Form-aware Poetic Generation for Bangla

Amina, Abdullah, Mueeze Al Mushabbir, Sabbir Ahmed
Department of Computer Science and Engineering
Islamic University of Technology, Bangladesh
{amina, abdullah20, almushabbir, sabbirahmed}@iut-dhaka.edu

Abstract

Poetry generation in low-resource languages such as Bangla is particularly challenging due to the scarcity of structured poetic corpora and the complexity of its metrical system (*matra*). We present a structure-aware framework for Bangla poetry generation using pretrained Bangla large language models (LLMs)–TigerLLM, TituLLM, and BanglaT5–trained on general non-poetic text corpora augmented with rich structural control tokens. These tokens capture rhyme, meter, word count, and line boundaries, enabling unsupervised modeling of poetic form without curated poetry datasets. Unlike prior fixed-pattern approaches, our framework introduces variable control compositions, allowing models to generate flexible poetic structures. Experiments show that explicit structural conditioning improves rhyme consistency and metrical balance while maintaining semantic coherence. Our study provides the first systematic evaluation of Bangla LLMs for form-constrained creative generation, offering insights into structural representation in low-resource poetic modeling.

1 Introduction

Poetry generation balances semantic fluency with formal constraints such as rhyme, meter, and line length (Mahbub et al., 2023; Hu et al., 2024). For low-resource languages like Bangla, this is challenging due to scarce curated poetic corpora and a complex *matra*-based prosody (Pakray et al., 2025).

Prior work, such as PoeLM (Ormazabal et al., 2022), showed that transformer LMs can learn rhyme and metrical patterns from general corpora augmented with structural codes, but relied on fixed configurations and focused on Spanish and Basque.

Bangla's phonetic and morphological complexity, coupled with limited poetic datasets and pretrained LLMs evaluated mainly on prose, motivates our use of non-poetic text with variable control tokens (<RYM>, <MTR>, <WRD>, <STA>, <FIN>) to enable structure-aware poetry generation.

Our contributions are: (1) a framework for unsupervised Bangla poetry generation with variable control tokens; (2) systematic evaluation of pretrained Bangla LLMs (TigerLLM (Raihan and Zampieri, 2025), TituLLM (Nahin et al., 2025), BanglaT5 (Bhattacharjee et al., 2023)) on form-constrained generation; (3) insights into how structural conditioning improves adherence to poetic form while maintaining coherence.

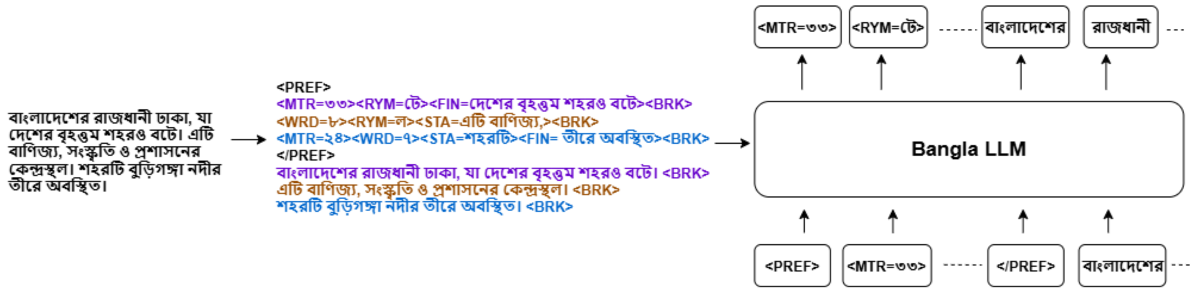
We show that pretrained Bangla LLMs can generate metrically consistent, rhymed, and semantically coherent verses under structural guidance, providing a foundation for further research in this domain.

