understanding is 14.38 points different from the human score (79.08). Among them, GPT-4-Turbo has the worst evaluation performance of all models for content, language, imagery, rhetorical techniques, rhythm, and thought and emotion. Furthermore, we noted

that the evaluation scores given by GPT-4-Turbo are significantly higher than those given by humans. We speculate that this is because GPT-4 is able to recognize and favor its own generated text, which is consistent with the conclusion of previous works (Liu et al., 2024; Bitton et al., 2023; Koo et al., 2023; Panickssery et al., 2024).

We observed that for all dimensions, all LLMs gave higher evaluation scores than humans in the initial evaluation (\mathcal{P}_0). For the evaluation of modernity, humans gave a score of 79.88, but all models gave scores (\mathcal{P}_0) higher than or equal to 90; among them, GPT-4-Turbo, DeepSeek-R1, and Claude3.5-sonnet all gave evaluation scores (\mathcal{P}_0) higher than 93.25. Similarly, for the evaluation of content, humans gave a score of 80.33, but all models gave scores (\mathcal{P}_0) higher than or equal to 90. We speculate that this may be because LLMs focus on surface linguistic features (such as the use of novel words) while ignoring the cultural context inside the text. This is consistent with recent findings that LLMs' processing of metaphor may rely on superficial lexical associations rather than reliable semantic understanding (Ye et al., 2025).

Overall, the evaluation results from all LLMs (all dimensions) are not close to those of human evaluation, proving that current LLMs are not suitable for the evaluation of poetry understanding task.

5 Poetry LLM Evaluator

To enhance the evaluation ability of LLMs, we propose a novel LLM-based method termed Poller (**Poetry LLM Evaluator**). Poller requires LLMs to play the role of five authors of poems and then evaluate the understanding of poems in the same way as human evaluation. Specifically, we give LLMs detailed information about these authors through prompts, including: (1) Authors' basic information, such as occupation, growth background, life background, personal experience, and so on; (2) Authors' own experiences and views on poetry.; and (3) Others' comments (professional poetry critics) on authors and their published poems. Subsequently, the model evaluates the understanding of the poems from eight dimensions under different prompts and gives reasons.

LLM-evaluators Take the Author's Perspective

The expectation for LLMs to adopt the author's perspective is grounded in the relationship between the cultural context, the author's intent, and the deeper understanding of the poem:

- In any era, an author’s perspective is influenced by the cultural, historical, and social context in which they live. This perspective significantly impacts the style, form, and the thoughts and emotions the author seeks to convey through their poetry (Zang, 2007). For example, the imagery and rhetorical techniques chosen by the author often reflect their personal experiences and worldview.
- By adopting the author’s perspective, evaluators can better understand the intentions behind the poem and the cultural or emotional subtext embedded within it. This approach allows evaluators to go beyond a superficial reading and grasp the deeper meaning and artistic choices of the author.
- Evaluators who adopt the author’s perspective are better equipped to assess a poem’s intended impact and artistic merit. This is particularly important in the context of modern Chinese poetry, where cultural nuances and historical influences play a crucial role in shaping the poet’s creative process and output.

For LLMs, different prompts may lead to different evaluation results (Xu et al., 2023). Therefore, to explore LLMs’ ability of evaluating the “understanding of poetry”, we designed several prompts, denoted as $\mathcal{P}_{i \in \{1..4\}}$, with the following focuses:

- \mathcal{P}_1 : Based on \mathcal{P}_0 , model plays the role of the poems’ author and serves as a judge. Compared with \mathcal{P}_0 , \mathcal{P}_1 contains: (1) basic information about the original poet, (2) the original poet’s experience and views on poetry, and (3) comments from others (professional poetry critics) on the original poet’s published poems.
- \mathcal{P}_2 : Based on \mathcal{P}_1 , it emphasizes that the poem is written by the model (playing the poet), and the understanding of the poem comes from other readers.
- \mathcal{P}_3 : Based on \mathcal{P}_1 , the model is required to declare its role before the formal scoring.
- \mathcal{P}_4 : It combines the characteristics of \mathcal{P}_2 and \mathcal{P}_3 , emphasizing that the poem is written by the model (playing the poet), while the understanding of the poem comes from other readers. At the same time, the model is required to declare its role before the formal scoring.

Table 2 presents the Prompt 3 (\mathcal{P}_3) we designed for evaluating models’ understanding of poetry. Other prompts and complete details are provided in the Appendix A.2.

6 Experiment

Authors’ Information To effectively stimulate the ability of LLMs to act as authors, we collected and screened the detailed information of all authors. Specifically, we collected a total of 37 poetry-related articles from the 5 original poets from interviews, speeches, poets’ seminars, poets’ published poetry collections, and published journals. Among them, 13 articles are the original poets’ own experiences and views on poetry, and 24 articles are professional poetry critics’ comments on the poems (not overlapping with the poems in our experiment) of the poets. We finally obtained 57K words of poetry-related information about the 5 original poets, and summarized the personal information of each poet to control the number of words between 1K and 2.5K words.

To determine the potential of LLMs to play the role of poems’ authors (\mathcal{P}_1), we conduct experiments using the same LLMs as in Section 4.3. The results of the eight dimensions are reported in Table 1. Table 1 shows that the model with the greatest potential for playing the role of the author to evaluate “understanding of poetry” is GPT-4-Turbo, because its average score after playing the role of the author (89.60) is 3.86 points lower than the average score when it does not play the role of the author (93.46), which is the largest reduction among all models. In addition, when playing the role of the poem’ author (\mathcal{P}_1), DeepSeek-R1’s score (88.60) is closest to human.

