System Report for CCL25-Eval Task 5: New Dataset and LoRA-Fine-Tuned Qwen2.5

Haotao Xie
The Hangzhou International Innovation Institute
Beihang University
China
1571855546@qq.com

Abstract

Recently, large language models (LLMs) have achieved promising progress in the fields of classical Chinese translation and the generation of classical poetry. However, domain-specific research on precise translation and affective-semantic understanding of classical poetry remains limited. The main challenge is that most studies treat the poetic appreciation task as a general-domain problem, neglecting the distinctive features of poetic appreciation, while high-quality and domain-specific datasets are extremely limited. To address this limitation, we decompose the task into three subtasks: term interpretation, semantic interpretation, and emotional inference. Based on multiple open-source datasets, we perform data cleansing and alignment to construct the Classical Chinese Poetry Instruction Pair Dataset (**CCPoetry-49K**), which comprises **49,404** high-quality instruction–response pairs explicitly optimized for this domain. We then propose a domain-specialized LLM, called **PoetryQwen**, by applying Low-Rank Adaptation (**LoRA**) to fine-tune the Qwen2.5-14B model. Experimental results on the CCL25-Eval Task 5 benchmark demonstrate that PoetryQwen achieves a score of **0.757**, representing a **9.7%** improvement over the Qwen2.5-14B-Instruct baseline (**0.690**). These findings clearly indicate that PoetryQwen significantly enhances performance in precise translation and emotional understanding of classical poetry. We present new dataset and methodological considerations intended to support the domain-specific optimization of LLMs.

Keywords: Classical Chinese Poetry , Instruction Dataset , Low-Rank Adaptation , Qwen

1 Introduction

Classical Chinese poetry is deeply rooted in historical and cultural contexts, often requiring knowledge of dynasties, the lives of poets, and historical events to be meaningful. The advent of LLMs offers innovative and effective solutions to this complexity, enabling more accessible understanding and lowering the barrier to learning classical poetry. However, the majority of existing study focus on classical Chinese translation(Liu et al., 2019) and poetry generation(Zhipeng et al., 2019; Liu et al., 2020; Zhang and Eger, 2024; Yu et al., 2024), while largely overlooking the importance of systematic evaluation and interpretive assessment of classical poetry.

Furthermore, the evaluation of classical poetry is often approached as a generic NLP problem, overlooking its domain-specific challenges. This, combined with the limited availability of specialized datasets, has largely hindered the development of effective evaluation models in the field. Therefore, it is necessary to construct domain-specific datasets and develop dedicated LLMs tailored to the unique characteristics and interpretive demands of classical Chinese poetry. Such efforts would better align with the intrinsic needs of this domain and enable more accurate, context-aware understanding and evaluation of classical poetry. In this study, inspired by the evaluation dimensions outlined in the CCL25-Eval Task 5 benchmark(Chen et al., 2024), we frame classical Chinese poetry appreciation as an evaluation-oriented task and decompose it into three subtasks. To support this, we construct CCPoetry-49K, a

Name	Number of Term Interpretation	Number of Semantic Interpretation	Number of Emotional Inference	Total
CCPoetry-49K	19,323	21,799	8,282	49,404

Table 1: Statistical Distribution of Multi-task Annotations in CCPoetry-49K

high-quality instruction dataset with 49K aligned samples derived from multiple open-source sources. And we apply LoRA(Hu et al., 2022) to fine-tune the Qwen2.5-14B model[1], resulting in PoetryQwen, a domain-adapted model specifically designed for classical poetry understanding. Experimental results show substantial gains in both translation accuracy and affective-semantic interpretation.

In summary, our main contributions are as follows:

- **Task Formulation:** Building upon the structure of CCL25-Eval Task 5, we adopt an evaluation framework for classical Chinese poetry that comprises three subtasks: term interpretation, semantic interpretation, and emotional inference.
- **Dataset Construction:** We build CCPoetry-49K, a 49K-sample instruction dataset tailored to poetry evaluation, through large-scale data cleaning and alignment from open-source corpora.
- **Model Development:** We propose PoetryQwen, a domain-specific model based on Qwen2.5-14B using LoRA fine-tuning, which achieves strong performance on classical poetry appreciation tasks.