2 Related Works

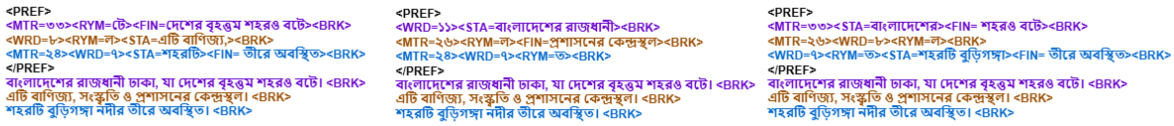
Modern advances in large language models (LLMs) have significantly reshaped natural language processing, enabling systems to achieve strong performance across tasks ranging from text generation and summarization to reasoning and dialogue (Abrar et al., 2024; Arif et al., 2025; Khan et al., 2023b; Ahmed et al., 2024; Khan et al., 2023a). These developments have also opened new possibilities for creative text generation, where models learn not only linguistic fluency but also stylistic, structural, and domain-specific patterns (Shedko, 2018; Gonçalo Oliveira, 2024).

Poetry generation in NLP has evolved from rule-based and template systems (Das, 2014), which enforced form but lacked fluency, to neural approaches using LSTMs and transformers for both free and structured verse (Ghazvininejad et al.,

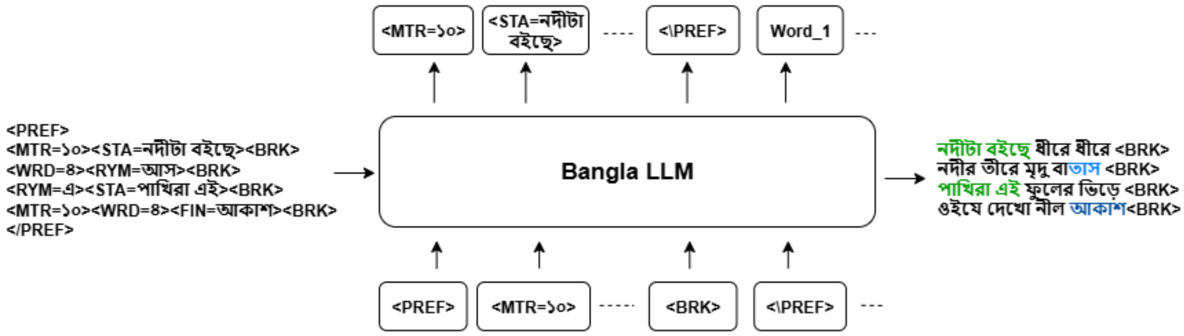
Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 366–372
December 23, 2025



(a) Training on regular, non-poetic Bangla text augmented with control tokens.



(b) Applying different control token combinations for the same sample.



(c) Controlled generation under mixed guidance tokens.

Figure 1: Overview of the proposed Bangla poetry generation framework: (a) During training, regular Bangla text is segmented and augmented with structural control tokens, enabling the model to associate textual patterns with explicit poetic constraints. (b) At inference time, multiple combinations of control tokens are applied to the same input, allowing flexible specification of desired poetic form. (c) The fine-tuned model generates verse conditioned on these tokens, producing lines that follow the requested structural layout while maintaining semantic continuity across the poem.

2016; Yang et al., 2023). These methods typically require large annotated poem corpora, limiting applicability in low-resource languages (Zhong et al., 2025).

In Bangla, Murad and Rahman (2023) and Chy et al. (2020) leveraged deep learning-based techniques to generate poet-style free-form poems but did not systematically enforce meter or rhyme. Ormazabal et al. (2022) introduced 'PoeLM', an unsupervised controlled verse generation with structural tokens for Spanish and Basque, yet relied on fixed patterns and simpler poetic attributes. Recent LLMs tuned for Bangla language, including TigerLLM (Raihan and Zampieri, 2025), TituLLMs (Nahin et al., 2025), and BanglaT5 (Bhattacharjee et al., 2023), have mostly been evaluated on prose. Studies on Bangla meter (Ahmed et al., 2023) provide a basis for controlled generation.

Our work extends these ideas by using variable control tokens for rhyme, meter, word count, and line boundaries, enabling flexible, unsupervised Bangla poetry generation from non-poetic corpora and evaluation of pretrained LLMs in this setting.

