

Jiuge-Tuiqiao: An Interpretable Human-AI System for Classical Chinese Poetry Refinement

Yufeng Han^{1,2,3*}, Lifan Deng^{1,2,3,4*}, Cunliang Kong^{1,2,3},

Wenhao Li^{1,2,3}, Xin Cong⁵, Yuzhuo Bai^{1,2,3}, Kangyang Luo^{1,2,3}, Maosong Sun^{1,2,3,6†}

¹Department of Computer Science and Technology, Tsinghua University, Beijing

²Beijing National Research Center For Information Science And Technology

³Institute for Artificial Intelligence, Tsinghua University, Beijing

⁴Rixin College, Tsinghua University ⁵Department of Statistics and Data Science, Tsinghua University

⁶Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University, Xuzhou
hanyufeng@mail.tsinghua.edu.cn, dlf24@mails.tsinghua.edu.cn

Abstract

Classical Chinese poetry composition has long valued *Tuiqiao*, the iterative refinement of words, imagery, and prosody. However, many current AI poetry systems follow a one-shot generation paradigm, which reduces users to prompt providers and weakens their creative agency. We present *Jiuge-Tuiqiao*¹, an interactive human-AI collaborative system for classical Chinese poetry composition. The system is designed around a triadic model: user-driven control, ancient-guided evidence, and AI-assisted generation. Users can lock characters or lines, receive real-time prosody feedback, and obtain interpretable refinement suggestions grounded in high-frequency collocations, PPL-ranked classical lines, and structured knowledge extracted from classical encyclopedias. This design turns AI from an autonomous generator into a background assistant that supports the user’s own process of poetic refinement. Preliminary experiments and user feedback indicate that *Jiuge-Tuiqiao* provides controllable refinement mechanisms, traceable literary evidence, and a positively received interactive experience for classical Chinese poetry composition.

formal integrity. Therefore, *Tuiqiao* sharpens poetic language, and the process of *Tuiqiao* embodies the creator’s subjectivity — including linguistic competence, cultural literacy, and stylistic voice.

Although one-shot generation is useful for rapidly producing a draft, it offers limited support for users who wish to preserve selected local decisions, inspect literary evidence, and iteratively revise the remaining text:

- **Lack of a *Tuiqiao* process:** Existing “one-click generation” paradigms generate complete poems from prompts but offer no mechanism for users to make precise revisions to unsatisfactory local elements. In particular, many one-shot workflows return a complete poem in response to a prompt but do not allow users to preserve selected local decisions while regenerating the remaining positions.
- **Lack of traceable evidence:** Many model-based systems present candidate words without traceable literary or linguistic evidence, making it difficult for users to understand why a suggestion fits the context or to verify it against classical sources.
- **Lack of creator autonomy:** Most critically, many model-centered workflows do not give users complete decision-making authority over how a poem evolves. Users may provide prompts or select outputs, but cannot determine at every refinement step which characters or lines should be preserved, revised, regenerated, accepted, or rejected. This may weaken users’ sense of authorship and make the interaction closer to content consumption than to creative composition.

To bridge the aforementioned gaps, we present *Jiuge-Tuiqiao*. The system returns creative initia-

1 Introduction

Classical Chinese poetry composition emphasizes *Tuiqiao*: a constrained iterative refinement where poets balance semantic precision, tonal prosody, and aesthetic coherence. The well-known case of *Jia Dao* deliberating between “push the moonlit gate” (*Seng Tui Yue Xia Men*) and “knock the moonlit gate” (*Seng Qiao Yue Xia Men*) shows how poets weigh linguistic accuracy, emotional nuance and

*Equal contribution.

†Corresponding author.

¹You can access our services at <https://github.com/jiangli-va/Jiuge-Tuiqiao>.

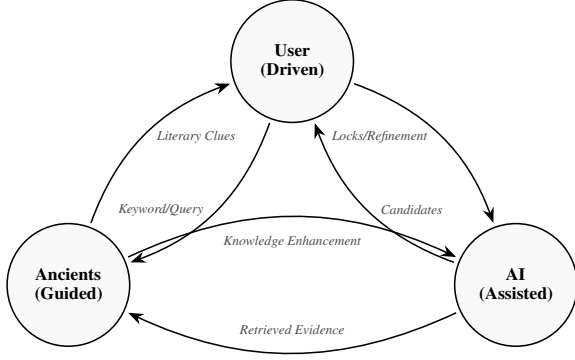


Figure 1: The collaborative workflow of *Jiuge-Tuiqiao*. It shows the relationship where the User maintains creative subjectivity, Ancient Knowledge provides knowledge guidance, and AI offers generative assistance.

tive to the poet by modeling a deep tripartite interaction among the user, the ancients, and AI (Fig1):

- **User-driven:** Users retain full control, can lock any satisfying character or line, and receive revision suggestions under those constraints — enabling true iterative refinement.
- **Ancient-guided:** A rich knowledge base (high-frequency collocations, famous lines and cues from ancient encyclopedias) provides literary or linguistically grounded suggestions, realizing “refinement following the ancients’ thinking”.
- **AI-assisted:** AI steps to the background as a powerful language generation engine and knowledge retrieval tool, optimizing its outputs based on user feedback and knowledge constraints, rather than replacing the user’s creative agency.

We also integrate a dynamic prosody checking mechanism that provides real-time feedback on tonal patterns and rhyme using color coding, helping users attend to both formal prosody and semantic expression during the *Tuiqiao* process.

