

From Poetry Parties to Poetry Critics: Benchmarking Classical Chinese Poetry Generation and Evaluation

Zihao Yi^{1,2,3}, Zhaopeng Tu², Ruotian Ma², Qingxuan Jiang²,
Zhe Xu^{1,2}, Qingyu Xiao³, Yixue Li^{3,5}, Ying Shen^{1,4,*}, Liefeng Bo²

¹Sun Yat-Sen University, ²Tencent Hunyuan Multimodal Department

³Guangzhou National Laboratory, ⁴Peng Cheng Laboratory

⁵GMU-GIBH Joint School of Life Sciences, The Guangdong-Hong Kong-Macao
Joint Laboratory for Cell Fate Regulation and Diseases, Guangzhou Medical University

Abstract

Classical Chinese poetry generation and evaluation with Large Language Models (LLMs) face two intertwined challenges: existing generation benchmarks rely on simplistic keyword inputs devoid of literary context, and no ground-truth benchmark exists for objectively validating poetry evaluation methods. We present a comprehensive framework addressing both challenges. On the **generation** side, we construct **Honglou-Poem**, the first narrative-driven poetry generation benchmark derived from *Dream of the Red Chamber*, where each entry provides rich author personas and narrative backgrounds—moving far beyond traditional keyword-to-poem paradigms. A “Digital Poetry Party” pipeline evaluates 12 state-of-the-art LLMs in this challenging setting. On the **evaluation** side, we introduce **Poetry-Tree**, a benchmark with axiomatically ordered quality levels created through systematic merit degradation, providing objective ground truth for evaluator validation. Leveraging Poetry-Tree as a training signal, we employ **GRPO** reinforcement learning with Spearman’s correlation as the reward to train a dedicated poetry evaluation model, **Poet-4B**. Our Poet-4B achieves a Spearman’s ρ of **0.8253**, dramatically improving from the base model’s 0.2929 and closely approaching the best closed-source model GLM-4.6 (0.8764) with over $100\times$ fewer parameters—a gap of only 0.051 in Spearman’s ρ . We release all benchmarks, trained models, and a leaderboard to facilitate future research. Our code and data are available at <https://github.com/Tencent/digitalhuman/tree/main/POET>

1 Introduction

Poetry parties (诗会) were a cornerstone of classical Chinese literary culture, where scholars gathered to compose poetry inspired by shared themes,

landscapes, or narratives. These gatherings were sophisticated intellectual exchanges involving composition, critique, and appreciation, with each poem emerging from a rich tapestry of context—the poet’s personal experiences, immediate circumstances, and cultural milieu. Both the *creation* and *judgment* of poetry were integral to these events, and we argue that the same dual perspective is essential for advancing LLM-based poetry research.

The advent of Large Language Models (LLMs) has opened new possibilities for classical poetry generation (Tan et al., 2025; Song et al., 2025). However, two fundamental bottlenecks persist.

On the generation side, existing benchmarks reduce the creative act of poetry composition to simplistic inputs—a single keyword, a topic phrase, or a fixed template (Liu et al., 2018; Ma et al., 2023). This stands in stark contrast to how poetry was actually composed: grounded in rich narrative contexts, authored by individuals with distinct personalities and emotional states. No benchmark captures this contextual richness, making it impossible to meaningfully assess whether LLMs can truly compose poetry in the way human poets once did.

On the evaluation side, automated metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) fail to capture the aesthetic value central to poetic artistry (Gonalo Oliveira, 2017), while human evaluation is costly and unreliable (Sawicki et al., 2025). More fundamentally, *no ground-truth benchmark exists* with axiomatically defined quality levels, making it impossible to rigorously validate any evaluation method. Recent work has explored using LLMs as evaluators, but open-source models perform substantially below state-of-the-art closed-source models in poetry evaluation (e.g., the base Qwen3-4B achieves only 0.2929 Spearman’s ρ on our benchmark, compared to 0.8764 for the best closed-source model).

These two challenges are deeply connected:

*Correspondence to: sheny76@mail.sysu.edu.cn.

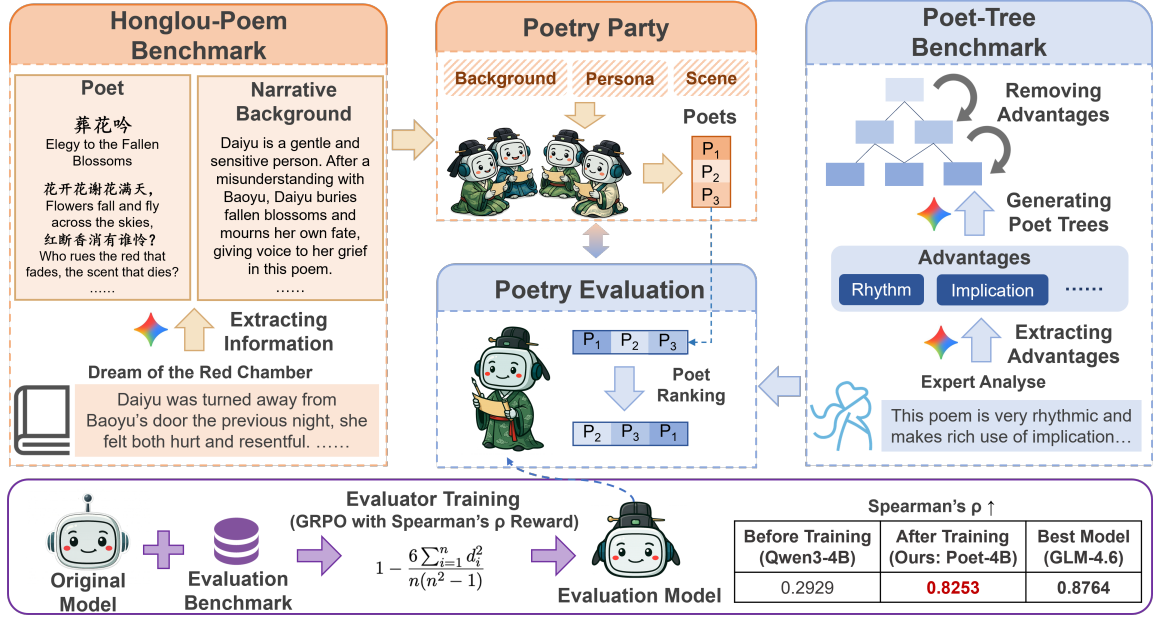


Figure 1: Overview of our framework. **Left**: The Honglou-Poem benchmark extracts poems from *Dream of the Red Chamber* and enriches each entry with narrative backgrounds, author personas, and scene descriptions. **Right**: Poetry-Tree is constructed by first extracting advantages from expert analyse, then progressively removing each merit via LLM, forming a quality-ordered “poetry tree”. **Middle**: The concrete setting bridging generation and evaluation—LLMs compose poems under the same narrative context, and the poems are ranked by quality (darker poem nodes P_i indicate lower quality). **Bottom**: The training pipeline for **Poet-4B**, our dedicated evaluation model trained with GRPO on Poetry-Tree, achieving performance rivaling top closed-source models.