3 Methodology

3.1 Overview

Our work extends PoeLM (Ormazabal et al., 2022) to Bangla by developing a structure-aware poetry generation framework, shown in Figure 1, that uses control tokens for rhyme, meter, word count, and line boundaries. Instead of curated poetry datasets, we repurpose general Bangla text into poem-like training samples with these tokens, enabling unsupervised learning of poetic form and generation of verses that preserve rhythmic consistency and semantic coherence under flexible structural guidance.

3.2 Corpus Acquisition and Preprocessing

Due to the scarcity of annotated Bangla poetry corpora, we used the Bangla portion of the OSCAR 2019 corpus (OSCAR bn) (Ortiz Suárez et al., 2020). OSCAR 2019 was generated from the Common Crawl using the goclassy architecture, containing noisy text from multiple languages. The Bangla sub-corpus contains both original and deduplicated versions. We use the deduplicated portion, which has 363,766,143 words (5.8 GB), to ensure that repeated lines do not bias model training.

The raw text is further cleaned to remove non-Bangla content, extraneous symbols, and malformed sentences. From the cleaned corpus, we extract coherent multi-line segments to serve as poem-like units. Sample creation follows these steps:

1. Choose a random starting point in corpus.
2. Select a variable number of consecutive lines for the segment.
3. Vary word lengths within each line randomly, avoiding single-word lines.
4. Assign structural control tokens to each multi-line segment.

This process preserves local context and allows the model to learn metrically and rhymically coherent text across multiple lines.

3.3 Control Tokens

Each multiline segment is annotated with control tokens to provide explicit structural guidance:

- **Rhyme tokens** (<RYM=...>): encode the rhyme class of the last syllable.
  Example: <RYM=অয়> → line must end with rhyme "অয়".
- **Meter tokens** (<MTR=...>): specify the number of meter/*matra* in a line.
  Example: <MTR=২৬> → line must contain 26 *matra*.
- **Word count tokens** (<WRD=...>): enforce the target number of words.
  Example: <WRD=৯> → line must contain exactly 9 words.
- **Start** (<STA=...>) **and End** (<FIN=...>) **tokens**: encode line fragments for guidance.
  Example: <STA= আমি আজ> → line must start with "আমি আজ ...",
  <FIN= ভোরের আলোয়> → line must end with "... ভোরের আলোয়".

3.4 Adaptive Control Token Schema

Control tokens are embedded directly into the training text, allowing the model to learn them as part of its vocabulary (Figure 1a). Unlike fixed-pattern approaches such as PoeLM (Ormazabal et al., 2022), each training instance can include any subset of available tokens–rhyme, meter, word count, start, and end; covering all possible combinations.

Exposure to diverse token combinations encourages generalization across varying levels of structural guidance (Figure 1b). This design supports generation under full, partial, or mixed prompts while preserving semantic coherence and poetic form.

3.5 Model Training

We adopt a next-token prediction approach where the model is conceptually conditioned on control tokens and optional start/end fragments. This allows the model to learn form-aware generation and handle flexible prompts, supporting rhyme, meter, and word count constraints at a high level.

3.6 Generation with Filtered Re-ranking

The generation process happens in three steps to ensure structural adherence and fluency:

Candidate Generation: The trained model generates multiple candidate verses conditioned on user-specified control tokens. Sampling strategies like top-k or nucleus sampling are used to ensure diversity.

Structural Filtering: Candidates are filtered to retain only those satisfying the structural constraints:

- Word count matches the target.
- Meter (*matra*) approximates the specified value.
- Rhyme ends with the required syllable.
- Start and end fragments are respected.
- Repetition of rhyming words is avoided.

Re-ranking: Filtered candidates are scored for fluency using model likelihoods. The highest-scoring candidate, balancing structural correctness and semantic coherence, is selected as the final output.