Jiuge-Tuiqiao is fully implemented and accessible via a WeChat mini-program, open to poetry enthusiasts, teaching, and cultural dissemination, offering a complete workflow from generation, refinement to feedback. Preliminary user ratings indicate positive perceptions of agency, prosody guidance, trustworthiness, and overall satisfaction. Through this digital workflow, the system carries forward the traditional wisdom of *Tuiqiao* in the AI era.

2 Related Work

Classical Chinese Poetry Generation. Automated poetry generation has evolved from rule-based methods to neural architectures optimizing coherence and style (Yi et al., 2018a; Yang et al., 2018; Yi et al., 2018b). Recent research leverages Large Language Models (LLMs) for structural and imagery control, including unified GPT-2 frameworks (Hu and Sun, 2020), imagery-focused masking in PoemBERT (Huang and Shen, 2025), token-free character control in CharPoet (Yu et al., 2024), and domain-specific LoRA fine-tuning (Xie, 2025).

Interactive Poetry Systems. Moving beyond one-shot generation, interactive systems increasingly support human-AI collaboration in poetry composition. *Jiuge* (Guo et al., 2019) enables repeated revision and dynamic regeneration of unsatisfactory parts, while *Yu Sheng* (Ma et al., 2023) provides constrained generation and fine-grained polishing. Building on these interactive approaches, *Jiuge-Tuiqiao* further makes the evidence behind refinement suggestions visible to users. It combines character- and line-level locking and real-time prosodic feedback with traceable evidence from high-frequency collocations, attributed historical lines, and ancient encyclopedia entries, allowing users to inspect retrieved evidence while retaining final control over each revision decision.

Interactive AI in Other Domains. Interactive refinement and human-in-the-loop feedback have also enhanced argumentative writing training (Ding et al., 2025) and narrative coherence in storytelling (Mayer Martins et al., 2025). We extend this philosophy of collaborative, feedback-driven intervention to the highly constrained domain of classical Chinese poetry.

3 System Architecture

We formalize *Jiuge-Tuiqiao* as a human-AI collaborative, constrained iterative refinement process. This section presents its mathematical formalization, introduces a three-layer decoupled architecture, and models single refinement iterations via a finite state machine (FSM).

3.1 Problem Formulation

We model a poem as a partially observable character matrix M , with dimensions determined by its poetic form (e.g., $M_{8 \times 5}$ for a 5-char *lvshi*). Each entry $M_{i,j}$ contains either a Chinese character or a placeholder 0 awaiting completion or refinement.

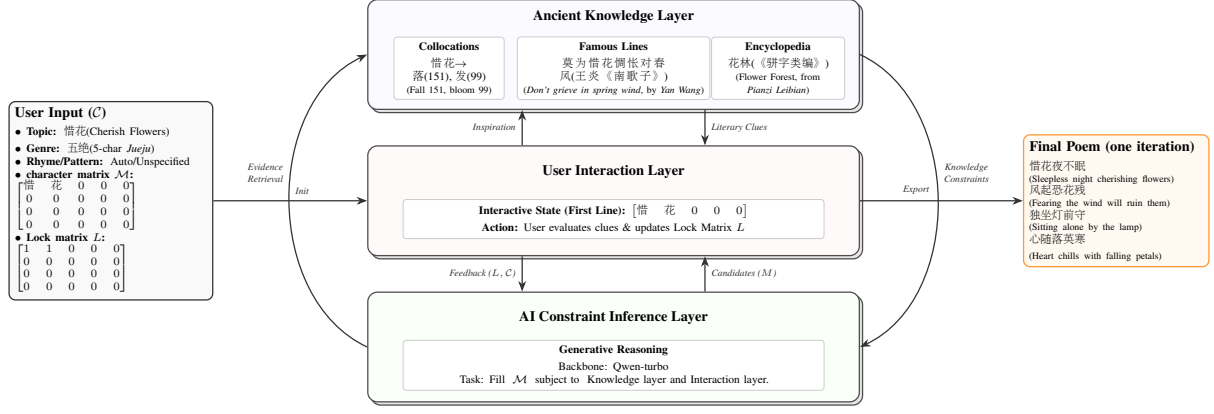


Figure 2: System architecture and the triadic collaborative iterative *Tuiqiao* process. The core engine coordinates the Ancient Heritage, User Subjectivity, and AI Assistant layers through continuous feedback and evidence retrieval.

Input Constraint Set $\mathcal{C}$. The user defines the constraints as a set $\mathcal{C} = \{\mathcal{K}, d, \mathcal{Y}, \mathcal{P}, T\}$, comprising topic keywords $\mathcal{K}$, artistic description d , a target *Pingshui Yun* rhyme category $\mathcal{Y}$, an optional first-line tonal pattern $\mathcal{P}$ (Appendix A), and a verse-derived tonal matrix $T \in \{P, Z, A\}^{n \times m}$, where P, Z, A denote level, oblique, and flexible tones, respectively.

