PoeTone: A Framework for Constrained Generation of Structured Chinese Songci with LLMs

Zhan Qu, Shuzhou Yuan, Michael Färber
TU Dresden
ScaDS.AI
Dresden, Saxony, Germany
zhan.qu@tu-dresden.de

Abstract

This paper presents a systematic investigation into the constrained generation capabilities of large language models (LLMs) in producing *Songci*, a classical Chinese poetry form characterized by strict structural, tonal, and rhyme constraints defined by Cipai templates. We first develop a comprehensive, multi-faceted evaluation framework that includes: (i) a formal conformity score, (ii) automated quality assessment using LLMs, (iii) human evaluation, and (iv) classification-based probing tasks. Using this framework, we evaluate the generative performance of **18 LLMs**, including 3 proprietary models and 15 open-source models across 4 families, under **five prompting strategies**: zero-shot, one-shot, completion-based, instruction-based, and chain-of-thought. Finally, we propose a Generate-Critic architecture in which the evaluation framework functions as an automated critic. Leveraging the critic's feedback as a scoring function for best-of-N selection, we fine-tune 3 lightweight open-source LLMs via supervised fine-tuning (SFT), resulting in improvements of up to **5.88%** in formal conformity. Our findings offer new insights into the generative strengths and limitations of LLMs in producing culturally significant and formally constrained literary texts.

Code — https://github.com/ZhanQu945/PoeTone

Introduction

Recent advancements in large language models (LLMs) have demonstrated remarkable capabilities in generating fluent, coherent, and contextually appropriate text across diverse domains (Brown et al. 2020; Touvron et al. 2023). From drafting emails and summarizing documents to composing fiction and poetry, models such as `GPT-4o`, `Gemini 2.5 Pro`, and `DeepSeek-R1` have significantly expanded the creative potential of AI systems (OpenAI et al. 2024; Comanici et al. 2025; DeepSeek-AI 2025). Yet, tasks requiring both expressive fluency and strict adherence to formal rules, such as generating classical poetry, remain a significant challenge (Yu et al. 2024).

This paper focuses on *Songci*, a prominent form of Chinese lyric poetry from the Song dynasty (960–1279 CE). Unlike Tang poetry, *Songci* follows fixed *Cipai*, predefined tune patterns that govern stanza count, line count, line length, tonal flow (píng and zè), rhyme type, and rhyme positions (Ge 2009). Effective *Songci* must not only conform to these structural constraints but also express coherent and emotionally resonant themes, typically involving longing, solitude, patriotism, or nature. This blend of metrical rigidity and artistic depth makes *Songci* a compelling testbed for constrained text generation. A common failure case in generating *Songci* is shown in Figure 1.

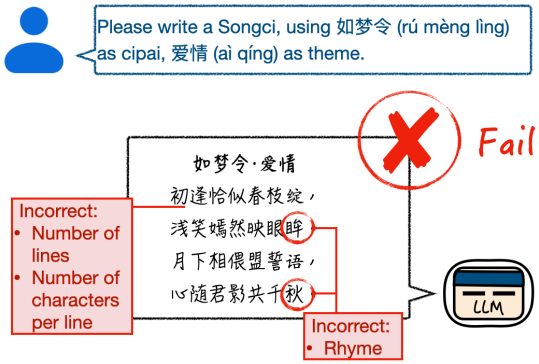

Figure 1: The multi-layered challenge of generating high-quality *Songci*, highlighting the formal constraints of the Cipai as the primary bottleneck.

While prior work has applied deep learning to classical Chinese poetry, most focus on stylistic imitation or surface-level coherence (Liao et al. 2019; Tang et al. 2025). Whether modern LLMs can generate *Songci* that satisfy its structural and tonal requirements remains largely unexplored (Song 2022). Moreover, existing benchmarks lack fine-grained tools for evaluating formal correctness, and few approaches aim to improve generation under these constraints.

To bridge these gaps, we introduce **PoeTone**, a systematic investigation into the capacity of LLMs to generate *Songci* while adhering to its strict metrical constraints. Our complete research pipeline is detailed in Figure 2:

First, we develop a comprehensive evaluation framework that includes: (i) a metadata resource detailing the structure and tonal patterns of 20 widely used *Cipai*, (ii) a curated corpus of canonical *Songci* poems organized by theme and *Cipai*, and (iii) a multi-dimensional evaluation protocol combining formal conformity scores, automated quality as-

sessment, human ratings, and classification-based probing.

Second, we benchmark 18 state-of-the-art LLMs, including 3 proprietary models (`GPT-4o`, `Gemini 2.5 Pro`, `ERNIE 4.5 Turbo`) and 15 open-source models from 4 families (`LLaMA`, `Mistral`, `Qwen`, `DeepSeek`), across five prompting strategies: zero-shot, one-shot, completion, instruction, and chain-of-thought.

Third, we propose a Generate-Critic architecture that uses rule-based conformity scoring to guide fine-tuning. Applied to 3 open-source LLMs via best-of-N rejection sampling and LoRA-based supervised fine-tuning (SFT), this method improves adherence to structural and tonal constraints by up to *5.88%*.

In summary, our key contributions are:

- A structured evaluation framework for *Songci* generation, including a metadata resource, a canonical corpus, and a multi-faceted evaluation framework.
- A benchmark of 18 LLMs under five prompting strategies, revealing model limitations in conforming to formal poetic constraints.
- A Generate-Critic method that improves constrained generation through automated rule-based feedback.

