

Feedback as a Guide: Enhancing LLM Comprehension of Modern Chinese Poetry

Shanshan Wang¹ Derek F. Wong^{1*} Jiebo Luo² Jingming Yao³
Lidia S. Chao¹

¹NLP²CT Lab, University of Macau

²University of Rochester

³Department of Portuguese, University of Macau

nlp2ct.shanshan@gmail.com, {derekfw, jmyao, lidiasc}@um.edu.mo
jluo@cs.rochester.edu

Abstract

The performance of large language models (LLMs) on poetry-related tasks has drawn increasing attention from researchers. Since understanding of poetry is fundamental to all poetry-related tasks, enhancing LLMs' comprehension of poetry is crucial for advancing this field. In this study, we propose a novel method to improve LLMs' comprehension of modern Chinese poetry. This method utilizes feedback from reliable LLM evaluators as guidance to enhance other LLMs' understanding of poetry in specific dimensions. To identify reliable LLM evaluators, we evaluate the initial performance of several mainstream LLMs on poetry comprehension tasks and further refine their evaluation capabilities until they approach human (expert) performance. These reliable LLMs are then employed as evaluators to simulate poet's perspective in assessing the poetry comprehension of other LLMs, providing reasons and suggestions for improvement. Experimental results show that our method effectively enhances LLMs' comprehension of poetry across multiple dimensions.

1 Introduction

The performance of large language models (LLMs) on poetry-related tasks has caught the attention of researchers, especially in poetry generation (Antar, 2023; Deng et al., 2024; Hutson and Schnellmann, 2023), translation (Virvou et al., 2023; Wang et al., 2024b), and detection (Wang et al., 2025; Li et al., 2026; Wang et al., 2026c). However, previous studies have been limited to evaluating and analyzing the experimental outputs of LLMs on poetry-related tasks, without addressing the problem of poetry understanding from the source. For example, Virvou et al. (2023) conducted an evaluation study examining the analysis of Greek poetry generated by ChatGPT. They believe ChatGPT has

the potential to understand the content of poetry in depth. However, the fundamental question remains: Do LLMs really understand poetry? Poetry, especially modern poetry, has long faced the challenge of being understood due to its unique characteristics (Zhang, 2022; Peskin, 1998), such as its use of figurative language, rhythm, and imagery (Wang et al., 2024b; Refaee, 2023). Despite its importance, the problem of understanding poetry has not been sufficiently addressed in previous studies.

The accurate understanding of poetry is a fundamental prerequisite for all poetry-related tasks (Tate, 1940; Pierce, 2003; McGrath, 2018). Without this foundation, progress in areas such as poetry generation (Zhang and Lapata, 2014; Yi et al., 2018; Van de Cruys, 2020) and translation (Ghazvininejad et al., 2018; Chakrabarty et al., 2021) remains inherently constrained. Therefore, addressing LLMs' comprehension of poetry is essential to making advancements in these tasks.

Recent feedback-driven methods have shown promise in different domains. Zhang et al. (2024a) utilized feedback to relay error information back to the LLM, enabling it to reflect on the shortcomings of existing prompts and subsequently generate refined prompts. Chen et al. (2024) summarized the interaction data between the LLM agent and the feedback generator to identify recurring errors or causes of failure, and then optimizes the steps in the prompts. In the work of Zhu et al. (2024), LLMs generate new problems of different types based on the feedback (i.e., the evaluation of the difficulty of the problem) from small language models (SLMs). Wang et al. (2024a) utilized the complete text bug report to ensure that the description of the original bug report remains intact. Zhang et al. (2024b) used feedback such as running errors and incorrect answers to improve code generation.

While existing feedback-driven methods have demonstrated success in objective domains, such as mathematical reasoning and code generation, their

*Corresponding Author

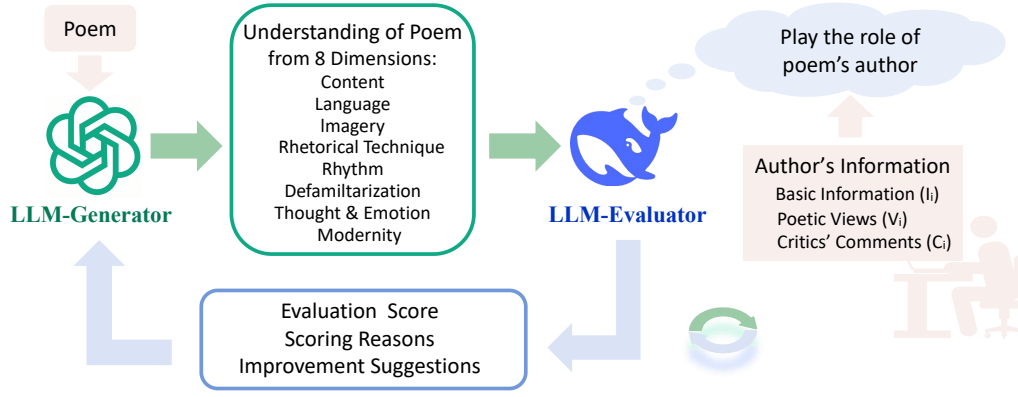


Figure 1: The framework of our proposed method.

effectiveness remains unexplored when applied to creative text comprehension tasks, particularly in understanding poetry. Unlike technical tasks with verifiable solutions, poetry interpretation is subjective and usually involves domain-specific expertise.

In this paper, we propose a novel method called **FIGURE** to enhance LLMs’ comprehension of modern Chinese poetry. **FIGURE** uses feedback from reliable evaluators as guidance to enhance the model’s understanding of poetry in specific dimensions. Unlike previous feedback-driven approaches that use error information (Zhang et al., 2024a) or difficulty scores (Zhu et al., 2024) as feedback, the feedback in our method incorporates three specialized components from reliable evaluators: (1) evaluation scores in eight dimensions, (2) reasons for each score, and (3) actionable improvement suggestions for enhancing the comprehension of the poem. The evaluators in our method are LLMs with evaluation capabilities close to humans.

Specifically, we first explore LLM’s initial comprehension of modern Chinese poetry. Reading poetry is inherently subjective, as interpretations often vary significantly among readers (Jiang, 2012; Zhang, 2022). However, the author’s original intent provides a unique benchmark for evaluation. Therefore, to evaluate the effectiveness of the understanding of LLMs, we invite all poetry authors to evaluate the understanding in terms of multiple dimensions. Human evaluation results prove that LLM’s comprehension of modern Chinese poetry is not satisfactory. For example, the error rate of LLMs’ understanding of rhetorical techniques in poetry is 22.17%. Subsequently, to enhance LLM’s comprehension, we use feedback from evaluators to guide the model to regenerate its understanding of poetry. Inspired by previous studies (Chiang and

Lee, 2023a; Wang et al., 2023; Liu et al., 2023a; Chu et al., 2024; Chan et al., 2023) using LLMs as evaluation tools, we use LLMs as evaluators for our poetry understanding task. Specifically, we prompt multiple mainstream LLMs to evaluate the understanding of poetry by playing the role of specific poetry authors, and enhance the evaluation capabilities of these LLMs. Finally, we identify the LLM with the evaluation performance closest to the human evaluation as the new evaluator.

In our method, the reliable LLM-Evaluator simulates the poet’s perspective to systematically evaluate the understanding of poetry from the LLM-Generator and give reasons and improvement suggestions. Next, the LLM-Generator iteratively regenerates its understanding of poetry using the multi-dimensional feedback from LLM-Evaluator as guidance. Experimental results show that our method enhances LLMs’ comprehension of poetry, proving the effectiveness of our work.

The main contributions of our work include:

- We are the first to propose a novel method to effectively enhance LLMs’ comprehension of modern Chinese poetry.
- We collect a set of high-quality, unpublished modern Chinese poems, and recruit several professional poets to evaluate LLMs’ comprehension to ensure a reliable evaluation.
- We explore and improve LLMs’ evaluation performance as evaluators on poetry comprehension, and identify models and prompts that are reliable in evaluating specific dimensions.

2 Related Work

Previous works explored the application of LLMs to poetry-related tasks, primarily focusing on

poetry generation and translation. [Hutson and Schnellmann \(2023\)](#) evaluated ChatGPT’s ability to create poetry by analyzing the limitations of poems generated by ChatGPT. They found that ChatGPT was able to imitate the poet’s language style, but the content created lacked emotional depth, highlighting the shortcomings of ChatGPT in poetry creation tasks. [Antar \(2023\)](#) studied the performance of GPT-4 in Arabic poetry creation, and the results showed that students could easily distinguish between texts created by GPT-4 and texts created by humans. [Deng et al. \(2024\)](#) proposed a framework for evaluating the performance of GPT-3.5 and GPT-4 in generating Chinese classical poetry. The experimental results show that most Chinese classical poems generated by LLM still have gaps compared with human poems. In an evaluative study ([Virvou et al., 2023](#)), professionals evaluated the English translation of Greek poetry generated by ChatGPT. They concluded that while ChatGPT could convey the basic meaning, it struggles to preserve the style of the original poem. [Wang et al. \(2024b\)](#) studied the translation of modern Chinese poetry by ChatGPT and found that there are some problems in ChatGPT’s translation that cannot be ignored. For example, ChatGPT’s translation is not accurate, and poeticity of the poem are lost in the process of translation.