without rigorous evaluation, we cannot measure generation progress; without challenging generation benchmarks, evaluation research lacks meaningful test scenarios. To address both challenges in a unified framework, we make the following contributions:

- We construct **Honglou-Poem**, the first narrative-driven poetry generation benchmark from *Dream of the Red Chamber*, providing rich author personas and narrative backgrounds that move far beyond simplistic keyword inputs. We design a “Digital Poetry Party” pipeline and benchmark 12 state-of-the-art LLMs.
- We introduce **Poetry-Tree**, a novel evaluation benchmark with axiomatically ordered quality levels created by systematically degrading specific poetic merits, providing the first objective ground truth for poetry evaluator assessment.
- We propose a **GRPO-based training method** that leverages Poetry-Tree as a training signal with Spearman’s correlation as the reward function, yielding **Poet-4B**, our dedicated poetry evaluation model. Poet-4B achieves Spearman’s $\rho=0.8253$, closely approaching the best closed-

source model GLM-4.6 ($\rho=0.8764$, a gap of only 0.051) with over $100\times$ fewer parameters.

2 Poetry Generation

2.1 Honglou-Poem: Dataset Construction

Traditional poetry generation datasets provide only isolated poems or simple keyword annotations such as a single topic word or a seasonal phrase. While such settings are useful for studying surface-level fluency, they fundamentally fail to capture the conditions under which classical Chinese poetry was actually composed: deeply embedded in narrative context, authored by individuals with specific personalities, social standings, and emotional states.

To bridge this gap, we construct the **Honglou-Poem** benchmark from *Dream of the Red Chamber* (红楼梦), an 18th-century Chinese masterpiece widely regarded as the pinnacle of Chinese literary fiction. The novel is uniquely suited for this purpose: it contains 160 classical poems (after filtering) composed by more than 20 distinct fictional characters, each with richly developed personalities, and every poem arises from a specific narrative occasion—a poetry party, a seasonal celebration, or a moment of personal reflection. Our construction

pipeline has three stages:

Stage 1: Poem Extraction. Let the novel be partitioned into chapters $\mathcal{T} = \{T_1, \dots, T_N\}$ ($N=120$). For each chapter T_i , an LLM $\mathcal{M}_{\text{extract}}$ extracts all embedded poems:

$$\mathcal{P}_i = \mathcal{M}_{\text{extract}}(T_i) = \{p_1^{(i)}, p_2^{(i)}, \dots, p_{k_i}^{(i)}\} \quad (1)$$

We apply rule-based post-processing to filter prose passages misidentified as poetry and to merge poems split across paragraphs, yielding a clean set of 160 poems.

Stage 2: Context Enrichment. Since the poems in *Dream of the Red Chamber* already exist within the novel, our task in this stage is *retrospective*: given the known poem $p_j^{(i)}$ and its surrounding narrative window, we reconstruct the authorial and situational context that would have motivated its composition. For each poem $p_j^{(i)}$, we provide $\mathcal{M}_{\text{context}}$ with a context window $\mathcal{C}_j^{(i)} = \{T_{i-1}, T_i, T_{i+1}\}$ spanning the current and adjacent chapters to extract the **Narrative Background** $b_j^{(i)}$:

$$b_j^{(i)} = \mathcal{M}_{\text{context}}(p_j^{(i)}, \mathcal{C}_j^{(i)}) \quad (2)$$

The Narrative Background $b_j^{(i)}$ contains the information necessary for poetry composition, which mainly consists of the following three parts:

- **Author Persona:** the character’s personality traits, social status, literary style, and emotional state at the moment of composition.
- **Background:** the broader narrative context, including ongoing events, interpersonal dynamics, and the social atmosphere surrounding the poem’s occasion.
- **Scene:** the immediate situational setting that precipitates the poem.

Stage 3: Benchmark Compilation. The complete benchmark is $\mathcal{D}_{\text{Honglou}} = \{(p, b) \mid p \in \mathcal{P}\}$, containing $|\mathcal{D}_{\text{Honglou}}| = 160$ entries. These 160 entries comprise *all* classical poems remaining after rule-based filtering of the novel—not an arbitrary subsample—and each carries a manually costly narrative context (persona, background, and scene). Compared to existing datasets, Honglou-Poem uniquely requires models to generate poetry *in character*—respecting the persona’s literary style, emotional state, and narrative context simultaneously. This setting is substantially more challenging than keyword-conditioned generation, as

models must integrate diverse contextual signals into coherent poetic expression.

2.2 Digital Poetry Party: Multi-Agent Generation Pipeline

Inspired by the literary gatherings depicted in *Dream of the Red Chamber*, we design a “Digital Poetry Party” where multiple LLMs act as competing poets, each assuming the same character persona and narrative context. For each entry $(p, b) \in \mathcal{D}_{\text{Honglou}}$, each model $\mathcal{L}_k$ generates a poem $\hat{p}_k = \mathcal{L}_k(\pi(b))$ conditioned on identical contextual information, where $\pi(\cdot)$ denotes the structured generation prompt.

This design enables controlled comparison: all models receive exactly the same rich contextual input, isolating differences in their poetry composition capabilities. The generation prompts instruct models to: (1) adopt the specified persona, (2) compose poetry fitting the narrative occasion, and (3) adhere to the formal constraints of the target poetic form (regulated verse, quatrain, or *ci* lyrics). Detailed prompts are provided in Appendix D.

We evaluate 12 state-of-the-art LLMs spanning both closed-source (GLM-4.6, Kimi-k2, Gemini-2.5-Pro, o3, GPT-5.2, etc.) and open-source categories. Each model generates poems for all 160 benchmark entries, producing a corpus of 1,920 poems for comparative evaluation. These 1,920 outputs are used *only* as generation-leaderboard targets; they are never used to construct Poetry-Tree, whose roots are the canonical novel poems (Section 3.1). The generation results are presented in Section 4.2.

3 Poetry Evaluation

3.1 Poetry-Tree: Dataset Construction

A fundamental challenge in developing poetry evaluation metrics is establishing validity—no gold-standard benchmark exists for poetry with axiomatically defined quality levels. We introduce **Poetry-Tree**, which fills this gap through controlled degradation.