Related Work

Automatic generation of Chinese poetry has evolved from statistical and rule-based methods to neural networks (Wang et al. 2016; Chen and Cao 2024). Initial approaches used RNNs and LSTMs (Zhang and Lapata 2014; Wang et al. 2016), later enhanced by attention mechanisms to improve coherence and context modeling (Yi, Li, and Sun 2017).

A major research focus has been controlling poetic form and content. To handle strict structural and rhyming constraints, methods include dual-encoder models conditioned on rhythmic patterns (Luo et al. 2021) and form-aware generation guided by stressed weighting (Hu and Sun 2020). Thematic coherence has been addressed through planning-based structures (Wang et al. 2016), working memory models (Yi et al. 2018), salient clue guidance (Yi, Li, and Sun 2018; Gao et al. 2021), adversarial training for title consistency (Li et al. 2018), and graph neural networks for topic modeling (Yan et al. 2023). Sentiment and style control have been explored via multi-tag classification and conditional variational autoencoders (Shao et al. 2021; Chen et al. 2019; Song et al. 2025).

The rise of large language models (LLMs) has transformed Chinese poetry generation. An early study showed that a simple Generative Pre-trained Transformer (GPT) could produce high-quality classical poetry without extensive feature engineering (Liao et al. 2019). In further studies, researchers have fine-tuned pre-trained models on specialized poetry corpora, finding that smaller, domain-specific models can sometimes outperform larger, general-purpose LLMs like GPT-4o on poetic tasks (Tang et al. 2025; Wang, Guan, and Liu 2022). Innovations like CharPoet introduce token-free, character-level control to improve format accuracy (Yu et al. 2024), reflecting a trend to harness LLM capabilities while imposing classical poetic constraints.

Beyond text generation, interactive systems such as Jiuge (Zhipeng et al. 2019) and Yu Sheng (Ma, Zhan, and Wong 2023) incorporate human-in-the-loop frameworks for collaborative poem refinement. SongSong (Hu et al. 2025) generates music from *Songci* lyrics, addressing musical restoration, but leaving high-quality, metrically-correct *Songci* generation as an open challenge. Our work focuses on this core task, leveraging LLMs to generate formal *Songci* text.

Songci Generation and Evaluation Framework

Evaluating the ability of LLMs to generate valid *Songci* requires two key resources that, to our knowledge, are unavailable in structured, machine-readable form. First, a formal constraint specification for each *Cipai* is necessary for automated verification. Second, a high-quality corpus of canonical *Songci*, thematically categorized and representative of diverse *Cipai*, is essential for model prompting and evaluation. This section details the creation of both resources and introduces our multi-faceted evaluation protocol.

Cipai Formal Constraint Resource

The primary challenge in *Songci* generation lies in adhering to the strict structural and phonological rules defined by each *Cipai*. To enable automated verification of these constraints, we constructed a structured metadata resource covering the 20 most commonly used two-stanza *Cipai*, selected for their frequency in classical anthologies and representational diversity (Ge 2009). Each *Cipai* may have up to 15 different variants, and all of them are included in the metadata.

Each *Cipai* is encoded in machine-readable JSON format, capturing the following elements:

- **Structure**: Specifies the number of stanzas, total line count, and exact character count for each line (with segmentation based on punctuation, i.e., commas, periods, and question marks).
- **Tonal Pattern**: Encodes the expected tonal category, píng (level) or zè (oblique), at key positions per line, using modern Mandarin tonal approximations.
- **Rhyme Scheme**: Identifies lines and positions that must rhyme, with rhyme group constraints (píng or zè) derived from classical rhyme dictionaries such as *Cilin Zhengyun* (Ge 2009).

Thematic Canonical Corpus

We compiled a curated corpus of 120 human-authored *Songci* as exemplars for prompting and evaluation. The *Songci* were selected from authoritative digital archives based on author prominence (e.g., Su Shi, Li Qingzhao) and fidelity to canonical *Cipai* forms (Ge 2009).

To support thematic generation and classification, each *Songci* was manually annotated with one of six recurring themes in classical *Songci*:

- **Love & Longing**: Romantic and delicate emotions.
- **Patriotism & Heroism**: Courageous expressions.
- **Nature & Landscapes**: Scenery, seasons, pastoral life.
- **History & Nostalgia**: Reflection on past eras or figures.

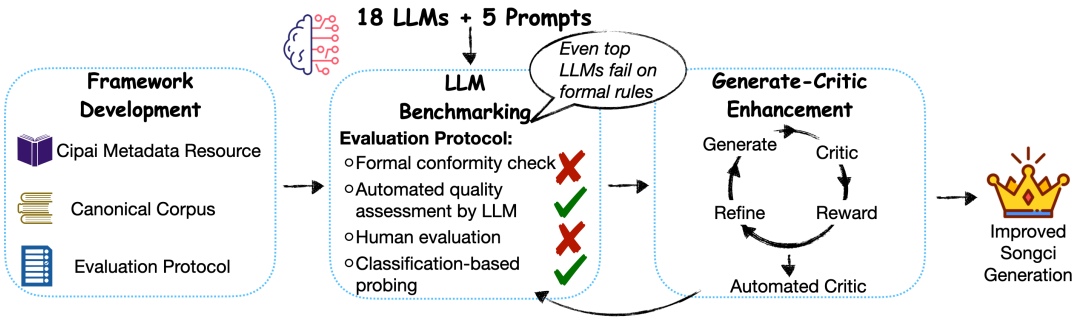

Figure 2: An overview of our research pipeline, from framework development and benchmarking to model enhancement.