4 Experimental Setup

4.1 Training Strategy

We fine-tuned three pretrained Bangla LLMs–TigerLLM (Raihan and Zampieri, 2025), TituLLM (Nahin et al., 2025), and BanglaT5 (Bhattacharjee et al., 2023)–on descriptor-augmented multi-line segments. Each training sample consisted of a prefix of control tokens and optional start/end fragments, with the model predicting the remaining tokens.

4.2 Evaluation Metrics

We measured both structural adherence and semantic coherence:

- **Rhyme correctness**: whether generated lines end with the specified rhyme.
- **Word-count correctness**: whether the number of words matches the target.
- **Meter correctness**: whether syllable counts match the specified *matra*.
- **Start/End correctness**: whether start and end fragments are respected.
- **Overall structure following**: whether all structural constraints are consistently met across the poem.
- **Text coherence**: grammatical validity and semantic consistency.

Rhyme, meter, word-count, start/end correctness, and overall structure following were evaluated using automatic validators, while coherence was assessed through manual qualitative inspection.

4.3 Human Evaluation Criteria

To complement the large-scale automatic structure estimates, we conducted a focused human evaluation to assess the coherence of generated poems. Three native Bangla speakers with experience reading both contemporary and classical Bangla poetry served as annotators. A stratified sample of model outputs was constructed to cover all model variants. Annotators independently rated each poem along the two components of coherence defined in our evaluation protocol:

- **Grammatical validity**: syntactic well-formedness and naturalness of phrasing.
- **Semantic consistency**: clarity of meaning within and across lines, and whether poem maintains a coherent thematic flow.

Both criteria were rated on a 5-point Likert scale. Written guidelines ensured a shared interpretation of the criteria. Samples were divided among annotators with partial overlap to allow reliability estimation; ratings were collected independently and averaged. Inter-annotator consistency was checked using a standard reliability measure for ordinal judgments and showed reasonable agreement. For reporting, we compute a **combined human coherence score** by averaging the grammatical and semantic ratings.

5 Results and Analysis

5.1 Quantitative Evaluation

Table 1 reports both baseline and finetuned performance of pretrained Bangla LLMs across structural and semantic metrics. Scores denote the percentage of generated lines satisfying each constraint (rhyme, word-count, meter, and boundary tokens), along with overall textual coherence.

Across all models, baseline scores remain low for structural metrics, reflecting that pretrained LLMs–without control-token finetuning–cannot reliably follow poetic form. However, coherence remains relatively high even in baseline generations, indicating that the models already possess strong linguistic fluency. Finetuning with structural control tokens yields substantial improvements: TigerLLM shows the strongest gains, achieving over 90% accuracy on all structural categories. TituLLM also improves consistently, though with slightly lower meter precision. BanglaT5 benefits from finetuning as well, but lags behind the larger models in structure fidelity. Overall, these results demonstrate that explicit structural conditioning is highly effective in enabling Bangla LLMs to generate form-consistent poetry.

5.2 Qualitative Evaluation

Qualitative inspection shows that variable control-token fine-tuning helps maintain coherence across lines, ensuring each verse flows naturally while respecting the specified rhyme, meter, word-count, and boundary constraints. Baseline generations, by contrast, often produce disjointed lines that, although locally fluent, fail to maintain continuity across the segment. Table 2 presents high-ranked

Table 1: Comparison of baseline (no control-token fine-tuning) and control-token–fine-tuned versions of each pretrained Bangla LLM.

Model / Setting	**Rhyme**	**Word-count**	**Meter**	**Start/End**	**Structure**	**Coherence**
TigerLLM (baseline)	42%	61%	45%	50%	51%	86%
TigerLLM (finetuned)	95%	97%	93%	96%	95%	94%
TituLLM (baseline)	28%	48%	33%	36%	36%	77%
TituLLM (finetuned)	93%	95%	88%	94%	91%	90%
BanglaT5 (baseline)	25%	42%	28%	32%	31%	72%
BanglaT5 (finetuned)	85%	87%	80%	82%	83%	88%

verse generated by each model after filtered re-ranking from prompts illustrated in Figure 2.