Interaction State & Constraint Handling. The interaction state uses a current poem matrix M_t and a lock matrix $L \in \{0, 1\}^{n \times m}$, where $L_{i,j} = 1$ freezes a character and $L_{i,j} = 0$ keeps it editable. One *Tuiqiao* step ranks returned candidates using an implicit scoring function $\mathcal{S}$:

$$\begin{aligned} M^* &= \arg \max_M \mathcal{S}(M \mid M_t, L, \mathcal{C}, E), \\ \text{s.t. } M_{i,j}^* &= (M_t)_{i,j}, \quad \forall (i, j) \text{ where } L_{i,j} = 1, \\ \text{Format}(M^*) &= 1, \\ \text{Rhyme}(M^*, \mathcal{Y}) &= 1. \end{aligned}$$

Here, E denotes retrieved literary evidence. The formula characterizes candidates returned after deterministic post-generation processing rather than raw model outputs. Locked characters are restored and verified by coordinates; candidates with invalid line or character counts are filtered out; and rhyme violations are rejected, with validation feedback passed to the next generation round for up to three rounds.

Tonal-pattern matching is handled separately and is not a hard acceptance condition. The target pattern T is included in the prompt and used to prioritize better-matching candidates. Candidates with at most two tonal mismatches enter the preferred pool; if this pool is insufficient, basic candidates

that pass the format, rhyme, and lock checks may still be returned with explicit mismatch warnings. The user retains final authority to accept, reject, lock, or manually revise every candidate.

3.2 Triadic Collaborative Architecture

We implement a three-layer decoupled architecture to support user-driven, ancient-guided, and AI-assisted collaboration (Figure 2).

Interaction Layer. Captures fine-grained user operations (text input, character locking, suggestion selection) and translates them into system events. It dynamically updates the lock matrix L and constraint set $\mathcal{C}$, maintaining full user control at each step.

Knowledge Layer. Grounded in classical scholarship, this layer provides literary justifications by integrating three types of knowledge (detailed in §4): (1) high-frequency collocations extracted from 320,000 lines of Tang and Song poetry; (2) famous line references ranked via PPL using Qwen3-32B on 1.8M historical poems; and (3) 11 classical encyclopedias (e.g., *Hailu Suishi*, *Baikong Liutie*) supplying imagery, antithesis, and rhyme words.

Constraint Inference Layer. Powered by Qwen-turbo, this core engine aggregates hard constraints ($L, \mathcal{C}$) from the interaction layer and soft constraints (literary clues) from the knowledge layer. Through tailored prompt engineering and decoding strategies, it generates compliant poem candidates while positioning AI strictly as a background assistant.

3.3 Iterative Refinement Cycle

The continuous, user-led *Tuiqiao* process is modeled as a finite state machine (FSM) loop:

- (1) **Initialization:** The user inputs constraints $\mathcal{C}$ (and optionally a draft), and the system generates an initial poem matrix M .
- (2) **AI Generation:** When the user triggers *Tuiqiao*, the system populates placeholders in M to produce candidate solutions.
- (3) **Review & Locking:** The user evaluates candidates and either locks satisfactory tokens (updating the lock matrix L) or manually edits the text.
- (4) **Iterative Loop:** Steps (2) and (3) repeat dynamically with the updated L until the user is satisfied.

Throughout this cycle, the AI engine is strictly confined to generating interpretable options, while the knowledge layer supplies historical precedents as evidence. Ultimate creative agency and editing rights remain entirely with the human user, fully operationalizing our triadic collaborative philosophy.

4 Methods

To ensure interpretability, *Jiuge-Tuiqiao* systematically integrates classical poetry knowledge with neural generation. For any refinement position, the system retrieves evidence from three knowledge sources: frequent co-occurrences, famous couplet matching, and ancient encyclopedia clues. Below, we detail the computation of each source and our joint fusion strategy.

4.1 Frequent Co-occurrence Retrieval

To discover habitual word pairings, we apply a t -test to assess collocation significance, filtering out spurious high-frequency pairings.

Preprocessing & Counting. We utilize 320,000 lines from Tang-Song *jueju* and *lvshi* poems from Sou-Yun². Lines are segmented using a dictionary-plus-metrical-foot method (Appendix B). After converting to simplified Chinese, filtering function words, and discarding short lines, we compute the global frequency $f(w)$ for each word w , and the co-occurrence frequency $f(w_i, w_j)$ for ordered pairs within the same line.

Significance Testing. Under the null hypothesis that words w_i and w_j appear independently within a line, their individual and joint probabilities are defined as $P(w) = f(w)/N$ and $P(w_i, w_j) = f(w_i, w_j)/N$, where N is the total line count. The t -statistic is calculated as:

$$T(w_i, w_j) = \frac{P(w_i, w_j) - P(w_i)P(w_j)}{\sqrt{P(w_i, w_j)/N}}.$$

²<https://www.sou-yun.cn/>.

We filter for pairs with $f(w_i, w_j) \geq 3$ and $T > 1.96$ (95% confidence), sorting by descending T to yield 28,739 high-frequency collocations.

Suggestion Generation. For a target position (i, j) with context words $\mathcal{W}_{ctx}$, candidate characters c are ranked by their highest t -value with any anchor word $w \in \mathcal{W}_{ctx}$, outputting the top candidates alongside their traceable collocation evidence.

4.2 Famous Couplet Matching via Perplexity

We employ Perplexity (PPL) as a proxy for canonical familiarity, assuming that lines with lower PPL are better predicted by a pretrained model and thus more likely to reflect canonical historical expressions. Low PPL is only a heuristic proxy: it may reflect canonical familiarity, pretraining memorization, or formulaic language.

Corpus Construction. From 1.8M+ historical poems (Pre-Qin to Qing dynasty) sourced from Sou-Yun, we split texts by standard punctuation, filter out fragments shorter than 7 characters, and deduplicate to construct the candidate famous-line collection $\mathcal{D}_{raw}$.