- **Sorrow & Grief**: Farewell, exile, homesickness.
- **Philosophical Reflection**: Tranquil and spiritual tones.

This dataset serves as reference and training data for some of our analyses.

Multi-Faceted Evaluation Protocol

Evaluating the quality of generated *Songci* requires more than assessing fluency or content relevance. A *Songci* that reads naturally may still violate strict structural, tonal, or rhyming constraints, making it invalid under its specified *Cipai*. Conversely, a *Songci* that rigidly conforms to format may fail to achieve poetic elegance or thematic depth. To address this duality, we design a layered evaluation framework comprising both objective rule-based metrics and subjective assessments of fluency and artistic merit.

Formal Conformity Score At the core of our evaluation is the **Formal Conformity Score**, which quantifies adherence to the structural, tonal, and rhyming rules encoded in the *Cipai* metadata. The score is computed automatically using a rule-checking script that evaluates each generated *Songci* along three dimensions:

- **Structural Integrity**: It verifies that the *Songci* has the correct number of lines and that each line has the precise number of characters required by the Cipai.
- **Tonal Adherence**: Using a standard modern Mandarin Chinese pronunciation lexicon, it checks if the characters at key positions match the mandatory píng or zè tonal patterns (or zhōng means both píng and zè are allowed).
- **Rhyme Scheme Compliance**: It identifies the characters at the designated rhyming positions and verifies that they belong to the same rhyme group according to a classical Chinese rhyme dictionary (Ge 2009).

The quality of a generated *Songci* is evaluated using a normalized, weighted scoring system. The final score is determined by calculating a 'best-fit' against all known variants of a given *Cipai* pattern. Formally, the total score S_{total} for a generated *Songci* is given by:

$$S_{\text{total}} = \max_{v \in V} \left(w_S \cdot S^{(v)}_{\text{structure}} + w_T \cdot S^{(v)}_{\text{tonal}} + w_R \cdot S^{(v)}_{\text{rhyme}} \right) \tag{1}$$

Where V is the set of all known variants of the target *Cipai*, $S^{(v)}_{\text{structure}}$, $S^{(v)}_{\text{tonal}}$, and $S^{(v)}_{\text{rhyme}}$ are component scores against variant v; w_S, w_T, and w_R are non-negative weights summing to 1.0. We use $w_S = 0.4$, $w_T = 0.3$, and $w_R = 0.3$.

The **structure score**, $S_{\text{structure}}$, measures completeness and structural correctness of a generated *Songci*. Let N_g be the number of generated lines and N_v be the number of template lines. The number of correctly structured lines, N_{match}, is found by checking the character count of a generated line, $|L_{g,i}|$, against the template's required count, $C_{v,i}$, for all lines up to the minimum of the two lengths.

$$N_{\text{match}} = \sum_{i=1}^{\min(N_g, N_v)} \mathbb{I}(|L_{g,i}| = C_{v,i}) \tag{2}$$

The final score is N_{match} normalized by the maximum of the two lengths, which penalizes both missing and superfluous lines.

$$S_{\text{structure}} = \frac{N_{\text{match}}}{\max(N_g, N_v)} \tag{3}$$

The **tonal score**, S_{tonal}, measures the quality of the well-formed parts of the *Songci*. Let I_{match} be the set of line indices that are structurally correct. Let the function $\text{tone}(c)$ return the tonal category of a character c, and let $T_{v,ij}$ be the required tone for the j-th character of the i-th line in the template. The equality $\text{tone}(c_{g,ij}) = T_{v,ij}$ is also considered true if $T_{v,ij}$ is marked as 'zhōng' in the template. The score is the ratio of matching tones to the total characters in the evaluated lines.

$$S_{\text{tonal}} = \frac{\sum_{i \in I_{\text{match}}} \sum_{j=1}^{|L_{g,i}|} \mathbb{I}(\text{tone}(c_{g,ij}) = T_{v,ij})}{\sum_{i \in I_{\text{match}}} |L_{g,i}|} \tag{4}$$

The **rhyme score**, S_{rhyme}, quantifies the internal consistency of rhyme among the generated rhyming lines. Let L_{g,R_v} denote the set of final characters from the generated *Songci* corresponding to the rhyming positions specified by variant v of the target *Cipai*. Let G be the set of rhyme groups defined in the metadata, and let $\text{rhyme}(c)$ return the group of character c. The score is defined as the proportion of rhyming characters that belong to the largest single rhyme group:

$$S_{\text{rhyme}} = \frac{\max_{g \in G} |\{c \in L_{g,R_v} \mid \text{rhyme}(c) = g\}|}{|L_{g,R_v}|} \tag{5}$$

Automated Quality Assessment To complement rule-based formal evaluation, we introduce an LLM-based assessment protocol to estimate the semantic and aesthetic quality of generated *Songci* at scale. We use two powerful proprietary models (`GPT-4o` and `ERNIE 4.5 Turbo`), and calculate the average score to minimize bias. Each generated *Songci* is presented to the judge model with its associated *Cipai* and theme, and scored along three dimensions:

- **Fluency**: Grammatical correctness and natural phrasing.
- **Coherence**: Logical and thematic consistency.
- **Poetic Quality (Yì Jìng)**: The overall artistic effect, emotional resonance, and imagistic depth.

