System Report for CCL25-Eval Task 5: Data Augmentation and Large Language Model Fine Tuning for Chinese Ancient Poetry Comprehension and Inference

Chengfei Li[a,c], Chunyu Wang[b,*], Bin Liu[a,c], Hanlin Li[a], Wenya Zhang[a], Hui Gao[a], Yue Wu[a,c]
[a]Artificial Intelligence Research Institute on Education, Qilu Normal University
[b]School of Geography and Tourism, Qilu Normal University
[c]China Language Promotion Base , Qilu Normal University
cywang@qlnu.edu.cn

Abstract

This paper introduces the CCL25-Eval evaluation task for ancient poetry comprehension and inference, which aims to enhance the capabilities of large language models(LLMs) in processing context-dependent texts with strong cultural backgrounds. Addressing the dual challenges of semantic analysis and emotional inferencein ancient poetry, we propose a solution that integrates Qwen-series LLMs with systematic data augmentation and LoRA-based parameter-efficient fine-tuning. We construct a high-quality dataset and design multi-phase training and inference strategies. Particularly in emotional inference tasks, we explore two approaches: emotion lexicon-based indirect matching and emotion appreciation-based direct judgment of emotional lexicon options. Experimental results indicated that: 1) Data augmentation significantly improves the model's overall performance; 2) The result of emotion appreciation-based direct judgment approach achieves an accuracy of 0.865, ranking first in Task A; 3) Attempts with Qwen3 and reinforcement learning approaches do not significantly improve Task B results, but demonstrated good performance in sentence semantic similarity scores and format stability.

Keywords: Chinese Ancient Poetry , Large Language Model , Data Augmentation , Instruction Fine-tuning

1 Introduction

As an important part of Chinese culture, ancient poetry carries the cultural essence of the Chinese nation. In recent years, natural language processing (NLP) technologies such as machine learning and deep learning have made certain progress in the artistic conception, theme and emotion analysis of ancient poetrys(Yeh et al., 2019; Cui, 2022; Liu, 2024). However, the ability of the Large Language Model (LLM) in the field of ancient poetry analysis has not been fully developed (Cao et al., 2024). Based on the competition process and results of the first Chinese ancient poetry comprehension and inference task of Alibaba Cloud Tianchi AI Massage Competition (CCL25-Eval), this study introduces the technical solution combining the Qwen2.5/3 series of models and data augmentation (knowledge distillation, data distillation)(Shi et al., 2025; Li et al., 2025), and discusses the experience and lessons learned.

2 Model Introduction

The official baseline model Qwen2.5-7B was evaluated using a zero-shot approach, where the model performed direct inference without any task-specific fine-tuning. This assesses the model's native ability for ancient poetry tasks. It also provides a unified performance benchmark for all participants. In the CCL25-Eval evaluation task for ancient poetry comprehension and inference, we employ the Qwen series models as base models to investigate the impacts of different model hyperparameter configurations, fine-tuning techniques, and data augmentation strategies.This approach specifically addresses the characteristics of ancient poetry that exhibit high-context dependency and strong cultural background. The models selected in this study are as follows:

* Corresponding author

2.1 Qwen2.5-7B-Instruct

The Qwen2.5-7B-Instruct is a 7.5 billion parameters instruction fine-tuned model from the Tongyi Qwen series. As a lightweight model, it maintains satisfactory performance while requiring relatively low computational resources.In ancient poetry comprehension tasks, despite its smaller parameter size, it demonstrates considerable potential in linguistic understanding and knowledge representation. We trained it using the initial dataset as a performance benchmark against larger models, aiming to evaluate the impact of model size on ancient poetry comprehension and inference tasks.

2.2 Qwen2.5-14B-Instruct

Compared with the 7B model, the 14B model exhibits superior expressive capabilities and a richer knowledge base, enabling better capture of complex semantics and deep emotions in ancient poetry. This model also employs instruction fine-tuning. With its greater model size, it demonstrates superior performance in contextual comprehension and diverse task execution(Yang et al., 2024).