PPL Evaluation. Using Qwen3-32B, we compute the PPL for each line $s = (x_1, \dots, x_{|s|})$ to obtain a model-based ranking score for retrieval:

$$\text{PPL}(s) = \exp \left(-\frac{1}{|s|} \sum_{t=1}^{|s|} \log P(x_t \mid x_{<t}; \theta) \right).$$

After ranking the lines by ascending PPL, we manually reviewed the top-ranked candidates and retained 20,000 verified lines to form the final famous-line index $\mathcal{D}_{FL}$.

Binning & Suggestion Generation. To eliminate length bias (as longer lines inherently alter average PPL distributions), we stratify $\mathcal{D}_{FL}$ into distinct length bins using thresholds at 12 and 16 characters. For a target position (i, j) and candidate character c , the system retrieves lines containing c that overlap with the context $\mathcal{W}_{ctx}$. Candidates are then ranked by ascending PPL within their respective length bins and presented as curated, attributed historical examples.

4.3 Ancient Encyclopedia Clue Retrieval

We extract structured textual data from 11 ancient Chinese encyclopedias—including topic-organized works (e.g., *Hailu Suishi*, *Baikong Liutie*) and rhyme-organized works (e.g., *Yunfu Qunyu*)—to construct three specialized databases (detailed in



Figure 3: Overview of system states: (a) metadata constraint initialization panel; (b–e) real-time feedback in the Character Grid Editor showing (b) valid compliance (green) and acceptable variants (yellow), (c) rhyme mismatch (purple), (d) tonal violation (red), and (e) user-locked tokens (gold fill).

Appendix C:

- (1) **Imagery Words**, storing topic-specific poetic images and descriptions;
- (2) **Antithesis Words**, containing parallel or antithetical word pairs;
- (3) **Rhyme Words**, categorizing terms by their prescriptive rhyme categories.

Suggestion Generation. For a target position (i, j) given the current word w and user keywords $\mathcal{K}$, the system dynamically queries these databases. It retrieves relevant imagery clues based on $\mathcal{K}$, antithesis clues based on w (if parallel constraints apply), or rhyme clues matching the target rhyme $\mathcal{Y}$. These retrieved entries are returned directly to the user as explicit literary justifications.

4.4 Suggestion Organization and Presentation

The interface displays the suggestion lists from three knowledge sources in parallel as categorized cards. Users can browse this traceable evidence to either adopt a candidate or perform manual edits, preserving ultimate human agency over the final text.

5 User Interface

The *Jiuge-Tuiqiao* WeChat mini-program (Taro 4, React, FastAPI) features a human-centric interface that balances direct user control with multi-channel poetic knowledge.

Character Grid Editor. As the central workspace, this grid dynamically renders characters alongside multi-dimensional states (Figure 3). With a 500 ms debounce, it provides real-time feedback: **green** borders denote valid compliance and **yellow** marks acceptable variants (Figure 3b); **purple** signals rhyme mismatches (Figure 3c); and **red** indicates tonal violations (Figure 3d). A **gold fill** highlights user-locked tokens (Figure 3e). Explicit justifications accompany violations to streamline error correction.



Figure 4: Multi-channel evidence in the Interpretable *Tuiqiao* Panel: (a–b) statistical high-frequency co-occurrence candidates, (c) ranked famous-line samples with metadata, and (d) textual clues retrieved from ancient encyclopedias.

Hierarchical Locking Mechanism. To preserve human creative agency, users can restrict AI generation at two granularities: (1) *Character-level lock* freezes individual cells ($L_{i,j} \leftarrow 1$, gold fill; Figure 3e) to preserve user tokens; (2) *Line-level lock* fixes an entire row, forcing the system to generate alternatives strictly for the remaining lines.

Interpretable *Tuiqiao* Panel. Clicking an unlocked cell activates a side panel that first aggregates evidence from three provenance-backed channels (Figure 4): (1) the *High-frequency co-occurrence* channel displays top collocates from a 28,739-entry table (Section 4.1, Figure 4(a–b)); (2) the *Famous-line samples* channel displays up to five lines ranked by ascending PPL with full source metadata (Section 4.2, Figure 4(c)); and (3) the *Ancient encyclopedia clues* channel provides imagery, antithesis, or rhyme clues mined from 11 historical encyclopedias (Section 4.3, Figure 4(d)). When none of these sources yields a match, the system invokes an LLM to generate a fallback rationale, which is served asynchronously through an 8,000-capacity LFU cache to reduce latency. This fallback is explicitly presented as model-generated and is not treated as provenance-backed literary evidence. Because the in-system famous-line index is limited to 20,000 lines retained after ranking by ascending PPL and manual verification, the active *Sou-Yun* integration hyperlinks each grid character to the corresponding search results, allowing users to inspect the full set of relevant historical poetic lines indexed by *Sou-Yun*. This extends evidence inspection beyond the local index while leaving the final decision to the poet.

Revision History & Workflow. Each line maintains an independent undo/redo stack for single-click version restoration. This supports an iterative co-creation loop: users initialize metadata (Figure 3a), generate candidates, lock satisfactory

segments, and progressively refine tokens via the *Tuiqiao* panel under real-time validation until completion.

6 Evaluation

We evaluate Jiuge-Tuiqiao across three dimensions: (1) a knowledge-source ablation study on cloze-filling; (2) automatic prosodic metrics on generated poems; and (3) a human evaluation assessing creative experiences.