Each criterion is rated on a 1–5 scale. This method provides a fast, repeatable, and cost-effective proxy for human evaluation. While the model-based judge may not fully replicate human literary sensibility, it enables consistent comparisons across thousands of outputs, and serves as a complementary lens to human judgment.

Human Evaluation and Poetic Turing Test To evaluate both the literary quality and perceived human-likeness of generated *Songci*, we design a two-stage human evaluation study. We select a subset of generated *Songci* from each LLM, chosen based on their highest combined Formal Conformity and LLM-judge scores. For each selected sample, we retrieve a matched human-written *Songci* from the canonical corpus with the same *Cipai* and theme, resulting in paired comparisons per model.

Each evaluation trial followed a two-stage procedure:

(1) Poetic Turing Test: Evaluators were shown a pair of anonymized *Songci* (one generated, one human-written) and asked to identify which *Songci* they believed was written by a human, followed by a 1–5 confidence score.

(2) Qualitative Scoring: After the true authorship was revealed, evaluators rated each LLM-generated *Songci* individually across three dimensions on a 1–5 Likert scale:

- **Thematic Faithfulness**: Alignment with the assigned theme in mood and content.
- **Artistic Merit**: Aesthetic impact, use of imagery, and literary creativity.
- **Overall Quality**: A holistic assessment of form, expression, and emotional depth.

Classification-based Probing To investigate whether LLMs implicitly encode stylistic and thematic features in their outputs, we conduct three probing tasks using classifiers trained on our canonical corpus of 120 human-written *Songci*: **Cipai Identification**, **Theme Classification**, and **Source Attribution (Human vs. LLM)**.

We experiment with two lightweight classification pipelines: (1) a Support Vector Machine (SVM) trained on character-level embeddings from a pretrained Chinese BERT model, and (2) a Multinomial Naive Bayes classifier using unigram TF-IDF features.

Benchmarking LLMs for *Songci* Generation

We designed a large-scale benchmark to evaluate the ability of contemporary LLMs to generate *Songci* under formal and thematic constraints. Our evaluation spans a broad range of models and prompting strategies, enabling analysis of both intrinsic capabilities and responsiveness to guidance.

Models Under Evaluation

To facilitate comparison across model sizes, training regimes, and linguistic specifications, we selected 18 large language models for evaluation, covering both proprietary and open-source LLMs from multiple architectural families. We include 3 **proprietary models**:

- **GPT-4o** (OpenAI): A leading multimodal model with strong general-purpose reasoning and instruction-following abilities (OpenAI et al. 2024).
- **Gemini 2.5 Pro** (Google): A flagship model with multilingual support and a long context window for structured prompts (Comanici et al. 2025).
- **ERNIE 4.5 Turbo** (Baidu): A Chinese-optimized model enhanced with external knowledge graphs for improved cultural grounding (Baidu-ERNIE-Team 2025).

We include 15 **open-source LLMs** from 4 major families:

- **LLaMA (Meta)**: Strong general-purpose baselines, though English-centric tokenization may limit performance on Chinese text (Grattafiori et al. 2024).
- **Mistral/Mixtral (Mistral AI)**: Efficient sparse mixture-of-experts (MoE) model (Mixtral) with high performance per parameter, but similarly limited in Chinese-specific pretraining (Jiang et al. 2024).
- **Qwen (Alibaba)**: Multilingual models with character-aware tokenization and strong performance on Chinese tasks (Yang et al. 2025).
- **DeepSeek (DeepSeek-AI)**: Chinese-based models optimized for reasoning and Chinese language understanding (DeepSeek-AI 2025).

Prompting Strategies

We designed five prompting strategies to assess how different forms of instruction impact generation quality. For each of the 20 *Cipai* and 6 themes, every model generated *Songci* using all five different prompts:

- **Zero-shot**: Minimal instruction testing the model's latent knowledge of *Songci*.
- **One-shot**: Provides a single example from our curated corpus to guide generation.
- **Completion**: The prompt includes the first stanza of a canonical *Songci* to establish rhythm and style. The model completes the poem by generating the second stanza. Evaluation focuses solely on the generated half, and we ensure that it is not a direct reproduction of the original second stanza.
- **Instruction**: Presents explicit structural, tonal, and rhyme rules for the target *Cipai*.
- **Chain-of-Thought (CoT)**: The model is prompted to first articulate the formal rules before generating the *Songci*, allowing us to test its capacity for self-directed reasoning. The reasoning text of the output is evaluated separately from the *Songci* itself.