In addition, ([Lu et al., 2024b](#)) constructed the CKnowEdit benchmark for Chinese knowledge editing, revealing challenges in adapting existing methods to Chinese-specific knowledge like Chinese ancient poetry and idioms. Complementing this, ([Gao et al., 2024](#)) demonstrated ChatGPT’s strong performance in translating classical Chinese poetry, especially with carefully designed prompts.

These works proved that LLMs have significant challenges in poetry-related tasks. Their conclusions and analyses are all derived from LLMs’ output. However, the basic premise of tasks such as poetry generation and translation should be the understanding of poetry. Different from previous studies, our work focuses on solving the problem of poetry understanding from the source.

3 Task Definition

In our work, understanding poetry means that the model generates the interpretation of a poem from different dimensions. Our task is to enhance the model’s comprehension of poetry, which is formalized as a multi-dimensional iterative optimiza-

tion problem. Given a poem P and its understanding U , our objective is to improve the evaluation score of the understanding of the poem after T iterations. The iterative refinement integrates role-playing driven feedback:

$$U^{(t+1)} = M_g \left(P, U^{(t)}, \mathcal{F}_D^{(t)} \right) \quad (1)$$

$$\mathcal{F}_D^{(t)} = M_e \left(P, U^{(t)}; \mathcal{R}(I_i, V_i, C_i) \right) \quad (2)$$

where $\mathcal{D}$ denotes the target dimension subset (Cont, Lang, Imag, Rhet, Rhyt, Defa, Thou, Mode). The multi-dimensional feedback $\mathcal{F}_D^{(t)}$ comprises three elements per dimension d : a quantitative score S_d , the evaluation reasons R_d , and actionable improvement suggestions $\mathcal{A}_d$. $S_d \in [0, 100]$ is the evaluation score for specific dimension d , and w_d is the dimension weight coefficient controlling the optimization objective type. When optimizing the average score across all dimensions (e.g., Tables 3 and 4), $w_d = \frac{1}{|\mathcal{D}|}$ with $\mathcal{D}$ containing eight dimensions. When optimizing a specific dimension (e.g., Tables 5 and 6), $w_d = 1$ with $\mathcal{D}$ containing only one dimension. This feedback is generated through the poet role-playing mechanism $\mathcal{R}(I_i, V_i, C_i)$, which synthesizes the detailed information of the poem’s author, including basic information I_i , poetic views V_i , and critics’ comments C_i . Simulating the poet’s perspective enables the evaluator model (LLM-Evaluator) M_e to approximate authentic authorial judgment.

Subsequently, the generator model (LLM-Generator) M_g produces optimized comprehension $U^{(t+1)}$ using $\mathcal{F}_D^{(t)}$, while M_e generates new feedback through $\mathcal{R}$. The comprehension output $U^{(t)}$ at iteration t begins with the initial understanding $U^{(0)} = M_g(P)$ generated without feedback ($\mathcal{F}_D^{(-1)} = \emptyset$), then iteratively improves through the described feedback loop. Experimental results (Section 7) demonstrate this framework with $T = 2$ iterations, where $\mathcal{R}$ is implemented with the carefully designed prompt.

4 LLM’s Initial Comprehension

4.1 Data

Referring to the scale of poetry datasets used in previous studies ([Wang et al., 2024b](#); [Virvou et al., 2023](#); [Hutson and Schnellmann, 2023](#); [Antar, 2023](#)), our work uses a larger amount of data. Specifically, we collected 40 high-quality **unpublished** modern Chinese poems by five professional

modern Chinese poets, of which eight poems are written by the same poet. These poems total 110 stanzas, 600 lines, and 8K words.

Professional Poets The poets involved in our work are all professional poets who write modern Chinese poetry, with rich experience in poetry, including senior editors of famous poetry journals, university professors, members of the Chinese Writers Association, etc. They have published excellent poetry collections, published influential research papers on poetry, and won many high-quality poetry awards. They are not only professional poetry creators, but also senior researchers of poetry theory. Therefore, the poetry data and evaluation results from them in our work are precious and valuable for promoting the development of the related domain.

Diversity of Poems Although our dataset contains only 40 poems, it ensures stylistic diversity. These poems come from five professional poets with different writing styles. It is worth noting that these poets were born in different eras and have different professions and specific regional life experiences. This has resulted in a rich diversity of styles, themes, and forms in their poems. For example, the dataset includes various forms of poetry, including single-stanza and multi-stanza poems, as well as poems with uniform or non-uniform line counts. The themes of these poems are also very broad, covering emotions, daily life, social issues, philosophy, and more. The language styles range from straightforward and natural to subtle and implicit, calm and restrained, and so on.

4.2 Dimensions of Poetry

Referring to previous work on understanding modern poetry (Chang, 2012; Wang et al., 2026a,b), and based on the recommendations of eight professional modern poets, we first identified eight critical dimensions (Yu, 2001; Wang, 1998; Xi, 2019; Wang, 2006, 2019; Cheng, 2005) for understanding modern poetry: 1) Content (Cont): Refers to the subject matter or theme of the poem, i.e., what the poem describes. 2) Language (Lang): The linguistic features of the poem. 3) Imagery (Imag): The use of images within the poem. 4) Rhetorical Techniques (Rhet): The rhetorical strategies employed in the poem. 5) Rhythm (Rhyt): The internal rhythm of the poem, which often manifests in external forms such as pauses, lines, and stanzas. 6) Defamiliarization (Defa): A common

method of expression in modern poetry that highlights the poem’s innovation. 7) Thought and Emotion (Thou): The thought and emotion conveyed by the poem. 8) Modernity (Mode): The modernity and contemporary consciousness of the poem. Each dimension has its own unique focus. These dimensions are clearly delineated to ensure clarity and precision in the evaluation process.

Subsequently, GPT-4 (gpt-4-turbo) (Achiam et al., 2023) generates the initial understanding of all poems from eight dimensions using the prompts in Appendix A.1. 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.

4.3 Human Evaluation

Human evaluation is often regarded as the gold standard (Resnik and Lin, 2010), particularly for creative tasks like poetry. Previous poetry-related works relied heavily on human evaluation (Manurung et al., 2012; Zhang and Lapata, 2014; Yan, 2016; Yi et al., 2018). In our work, human evaluation serves as the reference for other evaluation methods. Reading poetry is inherently subjective, as interpretations often vary significantly among readers (Brodsky, 1999; Jiang, 2012; Zhang, 2022). However, the author’s original creative intent is unique, providing a distinctive benchmark for evaluation. Therefore, to accurately confirm LLM’s capability to understand poetry, we invited the five authors of poems to evaluate the experimental results from all evaluation dimensions. As information provided in the Section 4.1, the authors (evaluator) are experienced professional poets and seasoned scholars of poetry theory. They are familiar with and agree with each of our proposed evaluation dimensions and their definitions.

Evaluation Dimensions The evaluation dimensions are exactly the same as *Dimensions of Poetry* presented in Section 4.

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. The content of the human evaluation questionnaire we provided to the evaluators is shown in Table 1. To avoid bias

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.

Meaning of Dimension: 1) Content: Refers to the subject matter or theme of the poem, i.e., what the poem describes. 2) Language: The linguistic features of the poem. 3) Imagery: The use of images within the poem. 4) Rhetorical Techniques: The rhetorical strategies employed in the poem. 5) Rhythm: The internal rhythm of the poem, which often manifests in external forms such as pauses, lines, and stanzas. 6) Defamiliarization: A common method of expression in modern poetry that highlights the poem’s innovation. 7) Thought and Emotion: The thought and emotion conveyed by the poem. 8) Modernity: The modernity and contemporary consciousness of the poem.

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 1: The content of the human evaluation questionnaire we provided to the authors (evaluators).

and ensure the validity of the evaluation, we did not tell the evaluators the source of the “readers’ understanding of the poem”.

Evaluation results are reported in the second row of Table 2. Human evaluation results prove that LLM’s comprehension of modern Chinese poetry is not satisfactory. For example, LLM scored 77.83 points for its understanding of rhetorical techniques (Rhet) in poems, which means that its error rate in understanding rhetorical techniques is 22.17%.

5 LLM as Evaluator

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., 2023b). While automatic evaluation is prevalent in Natural Language Processing (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., 2024b; Novikova et al., 2017).