6.1 Ablation Study: Famous-Line and Co-occurrence Contributions

Task and Setup. To assess the individual and combined contributions of the famous-line and co-occurrence retrieval sources, we design a controlled cloze-filling task using 50 poem lines stratified by poem type. For each instance, we mask a contiguous two-character content word and prompt the backbone LLM (Qwen-turbo) to predict the blank under four conditions: (1) **BASE**: context only; (2) **+FL**: adds up to five PPL-ranked famous lines containing the anchor term; (3) **+CO**: adds top-20 co-occurrence phrases; and (4) **BOTH**: provides both sources simultaneously. For each instance, we excluded the original source line from the famous-line index before retrieval. Then, we evaluate performance using **Hit@1** (whether the original ground-truth word is the top-ranked candidate) and **Hit@5** (whether the ground truth appears anywhere within the top five suggestions).

Results. As shown in Table 1, auxiliary knowledge largely boosts performance over the BASE condition. While **+FL** yields the highest Hit@1 (60%) due to historical exact-matches, introducing co-occurrences (**+CO** and **BOTH**) expands the lexical search space. Notably, **BOTH** maintains the peak Hit@5 (60%) while slightly softening Hit@1 to 56%. This trade-off is ideal for an interactive assistant: famous lines preserve classical diction, while co-occurrence phrases inject stylistic diversity rather than merely forcing the reproduction of the ground truth.

6.2 Automatic Prosody Evaluation

Setup & Metrics. We evaluate system-generated poems using 80 requests stratified across four forms (5/7-character *jueju* and *lvshi*) and four historical title-frequency tiers: *high* (top 10%), *mid* (10–50%), *low* (50–90%), and *rare* (bottom 10%). We report rule-based correctness via **Format** (struc-

Condition	Hit@1	Hit@5
BASE	12%	12%
+FL	60%	60%
+CO	36%	42%
BOTH	56%	60%

Table 1: Ablation results on cloze-filling. Combining both sources (**BOTH**) matches the highest Hit@5 (60%) while yielding a slightly lower Hit@1 (56%) than **+FL** alone (60%).

tural length), **L-Pattern** (line-level tonal accuracy), **P-Pattern** (global poem-level pattern adherence including *nian-dui* constraints), and **Rhyme** (*Pingshui Yun* consistency). Output diversity is quantified using **Distinct- n** and **Self-BLEU** ($n \in \{1, 2, 4\}$).

Results. As shown in Table 2, the system achieves perfect Format and Rhyme accuracy across all forms, alongside high line-level alignment (L-Pattern $\geq 91.25\%$). However, global constraints impose a strict bottleneck: poem-level adherence (P-Pattern) drops significantly for *lvshi* (15%–25%) compared to *jueju* (50%–65%), reflecting the complexity of compounding eight-line *nian-dui* requirements. Crucially, Table 3 confirms that these structural constraints do not induce mode collapse; lexical diversity remains high, with Distinct-4 reaching 1.0000 and Self-BLEU-4 at a low 0.1855.

Form	Format	L-Pattern	P-Pattern	Rhyme
5-char <i>jueju</i>	100%	96.25%	65%	100%
7-char <i>jueju</i>	100%	91.25%	50%	100%
5-char <i>lvshi</i>	100%	93.12%	25%	100%
7-char <i>lvshi</i>	100%	93.12%	15%	100%

Table 2: Automatic prosody accuracy by verse form.

	1-gram	2-gram	4-gram
Distinct	0.9804	0.9995	1.0000
Self-BLEU	0.8983	0.5717	0.1855

Table 3: Output diversity metrics across generations.

6.3 User Evaluation

Setup. We conducted a preliminary user evaluation with 10 participants (poetry enthusiasts and researchers in classical Chinese literature). After composing at least one poem, each participant completed a 9-point Likert scale questionnaire (1 = strongly disagree, 9 = strongly agree) covering four dimensions: (1) **Agency** (active creative control), (2) **Trustworthiness** (suggestion credibility), (3) **Prosody guidance** (color feedback utility), and

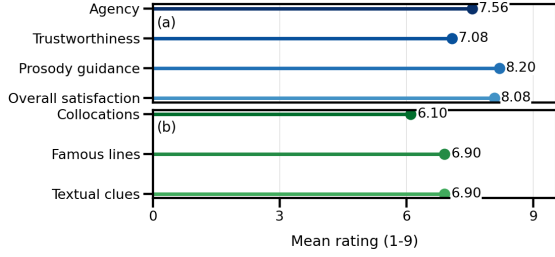


Figure 5: Human evaluation ratings (9-point Likert scale) showing (a) four evaluation dimensions and (b) perceived utility of three knowledge sources.

(4) **Overall satisfaction.** Open-ended qualitative feedback was also collected.

Results. As shown in Figure 5, participants rated the system positively overall. Prosody guidance scored highest (8.20/9), demonstrating that real-time color feedback effectively aided metrical revisions. Overall satisfaction (8.08/9) and agency (7.56/9) were high, confirming active user control over AI outputs. Trustworthiness and evidence utility scored 7.08/9. Among knowledge sources, users preferred famous lines and ancient textual clues over frequent collocations, prioritizing contextual literary evidence over statistical signals. Open-ended feedback highlighted that evidence panels (source lines, author attributions, citations) enhanced credibility and facilitated in-context learning.