Model	Zero-shot	One-shot	Completion	Instruction	Chain-of-Thought	Average	Best Score
Proprietary Models							
GPT-4o	**78.62**	76.59	75.79	73.70	71.11	75.16	78.62 (zero-shot)
Gemini 2.5 Pro	78.21	77.15	76.50	74.90	72.34	76.52	78.21 (zero-shot)
ERNIE 4.5 Turbo	77.36	**80.72**	**82.38**	**81.54**	**78.57**	**79.71**	**82.38 (completion)**
Open-Source Models							
DeepSeek-R1-Distill-Llama-70B	33.75	41.71	38.10	35.95	35.58	37.82	41.71 (one-shot)
DeepSeek-R1-0528-Qwen3-8B	51.34	49.09	51.18	42.50	27.61	44.74	51.34 (zero-shot)
Llama-3-8B-Instruct	34.13	58.21	47.74	39.69	35.70	43.89	58.21 (one-shot)
Llama-3.1-8B-Instruct	35.20	54.42	53.40	38.72	40.13	44.77	54.42 (one-shot)
Llama-3.2-1B-Instruct	33.46	35.30	34.04	35.32	33.46	34.32	35.32 (instruction)
Llama-3.2-3B-Instruct	36.15	46.02	45.79	35.35	36.47	39.96	46.02 (one-shot)
Llama-3.3-70B-Instruct	35.88	63.07	51.16	37.12	35.31	44.91	63.07 (one-shot)
Mixtral-8x7B-Instruct-v0.1	28.04	41.10	35.63	30.34	31.11	33.24	41.10 (one-shot)
Mistral-7B-Instruct-v0.3	33.60	43.13	39.84	31.87	35.09	36.71	43.13 (one-shot)
Mistral-Small-Instruct-2409	37.11	54.41	53.42	40.42	35.26	44.12	54.41 (one-shot)
Qwen3-0.6B	38.76	49.74	54.11	40.52	39.77	44.58	54.11 (completion)
Qwen3-1.7B	39.88	48.56	47.20	44.20	40.72	44.51	48.56 (one-shot)
Qwen3-4B	53.61	58.85	49.54	46.45	50.44	51.78	58.85 (one-shot)
Qwen3-8B	60.19	64.53	**63.75**	53.98	**62.58**	60.61	64.53 (one-shot)
Qwen3-32B	**64.93**	**68.55**	61.52	**56.51**	62.03	**62.71**	**68.55 (one-shot)**

Table 1: Formal Conformity Scores (%) Across All Models and Prompting Strategies.

Benchmarking Results

Formal Conformity Table 1 presents the Formal Conformity Scores across all models and prompting strategies. Overall, proprietary models outperform open-source counterparts by a large margin. The best-performing proprietary model, `ERNIE 4.5 Turbo`, achieves the highest score of 82.38% under the Completion prompt, followed by `Gemini 2.5 Pro` and `GPT-4o`, both of which show strong performance under Zero-shot prompting. Interestingly, `GPT-4o` performs best when given minimal guidance, suggesting its strong internal prior knowledge for classical poetic form.

Among open-source models, the `Qwen3` models are the top performers, with `Qwen3-32B` as the strongest, achieving a best score of 68.55% under the One-shot prompt. Other open-source families (DeepSeek, LLaMA, Mistral) generally show lower average scores, indicating weaker internalization of *Songci*'s structural constraints, likely due to tokenizer mismatch and lack of domain-specific pretraining.

Prompting strategies also influence performance significantly. One-shot prompting yields the highest average scores across open-source models, while Completion and Instruction perform best for `ERNIE`. Chain-of-Thought prompts were generally less effective. Manual inspection of the generated reasoning reveals that models often correctly identify the formal requirements of each *Cipai*; however, the increased generation length may dilute adherence to structural constraints in the actual *Songci*.

Poetic Quality and Human Judgement To assess the stylistic expressiveness and thematic alignment of generated *Songci* beyond structural correctness, we selected a representative subset of models for deeper evaluation. From the proprietary group, we chose `GPT-4o` and `ERNIE 4.5 Turbo`, both of which demonstrated the highest average Formal Conformity Scores. From the open-source group, we selected one top model from each major family: `DeepSeek-R1-0528-Qwen3-8B`, `Llama-3.3-70B-Instruct`, `Mistral-Small-Instruct-2409`, and `Qwen3-32B`. For each model, we used its best-performing prompt as determined by the conformity evaluation.

Automated poetic quality assessment was conducted using both `GPT-4o` and `ERNIE 4.5 Turbo` as judges. Each model rated all 120 generated *Songci* (covering 20 *Cipai* and 6 themes) independently across fluency, coherence, and poetic quality. The final score was computed by averaging the two judge ratings. This approach provides a scalable and consistent approximation of human judgment over a large sample.

To complement the automated evaluation, we conducted a human study on a smaller, high-quality subset. For each model, we selected five *Songci* with the highest combined Formal Conformity and LLM-judge scores. Each *Songci* was paired with a human-written *Songci* from our canonical corpus that matched the same *Cipai* and theme, resulting in 30 LLM–human *Songci* pairs.

Five native Chinese speakers participated in the evaluation. Each had received 19 years of Chinese-medium education, including sustained exposure to classical poetry throughout their schooling. While not literary experts, all participants demonstrated strong familiarity with the formal and aesthetic conventions of *Songci*.