Root Poems and Expert Critiques. Each root poem p_0 is a canonical poem from *Dream of the Red Chamber*, paired with a corresponding expert critique e_0 from Cai Yijiang’s *Essential Appreciation of the Poetry of Dream of the Red Chamber* (757 characters on average). These critiques analyze concrete literary merits such as imagery, diction, emotion, and artistic conception. We write e_0

Algorithm 1 POETRYTREE: Recursive Merit Degradation with Axiomatic Ordering

Require: Root poem p_0 with expert critique e_0 ; max depth $d_{\max}$; degradation budget $n_{\text{deg}}=5$;

Ensure: A tree $\mathcal{T}$ and a set of pairwise orderings $\mathcal{O}$ such that parent $\succ$ child.

```

1: function BUILDTREE( $p, M_0, d, R$ )
2:   if  $d \geq d_{\max}$  then
3:     return NULL
4:    $M \leftarrow M_0 \setminus R$   $\triangleright$  Exclude already-removed merits
     from candidates
5:   if  $|M| = 0$  then
6:     return NULL
7:    $v \leftarrow \text{CREATE\_NODE}(p, d)$ 
8:    $M_{\text{sample}} \leftarrow \text{RANDOM\_SAMPLE}(M, \min(n_{\text{deg}}, |M|))$ 
9:   for all  $m \in M_{\text{sample}}$  do
10:     $p' \leftarrow \text{DEGRADE\_POEM}(p, m)$   $\triangleright$  LLM removes
        merit  $m$  with minimal edits ( $\leq 15\%$ )
11:     $v' \leftarrow \text{BUILDTREE}(p', M_0, d+1, R \cup \{m\})$ 
12:    if  $v' \neq \text{NULL}$  then
13:       $\text{ADDCHILD}(v, v')$ 
14:       $\mathcal{O} \leftarrow \mathcal{O} \cup \{(p \succ p')\}$ 
15:   return  $v$ 
16:  $\mathcal{O} \leftarrow \emptyset$ 
17:  $M_0 \leftarrow \text{EXTRACT\_MERITS}(p_0, e_0)$   $\triangleright$  LLM extracts
    merits from expert critique  $e_0$  once
18:  $\mathcal{T} \leftarrow \text{BUILDTREE}(p_0, M_0, 0, \emptyset)$ 
19: return  $(\mathcal{T}, \mathcal{O})$ 

```

rather than c_0 to avoid collision with the narrative context window $\mathcal{C}$ in Section 2.1.

Merit m . A merit m is a specific literary quality criterion identified from expert analysis. Merits are grounded in e_0 , not generated arbitrarily. Degradation is criterion-specific: to degrade emotion, the LLM is instructed to weaken emotional expression while preserving other qualities; to degrade rhythm, it modifies tonal or rhyming patterns. Concrete prompts appear in Appendix Figures 6–7.

Construction. Algorithm 1 describes the recursive construction. Starting from a high-quality root poem p_0 with its expert critique e_0 , we first extract the full merit set $M_0 = \text{EXTRACTMERITS}(p_0, e_0)$ *once* using an LLM grounded in the expert critique. We then build a tree $\mathcal{T}$ where each parent–child edge corresponds to a controlled decrease in literary quality. At each recursive node, we reuse the fixed M_0 , subtract the already-removed merit set R to obtain the remaining candidates $M = M_0 \setminus R$, and sample up to $n_{\text{deg}} = 5$ merits from M . For each sampled merit m , we generate a degraded poem p' by removing *exactly one* merit m while requiring no more than $\sim 15\%$ character difference. The recursive construction produces 9,468 root-to-leaf paths in total. A concrete example of a Poetry-

Tree, including the character-level degradation that breaks a rhyme scheme, is provided in Appendix C.

Axiomatic Quality Ordering. Quality is not independently scored by humans at each construction step. It is *axiomatically defined by construction*: removing one literary merit while minimally editing the text produces a strictly lower-quality variant. This yields objective pairwise orderings without expensive per-poem human ratings at scale. The tree induces a partial order: for every parent–child pair (p, p') , $p \succ p'$ by construction. Collecting root-to-leaf paths yields ordered groups; our final Poetry-Tree contains 9,468 such groups.

Human Expert Validation. To verify that merit removal produces *perceptually* meaningful quality differences, we conducted a literary expert study. We sampled 50 Poetry-Tree groups, randomly selected a parent–child pair from each group, and asked two graduate students specializing in classical Chinese literature which poem has higher literary quality. Experts prefer the parent poem **86%** of the time (Cohen’s $\kappa=0.70$ with the Poetry-Tree ordering; inter-annotator $\kappa=0.722$). Disagreements concentrate on artistic conception and diction—the higher-order dimensions where trained evaluators score highest (Table 4). Human experts find subtle high-level degradations harder to articulate, whereas models trained on expert critiques capture these dimensions more reliably.

Data Usage and Splits. Poetry-Tree is built from canonical novel poems with expert critiques, *not* from the 1,920 LLM-generated poems. The 8,468 / 1,000 split has no overlap: Poet-4B is trained only on the training partition and validated only on the held-out test set. Table 6 in Appendix A lists each resource and its role.

3.2 Expert-Persona Evaluation via ICL

To enable LLM-based evaluation that captures the nuanced aesthetic qualities of poetry, we propose an expert-persona approach grounded in professional literary criticism.

Literary Expert Critic Dataset. We curate a dataset $\mathcal{E} = \{(p_i^{\text{ref}}, e_i)\}_{i=1}^K$ of expert poetry critiques from *Dream of the Red Chamber*. Each entry pairs a poem p_i^{ref} with a detailed critique e_i by renowned literary scholars, covering merits, flaws, imagery, emotional expression, and technical craftsmanship. The average critique length is 757 characters.

In-Context Learning. Given candidate poems $\mathcal{G} = \{\hat{p}_1, \dots, \hat{p}_M\}$, we construct prompt $\Psi(\mathcal{G}, \mathcal{E}_{\text{sample}})$ by sampling $n=5$ expert critiques as demonstrations. We implement **Batch Evaluation** where the judge receives all poems simultaneously for comparative ranking $\mathcal{R} = \mathcal{J}(\Psi(\mathcal{G}, \mathcal{E}_{\text{sample}}))$, with poems in randomized order to mitigate position bias. The expert-persona prompt (Appendix Figure 8) instructs the judge to compare candidates across the eight Poetry-Tree dimensions simultaneously—rhythm, emotion, imagery, structure, diction, implication, rhetoric, and artistic conception—rather than a single surface metric.

3.3 GRPO-Based Poetry Evaluation Model

While the expert-persona ICL approach achieves strong results with powerful closed-source models (e.g., GLM-4.6: $\rho=0.8764$), open-source models perform poorly in this setting (Table 1). This motivates training dedicated evaluation models.

Motivation. Our experiments reveal a striking evaluation gap: most open-source models achieve near-zero or even negative Spearman’s correlation on Poetry-Tree (e.g., base Qwen3-4B: 0.2929, LLaMA-3.1-70B: -0.0804). This indicates that general-purpose LLMs fundamentally lack the ability to discriminate poetry quality—a specialized capability that must be explicitly trained.

Training Data Construction. We leverage the training partition of Poetry-Tree (8,468 groups) as training data. Each sample consists of a group of 5 poems at different degradation depths sharing the same narrative context. The input presents these poems (in randomized order) with instructions to produce a quality ranking. The ground-truth ranking is determined by the axiomatic depth ordering (lower depth = higher quality).