7 Conclusion and Future Work

We presented *Jiuge-Tuiqiao*, a human-AI collaborative poetry composition system that re-centers human agency through classical aesthetics. The framework implements a triadic design—integrating hierarchical locking, real-time prosody validation, and multi-channel traceable recommendations. Its central contribution is to operationalize interpretability as provenance at the interface level: for retrieved suggestions, users can inspect the supporting statistical or literary sources while retaining final authority to accept, reject, lock, or manually revise the text.

Future work will focus on four directions: (1) **Personalized modeling** to adapt recommendations to individual stylistic preferences across sessions; (2) **Broader poetic forms** extending constraints to irregular meters like *ci* (song-lyrics) and *sanqu* (aria); (3) **Comprehensive user studies** engaging larger, diverse participant pools to rigorously validate the system’s usability and long-term creative

impact; and (4) **Preference alignment** utilizing interaction logs as training signals to fine-tune LLMs toward both high poetic quality and user agency.

Ethics and Impact

Creative authorship and attribution. *Jiuge-Tuiqiao* is designed so that the human poet retains full creative initiative: the system provides options and evidence, never unilaterally inserting text without user acceptance. Nevertheless, AI-assisted composition raises genuine questions about authorship. We recommend that users who publish AI-assisted poems disclose the assistance; the system’s export functionality includes a disclosure template for this purpose. We do not claim that AI-generated candidates carry the same creative authorship as wholly human-written verse.

Cultural preservation and accessibility. Classical Chinese poetry encodes millennia of cultural knowledge — prosodic conventions, canonical imagery, allusive traditions. By embedding the *Pingshui Yun* system, ancient encyclopedias, and a vast famous-line corpus into the interaction loop, *Jiuge-Tuiqiao* lowers the barrier to engaging with and practising this tradition, potentially reaching learners of Chinese as a second language and heritage communities outside China.

Potential biases. The knowledge base is constructed from historical corpora that skew heavily toward the Tang and Song dynasties and toward elite male authorship. As a result, recommended collocations and famous lines may systematically under-represent the poetic traditions of women poets, border literati, and minority-language contributors to the Chinese literary canon. Future versions should apply diversity-aware retrieval to broaden the representational range of suggestions.

Risk of creative dependency. A persistently available refinement tool may, over time, discourage users from developing independent prosodic intuition. The system mitigates this risk by displaying explanatory feedback for each prosodic violation rather than silently correcting it, and by making all knowledge sources and their justifications explicitly visible — encouraging users to learn *why* a suggestion is grounded in classical practice, not merely to accept it.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. We are also grateful to the

volunteers who participated in the user evaluation for their ratings and valuable feedback on the system. This work was supported by the National Natural Science Foundation of China (Grant No. 62236011), the Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of the Ministry of Education at Minzu University of China, Beijing, China, and the Shenzhen–Tsinghua Special Program for Fundamental and Frontier Research in Artificial Intelligence, under the project “Development of an AI-Driven Next-Generation Digital Humanities Research Platform” (No. AI2026027).

References

- Yuning Ding, Franziska Wehrhahn, and Andrea Horbach. 2025. [FEAT-writing: An interactive training system for argumentative writing](#). In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 217–225, Abu Dhabi, UAE. Association for Computational Linguistics.
- Zhipeng Guo, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, Jiannan Liang, Huimin Chen, Yuhui Zhang, and Ruoyu Li. 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*, pages 25–30, Florence, Italy. Association for Computational Linguistics.
- Jinyi Hu and Maosong Sun. 2020. [Generating major types of Chinese classical poetry in a uniformed framework](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4658–4663, Marseille, France. European Language Resources Association.
- 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.
- 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.
- Jonas Mayer Martins, Ali Hamza Bashir, Muhammad Rehan Khalid, and Lisa Beinborn. 2025. [Once upon a time: Interactive learning for storytelling with small language models](#). In *Proceedings of the First BabyLM Workshop*, pages 454–468, Suzhou, China. Association for Computational Linguistics.
- Haotao Xie. 2025. [System report for CCL25-eval task 5: New dataset and LoRA-fine-tuned qwen2.5](#). In *Proceedings of the 24th China National Conference on Computational Linguistics (CCL 2025)*, pages 200–205, Jinan, China. Chinese Information Processing Society of China.
- 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, and Maosong Sun. 2018a. [Chinese poetry generation with a salient-clue mechanism](#). In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 241–250, Brussels, Belgium. Association for Computational Linguistics.
- Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018b. [Automatic poetry generation with mutual reinforcement learning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3143–3153, Brussels, Belgium. Association for Computational Linguistics.
- Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. 2024. [CharPoet: A Chinese classical poetry generation system based on token-free LLM](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–325, Bangkok, Thailand. Association for Computational Linguistics.

A Rhyme Categories and First-line Tonal Patterns

Pingshui Yun Rhyme Categories (30 Level Tones). The system supports 30 Level Tone categories from the *Pingshui Yun* system.

Upper Level Tones (上平声): 东Dong, 冬Dong, 江Jiang, 支Zhi, 微Wei, 鱼Yu, 虞Yu, 齐Qi, 佳Jia, 灰Hui, 真Zhen, 文Wen, 元Yuan, 寒Han, 删Shan.

Lower Level Tones (下平声): 先Xian, 萧Xiao, 肴Yao, 豪Hao, 歌Ge, 麻Ma, 阳Yang, 庚Geng, 青Qing, 蒸Zheng, 尤You, 侵Qin, 覃Tan, 盐Yan, 咸Xian.

Tonal Pattern Matrices (7-character Jueju).