Results are presented in Table 2. Proprietary models `GPT-4o` and `ERNIE 4.5 Turbo` received the highest scores in all cases from both automated and human evaluations. Turing Test scores indicate that even the strongest models remain somewhat distinguishable from human poets, though the distinction is becoming increasingly subtle. Among open-source models, `DeepSeek-R1-0528-Qwen3-8B` performed best. Interestingly, while `Qwen3-32B` achieved the highest formal conformity among open-source models, it ranked among the lowest in terms of poetic quality and human judgment. This

Model	Best Prompt	LLM-as-Judge Scores			Human Evaluation Scores			
		Fluency	Coherence	Poetic Quality	Turing Test	Thematic Faith.	Artistic Merit	Overall Quality
Proprietary Models								
GPT-4o	zero-shot	4.72	**4.26**	**4.35**	2.10	**3.60**	**3.20**	3.50
ERNIE 4.5 Turbo	completion	**4.75**	4.05	4.10	**2.15**	3.45	3.10	**3.53**
Open-Source Models								
DeepSeek-R1-0528-Qwen3-8B	zero-shot	4.10	3.80	3.60	1.80	2.90	2.70	2.85
Llama-3.3-70B-Instruct	one-shot	4.05	3.60	3.40	1.75	2.80	2.55	2.70
Mistral-Small-Instruct-2409	one-shot	3.90	3.30	3.10	1.60	2.50	2.30	2.45
Qwen3-32B	one-shot	4.00	3.40	3.20	1.65	2.60	2.35	2.50

Table 2: Poetic Quality Scores (1–5 scale) from LLM-as-Judge and Human Evaluation.

suggests that achieving strict adherence to formal constraints may come at the cost of expressive quality.

Model	Cipai ID Accuracy (%)	Theme Class. Accuracy (%)	Source Attr. Accuracy (%)
Proprietary Models			
GPT-4o	12.61	**75.65**	97.67
ERNIE 4.5 Turbo	18.09	74.37	**80.18**
Open-Source Models			
DeepSeek-R1-0528-Qwen3-8B	8.41	51.40	80.95
Llama-3.3-70B-Instruct	**18.80**	49.57	88.64
Mistral-Small-Instruct-2409	4.17	39.17	86.36
Qwen3-32B	10.17	52.54	86.36

Table 3: Results of Classification-Based Probing on each model's best-performing prompt.

Classification Tasks To probe whether LLM-generated *Songci* encode latent stylistic or thematic signals, we conducted three classification tasks using a combined dataset of 120 canonical *Songci* and model-generated samples. Table 3 reports the performance of the best pipeline, which combines character-level embeddings from `bert-base-chinese` with a support vector machine (SVM) classifier.

Cipai identification remains the most challenging task. Accuracy across models is modest due to the fine-grained 20-class label space and small training set. The best result comes from `Llama-3.3-70B-Instruct` (18.80%), with proprietary models like `ERNIE 4.5 Turbo` also performing relatively well (18.09%). These results suggest that stronger models better preserve stylistic cues related to *Cipai*, though the signal is still weak.

Theme classification shows wide variation across models. `GPT-4o` and `ERNIE 4.5 Turbo` reach 75.65% and 74.37% respectively, indicating strong thematic alignment in their generations. Open-source models perform substantially worse, with accuracies ranging from 39.17% to 52.54%. Given the balanced 6-class setup, this suggests that thematic fidelity is a key differentiator between proprietary and open-source systems.

A lower accuracy in **Source Attribution** is preferable, as it suggests the generated *Songci* are more human-like. We trained a binary classifier to distinguish between model-generated and canonical *Songci* using an 80/20 train-test split. Surprisingly, `GPT-4o` was the most easily identified (97.67% accuracy), indicating that its outputs carry detectable machine-like features. In contrast, `ERNIE 4.5 Turbo` (80.18%) and `DeepSeek-R1-0528-Qwen3-8B` (80.95%) were harder to classify, implying a closer stylistic alignment with human-written texts.

All classification tasks use balanced label distributions (20 *Cipai*, 6 themes, 2 sources). While limited by the small training set, these results provide exploratory insight into how different models encode formal and thematic signals. We interpret this analysis as complementary to the evaluation metrics presented earlier.

The Generate-Critic Framework for Constrained Generation

To improve the formal conformity of generated *Songci*, we adopt a Generate-Critic framework (see Figure 3 for expected improvements). This architecture consists of two core components: a Generator that produces candidate *Songci*, and a Critic that evaluates their conformity against rule-based constraints.

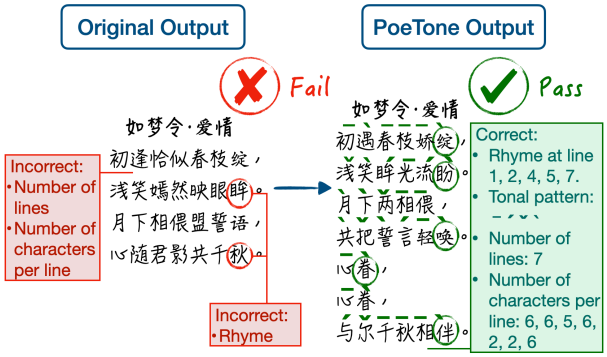

Figure 3: Expected formal improvements in generated *Songci* before and after fine-tuning.

The Generator The Generator G is a large language model with policy π_θ, parameterized by weights θ. Given an input prompt p (specifying the *Cipai* and theme), the Generator samples a sequence of tokens y representing a complete *Songci*:

$$y \sim \pi_\theta(p) \tag{6}$$

The Critic The Critic C is a deterministic, rule-based function that evaluates the formal validity of a *Songci* y un-

der the constraints c associated with the target *Cipai*. It outputs a scalar score S, computed as a weighted average of three sub-scores:

$$S = C(y, c) = w_S \cdot S_{\text{structure}} + w_T \cdot S_{\text{tonal}} + w_R \cdot S_{\text{rhyme}} \quad (7)$$

where $w_S + w_T + w_R = 1$.