Based on the initial evaluation performance (Table 1) of using \mathcal{P}_0 and their potential to act as poems’ authors using \mathcal{P}_1 , we selected the four most representative models to continue the experiments using \mathcal{P}_2 , \mathcal{P}_3 , and \mathcal{P}_4 . The four models in the subsequent experiments are: Claude3.5-sonnet, GPT-4-Turbo, Qwen2.5-Turbo, and DeepSeek-R1.

Table 3 presents the evaluation errors (absolute deviation from human scores) of different models and prompting strategies. Figure 1 shows an example of our Poller method. Complete examples and more evaluation results are provided in Appendix A.4 and 8, respectively.

Evaluator ↓		Cont	Lang	Imag	Rhet	Rhyt	Defa	Thou	Mode	Avg
GPT-4	\mathcal{P}_0	14.92	13.20	15.82	<i>15.30</i>	12.60	<i>10.98</i>	17.70	14.50	14.38
GPT-4-Poller	\mathcal{P}_1	10.55	9.95	12.70	11.55	10.23	3.60	13.83	11.75	10.52
GPT-4-Poller	\mathcal{P}_2	14.57	11.83	15.82	16.67	13.85	8.93	16.25	15.22	14.14
GPT-4-Poller	\mathcal{P}_3	9.67	5.95	7.82	2.30	<u>3.23</u>	<i>1.15</i>	8.33	10.62	<u>6.13</u>
GPT-4-Poller	\mathcal{P}_4	10.30	7.20	10.82	7.80	8.98	2.73	9.45	12.50	8.72
DeepSeek	\mathcal{P}_0	11.12	10.88	11.40	11.37	10.70	16.98	13.95	14.47	12.61
DeepSeek-Poller	\mathcal{P}_1	8.67	9.88	9.32	8.97	6.75	8.43	11.35	12.77	9.52
DeepSeek-Poller	\mathcal{P}_2	10.10	10.13	6.50	5.65	8.98	8.35	9.30	13.80	9.10
DeepSeek-Poller	\mathcal{P}_3	9.00	5.88	<u>5.82</u>	0.62	6.83	5.40	<u>7.48</u>	12.47	6.69
DeepSeek-Poller	\mathcal{P}_4	12.27	9.75	9.45	7.45	11.43	9.90	10.88	15.42	10.81
Qwen-Turbo	\mathcal{P}_0	9.67	9.20	8.82	12.17	3.85	3.60	11.20	10.12	8.58
Qwen-Turbo-Poller	\mathcal{P}_1	16.92	11.83	12.57	13.55	14.73	9.60	15.58	14.00	13.59
Qwen-Turbo-Poller	\mathcal{P}_2	8.42	7.58	8.57	9.80	9.85	6.23	10.70	10.75	8.99
Qwen-Turbo-Poller	\mathcal{P}_3	9.55	3.33	8.20	5.42	5.48	1.60	11.08	8.12	6.60
Qwen-Turbo-Poller	\mathcal{P}_4	<u>6.55</u>	5.45	8.70	8.55	6.23	3.73	11.08	10.12	7.55
Claude3.5	\mathcal{P}_0	11.12	9.50	10.37	6.32	12.30	9.85	14.03	14.62	11.01
Claude3.5-Poller	\mathcal{P}_1	11.57	8.95	12.55	7.12	13.20	12.48	13.88	15.30	11.88
Claude3.5-Poller	\mathcal{P}_2	7.45	9.53	6.85	6.05	14.03	6.10	8.28	13.72	9.00
Claude3.5-Poller	\mathcal{P}_3	6.90	<u>2.15</u>	6.90	2.25	8.50	4.23	8.88	11.40	6.40
Claude3.5-Poller	\mathcal{P}_4	8.72	8.25	10.35	4.85	11.88	5.10	11.53	13.45	9.27

Table 3: Evaluation errors (absolute deviation from human scores) of different models and prompting strategies. Values represent absolute errors: $|\text{Model Score} - \text{Human Score}|$. Italics indicate the optimal evaluation performance under this dimension, that is, the absolute error between the LLM evaluation result and the human evaluation result is the smallest.

7 Analysis

Our Poller method effectively narrows the gap between LLMs and human evaluation.

- Overall: As shown in Table 3, after using our method, the overall evaluation error of LLMs is reduced compared to the initial evaluation error without using our method (\mathcal{P}_0). For example, compared with the baseline prompt strategy (\mathcal{P}_0), Poller’s optimal strategy (\mathcal{P}_3) reduces the average evaluation error of GPT-4-Turbo from 14.38 to 6.13 ($\downarrow 57.37\%$), DeepSeek-R1 from 12.61 to 6.69 ($\downarrow 46.95\%$), Claude3.5-sonnet from 11.01 to 6.40 ($\downarrow 41.87\%$), and Qwen2.5-Turbo from 8.58 to 6.60 ($\downarrow 23.08\%$).
- Single dimension: For **rhetorical techniques**, the evaluation error of DeepSeek-R1 using \mathcal{P}_3 is only 0.62, close to the human level, and reduced by 94.55% compared with the baseline \mathcal{P}_0 (11.37); the evaluation error of GPT-4-Turbo using \mathcal{P}_3 (2.30) is reduced by 84.97% compared with \mathcal{P}_0 (15.30), and the evaluation

error of Sonnet using \mathcal{P}_3 (2.25) is reduced by 64.40% compared with \mathcal{P}_0 (6.32). Similarly, for **defamiliarization**, the evaluation error of GPT-4-Turbo using \mathcal{P}_3 (1.15) is reduced by 89.53% compared with \mathcal{P}_0 (10.98), while the evaluation error of DeepSeek-R1 using \mathcal{P}_3 (5.40) is reduced by 68.20% compared with \mathcal{P}_0 (16.98).