Inspired by previous work (Chiang and Lee, 2023a; Wang et al., 2023; Liu et al., 2023a) on using LLMs as evaluation tools for tasks such as open-domain conversations (Lin and Chen, 2023), text summarization (Chu et al., 2024), and open-ended question answering (Chan et al., 2023), we use LLMs as evaluators for the poetry understand-

ing task. We first explore the initial evaluation performance of LLMs for poetry understanding, and then enhance the LLMs’ evaluation performance to approach the human evaluation results. Finally, we identified the LLMs and prompts that are close to human evaluation as reliable evaluators.

5.1 LLMs Initial Evaluation

To explore the initial evaluation ability of LLMs in the poetry comprehension task, we prompted LLMs to evaluate the understanding of poems. Specifically, we experimented with four mainstream LLMs, including: GPT-4-Turbo (Achiam et al., 2023), Claude3.5-Sonnet (Anthropic, 2024), DeepSeek-R1 (Guo et al., 2025), Qwen2.5-Turbo (Yang et al., 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 8, denoted as $\mathcal{P}_0$.

Table 2 ($\mathcal{P}_0$) shows the evaluation results of LLMs on the “understanding of poems” from different dimensions. The smaller the difference between the model’s score and the human evaluation score, the closer its evaluation performance is to that of humans, proving that the model’s evaluation performance is better. To more intuitively observe the performance of different models, we present the evaluation difference between the LLMs’ scores and human scores in Table 14 of Appendix A.4.

LLMs’ evaluation results are unexpected, as we are disappointed to find that LLMs were not completely reliable in all dimensions. Specifically, Qwen2.5-Turbo’s score for defamiliarization (83.00) is closest to the human score (79.40), and its score for rhythm is only 3.85 points different from the human score. 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

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	$\mathcal{P}_1$	90.88	89.00	93.88	89.38	86.38	83.00	92.63	91.63	89.60
GPT-4	$\mathcal{P}_2$	90.00	85.00	89.00	80.13	79.38	78.25	87.13	90.50	84.92
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	$\mathcal{P}_1$	91.90	88.00	93.73	84.95	89.35	91.88	92.68	95.18	90.96
Claude3.5	$\mathcal{P}_2$	87.23	81.20	88.08	80.08	84.65	83.63	87.68	91.28	85.48
DeepSeek	$\mathcal{P}_0$	91.45	89.93	92.58	89.20	86.85	96.38	92.75	94.35	91.69
DeepSeek	$\mathcal{P}_1$	89.00	88.93	90.50	86.80	82.90	87.83	90.15	92.65	88.60
DeepSeek	$\mathcal{P}_2$	89.33	84.93	87.00	78.45	82.98	84.80	86.28	92.35	85.77
Qwen	$\mathcal{P}_0$	90.00	88.25	90.00	90.00	80.00	83.00	90.00	90.00	87.66
Qwen	$\mathcal{P}_1$	97.25	90.88	93.75	91.38	90.88	89.00	94.38	93.88	92.66
Qwen	$\mathcal{P}_2$	89.88	82.38	89.38	83.25	81.63	81.00	89.88	88.00	85.68

Table 2: Evaluation results from the human experts and different LLMs. The smaller the difference between the model’s score and the human evaluation score, the better the evaluation performance of the model. Cont, Lang, Imag, Rhet, Rhyt, Defa, Thou, Mode, and Avg represent content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thought and emotion, modernity, and the average score of all dimensions, respectively.

points (Claude3.5-Sonnet), different from the human score. And the score for specific dimensions is at least 6.32 points (Claude3.5-Sonnet), different from the human score. Overall, the evaluation results from all LLMs are not close to those of human evaluation, which proves that current LLMs are not suitable for the evaluation of poetry-related tasks.

5.2 Enhance LLM’s Evaluation

To enhance the evaluation ability of LLMs, we prompt LLMs to play the role of the 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, including: 1) Authors’ basic information, such as occupation, growth background, life background, personal experience, etc. 2) Authors’ own experiences and views on poetry. 3) Others’ comments (professional poetry critics) on authors and their published poems. Appendix A.2 shows the relationship between authors’ information and model’s comprehension.

Authors’ Information In order to effectively stimulate the ability of LLMs to act as authors, we collected and screened the detailed information of all real 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.

For LLMs, different prompts may lead to different evaluation results (Xu et al., 2023). Therefore, to enhance the evaluation ability of LLMs, we designed two different prompts, denoted as $\mathcal{P}_1$ and $\mathcal{P}_2$, 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. Details are shown in Table 9.
- $\mathcal{P}_2$: Based on $\mathcal{P}_1$, inspired by (Lu et al., 2024a), the model is required to declare its role before the formal evaluation. Details are shown in Table 10.

Subsequently, LLMs evaluate the understanding of the poems from all dimensions under different prompts and give reasons. Table 2 presents the evaluation results from humans and different LLMs.

Obviously, compared with $\mathcal{P}_0$, our specially designed $\mathcal{P}_1$ and $\mathcal{P}_2$ effectively improve the evaluation performance of all models on “understanding of poetry” in different dimensions. The improvement of model evaluation performance is reflected in the fact that the evaluation results using $\mathcal{P}_1$ and $\mathcal{P}_2$ are closer to the human evaluation results than the evaluation results using $\mathcal{P}_0$. Among them, the evaluation performance of DeepSeek-R1 on “understanding of poetry” in the dimension of

rhetorical techniques has the largest improvement. Specifically, for rhetoric, the evaluation result of DeepSeek-R1 using $\mathcal{P}_0$ is 89.20 points, which is 11.37 points different from the human evaluation result (77.83); while the evaluation result using $\mathcal{P}_2$ is 78.45 points, which is only 0.62 points different from the human evaluation. The improvement effect of the evaluation performance of defamiliarization is second only to rhetoric. Specifically, the evaluation result of GPT-4-Turbo using $\mathcal{P}_2$ on defamiliarization (78.25) is only 1.15 different from the manual evaluation result (79.40).

The evaluation results prove that playing the role of a detailed poem author can improve the model’s evaluation performance, and letting the model declare its role before evaluation can further improve the model’s evaluation ability. Based on the enhanced experimental results, we conclude that compared with other models, DeepSeek-R1’s evaluation results for rhetorical techniques when using $\mathcal{P}_2$ are reliable because it is closest to humans. Similarly, GPT-4-Turbo’s evaluation of defamiliarization using $\mathcal{P}_2$ is the most reliable.

6 Enhancing LLM Comprehension

To enhance LLM’s comprehension of modern poetry, we propose a novel method termed FIGURE (Feedback as a **G**uide: Enhancing LLM Comprehension of Modern Chinese Poetry). FIGURE method uses feedback from reliable evaluators as guidance to enhance the model’s understanding of poetry. The LLM that generates the understanding is denoted as LLM-Generator, while the LLM used for evaluation is denoted as LLM-Evaluator. Figure 1 presents the framework of our proposed method. Specifically, our FIGURE method is divided into three steps: 1) In the first step, LLM-Generator generates an initial understanding of the poem. 2) In the second step, the reliable LLM-Evaluator plays the role of the poem’s author, evaluates the understanding of the poem generated in the first step, and gives specific evaluation reasons and improvement suggestions. Among them, the evaluation capability of LLM-Evaluator is close to humans. 3) In the third step, LLM-Generator regenerates the understanding of the poem based on the evaluation and suggestions given by LLM-Evaluator in the second step.

7 Experiment

Based on the evaluation results and analysis in Section 5.2, DeepSeek-R1 ($\mathcal{P}_2$) is used as a reliable model for evaluating rhetorical techniques, and GPT-4-Turbo ($\mathcal{P}_2$) is used as a reliable model for evaluating defamiliarization. Using the proposed FIGURE method, we enhance the model’s understanding of poetry across all dimensions and individual dimensions through two iterative refinements, and evaluate these understandings. The evaluation needs to be objective and fair, so the temperature of the evaluation model is set to 0.1, and the top_p is set to 0.8. The prompt used by LLM-Evaluator is presented in Table 11. Other prompts and experimental examples are provided in Appendix A.1 and A.3, respectively.

DeepSeek-R1 as Evaluator When GPT-4-Turbo served as generator and DeepSeek-R1 served as evaluator, the experimental process is as follows:

1. GPT-4-Turbo first generates the initial understanding U_0 of poetry from multiple dimensions. Then, DeepSeek-R1 evaluates the initial understanding generated by GPT-4-Turbo and give a score, specific reasons, and improvement suggestions. The evaluation score of the initial understanding is denoted as E_0 .
2. In the first enhancement, GPT-4-Turbo regenerates its understanding U_1 of poetry based on the score E_0 , reasons, and suggestions given by DeepSeek-R1. Then, DeepSeek-R1 evaluates the re-understanding U_1 and gave a score, specific reasons, and suggestions for improvement. The evaluation score of the first re-understanding is denoted as E_1 .
3. In the second enhancement, GPT-4-Turbo regenerates its understanding U_2 of poetry based on the previous two scores (E_0 and E_1), reasons, and suggestions given by DeepSeek-R1. Finally, DeepSeek-R1 evaluates the second re-understanding and gives a score, specific reasons, and improvement suggestions. The evaluation score of the second re-understanding is denoted as E_2 .