2 Related works

2.1 Datasets for Classical Chinese Poetry

High-quality datasets are essential for advancing the study of classical Chinese poetry using LLMs. However, existing datasets largely focus on poetry generation or translation, with limited support for multi-faceted interpretation and evaluation.

For instance, Chen (2019) constructed a sentiment-labeled poetry corpus to support sentiment-controllable poetry generation, but the dataset is tailored to generation rather than comprehension or evaluation tasks.

In the domain of classical-modern Chinese translation, Liu (2019) developed a large parallel corpus that supports machine translation tasks. This work provides valuable aligned data, but lacks the interpretive and affective dimensions necessary for poetry understanding. The CCPM dataset(Li et al., 2021) focuses on poetry matching, aligning classical poems with their modern Chinese translations, which offers a semantic bridge, yet still not cover term-level or emotional interpretation.

More recently, WenMind(Cao et al., 2024) and WYWEB(Zhou et al., 2023) introduced comprehensive benchmarks spanning a variety of classical Chinese tasks. WenMind includes sub-domains such as ancient prose and poetry, while WYWEB encompasses nine tasks including translation, classification, and comprehension. These benchmarks are significant steps toward domain-wide evaluation but still lack a focused dataset specifically designed for poetry appreciation tasks.

To address this gap, our work presents CCPoetry-49K, a high-quality instruction–response dataset constructed via careful alignment and cleansing from multiple open-source datasets. It uniquely supports three critical subtasks: term interpretation, semantic interpretation, and emotional inference—inspired by the evaluation structure of CCL25-Eval Task 5[2]. This dataset aims to fill the void of fine-grained, domain-specific resources for classical Chinese poetry understanding and model evaluation.

2.2 LLMs for Classical Chinese Poetry appreciation Tasks

While recent studies have made promising progress in applying LLMs to classical Chinese poetry, their focus has primarily been on poetry generation. Systems such as Jiuge(Zhipeng et al., 2019), Deep

[1] https://huggingface.co/Qwen/Qwen2.5-14B

[2] https://tianchi.aliyun.com/competition/entrance/532345

Proceedings of the 24th China National Conference on Computational Linguistics, pages 200-205, Jinan, China, August 11-14, 2025.

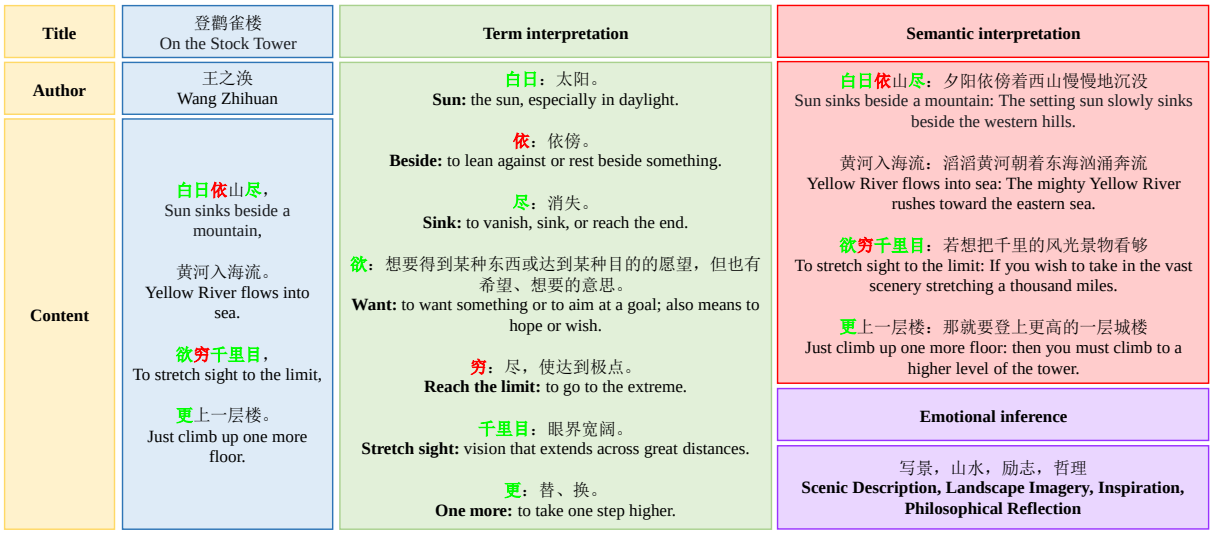

Figure 1: An example extracted from multiple open-source datasets after data cleansing and alignment, including the Title, Author, Content, Term Interpretation, Semantic Interpretation, and Emotional Inference.