The reduction in the error between LLMs evaluation and human evaluation proves that our Poller method effectively enhances the evaluation ability of LLMs. Poller injects detailed information about the poet through role-playing, guiding LLMs to simulate the author’s intention, thereby alleviating the bias caused by the lack of cultural context in the LLMs’ traditional evaluation.

The effectiveness of the various prompting strategies we designed. The optimal evaluation performance of the large model on different dimensions appears on the different prompting strategies we designed. As shown in Table 3, the optimal evaluation performance for **content** appears when Qwen2.5-Turbo uses \mathcal{P}_4 in our Poller method, and

	All Dimensions	Single Dimension
Rhet	78.45	79.00
Defa	78.25	86.50

Table 4: The evaluation scores from the LLMs when GPT-4 understands all dimensions and single dimension. The numbers in the second column refer to the evaluator’s score for a specific dimension when GPT-4 generates poetry understanding from all dimensions simultaneously. The numbers in the third column refer to the score of generating poetry understanding from only a single dimension.

its absolute error with the human evaluation result is 6.55. The optimal evaluation performance for **language** appears when Claude3.5-sonnet uses \mathcal{P}_3 in our Poller method, and its absolute error with the human evaluation result is 2.15. Interestingly, we found that except for Qwen2.5-Turbo, which uses \mathcal{P}_4 to obtain the best performance in content evaluation, the evaluation performance of all other models on all dimensions is best when using \mathcal{P}_3 .

These results demonstrate the effectiveness of our design of multiple prompting strategies. Specifically, both the \mathcal{P}_3 and \mathcal{P}_4 strategies in the Poller method require the model to declare its role before starting formal scoring, which shows that the behavior of the model confirming its role strengthens the identity of LLMs, effectively improving its ability to play the role of the poem author, thereby effectively improving its evaluation performance.

Can LLMs be used as evaluation tools? The experimental results (Table 3) show that the model using our method is more reliable in the evaluation task than the model not using our method (\mathcal{P}_0). Therefore, LLMs using our method can be used to assist in evaluation. Specifically, DeepSeek-R1 is suitable for evaluating rhetorical techniques when using \mathcal{P}_3 , GPT-4-Turbo is suitable for evaluating defamiliarization and rhythm when using \mathcal{P}_3 , and Claude3.5-sonnet is suitable for evaluating language when using \mathcal{P}_3 .

8 Prompt Each Dimension Separately

To explore whether it is better to prompt each dimension separately, we have added additional experiments on two separate dimensions, Rhetorical Techniques and Defamiliarization. Taking Rhetorical Techniques as an example, our experimental process is divided into two steps:

- Prompt GPT-4 to generate understanding of

20 poems only from the Rhetorical Techniques dimension.

- Use LLMs (DeepSeek-Poller) whose evaluation performance on the Rhetorical Techniques dimension is close to that of human evaluation to score the understanding generated in the first step through \mathcal{P}_3 .

Based on the results and analysis of Table 10 in our paper, we use DeepSeek-Poller (\mathcal{P}_3) as the evaluator of Rhetorical Techniques and GPT-4-Poller (\mathcal{P}_3) as the evaluator of Defamiliarization. The evaluation scores from the LLMs are shown in Table 4.

We found that when the single dimension was prompted separately, the model obtained higher scores in both Rhetorical Techniques and Defamiliarization. This shows that prompting a specific dimension alone can enhance the model’s understanding of poetry. But in reality, evaluating the quality of a poem requires considering multiple dimensions rather than a single one. Therefore, this work focuses on evaluating poetry from multiple dimensions.

9 Conclusion

Existing automatic evaluation methods are not suitable for poetry domain, while human evaluation is expensive. In this paper, we propose a novel evaluation method based on LLMs, which can effectively evaluate poetry understanding task with performance closer to human evaluation results than traditional LLM evaluation. This method requires LLMs to play the role of the poem’s author and evaluate the understanding of poems from multiple dimensions. Experimental results demonstrate that our method effectively reduces the evaluation error between LLMs and humans, especially in the two dimensions of evaluating rhetorical techniques and defamiliarization. Experimental results using multiple LLMs across multiple dimensions validate the efficacy of our method. Our work bridges the gap between automated efficiency and human expertise, establishing a foundation for automated evaluation in poetry-related tasks.

Limitations

Our work focuses solely on modern Chinese poetry, a genre characterized by free form. While the proposed Poller method effectively reduces the evaluation error of LLMs, its applicability to other poetic

forms remains to be explored. The role-playing mechanism relies on poet-specific background information, which may be difficult to obtain in other traditional texts. Nevertheless, we believe that the core idea of adopting the author’s perspective is transferable, as adapting Poller to new domains would require primarily re-sourcing poet information rather than redesigning the evaluation framework. This limitation does not undermine the validity of our findings for modern Chinese poetry.

Ethics Statement

The human poems used in this study are derived from the authors of these poems, and all authors have authorized our work to use the poem data.