Notably, during all experiments, LLM-Evaluator did not know that the second and third evaluated texts were enhanced. Specifically, we checked and deleted the texts containing words such as “re-understanding, correction” to ensure the fairness and effectiveness of the evaluation. Furthermore,

	Cont	Lang	Imag	Rhet	Rhyt	Defa	Thou	Mode	Avg
E_0	86.28	84.60	83.43	76.28	80.88	86.60	82.38	90.15	83.82
E_1	87.20	87.55	86.50	85.05	83.70	91.40	84.93	89.08	86.93
E_2	88.55	88.03	87.53	88.90	82.60	90.80	87.90	89.80	88.01

Table 3: Evaluation results of DeepSeek-R1 as an evaluator when GPT-4-Turbo as a generator generates understanding of poetry from multiple dimensions at the same time. E_0 in the first column represents the initial score of GPT-4-Turbo’s understanding of the poems, and E_1 and E_2 represent the scores after the first and second enhanced understanding, respectively.

	Cont	Lang	Imag	Rhet	Rhyt	Defa	Thou	Mode	Avg
E_0	90.75	86.25	89.90	85.40	79.83	83.38	88.75	89.08	86.67
E_1	89.75	86.25	88.78	83.05	82.10	87.25	87.80	88.58	86.69
E_2	89.00	85.50	88.78	81.48	81.05	87.80	86.90	88.15	86.08

Table 4: Evaluation results of GPT-4-Turbo as an evaluator when GPT-4-Turbo is used as a generator to generate an understanding of poetry from multiple dimensions at the same time.

	E_0'	E_1'	E_2'
Rhet	80.05	85.53	87.23

Table 5: Evaluation results of DeepSeek-R1 as evaluator when GPT-4-Turbo as generator only generates the understanding from rhetorical techniques.

	E_0'	E_1'	E_2'
Defa	87.00	90.58	92.08

Table 6: Evaluation results of GPT-4-Turbo as evaluator when GPT-4-Turbo as generator only generates the understanding from defamiliarization.

the evaluated model does not know that the “understanding of the poem” comes from LLM. Instead, we tell the evaluated model that the “understanding of the poem” comes from the reader. Table 3 presents the experimental results of GPT-4-Turbo as generator and DeepSeek-R1 as evaluator.

To determine the applicability of our method to a single dimension, GPT-4-Turbo is prompted to generate an understanding of poetry from the perspective of rhetorical techniques, and DeepSeek-R1 is used as an evaluation model to evaluate the understanding generated by GPT-4-Turbo and give improvement suggestions. This experimental process is the same as the experimental process of GPT-4-Turbo generating understanding of all dimensions at the same time. The experimental results are shown in Table 5.

GPT-4-Turbo as Evaluator Similarly, we use GPT-4-Turbo as the evaluation model to conduct experiments both for all dimensions and specifically for the defamiliarization dimension. The model for generating understanding is also GPT-4-Turbo. However, generation and evaluation are performed separately. The experimental results are

presented in Table 4 and Table 6.

8 Analysis

As shown in Table 3, when the model generates all dimensions at the same time, our method effectively enhances LLM’s understanding of poetry from the rhetorical techniques dimension. Specifically, the model’s initial understanding score (E_0) of poetry is 76.28 points, proving that the model’s accuracy in understanding rhetorical techniques is only 76.28%. However, after the first enhanced understanding using our method, the model’s accuracy in understanding poetry from the rhetorical techniques dimension reached 85.05%. After the second enhancement, the accuracy in understanding poetry from the rhetorical techniques dimension rose to 88.90%.

Similarly, when the model only generates the rhetoric dimension (Table 5), our method also effectively improves LLM’s accuracy in understanding poetry from rhetorical techniques. Specifically, after the first and second enhanced understanding using our method, the accuracy of the model’s understanding of poetry increased from 80.05% to 85.53% and 87.23%, respectively.

When using GPT-4-Turbo as a generator, which differs from the human evaluation results by 1.15 points in defamiliarization, our method can still improve the LLM’s comprehension of poetry. Specifically, as shown in Table 4, after using our method to enhance LLM’s understanding of poetry from the alienation dimension twice, LLM’s accuracy in understanding poetry increased from 83.38% to 87.25% and 87.80%, respectively. Table 6 once again verifies the effectiveness of our method in enhancing the model’s understanding performance.

Interestingly, as shown in Table 4, when GPT-4-Turbo is used as the evaluation model, except for the two dimensions of defamiliarization and rhythm, the evaluation model’s suggestions do not enhance LLM’s understanding of poetry and even have the opposite effect. We speculate that the reason for this result is that GPT-4-Turbo’s evaluation results for other dimensions except defamiliarization are unreliable. As shown in Table 2, GPT-4-Turbo’s evaluation of defamiliarization is relatively accurate, and its evaluation result (78.25) is only 1.15 points different from the human evaluation result (79.40). However, GPT-4-Turbo’s evaluation of other dimensions is not accurate because its evaluation results are far from the human evaluation results. For example, GPT-4-Turbo’s evaluation result of modernity (90.50) is 10.62 points different from the human evaluation result (79.88), and the evaluation result of content (90.00) is 9.67 points different from the human evaluation result (80.33).

Table 4 proves the necessity of determining a reliable evaluation model in Section 5.2, because a reliable evaluation model could provide evaluation results close to human evaluation and give correct improvement suggestions. This once again emphasizes the effectiveness of our proposed method.

8.1 Error Analysis

To ensure the interpretability and depth of this paper, we provided the granular error analysis, including both corrected and persist misunderstandings.

Our experimental results demonstrate that feedback-driven iteration enhances the model’s understanding of poetry in both rhetoric and defamiliarization dimensions. To analyze misunderstandings in more detail, we summarize the types of misunderstandings that were corrected and those that persist.

Corrected Misunderstandings General, superficial, and incomplete understandings. Specifically, in the model’s initial understanding, it consistently uses general descriptions to present an explicit understanding of the poem without analyzing it in conjunction with specific themes and content. However, guided by feedback, the model generates a more comprehensive and in-depth understanding, identifying the poem’s implicit meanings and conducting detailed analysis of specific dimensions based on the poem’s context. As illustrated in the example in Appendix A.3, in the initial understanding, the generative model (GPT-4-Turbo) can only

identify explicit metaphors and personification in the poem, but ignores implicit rhetorical devices such as synesthesia. After the evaluation model (DeepSeek-R1) explicitly points out these omissions and provides improvement suggestions in the feedback, the generative model is able to successfully incorporate these deeper rhetorical devices into its understanding in subsequent iterations.

Persist Misunderstandings Certain deep-seated cultural backgrounds and emotional tones closely related to the poet’s specific personal experiences or local memories, such as "manifestation of the urban-rural identity divide", remain difficult for the model to fully grasp even after two rounds of feedback. This reveals the model’s limitations in handling highly personalized and contextualized knowledge.

9 Conclusion

The understanding of poetry is the premise of all poetry-related tasks. In this paper, we propose a novel method that can effectively enhance LLMs’ comprehension of modern Chinese poetry. Our method utilizes feedback from reliable LLM-Evaluators as guidance to enhance other LLMs’ comprehension of poetry in specific dimensions. Specifically, to ensure the reliability of the evaluation, we invited several professional poets to evaluate LLMs’ comprehension of poetry. Furthermore, we identify reliable LLMs that can be used as evaluators to evaluate other LLMs’ comprehension and provide reasons and improvement suggestions. Experimental results demonstrate that our method effectively enhances LLMs’ comprehension of poetry across multiple dimensions. Our work provides a strong foundation for future research into poetry-related applications of LLMs.

Limitations

Referring to the scale of poetry datasets used in previous studies, we collected 40 high-quality unpublished modern Chinese poems written by several professional modern Chinese poets. These poems total 110 stanzas, 600 lines, and 8K words (excluding titles).

Compared to existing works on large language models (LLMs) and poetry, our dataset is not small. For example:

- [Antar \(2023\)](#) used 1 poem (10 lines) in their evaluation.

- [Hutson and Schnellmann \(2023\)](#) focused on 4 modern poems.
- [Wang et al. \(2024b\)](#) used 10 poems in their main experiments.
- [Virvou et al. \(2023\)](#) evaluated 30 poems.

Thus, our dataset is among the largest in studies related to LLMs and poetry. By including a larger and more diverse set of poems, we aim to enhance LLMs' evaluation performance on poetry understanding task.

However, compared to non-poetry related tasks, the amount of our data is relatively small, although our data is of high quality as it comes from professional poets.