Poetry(Liu et al., 2020), and CharPoet(Yu et al., 2024) enable creative generation via neural or token-free architectures, often with multimodal or user-guided input. Multi-agent generation approaches(Zhang and Eger, 2024) further enrich diversity and novelty.

However, poetry appreciation tasks remain underexplored. Most existing models treat poetry as a general NLP domain, overlooking its linguistic, historical, and affective complexity. Moreover, there is a notable lack of datasets and models specifically tailored for interpretive evaluation of classical poetry.

In contrast, our work focuses on evaluative understanding. Based on the structure of CCL25-Eval Task 5, we decompose the appreciation task into three subtasks and construct a high-quality instruction dataset (CCPoetry-49K). We also introduce PoetryQwen, a domain-specialized model fine-tuned via LoRA, marking a novel contribution to LLM-based poetry comprehension.

3 CCPoetry-49K Dataset

3.1 Source Datasets for Poetry Appreciation

To construct a domain-specific dataset for classical Chinese poetry appreciation, we began by collecting and analyzing several open-source datasets that contain partial annotations related to poetry understanding. These datasets were primarily crawled from public educational websites and open-access online platforms, and they provide information such as modern translations, cultural term explanations, and basic sentiment annotations. However, they vary significantly in structure, granularity, and annotation quality, posing challenges for direct use in instruction tuning.

The primary datasets utilized in this study include:

Poetry CN[3]**:** The dataset(He et al., 2024) is sourced from website[4], which compiles classical Chinese poetry along with translations, annotations, and commentary.

Chinese ancient poetry translation[5]**:** This dataset comprises aligned pairs of classical Chinese poetic lines and their corresponding human-authored modern Chinese translations.

poems-db[6]**:** Sourced from website[7], the dataset includes over 220,000 classical Chinese poems with annotations, commentaries, metadata on 10,000+ poets, 1,600+ poetic forms, 70+ dynasties, and nearly 200 thematic categories.

[3] https://opendatalab.com/ABear/Poetry_CN
[4] https://www.gushici.net/
[5] https://github.com/YuRuiii/chinese-ancient-poetry-translation
[6] https://github.com/yxcs/poems-db
[7] https://www.gushici.net/

Task	Example
Term Interpretation	{'instruction': '\n 我会给你提供一个 JSON 数据，格式如下： \n - **"title"**：古诗词的标题 \n - **"content"**：古诗词的内容 \n - **"qa_words"**：古诗词中需要翻译的词语 \n\n 这是我的数据：{\'title\': \'听弹琴\', \'content\': \'泠泠七弦上，静听松风寒。古调虽自爱，今人多不弹。\', \'qa_words\': \'**泠（líng）泠**\'}\n\n ### 你的任务：\n 请你根据提供的数据，生成如下格式的结果：\n - **"ans_qa_words"**：对 "qa_words" 词语的含义进行解释 \n\n ### **输出格式示例：**\n "ans_qa_words": \n "词语": "词语的含义"\n\n ', 'input': '', 'output': '**形容清凉、清淡，也形容声音清越。**'}
Semantic Interpretation	{'instruction': '\n 我会给你提供一个 JSON 数据，格式如下： \n - **"title"**：古诗词的标题 \n - **"content"**：古诗词的内容 \n - **"qa_sents"**：古诗词中需要提供白话文译文的句子 \n\n 这是我的数据：{\'title\': \'赠去婢\', \'content\': \'公子王孙逐后尘，绿珠垂泪滴罗巾。侯门一入深如海，从此萧郎是路人。\', \'qa_sents\': \'**侯门一入深如海**\'}\n\n ### 你的任务：\n 请你根据提供的数据，生成如下格式的结果：\n - **"ans_qa_sents"**：对 "qa_sents" 句子提供白话文译文 \n\n ### **输出格式示例：**\n "ans_qa_sents": \n "句子": "句子的白话文翻译"\n \n ###【输出要求】\n 注意 "ans_qa_sents"提供的古诗词句子是一整句还是半句，严格按照提供的句子进行翻译。\n\n ', 'input': '', 'output': '**一旦进入深幽如海的侯门**'}
Emotional Inference	{'instruction': '\n 我会给你提供一个 JSON 数据，格式如下：\n - **"title"**：古诗词的标题 \n - **"author"**：古诗词的作者 \n - **"content"**：古诗词的内容 \n\n 这是我的数据：{\'title\': \'京师得家书\', \'author\': \'袁凯\', \'content\': \'江水三千里，家书十五行。行行无别语，只道早还乡。\'}\n\n ### 你的任务：\n 请你根据提供的数据，生成如下格式的结果：\n [生成最符合该古诗词表达的情感的标签1,生成最符合该古诗词表达的情感的标签2,...]\n\n ### **json输出格式示例：**\n [\'怀古\', \'宫怨\']\n [\'思乡\', \'思亲\']\n [\'孤独\']\n\n ', 'input': '', 'output': ['**思乡**']}