2.3 Qwen3

Qwen3-14B is the latest 14 billion parameters LLM from the Tongyi Qwen series, representing a new generation upgrade in architectural design and capability enhancement for the Qwen family. Compared with Qwen2.5-14B-Instruct, Qwen3-14B demonstrates significant improvements across foundational model architecture, pre-training data scale, and instruction fine-tuning methodologies(Zheng et al., 2025).

3 Data Source and Data Augmentation

3.1 Data Source

The benchmark dataset consists of 200 training samples and 400 test samples provided by the official. Considering the high complexity and diversity of ancient poetry texts, to enhance model performance while adhering to competition rules, we expanded the dataset using public ancient poetry resources (`https://gitee.com/fatjay/knowledge_graph_practice`) to augment the official dataset. The final enhanced dataset (Base-Dataset) achieved a total data volume of approximately 8,000 samples.

3.2 Data Augmentation

Perform sample level null value detection on Base Dataset, removing null value samples from word annotations, sentence translations, and sentiment analysis to improve data integrity. Finally, 4741 data samples were retained as control dataset (Con-Dataset). Multi- phase processing was implemented for the Base-Dataset(Kamalloo et al., 2022) :

Stage1: First conducted null value detection on the Base-Dataset, utilizing Deepseek for automated null value completion , resulting in the Stage1-Dataset.

Stage2: Performed standardization processing on the format and content length of the Stage1-Dataset to ensure data consistency. The data at this stage is referred to as the Stage2-Dataset.

Stage3: This stage adopts two data construction schemes. The first is to construct indirect matching dataset for emotion lexicon generation based on Stage2-Dataset (Stage 3-Dataset 1), which uses DeepSeek model to generate 2-3 generalized emotion lexicons for training, predict emotion lexicons during inference, and then match options through embedding similarity; the second is to construct emotion appreciation-based direct judgment of emotional lexicon options dataset based on Stage2-Dataset (Stage 3-Dataset 2), design special emotion analysis prompt(Fu et al., 2022), and generate high-quality emotion summary data by employing Deepseek.

Stage4: Considering that the model based on Stage3-Dataset2 performed well on Task A, we continue to optimize Task B using this dataset. Specifically, we adjust the translation field from sentence level (sentence by sentence translation) to full text level (whole poetry translation) to enhance the model's ability to understand the overall context. This phase dataset is designated as Stage4-Dataset.

The specific experimental methods and procedures are illustrated in Figure 1.

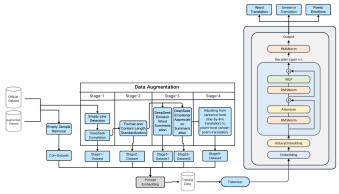

Figure 1: The Figure of specific experimental methods and procedures

4 Experimental Results and Analysis

4.1 Model Selection and Parameter Optimization

This experiment uses 8×K100 (64G) graphics card under DCU architecture and completes training based on LLaMA-Factory framework(Zheng et al., 2023). First, comparative experiments were conducted on the Qwen2.5-7B and Qwen2.5-14B models based on the Con-Dataset,with their fine-tuning hyperparameters shown in Table 1. Increasing the model parameter scale significantly enhances lexical comprehension capabilities but provides relatively limited improvement in emotional inference(Table 2). Considering comprehensive metric performance, the Qwen2.5- 14B with larger parameter scale is selected as the base model.