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 Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- 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.
- 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*.
- Weizhe Chen, Sven Koenig, and Bistra Dilkina. 2024. Reprompt: Planning by automatic prompt engineering for large language models agents. *arXiv preprint arXiv:2406.11132*.
- 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 automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.
- 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*.
- 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*.
- 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.
- Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2018. [Neural poetry translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 67–71.
- 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.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llm-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*.
- 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, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023b. Llm as narcissistic evaluators: When ego inflates evaluation scores. *arXiv preprint arXiv:2311.09766*.
- Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024a. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.
- Tianhe Lu, Jizhan Fang, Yunzhi Yao, Xin Xu, Ningyu Zhang, and Huajun Chen. 2024b. Benchmarking chinese knowledge rectification in large language models. *arXiv e-prints*, pages arXiv–2409.
- 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 Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.
- 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.
- Joan Peskin. 1998. Constructing meaning when reading poetry: An expert-novice study. *Cognition and Instruction*, 16(3):235–263.
- Robert B Pierce. 2003. [Defining" poetry"](#). *Philosophy and Literature*, 27(1):151–163.
- 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.
- Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. Finesure: Fine-grained summarization evaluation using llms. *arXiv preprint arXiv:2407.00908*.
- Allen Tate. 1940. [Understanding modern poetry](#). *The English Journal*, 29(4):263–274.
- Tim Van de Cruys. 2020. [Automatic poetry generation from prosaic text](#). In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2471–2480.
- Maria Virvou, George A Tsihrintzis, Dionisios N Sotiropoulos, Konstantina Chrysafiadi, Evangelos Sakopoulos, and Evangelia-Aikaterini Tsihrintzi. 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.
- Dingbang Wang, Yu Zhao, Sidong Feng, Zhaoxu Zhang, William GJ Halfond, Chunyang Chen, Xiaoxia Sun, Jiangfan Shi, and Tingting Yu. 2024a. Feedback-driven automated whole bug report reproduction for android apps. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 1048–1060.
- 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. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.
- 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. 2024b. [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. 2026a. [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, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2026b. [Poller: Are LLMs suitable for evaluating poetry understanding task?](#) In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 9720–9735. 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.
- Shanshan Wang, Fengying Ye, Hanjia Lyu, Caiwen Gou, Junchao Wu, Jingming Yao, Chengzhong Xu, Jiebo Luo, and Derek F Wong. 2026c. [Seeing the poem: Image-semantic detection of ai-generated modern chinese poetry with mllms.](#) *arXiv preprint arXiv:2605.22654*.
- Shuting Wang. 2006. [Giving shape to emotion: The theory and practice of rhythm and imagery in new poetry \(1917-1937\).](#) *World Literature Review*.
- 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*.
- 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.
- Haizhang Yu. 2001. [About the changes and analysis of new poetry expressions.](#) *writing*, 000(011):3–5.
- Chenrui Zhang, Lin Liu, Chuyuan Wang, Xiao Sun, Hongyu Wang, Jinpeng Wang, and Mingchen Cai. 2024a. [Prefer: Prompt ensemble learning via feedback-reflect-refine.](#) In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19525–19532.
- Huan Zhang, Wei Cheng, Yuhan Wu, and Wei Hu. 2024b. [A pair programming framework for code generation via multi-plan exploration and feedback-driven refinement.](#) In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 1319–1331.
- 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.
- Xunyu Zhu, Jian Li, Can Ma, and Weiping Wang. 2024. [Improving mathematical reasoning capabilities of small language models via feedback-driven distillation.](#) *arXiv preprint arXiv:2411.14698*.

A Appendix

A.1 Prompts Designed in Our Work

Table 7 shows the prompt we designed for LLMs to understand modern poetry.

Table 8, 10, and 10 respectively present the prompt 0 ($\mathcal{P}_0$), prompt 1 ($\mathcal{P}_1$), and prompt 2 ($\mathcal{P}_2$) we designed for evaluating models’ understanding of poetry.

The prompt used by LLM-Evaluator (enhanced) to evaluate and provide suggestions for improvement is shown in Table 11.

The prompts used by LLM-Generator to regenerate understanding in the first and second enhancements are presented in Table 12 and Table 13.

In our prompts, we use Chinese poems directly in the source language without translating them into English.

A.2 Relationship Between Author Information Model’s Comprehension

In our work, to enhance the evaluation ability of LLMs, we prompt LLMs to assume the role of the authors of poems. Specifically, we give LLMs detailed information about these authors, including:

- Biographical context (occupation, life background, personal experiences).
- Author’s own views on poetry.
- Critical commentary from professional reviewers.

Please understand the following modern poem titled T_i from the eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thought and emotion, and modernity, and finally give your understanding 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.
 Language: What are the characteristics of the language of this poem? Is it innovative?
 Imagery: What imageries are used in this poem, and what are their functions?
 Rhetorical Techniques: What rhetorical techniques are used in this poem? What is the function?
 Rhythm: How is the rhythm of this poem? Such as pauses, lines, and stanzas.
 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.
 Thought and Emotion: What thoughts and emotions does this poem express?
 Modernity: Is the poem modern or modern-minded?

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

Understanding of this poem: $\{\{U_i\}\}$

Table 7: The prompt we designed for LLMs to understand modern poetry.

In modern poetry, rhythm refers to internal cadence expressed through pauses, line breaks, and stanza structure. These choices reflect the poet’s intent and emotional state, which are deeply tied to personal background and artistic perspective. For example, abrupt, irregular line breaks may signal urban fragmentation, a stylistic decision rooted in the poet’s lived experience. Supplying this context enables the evaluator to interpret rhythm functionally rather than superficially, producing feedback that is both targeted and accurate. This feedback then guides the generator toward deeper, context-aware understanding.

In short, after understanding the author’s background information, the LLM evaluator can identify superficial interpretations and offer practical suggestions. The generator then incorporates this feedback into their analysis, thereby refining it.

A.3 Example of the Experiment

An experimental example with GPT-4-Turbo as the generator and DeepSeek-R1 as the evaluator is presented below:

1. Initial Understanding (from GPT-4-Turbo)

Poem: (Figure 2 presents the original Chinese version of this poem)

The Matter of Being Drunk (Title)

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”, not the poem itself.

Scoring Dimensions: Same as human evaluation

Scoring Rules: All dimensions range 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 8: The prompt 0 (P_0) we designed for evaluating models’ understanding of poetry, aiming to explore the initial evaluation ability of LLMs in the poetry comprehension task.

醉酒一事

在酒的国度里，鱼鳖可以
 离开海洋，鸟雀不必站在树上
 男人醉酒后，很放心地
 把自己摆到了街边。人间之事
 全都碎在杯里，一饮而尽后
 在肠胃里激荡，像巨浪冲刷着堤防

车辆在街灯的重影里轰鸣、拉扯
 犹如那些轻易给出的诺言，在黑暗中
 徒劳地撒网，没人会在意一个醉汉
 倒置的身体和语言，他只是马路上的一块礁石
 是礁石上一条搁浅的鱼，鱼腹里的一根刺

想到这里，他真的像一根鱼刺一样
 卡住了道路的喉咙

Figure 2: The original Chinese version of the poem *The Matter of Being Drunk*.

In the kingdom of wine, turtles might
 Depart the sea, and birds need not perch in trees
 A man, inebriated, with carefree trust
 Lay himself along the street’s edge, where worldly
 matters
 All shattered in the cup, drained in one gulp
 Surging in the stomach, like giant waves washing
 over the embankment

Vehicles roar and tug in the doubled shadows of
 streetlights

Like promises too lightly made, in dark
 Casting nets in vain. No one cares about a drunk
 Inverted body and language, he is just a rock on
 the road

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

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, 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.

Scoring Dimension: Same as Prompt 0

Scoring Rules: Same as Prompt 0

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

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

Table 9: The prompt 1 ($\mathcal{P}_1$) we designed for evaluating LLMs' comprehension of poetry.

Thinking of this, he really looks like a fishbone
Stuck in the throat of the road

GPT-4-Turbo's initial understanding of this poem:

- **(1) Content** This poem takes drunkenness as its theme and depicts a man's state and feelings after being drunk. The poem explores the relationship between man and society, as well as the psychological and physiological reactions of man after being drunk, through the state of drunkenness.
- **(2) Language** The language of the poem is concise and symbolic. By comparing drunkenness to the freedom of fish in the ocean and birds in the trees, it expresses the sense of relief brought by drunkenness. At the same time, the use of some everyday language (such as "the roar and pull of vehicles in the shadows of street lights") makes the poem more down-to-earth and enhances the intuitiveness and appeal of the expression.
- **(3) Imagery** The poem uses a variety of images, such as "fish claws", "birds", "huge waves", "reefs", and "fish bones". These images not only enrich the visual effect of the poem, but also deepen the theme of the poem, expressing the helplessness, confusion, and

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", not the poem itself.

The Meaning of Each Scoring Dimension: Same as Prompt 1 ($\mathcal{P}_1$)

Scoring Rules: Same as Prompt 1 ($\mathcal{P}_1$)

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

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

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

Table 10: The prompt 2 ($\mathcal{P}_2$) we designed for evaluating models' understanding of poetry.

isolation of people in society when they are drunk.