Reward Function. We use Spearman’s rank correlation coefficient ρ between the model’s predicted ranking and the ground-truth depth ordering as the reward signal:

$$r = \rho(\mathbf{r}_{\text{pred}}, \mathbf{r}_{\text{gold}}) = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (3)$$

where d_i is the difference between predicted and gold ranks for poem i , and $n=5$ is the group size. This reward ranges from -1 (perfectly inverted) to $+1$ (perfect match), directly optimizing the metric we care about.

GRPO Training. We employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to train the evaluation model. For each training sample x , GRPO samples a group of candidate responses $\{y_1, \dots, y_G\}$ from the current policy π_θ , computes the reward r_i for each response y_i using the Spearman-based reward defined above, and updates the policy to increase the probability of higher-reward responses. The training objective is:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E} \left[\sum_{i=1}^G A_i \log \pi_\theta(y_i | x) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right] \quad (4)$$

where β controls the KL penalty against the reference policy π_{ref} , and A_i is the group-relative advantage derived from the reward r_i :

$$A_i = \frac{r_i - \text{mean}(\{r_1, \dots, r_G\})}{\text{std}(\{r_1, \dots, r_G\})} \quad (5)$$

Intuitively, responses with above-average Spearman’s ρ (i.e., better poetry rankings) receive positive advantages and are reinforced, while responses with below-average correlation are suppressed. This directly connects our Spearman reward r to the policy gradient through the advantage A_i .

We train **Qwen3-4B-Instruct** as the base model, releasing the resulting model as **Poet-4B**, to demonstrate that even a compact model can achieve strong evaluation performance with targeted training.

4 Experiments

4.1 Poetry Evaluation Results

4.1.1 Experimental Setup

Test Set. We randomly sample 1,000 groups from the full set of 9,468 groups in the Poetry-Tree benchmark as the held-out test set, stratified by root poem to ensure diversity across different source poems and poetic forms. Each group contains 5 poems at different degradation depths sharing the same narrative context.

Evaluation Models. We evaluate a comprehensive set of closed-source models (GLM-4.6, DeepSeek-R1, Gemini-2.5-Pro, Kimi-k2), open-source models of varying scales (Qwen2.5-72B/14B/7B, Qwen3-8B/4B, GLM-4-9B, Mistral-7B, LLaMA-3.1-70B), and our GRPO-trained model **Poet-4B**. For all models, we apply the expert-persona ICL approach (batch mode, $n=5$ expert demonstrations) with randomized candidate order to mitigate position bias.

Table 1: Poetry evaluation results on Poetry-Tree benchmark (Spearman’s ρ). Our GRPO-trained models dramatically outperform their base versions and closely approach the best closed-source models.

Category	Model	Spearman’s ρ
Closed-Source	GLM-4.6	0.8764
	DeepSeek-R1	0.8260
	Gemini-2.5-Pro	0.6519
	Kimi-k2	0.6232
Open-Source	Qwen2.5-72B-Instruct	0.5282
	Qwen2.5-14B-Instruct	0.5118
	Qwen3-8B	0.4094
	Qwen3-4B-Instruct	0.2929
	GLM-4-9B-Chat	0.0843
	Qwen2.5-7B-Instruct	0.0499
	Mistral-7B-Instruct	−0.0174
	LLaMA-3.1-70B-Instruct	−0.0804
Ours	Poet-4B	0.8253

Metrics. We report (i) **Spearman’s ρ** between predicted ranks and gold depth-based ranks, which captures how well the judge preserves the monotonic quality trend even when minor swaps occur, and (ii) **Accuracy**, defined as the fraction of groups for which the predicted ranking exactly matches the ground-truth depth order.

4.1.2 Main Results

Table 1 presents the main results. Several key findings emerge:

Open-source models largely fail at poetry evaluation. Most open-source models achieve Spearman’s ρ below 0.5, with several showing near-zero or negative correlations. LLaMA-3.1-70B and Mistral-7B perform worse than random, indicating that general-purpose LLMs fundamentally lack poetry evaluation ability.

GRPO training dramatically improves a compact model to near closed-source performance. Poet-4B improves from the base Qwen3-4B’s 0.2929 to **0.8253** (a $2.8\times$ improvement), approaching GLM-4.6 (0.8764) within a gap of only 0.051 in Spearman’s ρ while using over $100\times$ fewer parameters. This demonstrates that targeted RL training with Poetry-Tree as a training signal can effectively distill poetry evaluation expertise into a compact, open-source model.

4.1.3 Ablation: Evaluation Strategy and Expert Knowledge

Table 2 presents the ablation results on GLM-4.6 with additional models in Appendix B.2.

Batch evaluation significantly outperforms

separate evaluation. Batch mode achieves 81.64% accuracy vs. 47.60% for separate mode, confirming that comparative assessment is far more reliable than isolated judgment. The depth-wise rankings further support this: batch-with-expert achieves the strongest discrimination (D1: 1.26 vs. D5: 4.73).

Expert knowledge substantially improves batch evaluation. Adding expert persona boosts batch accuracy from 70.40% to 81.64% and Spearman’s ρ from 0.7575 to 0.8764. Interestingly, expert persona provides minimal benefit in separate mode, suggesting that expert persona is most effective when enabling direct comparative reasoning.

4.1.4 Ablation: Effect of Expert Demonstrations on GRPO Training

Table 3 isolates a single factor. The *w/o expert* and *w/ expert* columns are both GRPO-trained; only the presence of expert demonstrations in the training prompt differs. Both GRPO variants outperform their base models (Qwen3-4B: 0.29→0.46; Qwen2.5-7B: 0.05→0.60), confirming that **GRPO is effective but bounded without external knowledge**: the *w/o-expert* variants plateau well below *w/ expert* (0.46 vs. 0.83 for Qwen3-4B). If the model had merely memorized Poetry-Tree’s degradation structure (e.g., “shorter edits = worse”), removing expert demonstrations would not cause such a large drop (0.83→0.46). The gap indicates that the model succeeds by **internalizing expert aesthetic criteria**—diction, imagery, and artistic conception—rather than surface-level degradation cues. Reward-signal feedback alone cannot instill such appreciation, highlighting the broader need to incorporate literary corpora into LLM pre-training.

4.1.5 Impact of Poem Length on Evaluation

We quantify evaluation difficulty using a per-poem score:

$$s(p) = \begin{cases} +1 & \text{if } r = r^* \\ -|r - r^*| & \text{if } r \neq r^* \end{cases} \quad (6)$$

where r^* is the true rank and r is the predicted rank. A higher average $s(p)$ indicates that the evaluator ranks poems in that group more correctly.

We further investigate whether poem length affects evaluation difficulty. Figure 2 reveals a clear positive correlation between poem length and evaluation performance: longer poems are consistently easier for the evaluator to rank correctly than shorter ones.

Table 2: Ablation study on GLM-4.6: impact of evaluation mode and expert persona on Poetry-Tree.