আমাকে তিন লাইনের একটি কবিতা লিখে দাও যার, প্রথম লাইনের মোট মাত্রা সংখ্যা ১০, লাইনের শেষ অংশ 'ঘাস'; দ্বিতীয় লাইনের মোট মাত্রা সংখ্যা ১০, লাইনের শুরুর অংশ 'চোখ', লাইনের শেষ অংশ 'চারপাশ'; তৃতীয় লাইনের শেষ ছন্দ 'আস', মোট শব্দ সংখ্যা ৫।

(a) Prompt format for baseline models

<PREF>
<MTR=১০><FIN=ঘাস><WRD=৫><BRK>
<MTR=১০><STA=চোখ><FIN=চারপাশ><BRK>
<RYM=আস><WRD=৫><BRK>
</PREF>

(b) Prompt format for fine-tuned models

Figure 2: Example of a test sample for qualitative evaluation.

These examples highlight that control tokens not only enforce formal constraints but also support semantic connectedness across lines, demonstrating their utility for producing coherent Bangla poetry beyond mere structural compliance.

6 Conclusion

We presented a framework to generate Bangla poetry guided by control tokens for rhyme, meter, word count, and line boundaries. Our models achieve high structural adherence while maintaining semantic coherence by fine-tuning pretrained Bangla LLMs on descriptor-augmented multi-line segments. Both quantitative and qualitative evaluations confirm the effectiveness of our approach, highlighting the potential of control-token-guided generation for creative text. Future work could focus on improving context-awareness, enabling finer-grained control, and better balancing structural fidelity with semantic coherence.

7 Limitations

While our framework successfully generates form-aware Bangla poetry that adheres to rhyme, meter, and word-count constraints, several limitations remain. Controllability is still imperfect: the model may drift from specified constraints when multiple or conflicting control tokens are applied, and longer generations can exhibit weakened semantic coherence or fragmented narrative flow. Moreover, although the system effectively enforces structural attributes, it does not explicitly capture deeper stylistic properties—such as tone, imagery, or poet-specific voice—which can lead to verses that feel less expressive or stylistically homogeneous. Finally, since our work builds incrementally on the ideas introduced in PoeLM (Ormazabal et al., 2022), a more comprehensive investigation

Table 2: High-ranked verse generated from each baseline (-bs) and fine-tuned (-ft) Bangla LLMs for sample prompts in Figure 2.

Model	**Sample Generated Verse**
TigerLLM-bs	আসে রোশনির চারপাশ করে সবুজ ঘাস, চোখ দিয়ে দেখি সূর্যের চারিপাশ, চারপাশের ঘাসে লুকিয়ে আছে আনন্দ।
TigerLLM-ft	সকালের রোদে ভিজে ওঠে ঘাস, চোখ রেখে দেখি নরম চারপাশ আজ বাতাসে ভাসে নতুন উচ্ছ্বাস।
TituLLM-bs	তিন লাইনের একটি কবিতা শেষে ঘাস, চোখ দিয়ে শুরু মাত্রা দশ শেষে ছন্দ আস আস আস।
TituLLM-ft	পাহাড়ের ঢালে সবুজ নরম ঘাস চোখ মেলে দেখি চারপাশ, হাওয়ায় ভাসে শান্ত নীল বিশ্বাস।
BanglaT5-bs	মাত্রা মাত্রা ঘাস লিখো চোখ লিখো চারপাশ লিখো আস ছন্দ শব্দ লিখো
BanglaT5-ft	সকালের শিশির ভেজা নরম ঘাসে। চোখে দেখি দূরের পাহাড়। আজ বাতাস খুব নরম ঘাস।

comparing stylistic behavior and constraint robustness across both approaches remains an important direction for future work.

References

Ajwad Abrar, Farzana Tabassum, and Sabbir Ahmed. 2024. Performance evaluation of large language models in bangla consumer health query summarization. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 2748–2753.