- **(4) Rhetorical Techniques** The poem uses metaphors (e.g., "like a huge wave washing over a dike") and personification (e.g., "the cars roared and dragged in the shadows of the streetlights"). This rhetoric enhances the expressiveness of the language, allowing the reader to more vividly experience the inner turmoil and the dynamics of the surrounding environment brought about by drunkenness.
- **(5) Rhythm** The rhythm of the poem is reflected through irregular line breaks and pauses. For example, the use of punctuation and a mixture of long and short sentences in the poem reflects the ups and downs of the drunk's psychology and emotions.
- **(6) Defamiliarization** The poem achieves a defamiliarization effect by comparing the drunk to "a reef on the road" and "a stranded fish on the reef." This expression makes the common scene of drunkenness more unique and thought-provoking, forcing the reader to examine the relationship between drunken-

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, ..., and modernity, and **give detailed reasons** for your score. Finally, please **give specific suggestions for improvement for each dimension**. Please note that you only need to score and give reasons for the "Readers' Understanding," not the poem itself.

Meaning of Dimension: Same as human evaluation

Scoring Rules: Same as human evaluation

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

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

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

Table 11: The prompt used by LLM-Evaluator to evaluate and provide suggestions for improvement.

ness and the individual from a new perspective.

- **(7) Thought & Emotion** By depicting the loneliness and helplessness of a drunkard, the poet expresses his concern for the living conditions of modern people and a profound revelation of the loneliness of human nature. The poem reveals a desire for freedom and helplessness in the face of the constraints of reality.
- **(8) Modernity** This poem is clearly modern, not only in the modern social context of its theme (drunkenness and the isolation of urban life), but also in the innovation of its language and expression. The poem explores the relationship between the individual and society from a modern perspective, reflecting the complexity and contradictions within modern people.

DeepSeek-R1's evaluation score, reasons, and suggestions for improvement of "GPT-4-Turbo's initial understanding of this poem":

Role Statement: As a modern Chinese poet, university professor, and designer, I will use my own creative philosophy and the perspective of a professional critic to combine the textual characteristics of the poem with the reader's understanding, score it from eight dimensions, and provide suggestions for improvement. The following is a detailed

The author of this poem (the original poet) served as a judge and scored your (the reader's) understanding of the above eight dimensions of content, language, imagery, rhetoric, rhythm, alienation, thoughts and emotions, and modernity, and gave detailed reasons for the scoring and suggestions for improvement. The scoring range is between 0 and 100. 0 points means that your understanding of a certain dimension is completely wrong, and 100 points means that your understanding of a certain dimension is completely correct; the more correct your understanding of a certain dimension, the higher the score you give. Now, please re-understand this modern poem titled T_i from the eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thought and emotion, and modernity based on the author's scoring, reasons, and suggestions for improvement.

The author of the poem's scoring, reasons, and suggestions for your understanding of the above eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thoughts and emotions, and modernity: $\{\{E_0, \text{reasons and suggestions}\}\}$

Table 12: The prompt used by LLM-Generator to regenerate understanding in the first enhancement.

analysis:

(1) Content

Score: 85 points

Reason: The reader accurately grasped the core content of "the state after being drunk" and pointed out the discussion of "the relationship between people and society." However, the reader did not delve into the philosophical nature of "drunkenness" as a metaphor in the poem - for example, "fish bones stuck in the throat of the road" implies the individual's subtle but sharp resistance to order, and this "feeling of powerlessness in resistance" was not pointed out by the reader.

Improvement Suggestions: More attention should be paid to the absurdity of the "body and space confrontation" implied in the poem (the drunkard becomes a reef or a fish bone), and the symbolic meaning of "drunkenness" as an escape from the real order.

(2) Language

Score: 80 points

Reason: The reader noticed the symbolism of the language (such as the freedom of fish claws and birds), but did not deeply analyze the tension between "daily language" and "poetic language". For example, "the roar and pull of vehicles in the shadows of street lights" is not a simple daily description, but a superposition of vision and hearing created by "shadows" and "pull", which implicitly corresponds to the sensory confusion of the drunk.

Improvement suggestions: It is necessary to

The author of this poem (the original poet) served as a judge and scored your (the reader's) **re-understanding** of the above eight dimensions of content, language, imagery, rhetoric, rhythm, alienation, thoughts and emotions, and modernity, and gave detailed reasons for the scoring and **new suggestions** for improvement. The scoring range is between 0 and 100. 0 points means that your understanding of a certain dimension is completely wrong, and 100 points means that your understanding of a certain dimension is completely correct; the more correct your understanding of a certain dimension, the higher the score you give. Now, please re-understand this modern poem titled T_i from the eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thoughts and emotions, and modernity based on the author's **previous two** scoring, reasons and suggestions for improvement.

The author of the poem's scoring, reasons, and suggestions for your **re-understanding** of the above eight dimensions of content, language, imagery, rhetorical techniques, rhythm, defamiliarization, thoughts and emotions, and modernity: $\{\{E_0, \text{reasons and suggestions}\}\}$

Table 13: The prompt used by LLM-Generator to regenerate understanding in the second enhancement.

emphasize how language can achieve the alienation of the drunk's perspective through "sensory dislocation" (such as sound being "pulled" and light being "shadowed").

(3) Imagery

Score: 85 points

Reason: The reader listed the main images, but did not reveal the progressive structure of the images: from "the fish claw leaves the ocean" (freedom) to "the reef and the stranded fish" (trapped), and finally to "the fish bone stuck in the throat" (sharp pain). This chain of "freedom-confinement-pain" is the deep logic of the imagery.

Improvement suggestions: Analyze the dynamic relationship of the image group rather than listing them in isolation. For example, how can "fish bone" as the ultimate image condense the outbreak of all the contradictions in the previous text?

(4) Rhetorical Techniques

Score: 75 points

Reason: The reader only recognized metaphors and personification, but ignored the key rhetoric - "paradox" (such as "inverted body and language") and "synaesthesia" (the "turbulent" stomach is like "huge waves"). In addition, "a fish bone stuck in the road's throat" is "the materialization of a body organ" and comparing the road to the throat is a surreal metaphor.

Improvement Suggestions: Attention should

be paid to the complexity of rhetoric (such as metaphors nested in personification) and the role of rhetoric in deepening the theme.

(5) Rhythm

Score: 80 points

Reason: The reader mentioned "irregular line breaks" but did not specifically analyze how the rhythm echoes the flow of consciousness of the drunkard. For example, "stirring in the stomach, like a huge wave washing over the embankment" uses a long sentence to simulate the prolonged feeling of physical discomfort, while "a thorn in the fish's belly" suddenly changes the line to create a short, stinging feeling.

Improvement Suggestions: Combine the physical length and emotional density of the poem to analyze how the rhythm imitates the drunkard's psychological state.

(6) Defamiliarization

Score: 90 points

Reason: The reader keenly grasps the defamiliarization chain of "reef-fish-fishbone", but fails to point out that its essence is "dimensionality reduction of perspective": from macroscopic (ocean) to microscopic (fishbone), the insignificance and sharpness of the individual are highlighted by reducing the scale.

Improvement Suggestions: The connection between defamiliarization and the theme needs to be supplemented. The more microscopic the image, the more it can pierce the hypocrisy of reality.

(7) Thought & Emotion

Score: 85 points

Reason: The reader extracts "loneliness" and "desire for freedom", but fails to touch upon the more complex dilemma of modernity in the poem: the promise of "casting a net in vain" points to the emptiness of interpersonal trust, while "fishbone stuck in the throat" implies the individual's retaliation against society.

Improvement Suggestions: The "cold lyricism" feature needs to be combined to reveal the multiple contradictions of emotions (such as the coexistence of relief and pain).

(8) Modernity

Score: 85 points

Reason: The reader correctly points out the modern background of "urban isolation", but does not delve into the modernity of the language structure. For example, "inverted body and language" alludes to the postmodern discussion of "subject dissolution", while the violent image of "fishbone stuck in the throat" points to Cioran's existential absurdity.

Improvement Suggestions: The poem should be linked to its metaphor of "body politics" (drunkard as disciplined flesh) and language experiments (such as paradoxical rhetoric).

Summary

The reader's understanding of the poem is generally accurate, but there is still room for improvement in the complexity of Rhetorical Techniques and Thought & Emotion. It is suggested to use "micro versus macro" as a clue to re-examine the intertextual relationship between imagery, rhetoric, and theme, and pay attention to the reflection of the "cold lyricism" in the poem on the spiritual dilemma of modern people.

2. First Enhanced Understanding

GPT-4-Turbo's first enhanced understanding of this poem:

Based on the detailed Scores, reasons, and improvement suggestions from the poet, we can conduct a deeper analysis of our understanding of "The Matter of Being Drunk":

(1) Content

This poem not only describes the state of being drunk, but also, through the metaphor of "fish bones stuck in the throat of the road," expresses the individual's subtle but sharp resistance to social order and the powerlessness of resistance. Drunkenness here becomes a symbol of escaping the real order, revealing the absurd conflict between body and space.