Mode	Expert	Accuracy	Spearman’s ρ	Average Ranking by Depth				
				D1	D2	D3	D4	D5
Batch	✗	0.7040	0.7575	1.60	2.08	2.93	3.73	4.55
Batch	✓	0.8164	0.8764	1.26	2.08	2.96	3.90	4.73
Separate	✗	0.4760	0.5794	1.98	2.35	2.83	3.56	4.21
Separate	✓	0.4752	0.5001	1.88	2.54	2.99	3.50	3.90

Table 3: Effect of expert demonstrations in GRPO training prompts (Spearman’s ρ). All three columns use the same data: *Base* is the untrained model; *w/o expert* and *w/ expert* are both GRPO-trained on training set, differing only in whether expert-critique demonstrations appear in the training prompts.

Model	Base	w/o expert	w/ expert
Qwen2.5-7B-Instruct	0.0499	0.5979	0.8162
Qwen3-4B-Instruct	0.2929	0.4579	0.8253

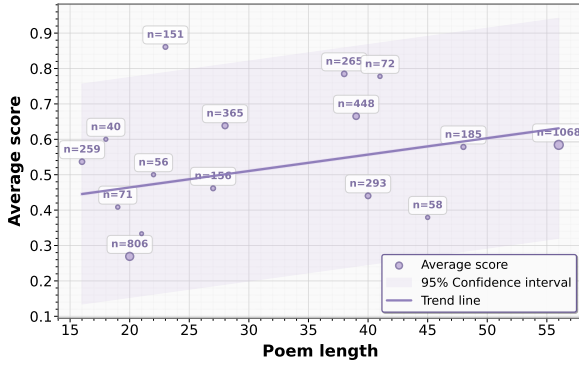


Figure 2: Evaluation performance by poem length. Each bubble represents a length group; the x -axis is poem length (number of characters), the y -axis is the average score $s(p)$, and bubble size indicates the number of poems in that group.

Shorter poems are harder to evaluate due to higher information density. In five-character quatrains (only 20 characters), each character carries dense semantic and aesthetic weight, so degrading a single merit produces subtle, diffuse changes that are difficult to detect. In contrast, longer poems (e.g., *ci* lyrics exceeding 50 characters) provide richer textual context, making merit removal produce more detectable local disruptions.

4.1.6 Merit-Specific Evaluation Difficulty

In Poetry-Tree, each child poem degrades a specific merit along one of eight dimensions, so evaluation is inherently multidimensional. We measure difficulty across these dimensions using the score $s(p)$

Table 4: Evaluation difficulty by poetic merit. Higher scores indicate that expert-grounded evaluators more reliably detect quality differences on that dimension—i.e., *expert aesthetic priority*, not absolute hardness.

Merit Type	Score
Rhythm (ya yun)	-0.2598
Emotion (qing gan)	-0.0264
Imagery (yi xiang)	0.0084
Structure (jie gou)	0.0178
Diction (zi ci)	0.1286
Implication (liu bai)	0.1337
Rhetoric (xiu ci)	0.1980
Artistic Conception (yi jing)	0.2130

defined in Eq. 6, aggregated per merit type.

Table 4 should *not* be read as claiming that rhythm is “harder to evaluate” in absolute terms. Rather, it reflects **expert aesthetic priorities**: expert-grounded evaluators weight artistic conception and diction more heavily than surface rhythm, because literary critics treat rhyme as a baseline expectation rather than a distinguishing merit.

Rhythm scores lowest (score: -0.2598) even though it is the dimension most readily identified by traditional automated metrics. Correspondingly, **higher-order dimensions—artistic conception, rhetoric, implication, and diction—receive the highest scores**. These dimensions involve mood beyond literal meaning, strategic understatement, and lexical resonance—features inaccessible to n-gram metrics. **The score distribution thus shows that the framework internalizes expert aesthetic priorities.**

This reading is corroborated by the human study in Section 3.1: overall agreement with Poetry-Tree is high (86%, $\kappa=0.70$), yet disagreements concentrate on artistic conception and diction—exactly the dimensions with the highest scores in Table 4. Deeper qualities are harder for humans to articu-

late, but are precisely what expert-trained models capture well.

4.2 Poetry Generation Results

4.2.1 Experimental Setup

Models Under Test. We evaluate 12 state-of-the-art LLMs (Hurst et al., 2024; Team et al., 2025a; Liu et al., 2024; Comanici et al., 2025; Anthropic, 2025; Team et al., 2025b; Zeng et al., 2025) spanning both Chinese-native models (GLM-4.6, Kimi-k2, Hunyuan-turbos, DeepSeek series) and internationally-oriented models (Gemini-2.5-Pro, Gemini-3-Pro, o3, GPT-5.2, GPT-4o, Claude-4.5-Sonnet, Grok-4).

Primary Judge: Poet-4B. Having validated Poet-4B on the held-out Poetry-Tree test set ($\rho=0.8253$; Table 1), we adopt it as the *sole* judge of all 1,920 generated poems. Poet-4B is a separately trained 4B evaluator that did not generate any candidate poem, so evaluator-generator overlap does not arise. We keep the expert-persona prompt ($n=5$ demonstrations) and randomize candidate order. Averaging four closed-source judges—two of which (Gemini $\rho=0.6519$, Kimi $\rho=0.6232$) are substantially less reliable than Poet-4B—would dilute rather than improve the ranking. The original ensemble, used before Poet-4B was trained, is reported in Appendix B.1; the two leaderboards are highly consistent (the same three models occupy the top three).

4.2.2 Results and Analysis

Table 5 presents the Poet-4B generation leaderboard alongside each model’s Chatbot Arena overall ranking. The overall ranking is highly consistent with the original ensemble: GLM-4.6, Gemini-2.5-Pro, and Kimi-k2 remain the top three, and Chinese-native models still occupy the lead (GLM-4.6). Two key findings continue to hold:

(1) Chinese-native models exhibit a clear advantage in poetry generation. Not every Chinese-native model excels—DeepSeek-R1 and DeepSeek-V3.1 remain near the bottom—but GLM-4.6 leads with a 27.9% Top-1 rate and Kimi-k2 stays in the top three (16.5%). This demonstrates that models with strong Chinese cultural and linguistic training possess inherent advantages in classical poetry composition, even though this advantage is not universal across all Chinese models.

(2) General reasoning performance does not predict poetry generation ability. Comparing

Table 5: LLM Leaderboard for narrative-driven poetry generation on Honglou-Poem. Rankings are produced by Poet-4B. Arena rank is included for comparison.