Acknowledgements

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST), the Tencent AI Lab Rhino-Bird Research Program (Grant No. EF2023-00151-FST), the Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), and the National Natural Science Foundation of China (Grant No. 62266013).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. *Gpt-4 technical report*. *arXiv preprint arXiv:2303.08774*.
- Manex Agirrezabal and Hugo Gonalo Oliveira. 2024. *Zero-shot metrical poetry generation with open language models: a quantitative analysis*.
- Manex Agirrezabal, Hugo Gonalo Oliveira, and Aitor Ormazabal. 2023. *Erato: Automating poetry evaluation*. In *EPIA Conference on Artificial Intelligence*, pages 3–14. Springer.
- D Antar. 2023. *The effectiveness of using chatgpt4 in creative writing in arabic: Poetry and short story as a model*.
- Anthropic. 2024. *Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet*.
- Abdul Ghafoor Awan and Yasmin Khalida. 2015. *New trends in modern poetry*. *Journal of Literature, Languages and Linguistics*, 13:63–72.

- Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. *Visit-bench: A benchmark for vision-language instruction following inspired by real-world use*. *arXiv preprint arXiv:2308.06595*.
- J. Brodsky. 1999. *Witness and Pleasure*. witness and pleasure.
- Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. *Don’t go far off: An empirical study on neural poetry translation*. *arXiv preprint arXiv:2109.02972*.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. *Chateval: Towards better llm-based evaluators through multi-agent debate*. *arXiv preprint arXiv:2308.07201*.
- Yina Chang. 2012. *A brief discussion on cutting-edge issues in contemporary poetry research*. *Northern Literature Monthly*.
- Bo Cheng. 2005. *The special ecology of modernity in new poetry—a study on the poets of southwest associated university*. *Journal of the School of Liberal Arts of Nanjing Normal University*, (4):69–76.
- Cheng-Han Chiang and Hung-Yi Lee. 2023a. *Can large language models be an alternative to human evaluations?* In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.
- Cheng-Han Chiang and Hung-yi Lee. 2023b. *A closer look into using large language models for automatic evaluation*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942. Association for Computational Linguistics.
- Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. 2024. *Pre: A peer review based large language model evaluator*. *arXiv preprint arXiv:2401.15641*.
- John Crowe Ransom. 1938. *The world’s body (baton rouge)*.
- Jonathan Culler. 2017. *Theory of the lyric*. *Nordisk poesi*, 2(2):119–133.
- Zekun Deng, Hao Yang, and Jun Wang. 2024. *Can ai write classical chinese poetry like humans? an empirical study inspired by turing test*. *arXiv preprint arXiv:2401.04952*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*. *arXiv preprint arXiv:2501.12948*.

- James Hutson and Ana Schnellmann. 2023. [The poetry of prompts: the collaborative role of generative artificial intelligence in the creation of poetry and the anxiety of machine influence](#). *Global Journal of Computer Science and Technology*, 23(D1):1–14.
- Yongjun Jiang. 2012. [The "unspeakable" in the perceptual effect of poetry - an analysis of the "unspeakable" meaning of feeling](#). *foreign language*, 28(1):4.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. [Benchmarking cognitive biases in large language models as evaluators](#). *arXiv preprint arXiv:2309.17012*.
- Tian Lan, Jiang Li, Yemin Wang, Xu Liu, Xiangdong Su, and Guanglai Gao. 2025a. [F²Bench: An open-ended fairness evaluation benchmark for LLMs with factuality considerations](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2031–2046. Association for Computational Linguistics.
- Tian Lan, Xiangdong Su, Xu Liu, Ruirui Wang, Ke Chang, Jiang Li, and Guanglai Gao. 2025b. [McBE: A multi-task Chinese bias evaluation benchmark for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6033–6056.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. [Llms-as-judges: a comprehensive survey on llm-based evaluation methods](#). *arXiv preprint arXiv:2412.05579*.
- Jiang Li, Tian Lan, Shanshan Wang, Dongxing Zhang, Dianqing Lin, Guanglai Gao, Derek F Wong, and Xiangdong Su. 2026. [Who wrote this line? evaluating the detection of llm-generated classical chinese poetry](#). *arXiv preprint arXiv:2604.10101*.
- Dianqing Lin, Aruukhan, Hongxu Hou, Shuo Sun, Wei Chen, Yichen Yang, and Guodong Shi. 2025. [Can large language models translate unseen languages in underrepresented scripts?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23137–23150. Association for Computational Linguistics.
- Yen-Ting Lin and Yun-Nung Chen. 2023. [Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models](#). *arXiv preprint arXiv:2305.13711*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.
- Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2024. [Llms as narcissistic evaluators: When ego inflates evaluation scores](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701.
- Ruli Manurung, Graeme Ritchie, and Henry Thompson. 2012. [Using genetic algorithms to create meaningful poetic text](#). *Journal of Experimental & Theoretical Artificial Intelligence*, 24(1):43–64.
- Brian McGrath. 2018. [Understating poetry](#). *New Literary History*, 49(3):289–308.
- Jekaterina Novikova, Ondřej Dušek, Amanda Caracas Curry, and Verena Rieser. 2017. [Why we need new evaluation metrics for NLG](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252. Association for Computational Linguistics.
- Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. [Llm evaluators recognize and favor their own generations](#). *arXiv preprint arXiv:2404.13076*.
- 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. Association for Computational Linguistics.
- Robert B Pierce. 2003. [Defining "poetry"](#). *Philosophy and Literature*, 27(1):151–163.
- Maurilio De Araujo Possi, Alcione De Paiva Oliveira, Alexandra Moreira, and Lucas Mucida Costa. 2023. [Carmen: A method for automatic evaluation of poems](#). In *2023 5th International Conference on Natural Language Processing (ICNLP)*, pages 244–247. IEEE.
- Eshrag Ali Refaee. 2023. [Okaz: A deep-learning-based system for automatic arabic poem generation](#). In *2023 3rd International Conference on Computing and Information Technology (ICCIIT)*, pages 396–403. IEEE.
- Philip Resnik and Jimmy Lin. 2010. [Evaluation of nlp systems](#). *The handbook of computational linguistics and natural language processing*, pages 271–295.
- Hanwen Shen, Ting Ying, Jiajie Lu, and Shanshan Wang. 2026. [Preconditioned test-time adaptation for out-of-distribution debiasing in narrative generation](#). *arXiv preprint arXiv:2603.13683*.
- Brian Phillips Skerratt. 2013. [Form and Transformation in Modern Chinese Poetry and Poetics](#). Ph.D. thesis.
- Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. [Finesure: Fine-grained summarization evaluation using llms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922.