(2) Language

The language of the poem creates an effect of sensory dislocation through the combination of everyday and poetic. For example, "vehicles roar and pull in the shadows of street lights" expresses the sensory confusion of the drunkard through the superposition of vision and hearing. This use of language not only enhances the alienation effect, but also deepens the reader's perception of the drunken state.

(3) Imagery

The imagery in the poem presents a progressive structure from freedom to confinement and then to acute pain. For example, "the fish claw leaves

the ocean" symbolizes freedom, "the stranded fish on the reef" implies confinement, and "a thorn in the fish's belly" turns to acute pain. These images are interrelated and together construct a complex narrative about bondage and resistance.

(4) Rhetorical Techniques

The poem uses rhetoric such as metaphor, personification, paradox, and synaesthesia. In particular, the surreal metaphor of "a fish bone stuck in the road's throat" adds to the symbolic depth and appeal of the poem by comparing the road to the throat, while also reflecting the deepening of the theme.

(5) Rhythm

The rhythm of the poem simulates the flow of consciousness and physiological feelings of the drunkard through varying sentence lengths and sudden line breaks. Long sentences express the persistent feeling of physical discomfort, while short sentences such as "a thorn in the fish's belly" highlight the momentary feeling of pain, echoing the drunkard's psychological state.

(6) Defamiliarization

The defamiliarization in the poem is achieved through a shift in perspective from macro to micro, giving readers a new understanding of the common drunken scene. This dimensionality reduction of perspective not only highlights the insignificance and sharpness of the individual, but also strengthens the criticism of social reality.

(7) Thought & Emotion

The poem expresses a complex dilemma of modernity, including the emptiness of interpersonal trust and the reciprocal existence of the individual against society. Through the "cold lyricism" technique, the poet reveals the emotional contradiction of the coexistence of relief and pain, reflecting the spiritual dilemma of modern people.

(8) Modernity

The modernity of this poem is reflected not only in its theme and language, but also in its metaphor and language experimentation with "body politics". The postmodern elements in the poem, such as the discussion of "subject dissolution" and the use of paradoxical rhetoric, emphasize the profound reflection and criticism of reality in modern poetry.

Through such analysis, we can more comprehensively understand the deep meaning and artistic achievements of this poem.

Evaluation results after the first enhancement (from DeepSeek-R1):

Role Statement: As a judge, I will use the perspective of a professional poet, combined with my own creative ideas and poetry criticism experience, to conduct a multi-dimensional evaluation of readers' understanding of "The Matter of Being Drunk" and provide improvement suggestions for their interpretation deviations.

(1) Content

Score: 88 points

Reason: Readers accurately capture the confrontational relationship between individuals and order under the appearance of "drunkenness", but the interpretation of "absurd conflict" is a bit vague. The "fishbone stuck in the throat" in the poem is not simply resistance, but also implies the tragedy of the individual's self-dissolution in a huge system - the sharpness of the thorn is eventually dissolved by the devouring nature of the road, which forms a closed loop with the "feeling of powerlessness to resist" and requires strengthening the text details to support it.

Improvement Suggestions: The progressive relationship between "reef" and "fishbone" can be supplemented with analysis - the reef symbolizes passive endurance, while the fishbone refers to active pain, but both are assimilated by the road (system) as "foreign objects", so as to deepen the interpretation of "failure of resistance".

(2) Language

Score: 92 points

Reason: The reader pointed out the language strategy of "sensory dislocation", but did not deeply analyze the ambiguity of the word "double image" - it is both the halo of the street lamp and the cognitive split of the drunk. For example, "pulling" not only describes the dynamics of the vehicle, but also metaphorically implies the falseness of promises. It needs to be analyzed in detail in the context.

Improvement suggestions: The tension between the liquid image of "agitation/washing" and the solid image of "reef/fish bone" can be disassembled to reveal how language reflects the mental collapse of the drunk through the transformation of physical state.

(3) Imagery

Score: 95 points

Reason: The reader grasped the progressive structure of the image accurately, but did not point out the dimensionality reduction chain of "fish claw-bird-drunkard": marine life → land life →

human, implying that drunkenness is a reverse escape from the "evolutionary order". In addition, the "thorn in the fish's belly" is actually a metaphor for self-harm, and the subjective critique needs to be strengthened.

Improvement Suggestions: The "fish claw that left the ocean" and the "fish bone stuck in the road" can be linked to analyze how the poet constructs the existential dilemma of "nowhere to escape" through the dislocation of organisms and space.

(4) Rhetorical Techniques

Score: 90 points

Reason: The reader recognizes the core metaphor, but ignores the personification of "the promise is cast in vain"—comparing the promise to the fishless fishing, which forms an intertextuality with "the drunkard is like a reef": both are abandoned and ineffective existences. In addition, "the stomach is churning like a huge wave" is actually a miniature tsunami in the body, and the epic treatment of its body space needs to be pointed out.

Improvement Suggestions: The nested metaphors of "drunkard-fish bone-road" can be explored to form a rhetorical box, making the tiny individual a symbol of system obstruction.

(5) Rhythm

Score: 85 points

Reason: The reader noticed the change in sentence length, but did not analyze the dramatic effect of the line-breaking strategy. For example, "He was just a reef on the road" is written on a separate line, which strengthens the sense of entity through visual weight; while "a thorn in the fish's belly" is quickly concluded, imitating the instantaneous pain of being pierced. It is also necessary to pay attention to the sense of disconnection between "after drinking it all" and the following text, which is actually a deliberate blank in the flow of consciousness.

Improvement Suggestions: The analysis can be combined with breathing rhythm, such as the extension of long sentences corresponds to alcohol paralysis, and the abruptness of short sentences corresponds to nerve tingling, revealing the isomorphism between form and content.

(6) Defamiliarization

Score: 93 points

Reason: The reader proposed "perspective dimension reduction", but did not reveal a deeper cognitive subversion. For example, the "drunkard's

inverted body" is not only a physical reversal, but also a perversion of the value order - the individual who is ignored when sober becomes a "heterogeneous existence that gets stuck in the throat of the system" after being drunk. It is necessary to strengthen the reversal of the power relationship between the subject and the environment.

Improvement Suggestions: The confrontation between "drunkard as reef" and "vehicle as wave" can be linked to analyze how alienation elevates the individual's dilemma into a modern allegory.

(7) Thought & Emotion

Score: 87 points

Reason: The reader mentioned "cold lyricism" but did not touch on the paradox of "self-objectification" in the poem - the drunkard gained temporary freedom by giving up his humanity (turning into a reef, a fishbone), which is an irony of the spiritual alienation of modern people. In addition, "promises cast a net in the dark" alludes to the performativity of interpersonal relationships, and the deconstruction of the trust mechanism needs to be supplemented.

Improvement Suggestions: The metaphor of the digestive system of the body and the city can be combined with "stomach turmoil" and "road throat" to reveal the terrifying warmth of modernity devouring individuals.

(8) Modernity

Score: 94 points

Reason: The reader keenly grasped the metaphor of "body politics", but did not delve into the critique of "drunkenness" as a modern survival strategy - actively creating chaos to resist discipline, but ultimately falling into deeper alienation. The "roaring of vehicles" in the poem symbolizes the violence of industrialized time, and the "fishbone stuck in the throat" is a micro-uprising of the individual against the system's backlash. It is necessary to strengthen the revelation of the paradox of modernity.

Improvement Suggestions: The "human affairs broken in the cup" can be linked to the fragmented existence of postmodernism, and how poetry deconstructs grand discourse through micro-narratives.

Summary

The reader's interpretation of the poem is highly professional, but there is still room for improvement in the close reading of the text and the in-

tegrity of the metaphor chain. It is recommended to anchor more closely to specific verses, avoid over-reliance on theoretical assumptions, and keep the interpretation rooted in the texture of language and the breath of imagery.

3. Second enhanced understanding (from GPT-4-Turbo)

GPT-4-Turbo's second enhanced understanding of this poem: Based on the detailed Scores, reasons, and improvement suggestions provided by the poet, we can further deepen our understanding of the modern poem "The Matter of Being Drunk":

(1) Content

This poem explores the confrontation between the individual and the social order, as well as the self-dissolution of the individual in a huge system, through the act of "drunkenness." In particular, the symbol of "fishbone stuck in the throat" reflects the sharpness of individual resistance and the devouring nature of the system, forming a sense of powerlessness in resistance and the tragedy of a closed loop.

(2) Language

The language in the poem enhances the expression through sensory dislocation and puns. For example, "double image" not only refers to the halo of street lights, but also symbolizes the cognitive split of the drunkard, while "pulling" implies dynamics and false promises. These language techniques reflect the spiritual collapse of the drunkard through the transformation of physical states.

(3) Imagery

The imagery, from "fish claw leaving the ocean" to "bone in the fish belly," constructs a progressive chain from freedom to predicament. This series of images not only reveals that drunkenness is an act of escaping the evolutionary order in reverse, but also implies a subjective critique of self-harm.