Figure 2: Illustrative Subtasks Examples from the CCPoetry-49K Dataset: Term Interpretation, Semantic Interpretation, and Emotional Inference.

3.2 Construction of CCPoetry-49K: A Domain-Specific Instruction Dataset

Building on the open-source datasets introduced in **Section** 3.1, we perform comprehensive data cleansing and alignment to obtain a high-quality foundation for downstream tasks. **Figure** 1 illustrates a representative example that includes key elements such as the poem's title, author, content, term interpretation, semantic interpretation, and emotional inference. Based on this processed data, we formulate three subtasks—Term Interpretation, Semantic Interpretation, and Emotional Inference—and construct an instruction-tuning dataset tailored to each. **Figure** 2 provides illustrative examples for these subtasks from the proposed CCPoetry-49K Dataset, while **Table** 1 summarizes the number of instruction–response pairs per subtask, totaling 49,404 instances.

4 Experiments

In this section, we present the experimental setup, including model configurations, the base model used for fine-tuning, and the overall training and evaluation pipeline. We also report and analyze the experimental results to assess the effectiveness of our proposed approach.

4.1 PoetryQwen Model

Base Model. We adopt Qwen2.5-14B[8] as base model, with 14.7 billion parameters, which supports long-context modeling up to 128K tokens and excels in instruction following, multilingual understanding, and structured output generation.

LoRA Fine-tuning Setup. We apply LoRA with the following configuration: target modules include **q-proj, k-proj, v-proj, o-proj, gate-proj, up-proj** and **down-proj**; the maximum input length is set to **1240**; LoRA rank is **16**, with LoRA alpha set to **32** and dropout to **0.1**. Training is conducted for **2** epochs using a learning rate of **2e-4** and a fixed random seed of **42** to ensure reproducibility.

PoetryQwen Model. To address the three subtasks: Term Interpretation, Semantic Interpretation, and Emotional Inference, we fine-tune the Qwen2.5-14B base model using LoRA, resulting in three task-specific LoRA adapters. Each adapter is trained independently on its corresponding instruction dataset and evaluated on the respective subtask. Collectively, the base model and the three adapters constitute our proposed system, termed PoetryQwen, as illustrated in Figure 3. For the Emotional Inference subtask,

[8] https://qwenlm.github.io/blog/qwen2.5/

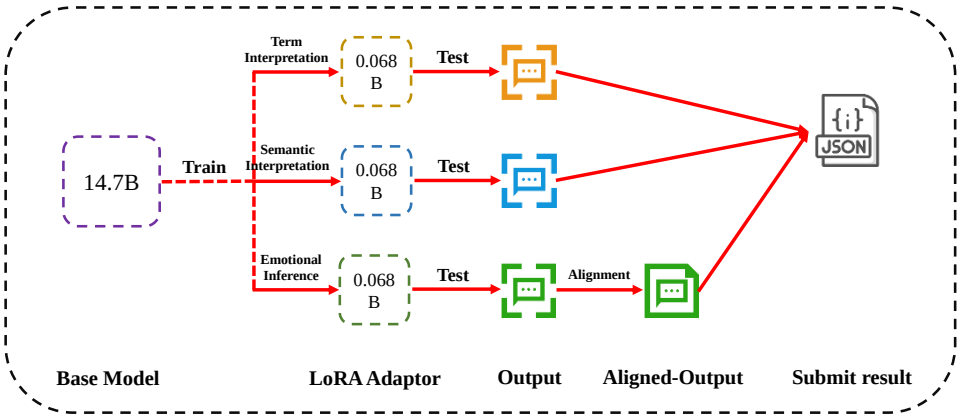

Figure 3: The Structure of PoetryQwen.