Model	Avg. Rank	Top-1 Rate(%)	Arena Rank
GLM-4.6	3.32	27.9	10
Gemini-2.5-Pro	3.39	27.3	1
Kimi-k2	3.86	16.5	11
GPT-5.2	4.64	5.5	8
o3	5.16	3.7	2
Gemini-3-Pro	5.63	14.5	7
Claude-4.5-Sonnet	6.33	2.5	2
Hunyuan-turbos	6.96	2.4	49
Deepseek-V3.1	9.40	0	11
Deepseek-R1	9.44	0	11
Grok-4	9.56	0	12
GPT-4o	9.89	0	2

our poetry leaderboard with the Chatbot Arena rankings reveals a striking discrepancy: models that dominate general-purpose benchmarks perform poorly on poetry generation, and vice versa. For instance, o3 (Arena #2) and Claude-4.5-Sonnet (Arena #2) rank in the middle or bottom tier for poetry, and GPT-4o (Arena #2) ranks last. Conversely, GLM-4.6 (Arena #10) and Kimi-k2 (Arena #11) remain at the top despite modest Arena positions. Gemini-2.5-Pro (Arena #1) is a partial exception—it rises to second under Poet-4B—yet the broader misalignment persists. This indicates that composing classical Chinese poetry—requiring literary conventions, persona modeling, and aesthetic expression—is a distinct capability that current general-purpose training does not adequately address.

5 Related Work

5.1 Poetry Generation

Early poetry generation relied on rule-based systems (Franky, 2013; Gervás, 2001) and neural approaches using RNNs (Wang et al., 2016) and LSTMs (Yang et al., 2018). Pre-trained language models (Vaswani et al., 2017; Radford et al., 2018) further elevated generation quality, with both general-purpose and poetry-specialized models (Tan et al., 2025; Huang and Shen, 2025; Song et al., 2025) showing strong capabilities. These support diverse control mechanisms including keyword-based, sentiment-conditioned, and

image-to-poetry synthesis (Ma et al., 2023; Yi et al., 2020; Liu et al., 2018). However, all existing generation benchmarks operate on shallow inputs—a single keyword, a topic phrase, or a visual prompt—without requiring models to integrate narrative context, author personas, or emotional states into the creative process. This stands in stark contrast to how classical poetry was actually composed.

5.2 Poetry Evaluation

Traditional metrics (BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), Jaccard (Jaccard, 1901)) are repeatedly shown inadequate for poetry (Gonçalo Oliveira, 2017). Human evaluation suffers from cost and unreliability (Sawicki et al., 2025). Recent work explores LLMs as evaluators, but existing approaches employ generic personas. Our work introduces expert-persona emulation and Poetry-Tree—the first benchmark with axiomatically ordered quality levels for objectively validating poetry evaluation methods.

6 Conclusion

In this paper, we presented a comprehensive framework addressing both generation and evaluation of classical Chinese poetry. On the generation side, **Honglou-Poem** is the first narrative-driven poetry generation benchmark, requiring models to compose poetry grounded in rich author personas and narrative contexts—a substantially more challenging setting than existing keyword-based paradigms. Our generation leaderboard across 12 LLMs, now judged by the validated Poet-4B evaluator, reveals that Chinese-native models with strong cultural understanding significantly outperform reasoning-oriented models. On the evaluation side, **Poetry-Tree** provides the first benchmark with axiomatically ordered quality levels for objectively validating poetry evaluators. Leveraging Poetry-Tree as a training signal, our **GRPO-based training method** produces **Poet-4B**, our dedicated poetry evaluation model, achieving Spearman’s $\rho=0.8253$ —closely approaching the best closed-source model (0.8764, gap of 0.051) and dramatically improving upon the base model’s 0.2929.

Our experiments reveal two key insights: (1) general-purpose LLMs largely lack poetry evaluation ability, but targeted reinforcement learning can bridge this gap—Poet-4B improves from 0.2929 to 0.8253; and (2) shorter poems are substantially harder to evaluate than longer ones due to higher

information density. We release all benchmarks, trained models, and code to facilitate future research in automated literary evaluation.

7 Limitations

Our work takes a step toward rigorous, scalable evaluation of classical Chinese poetry generation and introduces the first narrative-driven generation benchmark grounded in rich literary context. Nevertheless, due to the inherently subjective nature of literary works, several limitations remain.

Instantiation on a single literary source. Honglou-Poem and Poetry-Tree are currently built from *Dream of the Red Chamber*. The broader contribution is a complete pipeline—data synthesis, inference, and training—that applies to *any* literary work embedding poetry in narrative context, including *Flowers in the Mirror*, *Eugene Onegin*, and *The Tale of Genji*. Each source needs only (1) poems in narrative, (2) expert critiques, and (3) LLM-guided merit extraction with controlled degradation. The 160 Honglou-Poem entries are all filtered poems in the novel, not a subsample; Poetry-Tree already provides 9,468 ordered groups for training. We leave end-to-end replication on additional works to subsequent research.

Training-method comparisons. We train Poet-4B with GRPO and a Spearman- ρ reward because the task is inherently ranking-based (batch comparison of five poems). Comparisons with supervised fine-tuning and preference methods such as DPO would further situate this choice, but they require substantial additional compute. We will incorporate richer training methods and baselines in subsequent work.

Acknowledgments

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123003), and the National Natural Science Foundation of China Enterprise Innovation and Development Joint Fund (Artificial Intelligence Field) Key Support Projects (U25B2072).

References

Anthropic. 2025. System card: Claude opus 4 & claude sonnet 4. <https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf>.

- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Franky. 2013. [A rule-based approach for karmina generation](#). In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 24–31, Atlanta, Georgia. Association for Computational Linguistics.
- Pablo Gervás. 2001. An expert system for the composition of formal spanish poetry. *Knowledge-Based Systems*, 14(3-4):181–188.
- Hugo Gonalo Oliveira. 2017. [A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Chihan Huang and Xiaobo Shen. 2025. [PoemBERT: A dynamic masking content and ratio based semantic language model for Chinese poem generation](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 50–60, Abu Dhabi, UAE. Association for Computational Linguistics.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Dayiheng Liu, Quan Guo, Wubo Li, and Jiancheng Lv. 2018. [A multi-modal chinese poetry generation model](#). In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Jingkun Ma, Runzhe Zhan, and Derek F. Wong. 2023. [Yu sheng: Human-in-loop classical Chinese poetry generation system](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 57–66, Dubrovnik, Croatia. Association for Computational Linguistics.
- 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.
- Piotr Sawicki, Marek Grześ, Dan Brown, and Fabrício Góes. 2025. [Can large language models outperform non-experts in poetry evaluation? a comparative study using the consensual assessment technique](#). *Preprint*, arXiv:2502.19064.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Xinglong Song, Changlin Song, Haolu Yu, Yonghua Zhu, and Hong Yao. 2025. [Mixsong: Diverse and strictly formatted chinese poetry generation](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(4).
- Tao Tan, Quan Tu, Songhao Wu, Hongda Sun, Chuanqi Cheng, Weikai Xu, and Rui Yan. 2025. [Luoling: An immersive cross-modal interactive poem creation factory through multi-agent collaboration](#). In *Companion Proceedings of the ACM on Web Conference 2025, WWW ’25*, page 2911–2914, New York, NY, USA. Association for Computing Machinery.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025a. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.
- Tencent Hunyuan Team, Ao Liu, Botong Zhou, Can Xu, Chayse Zhou, ChenChen Zhang, Chengcheng Xu, Chenhao Wang, Decheng Wu, Dengpeng Wu, and 1 others. 2025b. Hunyuan-turbos: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. *arXiv preprint arXiv:2505.15431*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. [Chinese poetry generation with planning based neural network](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1051–1060, Osaka, Japan. The COLING 2016 Organizing Committee.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. [Stylistic Chinese poetry generation via unsupervised style disentanglement](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3960–3969, Brussels, Belgium. Association for Computational Linguistics.