Tonal constraints are modeled as matrices $T \in \{P, Z, A\}^{n \times m}$, where P : Level tone (平声), Z : Oblique tone (仄声), and A : Flexible. The four globally valid patterns are consolidated below.

1. Level-start, Level-end (平起平收)

$$T_{L-L} = \begin{bmatrix} A & P & A & Z & Z & P & P \\ A & Z & P & P & A & Z & P \\ A & Z & A & P & P & Z & Z \\ A & P & A & Z & Z & P & P \end{bmatrix}$$

2. Level-start, Oblique-end (平起仄收)

$$T_{L-O} = \begin{bmatrix} A & P & A & Z & A & P & Z \\ A & Z & P & P & A & Z & P \\ A & Z & A & P & P & Z & Z \\ A & P & A & Z & Z & P & P \end{bmatrix}$$

3. Oblique-start, Level-end (仄起平收)

$$T_{O-L} = \begin{bmatrix} A & Z & P & P & A & Z & P \\ A & P & A & Z & Z & P & P \\ A & P & A & Z & A & P & Z \\ A & Z & P & P & A & Z & P \end{bmatrix}$$

4. Oblique-start, Oblique-end (仄起仄收)

$$T_{O-O} = \begin{bmatrix} A & Z & A & P & P & Z & Z \\ A & P & A & Z & Z & P & P \\ A & P & A & Z & A & P & Z \\ A & Z & P & P & A & Z & P \end{bmatrix}$$

B Poem Segmentation Algorithm

Two-Stage Segmentation Strategy. The system employs a two-stage strategy to achieve grammar-based segmentation (G-Seg): (1) *Candidate Generation*: Based on a scoring dictionary and metrical rules, the system enumerates all possible paths via Depth-First Search (DFS) under maximum word length limits, globally selecting the top- K scoring candidate paths. (2) *LLM Reranking*: An LLM acts as a discriminator to select the optimal candidate that best adheres to grammatical norms from the top- K list.

Line Type	Boundary Rules	Length Rules
5-char	Pos 2: +0.8	2-char: +0.1
	Pos 3: -0.3	1-char: -0.05
7-char	Pos 2, 4: +0.5	2-char: +0.1
	Pos 3, 5: -0.3	1-char: -0.05
		≥ 4 -char: -0.1
General	-	2-char: +0.05
		1-char: -0.05

Table 4: Metrical rhythm reward and penalty rules.

Dictionary Scoring Formula. The total score for a segmentation path is defined as $\text{Score} = \sum_{w_i \in \text{path}} \text{Score}_{\text{dict}}(w_i) + \text{Bonus}_{\text{meter}}(w_i)$. The base dictionary score $\text{Score}_{\text{dict}}(w)$ for a word w of length L integrates multiple empirical lexical features:

$$\begin{aligned} \text{Score}_{\text{dict}}(w) = & \log(1 + \text{freq}(w)) + 0.8 \cdot \text{MI}_{\text{norm}}(w) \\ & + 0.3 \cdot \text{DictCount}(w) + \text{Allusion}(w) \\ & + \text{LenPref}(L) - \text{OOV_Penalty}, \end{aligned}$$

where MI_{norm} represents normalized mutual information, and DictCount tracks occurrence counts across historical dictionaries. $\text{Allusion}(w)$ adds +1.0 for established literary references. The length preference term is defined as $\text{LenPref}(L) = \{+0.5 \text{ for } L = 2, +0.2 \text{ for } L = 3, -0.2 \text{ for } L \geq 4\}$. Out-of-vocabulary (OOV) tokens receive a base score of -1.0 and an additional penalty of -0.1.

Metrical Constraints. To simulate natural classical Chinese recitation rhythms (2 + 3 for 5-character and 2 + 2 + 3 for 7-character lines), position- and length-specific bonuses ($\text{Bonus}_{\text{meter}}$) are applied directly following the operational rules defined in Table 4.

G-Seg Specification for LLM Reranking. In Stage 2, the LLM evaluates the top- K candidates via few-shot prompts under strict linguistic guidelines: (1) functional words (e.g., particles, adverbs, conjunctions) must stand alone as single characters whenever possible; (2) tightly coupled semantic entities (names like “Huang Siniang”, places) must be merged into single tokens; (3) contextual flexibility takes precedence over blindly preferring longer words. If no available candidate is grammatically sound, the LLM triggers a WARNING fallback and outputs its self-corrected proper segmentation sequence.

C Ancient Encyclopedias Database

To provide well-grounded philological and linguistic suggestions during the *Tuiqiao* process, our