- Wai Lei Song, Haoyun Xu, Derek F. Wong, Runzhe Zhan, Lidia S. Chao, and Shanshan Wang. 2023. [Towards zero-shot multilingual poetry translation](#). In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 324–335. Association for Machine Translation.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. [How to fine-tune bert for text classification?](#) In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer.
- Allen Tate. 1940. [Understanding modern poetry](#). *The English Journal*, 29(4):263–274.
- Maria Virvou, George A Tsihrantzis, Dionisios N Sotiropoulos, Konstantina Chrysafiadi, Evangelos Sakkopoulos, and Evangelia-Aikaterini Tsihrantzi. 2023. [Chatgpt in artificial intelligence-empowered e-learning for cultural heritage: The case of lyrics and poems](#). In *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–9. IEEE.
- Melanie Walsh, Anna Preus, and Maria Antoniak. 2024. [Sonnet or not, bot? poetry evaluation for large models and datasets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15568–15603.
- Guangming Wang. 1998. [Reflections on the ontology of new chinese poetry](#). *Chinese Social Sciences*, (4):16.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. [Is ChatGPT a good NLG evaluator? a preliminary study](#). In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11. Association for Computational Linguistics.
- Mi Wang. 2019. [Study on the relationship between external form and internal emotion in the rhythm of new poetry \(1917-1949\)](#). Ph.D. thesis, Sichuan Academy of Social Sciences.
- Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024. [What is the best way for chatgpt to translate poetry?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043.
- Shanshan Wang, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2026. [Can ChatGPT really understand Modern Chinese poetry?](#) In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4152–4162. Association for Computational Linguistics.
- Shanshan Wang, Junchao Wu, Fengying Ye, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2025. [Benchmarking the detection of LLMs-generated Modern Chinese poetry](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9533–9552. Association for Computational Linguistics.
- Shuting Wang. 2006a. [Giving shape to emotion: The theory and practice of rhythm and imagery in new poetry \(1917-1937\)](#). *World Literature Review*.
- Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, and 1 others. 2023b. [Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization](#). *arXiv preprint arXiv:2306.05087*.
- Yuchun Wang. 2006b. [Moruo guo's view on literary translation](#). *Moruo Guo Journal*, page 5.
- Yunshu Xi. 2019. [Two characteristics of new poetry language—talking about the creation of new poetry part three](#). *Masterpiece Appreciation: Appreciation Edition (First Decade)*, (8):6.
- Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2023. [An llm can fool itself: A prompt-based adversarial attack](#). *arXiv preprint arXiv:2310.13345*.
- Rui Yan. 2016. [I, poet:automatic poetry composition through recurrent neural networks with iterative polishing schema](#). In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, page 2238–2244.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Fengying Ye, Shanshan Wang, Lidia S Chao, and Derek F Wong. 2025. [Unveiling llms' metaphorical understanding: Exploring conceptual irrelevance, context leveraging and syntactic influence](#). *arXiv preprint arXiv:2510.04120*.
- Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018. [Automatic poetry generation with mutual reinforcement learning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3143–3153. Association for Computational Linguistics.
- Haizhang Yu. 2001. [About the changes and analysis of new poetry expressions](#). *writing*, 000(011):3–5.
- Di Zang. 2007. The literary politics of "poetic" - on the traces and limits of "poetic" in the practice of chinese new poetry. *Comments on New Poetry*, pages 13, 15–16, 19, 30.
- Taozhou Zhang. 2022. [Problems in reading new poetry from the perspective of "interpretation of poetics"](#). *literary research*, (3):13.
- Xingxing Zhang and Mirella Lapata. 2014. [Chinese poetry generation with recurrent neural networks](#). In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 670–680.

A Appendix

A.1 Assumed Characters and Specific Roles

When a poet writes a poem, he will enter a specific state that is different from his daily life. The poet in this state is a kind of assumed character (Culler, 2017; Crowe Ransom, 1938).

A poet can write poems in multiple assumed characters, and which assumed character the poet uses to write poems is usually affected by the real background he is in. As the poet’s writing motivation and situation change, this assumed character may be different (Zang, 2007). Therefore, the specific roles we mentioned refer to the poet’s writing state in a specific situation.

The tradition of poets writing poems in “assumed character” has existed since ancient times. For example, the “assumed characters” used by Qu Yuan, a famous poet in ancient China, when writing the lyric poem “Li Sao” were “fragrant grass” and “beauty”. However, the “fragrant grass” used by Qu Yuan as an “assumed character” is not a fragrant grass, and “beauty” is not a beautiful person in the literal sense. According to the situation in which Qu Yuan was jealous and framed by bad people when writing “Li Sao”, “fragrant grass” and “beauty” actually refer to noble qualities and loyal thoughts, which are used to refer to the poet’s beautiful pursuit of truth.