(4) Rhetorical Techniques

The rhetoric of the poem, such as the personification of "promises cast in vain" and the micro-tsunami in the body of "the stomach churns like huge waves", deepens the theme of the poem. These rhetorical techniques not only enhance the power of expression, but also make the tiny individual a symbol of systemic obstruction.

(5) Rhythm

The rhythm of the poem imitates the flow of consciousness and physiological feelings of the drunkard through changes in sentence length and line division strategies. In particular, the line divisions of "he is just a reef on the road" and "a thorn in the

belly of the fish" strengthen the sense of entity and instantaneous pain through visual weight and rapid conclusion.

(6) Defamiliarization

Defamiliarization in this poem achieves the subversion of cognition and the perversion of the value order through the "inverted body of the drunkard". This defamiliarization is not only a physical reversal, but also an allegory of modernity, reflecting the reversal of the power relationship between the individual and the environment.

(7) Thought & Emotion

The poem explores the spiritual alienation of modern people through the paradox of "self-objectification". The temporary freedom that the drunkard gains by giving up his humanity is actually an irony of the spiritual state of modern people. In addition, the poem also deconstructs the performative nature and trust mechanism of interpersonal relationships.

(8) Modernity

The poem criticizes the paradox of "drunkenness" as a modern survival strategy, which resists discipline through active chaos, but ultimately falls into deeper alienation. The "roaring of vehicles" and "fish bones stuck in the throat" in the poem symbolize the violence of industrialized time and the micro-uprising of individuals against the system, revealing the complex paradox of modernity.

Through this deeper analysis, we can more comprehensively understand the artistic achievements and deep meaning of the modern poem "The Matter of Being Drunk".

Evaluation results after the second enhancement (from DeepSeek-R1):

Role Statement: As a judge, I will score readers' understanding of "The Matter of Being Drunk" from eight dimensions based on my own poetry creation philosophy and the critics' summary of my poetry characteristics, and provide specific suggestions for improvement.

(1) Content

Score: 92 points

Reason: The reader accurately captured the confrontational relationship between the individual and the system in the poem, especially the interpretation of the sharpness of resistance and the tragic closed loop symbolized by the "fish bone stuck in the throat", which is consistent with my creative concept of "interweaving pain and poetry" in the poem. However, I did not notice the infiltration of the implicit "local memory" into the dilemma

of modernity in the poem (for example, the image of "reef" and "fish" may be a metaphor for the urban-rural divide), so I deduct points.

Improvement Suggestions: The content needs to be analyzed in combination with the poet's potential life experience (such as the sticky memory of the local area), revealing that "drunkenness" is not only a social confrontation, but also a concrete manifestation of the urban-rural identity divide.

(2) Language

Score: 88 points

Reason: The reader pointed out the punning of "double image" and "pull", which is consistent with my language strategy of "writing close to the subject". However, it did not point out that "the stomach is churning like a huge wave" is actually a "micro-tsunami in the body". This design thinking - transforming physiological feelings into geographical disasters is the key to language innovation.

Improvement Suggestions: It is necessary to emphasize how language can achieve sensory synaesthesia through "physical transformation" (such as elevating the feeling of drunken vomiting to natural violence), rather than just staying at the pun level.

(3) Imagery

Score: 85 points

Reason: The reader paid attention to the progressive chain of imagery, but did not analyze the deep connection between "fish claws leaving the ocean" and "birds do not have to stand on trees" - the two are actually the dissolution of the natural order, forming a logical closed loop with "drunkard inversion". In addition, "reef" as the boundary image between the city and the ocean was not included in the analysis.

Improvement suggestion: It is necessary to explore the topological relationship between images (e.g., "fish-reef-thorn" forms a spatial nesting), rather than just linear interpretation.

(4) Rhetorical Techniques

Score: 90 points

Reason: The reader's interpretation of the personification rhetoric (the promise of casting a net) and the internal tsunami metaphor is accurate, but he did not notice that "the drunkard is the reef" is actually a perspective shift of "viewing myself through the object" - this is the core technique of my "subject role switching".

Improvement suggestion: It should be pointed out how rhetoric achieves defamiliarization through perspective shift (person → reef → fish → thorn), rather than analyzing a single rhetoric in isolation.

(5) Rhythm

Score: 80 points

Reason: The reader mentioned the effect of line breaks on "visual weight", but did not find the breathing rhythm between stanzas: the first two long sentences describe the swelling feeling after being drunk, and the short sentences in the last stanza suddenly shrink to "fish bones stuck in the throat", imitating the physiological curve of the drunk from ecstasy to suffocation.

Improvement Suggestions: It is necessary to combine the synchronization of sentence length changes with physiological experience (such as the long sentences corresponding to "excitement" and the abrupt stop corresponding to "stuck").

(6) Defamiliarization

Score: 95 points

Reason: The reader accurately identified the cognitive subversion of "inverted body" and related it to the allegory of modernity, which is in line with my defamiliarization strategy of "subject materialization". However, it did not point out that "fish bones stuck in the road and throat" is actually the paradoxical aesthetic of "using small obstructions to fight against the huge system", which is exactly the high-level expression of defamiliarization.

Improvement Suggestions: It is necessary to further explain how alienation creates the tension between "the insignificance of resistance and the absurdity of the system".

(7) Thought & Emotion

Score: 85 points

Reason: The reader's interpretation of "self-objectification" and "interpersonal performativity" is profound, but the "lyrical undertone of sadness" in the poem is ignored. For example, the loneliness implied by "the roar of vehicles" and "the casting of promises" is actually a metaphor for the individual's wandering in the process of urbanization, which needs to be echoed in my poem "the pain that is rare among my peers".

Improvement Suggestions: It should combine the dual emotional structure of "sadness" and "absurdity" instead of just emphasizing criticality.

(8) Modernity

Score: 93 points

Reason: The reader interprets "drunkenness" as a modern survival strategy and reveals the paradox of industrialized violence and individual uprising, which is completely consistent with my pursuit of modernity of "combining real dilemmas with metaphysical thinking". However, the intertextuality between "active chaos" and "local memory" in the poem is not explored (e.g., drunkenness may be a mourning for the poetic spirit of farming).

Improvement Suggestions: The "nostalgic dimension" of modernity needs to be addressed, namely, drunkenness as a temporary antidote to the urban-rural divide.

Summary

The reader's overall understanding of "The Matter of Being Drunk" is profound, particularly in terms of defamiliarization and modernity, which approach the poetic spirit. However, there is still room for improvement in the mechanisms of linguistic innovation, the topological relationships of imagery, and the complexity of emotion. It is suggested to restructure the analytical framework using "design thinking," viewing the poem as a dynamic system and focusing on the interplay between elements rather than a one-way interpretation.

A.4 Evaluation Difference Between LLMs and Human

To more intuitively observe the performance of different models, Table 14 presents the evaluation difference between the LLMs' scores and human scores. The smaller the difference between the model's score and the human evaluation score, the closer its evaluation performance is to that of humans. Therefore, the smaller the score, the better the evaluation performance of the model.

Evaluator		Cont	Lang	Imag	Rhet	Rhyt	Defa	Thou	Mode	Avg
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	$\mathcal{P}_1$	11.85	9.00	12.52	7.45	13.45	12.48	14.00	15.35	12.01
Claude3.5	$\mathcal{P}_2$	6.90	2.15	6.90	2.25	8.50	4.23	8.88	11.40	6.40
GPT-4	$\mathcal{P}_0$	14.92	13.20	15.82	15.30	12.60	10.98	17.70	14.50	14.38
GPT-4	$\mathcal{P}_1$	10.55	9.95	12.70	11.55	10.23	3.60	13.83	11.75	10.52
GPT-4	$\mathcal{P}_2$	9.67	5.95	7.82	2.30	3.23	1.15	8.33	10.62	5.85
Qwen	$\mathcal{P}_0$	9.67	9.20	8.82	12.17	3.85	3.60	11.20	10.12	8.58
Qwen	$\mathcal{P}_1$	16.92	11.83	12.57	13.55	14.73	9.60	15.58	14.00	13.60
Qwen	$\mathcal{P}_2$	9.55	3.33	8.20	5.42	5.48	1.60	11.08	8.12	6.60
DeepSeek	$\mathcal{P}_0$	11.12	10.88	11.40	11.37	10.70	16.98	13.95	14.47	12.61
DeepSeek	$\mathcal{P}_1$	8.67	9.88	9.32	8.97	6.75	8.43	11.35	12.77	9.52
DeepSeek	$\mathcal{P}_2$	9.00	5.88	5.82	0.62	6.83	5.40	7.48	12.47	6.69

Table 14: Evaluation difference between the LLMs’ scores and human scores. The smaller the difference between the model’s score and the human evaluation score, the closer its evaluation performance is to that of humans. Therefore, the smaller the score, the better the evaluation performance of the model.