Fine-Tuning with Automated Rule-Based Feedback

To enhance the model's ability to internalize formal constraints, we apply a rejection sampling fine-tuning strategy known as **Best-of-N Sampling (BoN)**. This approach curates a high-quality dataset using the Critic as an automated filter and then fine-tunes the Generator on these filtered examples via standard supervised learning.

The iterative process proceeds as follows:

1. **Candidate Generation.** For each prompt p_i in the prompt set P, we use the base model's policy $\pi_{\theta_{\text{base}}}$ to generate N candidate *Songci*:

$$Y_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,N}\}, \quad \text{where } y_{i,j} \sim \pi_{\theta_{\text{base}}}(p_i) \quad (8)$$

2. **Critic Scoring.** Each candidate $y_{i,j} \in Y_i$ is scored by the Critic to produce a conformity score:

$$S_{i,j} = C(y_{i,j}, c_i) \quad (9)$$

3. **Best Sample Selection.** The highest-scoring *Songci* y_i^* from each set is selected:

$$y_i^* = \arg\max_{y \in Y_i} C(y, c_i) \quad (10)$$

4. **Dataset Construction.** The selected prompt-*Songci* pairs are collected into a curated dataset:

$$D_{\text{best}} = \{(p_1, y_1^*), (p_2, y_2^*), \ldots, (p_M, y_M^*)\} \quad (11)$$

where M is the total number of prompts.

5. **Supervised Fine-Tuning.** The Generator is fine-tuned on D_{best} using the standard negative log-likelihood objective:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{(p, y^*) \in D_{\text{best}}} \log \pi_\theta(y^* \mid p) \quad (12)$$

Training Details The method was tested on 3 models: `Qwen3-8B`, `Llama-3.1-8B-Instruct`, and `Mistral-Small-Instruct-2409`. We used the PEFT framework for Low-Rank Adaptation (LoRA), with a rank of 16 and scaling factor $\alpha = 32$. The model was quantized to 4-bit precision using `bitsandbytes` for memory-efficient training. The optimization used paged AdamW with 8-bit weights, a learning rate of 5×10^{-5}, batch size of 2, and trained for 3 epochs.

Generate-Critic Fine-Tuned Results

Table 4 presents the formal conformity scores after fine-tuning using the Generate-Critic framework for 3 iterations, along with the corresponding score changes relative to the base model. All three models demonstrate measurable gains across most prompting strategies, confirming the effectiveness of rule-based automated supervision.

`Qwen3-8B` achieved the highest overall scores, both before and after fine-tuning. It exhibited particularly strong gains in zero-shot (+5.70), one-shot (+5.88), and completion (+5.63) settings. These results suggest that models with stronger baseline alignment to Chinese linguistic structure can not only maintain their advantage post-fine-tuning, but also benefit more from targeted feedback. Furthermore, high-quality initial generations provide more informative signal for the critic, such models stand to gain the most from Generate-Critic fine-tuning.

`Mistral-Small-Instruct-2409` showed notable improvements in zero-shot (+2.30) and one-shot (+3.40) settings. However, its performance declined under instruction (–1.46) and CoT (–1.13) prompts. `Llama-3.1-8B-Instruct` demonstrated consistent, moderate gains across all prompting strategies, with improvements ranging from +1.11 to +1.63. These uniform gains indicate that the Generate-Critic approach is broadly applicable, even for models not explicitly trained on Chinese or poetic tasks. The method generalizes well across prompt types, providing robust structural enhancements without degrading output diversity.

	Llama-3.1-8B	**Mistral-Small**	**Qwen3-8B**
Zero-shot	36.83 (+1.63)	39.41 (+2.30)	65.89 (+5.70)
One-shot	55.53 (+1.11)	57.81 (+3.40)	70.41 (+5.88)
Completion	54.77 (+1.37)	55.31 (+1.89)	69.38 (+5.63)
Instruction	39.93 (+1.21)	38.96 (–1.46)	54.31 (+0.33)
CoT	41.30 (+1.17)	34.13 (–1.13)	59.47 (–3.11)

Table 4: Formal conformity scores after fine-tuning

Conclusion

This paper presents **PoeTone**, the first large-scale, multi-faceted evaluation of large language models (LLMs) in the context of constrained Chinese *Songci* generation. We introduce a structured evaluation framework grounded in traditional poetic principles, benchmark 18 proprietary and open-source models across five prompting strategies, and systematically assess both formal conformity and lyrical quality.

Our results reveal that proprietary models like `GPT-4o` and `ERNIE 4.5 Turbo` outperform open-source models across most metrics, especially under zero- and one-shot prompts. However, even the best models face trade-offs between structural accuracy and poetic expressiveness.

Additionally, we propose a novel Generate-Critic framework where rule-based feedback is used to fine-tune model outputs. Applied to 3 open-source LLMs, this method yields measurable improvements in formal conformity, validating the use of automated critics for constrained generation.