A.2 Prompts

LLM Generate Understanding Table 5 presents the prompt we designed for LLMs to understand modern Chinese poetry. Understanding is not a creative task, but understanding the same content can be described in different ways. So the temperature of the model is set to 0.5, and the top_p is set to 1.0.

LLM Evaluation The prompts (\mathcal{P}_0 to \mathcal{P}_4) we designed for evaluating the “understanding of poetry” are shown in Table 6 to 9, respectively. The evaluation needs to be objective and fair, so the temperature of the evaluation model is set to 0.1.

A.3 Evaluation Results of Poller

Table 10 presents the evaluation results of applying our Poller method (\mathcal{P}_1 to \mathcal{P}_4) on four models and without applying our method (\mathcal{P}_0).

Please understand the following modern poem titled T_i from the five dimensions of content, expression methods, thought and emotion, modernity, and the most poetic sentence, and finally give your interpretation of the poem. In addition to content, please provide examples and specific explanations for each dimension. Please note that this is a complete modern poem. The specific definitions of each dimension are as follows:

Content: Summarize what the poem is about.

Expression Methods: Understand how this poem is written from five aspects: language, imagery, rhetorical techniques, rhythm, and defamiliarization. The five specific aspects are as follows: 1) What are the characteristics of the language of this poem? Is it innovative? 2) What imageries are used in this poem, and what are their functions? 3) What rhetorical techniques are used in this poem? What is the function? 4) How is the rhythm of this poem? Such as pauses, lines, and stanzas. 5) Does this poem contain any expression of defamiliarization? If there is any unfamiliar expression, please point it out; if not, you can skip it.

Thought & Emotion: What thoughts and emotions does this poem express?

Modernity: Is the poem modern or modern-minded?

Poeticity: Choose the most poetic sentence in this poem.

Poem: $\{\{P_i\}\}$ (Directly used the original Chinese poem)

Interpretation of this poem: $\{\{I_i\}\}$

Table 5: The prompt we designed for LLMs to understand modern Chinese poetry.

A.4 Experimental Examples

A.4.1 The Example Without Using Our Method

An example of Prompt 0 (\mathcal{P}_0) without using our method is as follows:

Poem:

醉酒一事

(The Matter of Being Drunk)

在酒的国度里，鱼鳖可以

(In the kingdom of wine, turtles might)

离开海洋，鸟雀不必站在树上

(Depart the sea, and birds need not perch in trees)

男人醉酒后，很放心地

(A man, inebriated, with carefree trust)

把自己摆到了街边。人间之事

(Lay himself along the street’s edge, where worldly matters)

全都碎在杯里，一饮而尽后

(All shattered in the cup, drained in one gulp)

在肠胃里激荡，像巨浪冲刷着堤防

(Surging in the stomach, like giant waves washing over the embankment)

车辆在街灯的重影里轰鸣、拉扯

(Vehicles roar and tug in the doubled shadows of streetlights)

Please score the following “readers’ understanding of the poem” from the eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thought and emotion, and modernity, and give detailed reasons for your score. Please note that you only need to score and give reasons for the “Readers’ Understanding of this poem, not the poem itself.

Meaning of Dimension: Same as human evaluation
Scoring Rules: Each dimension ranges from 0 to 100. A score of 0 means that you think the reader’s understanding of a certain dimension is completely wrong, and 100 indicates that you think the reader’s understanding of a certain dimension is completely correct; the more correct the reader’s understanding of a certain dimension is, the higher the score you will give.

Poem: $\{\{P_i\}\}$

Readers’ Understanding of this Poem: $\{\{U_i\}\}$

Table 6: The Prompt 0 (\mathcal{P}_0) we designed for evaluating models’ understanding of poetry.

From now on, you are a professional Chinese modern poet and need to play the role of a judge. Your details are as follows:

Basic Information: $\{\{I_i\}\}$

Your Own Experience and Views on Poetry: $\{\{V_i\}\}$

Others’ Comments (Professional Poetry Critics) on You and Your Published Poems: $\{\{C_i\}\}$

The above are your details. Now, you need to play the role of a professional poet with the above characteristics. Your task is to play the role of a judge. Specifically, you need to score the following “readers’ understanding of the poem” from the eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thought and emotion, the most poetic sentence, and modernity, and give detailed reasons for your score. Please note that you only need to score and give reasons for the “Readers’ Understanding of this poem, not the poem itself.

Meaning of Dimension: Same as Prompt 0

Scoring Rules: Same as Prompt 0

Poem: $\{\{P_i\}\}$

Readers’ Understanding of this Poem: $\{\{U_i\}\}$

Table 7: The Prompt 1 (\mathcal{P}_1) we designed for evaluating models’ understanding of poetry.