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9450–9457.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Table 6: Data resources and usage. Poetry-Tree is *not* constructed from the 1,920 generated poems.

Resource	Size	Role
Honglou-Poem	160 entries	Generation inputs
Generated poems	1,920	Leaderboard targets only
Expert critiques	~160	Merit extraction for roots
Poetry-Tree train	8,468 groups	Poet-4B GRPO training
Poetry-Tree test	1,000 groups	Evaluator validation

Technical Appendix

A Implementation Details

Training Details. Poet-4B is trained on $8 \times H20$ GPUs using the VERL training platform. We employ GRPO as the training algorithm with a batch size of 512, a maximum context length of 4,096 tokens, and train for 4 epochs. The reward function is Spearman’s rank correlation coefficient ρ between the model’s predicted ranking and the ground-truth depth ordering, as described in Section 3.3.

Expert Critique Source. We extract expert critiques e_0 from the literary commentary collection by Cai Yijiang,¹ and use these critiques as the foundation for constructing the Poetry-Tree benchmark. Each critique provides detailed analysis of a poem’s literary merits, covering imagery, emotional expression, diction, and artistic conception, which grounds the merit extraction step in Algorithm 1. All datasets are included in the supplementary materials and will be open-sourced in a subsequent release.

Data Resources and Usage. Table 6 clarifies that Poetry-Tree is constructed from canonical novel poems, not from the 1,920 LLM-generated outputs, and that the train/test split has no overlap.

B Detailed Experiment Results

B.1 Results on Poem Generation

¹Cai, Yijiang. 红楼梦诗词精华鉴赏. Changjiang Literature & Art Press, 2021.

The tables below are the *original* generation rankings under the four-judge ensemble (Kimi-k2, GLM-4.6, Gemini-2.5-Pro, DeepSeek-R1), included for comparison with the Poet-4B leaderboard in Table 5. The top three models are the same under both protocols. Lower average ranks indicate better performance.

Table 7: LLM Leaderboard ranked by evaluator Gemini-2.5-Pro. Lower Average Rank is better.

Model	Average Rank
Kimi-k2	3.18
GLM-4.6	4.15
Gemini-2.5-Pro	4.34
GPT-5.2	4.35
o3	4.55
Gemini-3-Pro	5.50
Claude-4.5-Sonnet	6.41
Hunyuan-turbos	6.70
Grok-4	9.53
DeepSeek-V3.1	9.76
GPT-4o	9.76
DeepSeek-R1	9.79

Table 8: LLM Leaderboard ranked by evaluator GLM-4.6. Lower Average Rank is better.

Model	Average Rank
Kimi-k2	3.41
o3	4.17
GLM-4.6	4.60
GPT-5.2	4.67
Gemini-2.5-Pro	4.96
Gemini-3-Pro	5.62
Claude-4.5-Sonnet	6.45
Hunyuan-turbos	6.73
GPT-4o	8.99
DeepSeek-V3.1	9.36
Grok-4	9.40
DeepSeek-R1	9.45

B.2 Validation on the Poetry-Tree Benchmark

Tables 11–13 present the ablation study results for Kimi-k2, Gemini, and DeepSeek-R1 on the Poetry-Tree benchmark, respectively. Among the three models, DeepSeek-R1 demonstrates the strongest capability in poetry quality assessment, achieving 70.12% accuracy and 0.8260 Spearman’s correlation in the batch-with-expert setting. Across all

Table 9: LLM Leaderboard ranked by evaluator Kimi-k2. Lower Average Rank is better.

Model	Average Rank
Gemini-2.5-Pro	3.7
o3	3.7
Kimi-k2	4.0
GPT-5.2	4.9
GLM-4.6	5.6
Hunyuan-turbos	5.95
Claude-4.5-Sonnet	6.25
Gemini-3-Pro	8.6
DeepSeek-V3.1	8.65
GPT-4o	8.75
DeepSeek-R1	8.85
Grok-4	9.05

Table 10: LLM Leaderboard ranked by evaluator DeepSeek-R1. Lower Average Rank is better.

Model	Average Rank
Kimi-k2	3.69
o3	4.08
GLM-4.6	4.53
Gemini-2.5-Pro	4.68
GPT-5.2	4.74
Claude-4.5-Sonnet	6.43
Gemini-3-Pro	6.59
Hunyuan-turbos	6.59
Grok-4	8.41
GPT-4o	9.05
DeepSeek-R1	9.44
DeepSeek-V3.1	9.77

models, batch evaluation consistently outperforms separate evaluation, confirming that comparative assessment is more reliable than isolated judgment. However, a notable exception occurs in the separate evaluation setting for both Kimi-k2 and DeepSeek-R1, where adding expert persona leads to performance degradation. This suggests that when evaluating poems individually, the expert persona may cause the model to focus excessively on mimicking expert-like critique styles rather than objectively discriminating poetry quality, resulting in inconsistent and unreliable judgments.

C Examples

Figure 3 presents a concrete example of a Poetry-Tree. Each node in the tree represents a poem,

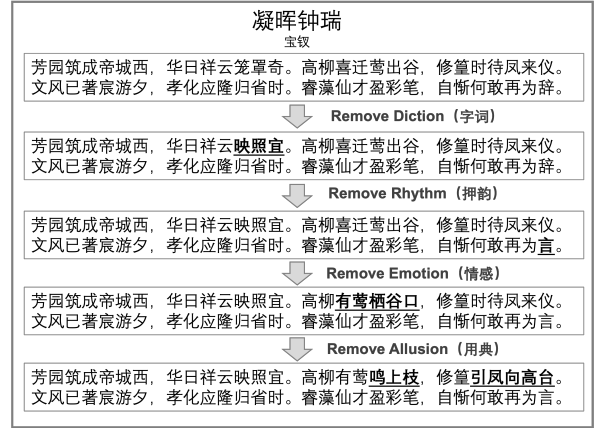


Figure 3: A concrete example of a Poetry-Tree. Starting from the root poem (top), each child node is derived from its parent by removing a specific poetic merit through minimal character-level edits ($\leq 15\%$ character difference), resulting in a strictly lower-quality poem. Each edge is annotated with the degraded merit, yielding an axiomatically ordered quality hierarchy from root to leaves.

and each parent–child edge corresponds to a single targeted degradation: one specific poetic merit is removed via minimal character-level edits while all other merits are held fixed. For instance, in the transition from the second to the third layer, the character “辞” (*cí*) at the end of a line is replaced by “言” (*yán*), which weakens the rhythmic quality of the verse while leaving its imagery, diction, and other literary dimensions intact.