Nasim Ahmed, Sheikh Taraque Aziz, Md. Abu Naser Mojumder, and Maruf Ahmed Mridul. 2023. Automatic classification of meter in bangla poems: A machine learning approach. In *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–5.

Tasnim Ahmed, Shahriar Ivan, Ahnaf Munir, and Sabbir Ahmed. 2024. Decoding depression: Analyzing social network insights for depression severity assessment with transformers and explainable ai. *Natural Language Processing Journal*, 7:100079.

Nokimul Hasan Arif, Shadman Rabby, Md Hefzul Hossain Papon, and Sabbir Ahmed. 2025. Preemptive hallucination reduction: An input-level approach for multimodal language model. *Preprint*, arXiv:2505.24007.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla.

Md. Kalim Amzad Chy, Md. Abdur Rahman, Abdul Kadar Muhammad Masum, Shayhan Ameen Chowdhury, Md. Golam Robiul Alam, and Md. Shahidul Islam Khan. 2020. Bengali poem generation using deep learning approach. In *Intelligent Computing Paradigm and Cutting-edge Technologies*, pages 148–157.

Amitava Das. 2014. Poetic machine: Computational creativity for automatic poetry generation. In *Proceedings of the 2014 International Conference on Computational Creativity*, pages 11–18.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191, Austin, Texas. Association for Computational Linguistics.

Hugo Gonçalo Oliveira. 2024. Automatic generation of creative text in portuguese: an overview. *Language Resources and Evaluation*, 58(1):7–41.

Zhiyuan Hu, Chumin Liu, Yue Feng, Anh Tuan Luu, and Bryan Hooi. 2024. Poetrydiffusion: Towards joint semantic and metrical manipulation in poetry generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18279–18288.

Alvi Khan, Fida Kamal, Mohammad Abrar Chowdhury, Tasnim Ahmed, Md Tahmid Rahman Laskar, and Sabbir Ahmed. 2023a. BanglaCHQ-summ: An abstractive summarization dataset for medical queries in Bangla conversational speech. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 85–93, Singapore. Association for Computational Linguistics.

Alvi Khan, Fida Kamal, Nuzhat Nower, Tasnim Ahmed, Sabbir Ahmed, and Tareque Chowdhury. 2023b. NERvous about my health: Constructing a Bengali medical named entity recognition dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5768–5774, Singapore. Association for Computational Linguistics.

Ridwan Mahbub, Ifrad Khan, Samiha Anuva, Md Shihab Shahriar, Md Tahmid Rahman Laskar, and Sabbir Ahmed. 2023. Unveiling the essence of poetry: Introducing a comprehensive dataset and benchmark for poem summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14878–14886, Singapore. Association for Computational Linguistics.

Hasan Murad and Rashik Rahman. 2023. Ai poet: A deep learning based approach to generate artificial poetry in bangla. In *Applied Intelligence for Industry 4.0*. CRC Press.

Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, and Firoj Alam. 2025. TituLLMs: A family of Bangla LLMs with comprehensive benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24922–24940, Vienna, Austria. Association for Computational Linguistics.

Aitor Ormazabal, Mikel Artetxe, Manex Agirrezabal, Aitor Soroa, and Eneko Agirre. 2022. Poelm: A meter- and rhyme-controllable language model for unsupervised poetry generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 268–278.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Nishat Raihan and Marcos Zampieri. 2025. TigerLLM - a family of Bangla large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 887–896, Vienna, Austria. Association for Computational Linguistics.

Andrey Y. Shedko. 2018. Semantic-map-based assistant for creative text generation. *Procedia Computer Science*, 123:446–450. 8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017 (Eighth Annual Meeting of the BICA Society), held August 1-6, 2017 in Moscow, Russia.

Liang Yang, Zhexu Shen, Fengqing Zhou, Hongfei Lin, and Junpeng Li. 2023. Tpoet: Topic-enhanced chinese poetry generation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2025. Opportunities and challenges of large language models for low-resource languages in humanities research. *Preprint*, arXiv:2412.04497.