Model Name	Score	Term Interpretation		Semantic Interpretation		Emotional Inference
		Blue	BertScore	Blue	BertScore	Accuracy
Qwen2.5-7B	0.667	0.230	0.873	0.241	0.911	0.771
Qwen2.5-14B-Instruct	0.690	0.169	0.865	0.251	0.910	0.832
GLM-4-9B	0.628	0.136	0.846	0.204	0.901	0.734
PoetryQwen	**0.757**	**0.405**	**0.909**	**0.436**	**0.914**	**0.847**

Table 2: Quantitative Comparison of Model Performance on Poetry Appreciation Tasks

we further align the output of PoetryQwen with the submission format required by CCL25-Eval Task 5, which involves mapping the generated emotional expressions to one of the predefined emotion labels **(A, B, C, or D)**. To this end, we apply Qwen2.5-14B-Instruct for post-hoc alignment, ensuring compatibility with the evaluation requirements.

4.2 Experiment result

To evaluate the performance of our proposed dataset and model in the domain of classical Chinese poetry appreciation, we conduct experiments on the CCL25-Eval Task 5 benchmark. The evaluation encompasses three subtasks: Term Interpretation, Semantic Interpretation, and Emotional Inference with using metrics: BLEU, BERTScore, and Accuracy.

We compare our system, PoetryQwen, with several strong baselines, including Qwen2.5-7B[9], Qwen2.5-14B-Instruct[10], and GLM-4-9B[11]. As shown in Table 2, PoetryQwen consistently outperforms all baselines across the three subtasks, demonstrating the effectiveness of our instruction-tuning strategy. These results also highlight the value of the CCPoetry-49K dataset in providing high-quality, task-specific supervision tailored for classical Chinese poetry appreciation.

5 Conclusion

In this work, we propose a domain-specific framework for classical Chinese poetry appreciation, including term interpretation, semantic interpretation, and emotional inference. Accordingly, we construct CCPoetry-49K, a high-quality and domain-specific instruction–response dataset with 49,404 examples. Leveraging this dataset, we fine-tune the Qwen2.5-14B model using LoRA to obtain PoetryQwen. Experimental results on the CCL25-Eval Task 5 benchmark show a 9.7% improvement over the base model. Our team AI4S, registered on the Tianchi platform, demonstrates that domain specialization significantly enhances LLM performance in classical poetry understanding. Dataset will be available at https://github.com/XieHaoTao/CCPotery.

[9] https://huggingface.co/Qwen/Qwen2.5-7B
[10] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct
[11] https://huggingface.co/THUDM/glm-4-9b

References

Jiahuan Cao, Yang Liu, Yongxin Shi, Kai Ding, and Lianwen Jin. 2024. Wenmind: A comprehensive benchmark for evaluating large language models in chinese classical literature and language arts. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 51358–51410. Curran Associates, Inc.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4925–4931. International Joint Conferences on Artificial Intelligence Organization, 7.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Large language models for classical chinese poetry translation: Benchmarking, evaluating, and improving.

Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. 2024. Opendatalab: Empowering general artificial intelligence with open datasets.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. Ccpm: A chinese classical poetry matching dataset.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2019. Ancient–modern chinese translation with a new large training dataset. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(1), May.

Yusen Liu, Dayiheng Liu, and Jiancheng Lv. 2020. Deep poetry: A chinese classical poetry generation system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13626–13627, Apr.

Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. 2024. Charpoet: A chinese classical poetry generation system based on token-free llm.

Ran Zhang and Steffen Eger. 2024. Llm-based multi-agent poetry generation in non-cooperative environments.

Guo Zhipeng, 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 Marta R. Costa-jussà and Enrique Alfonseca, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 25–30, Florence, Italy, July. Association for Computational Linguistics.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. Wyweb: A nlp evaluation benchmark for classical chinese.