To see why this single substitution disrupts rhythm, note that the line-ending characters of the other lines in the parent poem are “宜” (*yí*), “仪” (*yí*), and “时” (*shí*)—all sharing the *-i* vowel rhyme with “辞” (*cí*). Replacing “辞” with “言” (*yán*) breaks this rhyme scheme, as *-an* does not rhyme with *-i*, thereby degrading the poem’s rhythmic coherence. This targeted substitution strategy ensures that every child poem is strictly inferior to its parent in exactly one dimension—guaranteeing the axiomatic quality ordering that makes Poetry-Tree a reliable ground-truth benchmark for evaluator validation.

D Prompts

Figures 4–8 present the three types of prompts used in our framework. The background extraction prompt (Figure 4) summarizes narrative context from the source text, including key characters, events, and emotional states. The generation prompt (Figure 5) instantiates a Poetry Party set-

Table 11: Ablation study results on Kimi-k2. The model shows moderate performance with expert persona improving batch evaluation but struggling in separate mode.

Mode	Expert	Accuracy	Spearman’s ρ	Average Ranking by Depth				
				D1	D2	D3	D4	D5
Batch	✗	0.3769	0.3853	2.51	2.39	2.74	3.37	3.96
Batch	✓	0.5285	0.6232	1.87	2.26	2.91	3.66	4.30
Separate	✗	0.2880	0.2370	2.90	2.57	2.93	3.18	3.71
Separate	✓	0.3499	0.4054	2.11	2.72	2.95	3.36	3.80

Table 12: Ablation study results on Gemini 2.5 pro. The model demonstrates balanced performance with expert persona providing consistent improvements in both evaluation modes.

Mode	Expert	Accuracy	Spearman’s ρ	Average Ranking by Depth				
				D1	D2	D3	D4	D5
Batch	✗	0.4975	0.5188	2.15	2.26	2.78	3.55	4.11
Batch	✓	0.5497	0.6519	1.73	2.26	2.93	3.72	4.25
Separate	✗	0.4233	0.5321	2.18	2.40	2.93	3.57	4.15
Separate	✓	0.4570	0.4965	1.95	2.64	3.07	3.58	3.88

Table 13: Ablation study results on DeepSeek-R1. The model achieves strong performance with expert persona in batch mode, showing the largest improvement among all tested models.

Mode	Expert	Accuracy	Spearman’s ρ	Average Ranking by Depth				
				D1	D2	D3	D4	D5
Batch	✗	0.5824	0.6022	1.83	2.30	2.94	3.69	4.16
Batch	✓	0.7012	0.8260	1.32	2.16	3.02	3.89	4.57
Separate	✗	0.4776	0.5930	1.65	2.39	3.10	3.60	4.01
Separate	✓	0.3914	0.4036	2.10	2.68	3.13	3.41	3.71

ting, in which the model generates classical Chinese poetry grounded in the extracted background. The expert-persona prompt (Figure 8) instructs the model to act as a literary expert and perform comparative evaluation based on explicit aesthetic criteria.

Prompt:
你是《红楼梦》文献整理师，请从下面的章节中提取诗作的背景。
【章节】
【标题】
【诗歌原文】
【正文片段】
输出包括：
作者分析：这是一个字符串，请深入分析作者在作此诗时的状态。
作诗背景：这是一个字符串，请详细描述诗歌创作的直接背景。
输出：

Figure 4: Background extraction prompt for summarizing narrative context.

Prompt:
你正在参与一场古风诗会，多位文人雅客围绕同一题目轮流作诗，互相交流切磋，尽可能做出优秀的诗，取得诗会胜利。
请依据给定背景写一首原创古体诗（五言/七言/词皆可），题目可自拟或改写。
须兼顾意境与格律，可适度创新。
重要：必须返回 JSON 格式，格式为：
{ "title": "题目", "poem": "诗歌正文", "notes": "创作思路与手法" }
不要附加任何额外解释文字。
【诗会题旨】
【作诗背景 / 动机】
【作者信息】

Figure 5: Poetry Party generation prompt conditioned on narrative background.

Prompt:
你是一位资深的古典诗词鉴赏专家。请仔细阅读诗词及其解析，找出这首诗词的所有妙处。
妙处包括但不限于：意象妙用、修辞手法、音韵和谐、意境营造、情感表达、结构安排、字词锤炼、典故运用、巧妙留白
请以JSON字典格式返回，格式如下：
{
 "妙处1": "详细的解释说明",
 "妙处2": "详细的解释说明",
 ...
}
要求：
1. 必须包含所有发现的妙处
2. 每个妙处都要有简洁的简称和详细的解释
3. 从专家的【内容】、【说明】、【注释】、【评解】中找出妙处
4. 妙处的解释要具体，结合诗词中具体的词和字进行说明
.....

Figure 6: Merit extraction prompt for identifying literary merits from a poem grounded in expert critique evidence.

Prompt:

你是一位优秀的古典诗词创作专家。请根据要求重新创作诗词，要求：

1. 保持原诗的主题和基本意境
2. 保留指定的优势（妙处）
3. 去除指定的优势（妙处），使诗词在该方面有所削弱（尽可能只通过一两个字眼的修改实现削弱，以确保诗歌保持其他优势）
4. 保持诗词的完整性和可读性
5. 保持相同的体裁和格律
6. 尽量保持与原诗相同的时代背景、人物关系和叙事视角

格式示例：

`{"poem": "新创作的诗词内容","explanation": "详细的str解释，不要输出字典等其他格式"}`

【诗歌】

【需要保留的优势与去除的优势】

【写作背景与人物介绍】

Figure 7: Degradation prompt for generating a child poem node in Poetry-Tree.

Prompt:

你是金标准参考样本中那位专家。请先深入研读以下 5 个金标准评点案例，全面学习并内化专家的评审风格、评审结构、评审思路、语言表达、专业术语、分析角度、论述逻辑等所有特征，然后以完全相同的专家身份和口吻对下列所有作品一次性给出“专家点评”和“整体排名”。

核心要求：

1. **身份认同：**你就是那位专家本人，不是模仿者。你的每一句话都应该让人以为是专家亲笔所写。
2. **全面学习：**仔细分析参考样本中的以下方面：

语言风格：用词习惯、句式特点、修辞手法、专业术语的使用

评审结构：如何开头、如何展开、如何收尾、段落组织方式

分析角度：从哪些维度切入（思想内容、艺术手法、格律技巧、意象营造、情感表达等）

论述逻辑：论证思路、观点展开方式、例证引用方式

专业表述：古典诗词评论的专业术语、典故引用、比较分析等

3. **评审原则：**不要机械罗列维度，要像专家那样自然流畅地组织论述，使用与专家一致的专业术语和表达方式，保持与专家相同的学术深度和鉴赏水准

【待评审作品】

【统一背景】

【参考样本】

Figure 8: Expert-persona prompt for comparative poetry evaluation.