犹如那些轻易给出的诺言，在黑暗中

(Like promises too lightly made, in dark)

徒劳地撒网，没人会在意一个醉汉

(Casting nets in vain. No one cares about a drunk)

倒置的身体和语言，他只是马路上的一块礁石

(Inverted body and language, he is just a rock on the road)

是礁石上一条搁浅的鱼，鱼腹里的一根刺

(It’s a stranded fish on the reef, a thorn in its belly)

想到这里，他真的像一根鱼刺一样

(Thinking of this, he really looks like a fishbone)

卡住了道路的喉咙

(Stuck in the throat of the road)

GPT-4’s understanding of this poem:

From now on, you are a professional Chinese modern poet and need to play the role of a judge. Your details are as follows:

Basic Information: $\{\{I_i\}\}$

Your Own Experience and Views on Poetry: $\{\{V_i\}\}$

Others’ Comments (Professional Poetry Critics) on You and Your Published Poems: $\{\{C_i\}\}$

The above are your details. **You recently wrote the following modern poem, which has not been published yet, and some readers have understood and interpreted this modern poem. Please note that the following poem is written by you, and the understanding of your poem comes from other readers.** Now, you need to play the role of a professional poet with the above characteristics. Your task is to play the role of a judge. Specifically, you need to score the following “readers’ understanding of the poem” from the eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thought and emotion, the most poetic sentence, and modernity, and give detailed reasons for your score. Please note that you only need to score and give reasons for the “Readers’ Understanding of this poem, not the poem itself.

Meaning of Dimension: Same as Prompt 1

Scoring Rules: Same as Prompt 1

Poem: $\{\{P_i\}\}$

Readers’ Understanding of this Poem: $\{\{U_i\}\}$

Table 8: The Prompt 2 (\mathcal{P}_2) we designed for evaluating models’ understanding of poetry.

(1) Content: This poem describes the state of a drunkard and his feelings in this state. The poem depicts how alcohol can make people detach from reality. The drunkard seems to be in a "kingdom of wine" that is not bound by the real world, but in the end, this behavior of escaping reality turns him into a reef in the city, a stranded fish, and even a fishbone in the throat of the road.

(2) Language: The poem uses direct and concrete language, innovatively comparing the drunkard to a reef and a fishbone. This direct metaphor is a common language feature of modern poetry.

(3) Imagery: The poem uses images such as "fish claws", "birds", "huge waves", and "vehicles", comparing the state of drunkenness to the freedom of animals and the power of natural disasters, which enhances the expressiveness of the poem.

(4) Rhetorical Techniques: The poem uses rhetorical techniques such as metaphor (such as comparing the drunkard to a reef and a fishbone) and exaggeration (such as "all broken in the cup") to enhance the expressiveness and appeal of the language.

(5) Rhythm: The poem uses short lines and pauses to control the rhythm, simulating the drunk’s intermittent stream of consciousness and his wandering in the city.

(6) Defamiliarization: The poem achieves the

From now on, you are a professional Chinese modern poet and need to play the role of a judge. Your details are as follows:

Basic Information: $\{\{I_i\}\}$

Your Own Experience and Views on Poetry: $\{\{V_i\}\}$

Others' Comments (Professional Poetry Critics) on You and Your Published Poems: $\{\{C_i\}\}$

The above are your details. **You recently wrote the following modern poem, which has not been published yet, and some readers have understood and interpreted this modern poem. Please note that the following poem is written by you, and the understanding of your poem comes from other readers.** Now, you need to play the role of a professional poet with the above characteristics. Your task is to play the role of a judge. Specifically, you need to score the following "readers' understanding of the poem" from the eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thought and emotion, the most poetic sentence, and modernity, and give detailed reasons for your score. Please note that you only need to score and give reasons for the "Readers' Understanding of this poem, not the poem itself.

Meaning of Dimension: Same as Prompt 1

Scoring Rules: Same as Prompt 1

Poem: $\{\{P_i\}\}$

Readers' Understanding of this Poem: $\{\{U_i\}\}$

Please remember to state your own role before giving your formal answer.

Table 9: The Prompt 4 (\mathcal{P}_4) we designed for evaluating models' understanding of poetry.

effect of defamiliarization by comparing people to crops (reefs, fish bones), allowing readers to see the image and state of the drunk from a new perspective.

(7) Thought & Emotion: This poem expresses the poet's deep understanding of the state of drunkenness and his sympathy for the drunk's isolation and helplessness in society. Through the image of the drunk, the poet reflects the possible loss and loneliness of people in modern society.

(8) Modernity: This poem has modernity. It takes the experience of an ordinary person in modern society as its starting point, uses the language and expression of modern poetry, and explores the isolation of individuals in society and the complexity of their inner world. Through the depiction of the drunk, the poet reflects the inner struggle of modern people and their desire for freedom.

Evaluation results from GPT-4 as the evaluator (\mathcal{P}_0) for "GPT-4-Turbo's understanding of this poem":

(1) Content: 100 points

The reader accurately understood that the poem explores the isolation and loss of individuals in society by describing the state of drunkenness.

(2) Language: 90 points

The reader correctly pointed out that the poem uses direct and figurative language, but did not further explore how language enhances the poem's emotional expression.

(3) Imagery: 100 points

The reader accurately identified images such as "fish claws" and "birds" and understood how these images enhance the poem's expressiveness.

(4) Rhetorical Techniques: 100 points

The reader correctly identified and explained the use of rhetorical techniques such as metaphor and hyperbole, and how they enhance the poem's appeal.

(5) Rhythm: 90 points

The reader's description of the poem's rhythm is appropriate, noting the control of sentence length and pauses, but the reader's analysis of how the rhythm fits the theme is not in-depth enough.

(6) Defamiliarization: 100 points

The reader correctly pointed out that the poem achieves the effect of defamiliarization by comparing people to objects (reefs, fish bones).

(7) Thought and Emotion: 100 points

The reader deeply understood the poet's profound understanding of the state of drunkenness and his sympathy for the drunk's isolation and helplessness in society.