Book	Description	Content Example (Extracted Data)
《北堂书钞》 <i>Beitang Shuchao</i>	Compiled by Yu Shinan (Tang). An early extant encyclopedia focusing on politics and rituals, preserving numerous pre-Sui texts.	Keyword: 功业(Achievements) Volume: 4 Section: 帝王(Emperors) Content: “四本具即帝初立，举而措之事业，功业赫赫，功盛德厚，功侔太古...”
《白孔六帖》 <i>Baikong Liutie</i>	Compiled by Bai Juyi (Tang) and Kong Chuan (Song). Collects phrases and sentences as source materials for poetry and essays.	Keyword: 天(Heaven) Volume: 1 Content: “(白)高明柔克(明天也柔克寒暑不干)，阴鹭下民(言天定下民之命)...”
《太平御览》 <i>Taiping Yulan</i>	A massive state-sponsored encyclopedia from the early Song Dynasty. It is highly comprehensive, systematically categorized by heaven, earth, human, objects, etc.	Keyword: 太初(Vital Energy) Volume: 1 Section: 天部一(Heaven) Content: “《易乾鉴度》曰：太初者，气之始也。《帝王世纪》曰：元气始萌，谓之太初...”
《艺文类聚》 <i>Yiwen Leiju</i>	Compiled by Ouyang Xun (Tang). A state-sponsored encyclopedia intertwining historical facts and literature (poems and prose) under a clear Confucian orthodox viewpoint.	Keyword: 日(Sun) Volume: 1 Ref Word: 天部上(Heaven) Content: “《易》曰：日月丽乎天。又曰：离为日。又曰：日中则昃，月盈则食。天地盈虚，与时消息。而况于人乎，况于鬼神乎...”
《初学记》 <i>Chuxueji</i>	Compiled by Xu Jian (Tang). An introductory encyclopedia strictly organized by chapters (heaven, earth, etc.), drawing from historical literature and preserving ancient fragments.	Keyword: 日(Sun) Volume: 1 Section: 天部(Heaven) Content: 叙事(Narrative): “说文云日者实也...”; 事对(Events): 丽天&出地 (易曰日月丽乎天百谷草木丽乎地文子曰日出于地...); 诗文(Poetry): “梁简文帝咏朝日诗 (团团出天外煜煜上层峰光随浪高下影逐树轻浓) ...”
《骈字类编》 <i>Pianzi Leibian</i>	Compiled by Shen Zongjing et al. under imperial order (Qing). It collects disyllabic compound words (<i>Pianzi</i>) with sources for phrasing reference.	Keyword: 天地(Heaven and Earth) Volume: 1 Section: 天(Heaven) Content: “易干夫大人者与天地合其德，又坤天地变化草木蕃天地闭贤人隐...”
《海录碎事》 <i>Hailu Suishi</i>	Compiled by Ye Tinggui (Song). Supplements missing allusions from larger encyclopedias with concise entries of Tang poems and prose.	Keyword: 天末(Horizon) Volume: 1 Section: 天(Heaven) Content: “沧波渺川汜白日隐天末 (李白诗) ...”
《佩文韵府》 <i>Peiwen Yunfu</i>	An imperially compiled Qing rhyme dictionary arranged by 106 <i>Pingshui Yun</i> categories. It details characters in each group and provides phonological analysis for writers.	Keyword: 东(East) Volume: 01之一 Tones: 上平声(Shangping) Rhyme: 一东(Dong)(东德红切眷方也...) Words of Rhyme: 自东(诗我来自东又自西自东) 大东(诗遂荒大东)... Words of Pairs: 渭北&江东，日下&天东，河内&济东... Excerpt: 力障百川东，光升必自东...
《声律启蒙》 <i>Shenglü Qimeng</i>	A Qing rhyme primer by Che Wanyu focusing on antithesis training.	Antithesis Pairs: “云” (Cloud) vs. “雨” (Rain); “雪” (Snow) vs. “风” (Wind)
《笠翁对韵》 <i>Liweng Duiyun</i>	A Qing rhyme primer by Li Yu for antithesis and allusion training.	Antithesis Pairs: “天” (Heaven) vs. “地” (Earth); “雨” (Rain) vs. “风” (Wind)
《龙文鞭影》 <i>Longwen Bianying</i>	A Ming primer by Xiao Liangyou utilizing four-character sentences to catalog historical names, places, and allusions.	Antithesis Pairs: “粗成四字” vs. “诲尔童蒙”

Table 5: Overview and data samples of the 11 utilized classical sourcebooks.

system constructs a structured knowledge base by extracting textual data from over ten representative classical Chinese encyclopedias (*Leishu*) and rhyme dictionaries spanning the Tang to Qing dynasties (e.g., *Chuxueji*, *Taiping Yulan*, and *Peiwen Yunfu*). Table 5 outlines the basic profiles and concrete database entry examples of these primary resources.

Data Processing and Knowledge Base Construction. Texts from the *Siku Quanshu* editions and verified open-source assets were systematically parsed, cleaned, and rule-filtered to unify disparate historical formats. The consolidated pipeline yields 503,908 structured records written across three distinct SQLite index tables to serve as explicit etymological anchors:

(1) Image Word Database (tb_danci): Aggregates 6 sourcebooks (e.g., *Beitang Shuchao*, *Taiping Yulan*) to map a central thematic keyword (key_word) to extensive imagery tokens historically favored by poets, attaching original definitions and cited literature contexts (content) as

conceptual justifications during theme refinement.

(2) Antithesis Word Database (tb_duizhang): Combines 5 phonological books (e.g., *Chuxueji*, *Shenglü Qimeng*) by extracting all original antithetical couplets (gen_word_f and gen_word_b) into a parallel-pair database. The system dynamically recommends position-specific antithetical words calibrated against the corresponding characters in parallel lines, returning classical source texts as reference.

(3) Rhyme Word Database (tb_yunyu): Specially processes texts structured around rhyme schemas (primarily *Peiwen Yunfu*). It isolates the leading characters of *Pingshui Yun* categories (yun_word) and their corresponding phrases (gen_word) to supply rhyme-compliant generation candidates.

Ultimately, these verified text streams are not only injected into the language model as internal structure constraints (Constraint Inference Layer) but are also rendered explicitly to users via interactive cards, securing high interpretability and strict knowledge tracing for *Jiuge-Tuiqiao* framework.