Beyond *Songci*, our framework has broader implications for structured text generation in domains such as legal writing, classical verse composition (e.g., sonnets, qasida, shloka), and digital humanities. By operationalizing formal constraints as reusable feedback signals, PoeTone provides a scalable, annotation-free pathway to align LLMs with symbolic, cultural, or rule-based goals, bridging human creativity and machine generation in highly structured genres.

Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI.

Additionally, the project was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) via the Software Campus project (01—S23070).

The authors gratefully acknowledge the computing time made available to them on the high-performance computer at the NHR Center of TU Dresden. This center is jointly supported by the Federal Ministry of Research, Technology and Space of Germany and the state governments participating in the NHR (www.nhr-verein.de/unsere-partner).

References

Baidu-ERNIE-Team. 2025. ERNIE 4.5 Technical Report. https://yiyan.baidu.com/blog/publication/. Accessed: 2025-11-27.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, H.; Yi, X.; Sun, M.; Li, W.; Yang, C.; and Guo, Z. 2019. Sentiment-Controllable Chinese Poetry Generation. In *IJCAI*, 4925–4931.

Chen, Z.; and Cao, Y. 2024. A Polishing Model for Machine-Generated Ancient Chinese Poetry. *Neural Processing Letters*, 56(2): 77.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Gao, T.; Zhu, S.; Liu, J.; Shen, J.; Shen, J.; Yang, S.; and Xiong, P. 2021. A new context-aware approach for automatic Chinese poetry generation. *Knowledge-Based Systems*, 232: 107409.

Ge, Z. 2009. *Cilin Zhengyun [Cilin Rhyming Dictionary]*. Shanghai, China: Shanghai Classics Publishing House.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hu, J.; Li, J.; Pan, Z.; Chen, C.; Li, Z.; Wang, P.; and Zhang, L. 2025. SongSong: A Time Phonograph for Chinese SongCi Music from Thousand of Years Away. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(25), 26229–26237.

Hu, J.; and Sun, M. 2020. Generating major types of chinese classical poetry in a uniformed framework. *arXiv preprint arXiv:2003.11528*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2024. Mixtral of Experts. arXiv:2401.04088.

Li, J.; Song, Y.; Zhang, H.; Chen, D.; Shi, S.; Zhao, D.; and Yan, R. 2018. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 3890–3900.

Liao, Y.; Wang, Y.; Liu, Q.; and Jiang, X. 2019. GPT-based Generation for Classical Chinese Poetry. *CoRR*, abs/1907.00151.

Luo, Y.; Li, C.; Huang, C.; Xu, C.; Zeng, X.; Wei, B.; Xiao, T.; and Zhu, J. 2021. Chinese poetry generation with metrical constraints. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 377–388. Springer.

Ma, J.; Zhan, R.; and Wong, D. F. 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*, 57–66. Dubrovnik, Croatia: Association for Computational Linguistics.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Shao, Y.; Shao, T.; Wang, M.; Wang, P.; and Gao, J. 2021. A sentiment and style controllable approach for chinese poetry generation. In *Proceedings of the 30th ACM international conference on information & knowledge management*, 4784–4788.

Song, X.; Song, C.; Yu, H.; Zhu, Y.; and Yao, H. 2025. MixSong: Diverse and Strictly Formatted Chinese Poetry Generation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(4).

Song, Y. 2022. Composing Ci with Reinforced Non-autoregressive Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7219–7229. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Tang, Y.; Li, Z.; Yu, J.; and Yang, L. 2025. A Simple Approach of Chinese Poetry Generation Using Pre-trained LLMs. In *Proceedings of the 2025 7th Asia Pacific Information Technology Conference*, 106–112.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale,

S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, Z.; Guan, L.; and Liu, G. 2022. Generation of Chinese classical poetry based on pre-trained model. arXiv:2211.02541.

Wang, Z.; He, W.; Wu, H.; Wu, H.; Li, W.; Wang, H.; and Chen, E. 2016. Chinese Poetry Generation with Planning based Neural Network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1051–1060. Osaka, Japan: The COLING 2016 Organizing Committee.

Yan, Y.; Wen, D.; Yang, L.; Zhang, D.; and Lin, H. 2023. Poetry Generation Combining Poetry Theme Labels Representations. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, 1246–1255. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yi, X.; Li, R.; and Sun, M. 2017. Generating chinese classical poems with rnn encoder-decoder. In *China National Conference on Chinese Computational Linguistics*, 211–223. Springer.

Yi, X.; Li, R.; and Sun, M. 2018. Chinese Poetry Generation with a Salient-Clue Mechanism. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 241–250. Brussels, Belgium: Association for Computational Linguistics.

Yi, X.; Sun, M.; Li, R.; and Yang, Z. 2018. Chinese poetry generation with a working memory model. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4553–4559.

Yu, C.; Zang, L.; Wang, J.; Zhuang, C.; and Gu, J. 2024. CharPoet: A Chinese Classical Poetry Generation System Based on Token-free LLM. arXiv:2401.03512.

Zhang, X.; and Lapata, M. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 670–680. Association for Computational Linguistics.

Zhipeng, G.; Yi, X.; Sun, M.; Li, W.; Yang, C.; Liang, J.; Chen, H.; Zhang, Y.; and Li, R. 2019. Jiuge: A Human-Machine Collaborative Chinese Classical Poetry Generation System. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 25–30. Florence, Italy: Association for Computational Linguistics.