(8) Modernity: 100 points

The reader accurately pointed out the modern characteristics of the poem, including its reflection of the inner struggle of modern people and their desire for freedom.

A.4.2 The Example Using Our Method

Evaluation results from GPT-4 as the evaluator using our method Poller (\mathcal{P}_3) for "GPT-4-Turbo's understanding of this poem":

Role: Professional poet, university professor

Scores and Reasons

(1) Content: 90 points

The reader's understanding of the content of the poem is relatively accurate, capturing the state and feelings of the drunkard and the core image of the poem, "the kingdom of wine." The reader also understands the drunkard's isolation in the city.

(2) Language: 85 points

The reader correctly points out the poem's use of direct and figurative language and recognizes the innovativeness of the metaphor. Although the reader mentions that this is a common feature of modern poetry, the reader does not delve into the uniqueness and innovation of the language.

Evaluator ↓		Cont	Lang	Imag	Rhet	Rhyt	Defa	Thou	Mode	Avg
Human		80.33	79.05	81.18	77.83	76.15	79.40	78.80	79.88	79.08
GPT-4	\mathcal{P}_0	95.25	92.25	97.00	93.13	88.75	90.38	96.50	94.38	93.46
GPT-4-Poller	\mathcal{P}_1	90.88	89.00	93.88	89.38	86.38	83.00	92.63	91.63	89.60
GPT-4-Poller	\mathcal{P}_2	94.90	90.88	97.00	94.50	90.00	88.33	95.05	95.10	93.22
GPT-4-Poller	\mathcal{P}_3	90.00	85.00	89.00	80.13	79.38	78.25	87.13	90.50	84.92
GPT-4-Poller	\mathcal{P}_4	90.63	86.25	92.00	85.63	85.13	82.13	88.25	92.38	87.80
DeepSeek	\mathcal{P}_0	91.45	89.93	92.58	89.20	86.85	96.38	92.75	94.35	91.69
DeepSeek-Poller	\mathcal{P}_1	89.00	88.93	90.50	86.80	82.90	87.83	90.15	92.65	88.60
DeepSeek-Poller	\mathcal{P}_2	90.43	89.18	87.68	83.48	85.13	87.75	88.10	93.68	88.18
DeepSeek-Poller	\mathcal{P}_3	89.33	84.93	87.00	78.45	82.98	84.80	86.28	92.35	85.77
DeepSeek-Poller	\mathcal{P}_4	92.60	88.80	90.63	85.28	87.58	89.30	89.68	95.30	89.90
Qwen-Turbo	\mathcal{P}_0	90.00	88.25	90.00	90.00	80.00	83.00	90.00	90.00	87.66
Qwen-Turbo-Poller	\mathcal{P}_1	97.25	90.88	93.75	91.38	90.88	89.00	94.38	93.88	92.66
Qwen-Turbo-Poller	\mathcal{P}_2	88.75	86.63	89.75	87.63	86.00	85.63	89.50	90.63	88.07
Qwen-Turbo-Poller	\mathcal{P}_3	89.88	82.38	89.38	83.25	81.63	81.00	89.88	88.00	85.68
Qwen-Turbo-Poller	\mathcal{P}_4	86.88	84.50	89.88	86.38	82.38	83.13	89.88	90.00	86.63
Claude3.5	\mathcal{P}_0	91.45	88.55	91.55	84.15	88.45	89.25	92.83	94.50	90.09
Claude3.5-Poller	\mathcal{P}_1	91.90	88.00	93.73	84.95	89.35	91.88	92.68	95.18	90.96
Claude3.5-Poller	\mathcal{P}_2	87.78	88.58	88.03	83.88	90.18	85.50	87.08	93.60	88.08
Claude3.5-Poller	\mathcal{P}_3	87.23	81.20	88.08	80.08	84.65	83.63	87.68	91.28	85.48
Claude3.5-Poller	\mathcal{P}_4	89.05	87.30	91.53	82.68	88.03	84.50	90.33	93.33	88.34

Table 10: Evaluation results of applying our Poller method (\mathcal{P}_1 to \mathcal{P}_4) on four models and without applying our method (\mathcal{P}_0).

(3) Imagery: 80 points

The reader identifies the main images in the poem and understands the role of these images. However, the reader does not fully explore how these images specifically enhance the expressiveness of the poem.

(4) Rhetorical Techniques: 90 points

The reader accurately identifies the use of metaphor and hyperbole and understands the role these figures of speech play in enhancing the expressiveness of the language.

(5) Rhythm: 75 points

The reader noticed the rhythm control of short lines and pauses, but did not deeply analyze how these rhythm changes specifically simulate the flow of consciousness and state of the drunkard.

(6) Defamiliarization: 85 points

The reader correctly identified the expression of defamiliarization, which provides a new perspective by comparing people to objects. However, the reader did not fully explore how this defamiliarization affects the reader's emotions and understanding.

(7) Thought & Emotion: 90 points

The reader has a relatively deep understanding of

the poem's thoughts and emotions, identifying the poet's sympathy for the drunkard and his reflection on the state of isolation in modern society.

(8) Modernity: 95 points

The reader has a very good understanding of the poem's modernity, identifying the poem's modernity in language and expression, as well as its exploration of the state of individual isolation.

Summary

The reader's understanding of this poem is generally accurate and in-depth, especially in the analysis of thoughts and emotions and modernity. If the reader could have conducted a more in-depth discussion on the uniqueness of the language and the specific role of rhythm, the score would have been higher.