Hyperparamete	**Value**
per_device_train_batch_size	4
gradient_accumulation_steps	4
learning_rate	1.0e-4
num_train_epochs	3.0
lr_scheduler_type	cosine
warmup_ratio	0.1

Table 1: Hyperparameter settings in fine-tuning

models	dataset	score	sim_words	sim_sents	emo_acc
Qwen2.5-7B	Con-Dataset	0.6912	0.883	0.891	0.817
Qwen2.5-14B		0.7026	0.899	0.895	0.807
Relative Differences		+1.65%	+1.81%	+0.45%	-1.22%

Table 2: The performance of model architecture on various indicators

The inference temperature coefficient usually directly affects the diversity and accuracy of the generated text. A higher temperature coefficient helps to increase the diversity of the generated content, but may lead to a decrease in accuracy, while a lower temperature helps to increase the consistency and certainty of the output. Through a series of decode contrast experiments, we find that the model accuracy and stability are optimal when the temperature coefficient is 0.5, so the optimal temperature coefficient is determined to be 0.5(Table 3).

4.2 Experimental Results

We conducted comparative experiments on model performance under different data augmentation strategy from four dimensions: data expansion, format standardization, emotional inference strategy optimization, and translation strategy optimization.Experimental results demonstrate that with gradual optimization of data processing strategies, the model's comprehensive performance metric (score) shows a steady upward trend, improving from 0.703 with basic processing to 0.749 after translation strategy optimization. This indicates that data processing strategies exert significant positive impacts on model performance, particularly manifesting notable improvements in emotional inference accuracy and text

Temperature	score	sim_words	bleu_words	sim_sents	bleu_sents	emo_acc
1.0	0.7378	0.906	0.430	0.908	0.294	0.841
0.6	0.7340	0.903	0.416	0.878	0.240	0.859
0.5	0.7495	0.907	0.426	0.909	0.304	0.862

Table 3: Temperature Coefficient Impact on Performance Metrics

comprehension capabilities. Adjusting the translation field from sentence level to full text level further improves the model's contextual comprehension capability. We adopted the task-level optimal result integration method by selecting data processing strategies and model outputs that demonstrated optimal performance on their respective evaluation metrics, then integrated these optimal results to achieve a comprehensive score of 0.755(Table 4).

data augmentation solutions	score	sim_words	bleu_words	sim_sents	bleu_sents	emo_acc
Con-Dataset	0.703	0.899	0.334	0.895	0.264	0.807
Stage1-Dataset	0.728	0.900	0.367	0.876	0.231	0.862
Stage2-Dataset	0.732	0.904	0.393	0.876	0.231	0.862
Stage3-Dataset2	0.737	0.902	0.406	0.878	0.249	**0.865**
Stage4-Dataset	0.749	**0.912**	**0.450**	**0.909**	**0.304**	0.856

Table 4: Data augmentation strategies and their performance metrics.

It should be noted that we adopted two data construction schemes in Task A, namely emotion lexicon-based indirect matching and emotion appreciation-based direct judgment of emotional lexicon options.The result of latter data construction schemes (emo_acc=0.865) significantly outperforms the former (emo_acc=0.419), and it achieves the best performance in Task A.

4.3 Qwen3 Model Experimentation

To explore the potential of the latest LLM for the ancient poetry comprehension task, we fine-tuned the newly released Qwen3 model. Note that Qwen3 model provides thinking mode, but our dataset contains no thinking data. Therefore, we added a non-thinking mode template based on LLaMA-Factory template.py. On this basis, we tried the following experiments.

4.3.1 Qwen3 Fine-tuning

According to the inference results, Qwen3 generates high-quality, concise, and accurate content, achieving a lexical semantic similarity of 0.911, the highest among all models. Nevertheless, its output format is unstable, with approximately 7.5% (30) of test samples exhibiting null value issues. Consequently, as indicated in Table 5, its comprehensive score is 0.6801, considerably lower than the optimal results of Qwen2.5-14B.

4.3.2 Qwen3+DPO Optimization

The application of Direct Preference Optimization (DPO)(Rafailov et al., 2023), leveraging Human-Preference Dataset, has led to a significant enhancement in output format stability. This advancement has also elevated the sentence semantic similarity score to 0.910, marking an optimal level of performance. Nonetheless, the bleu score was abnormally low, with bleu_words at merely 0.110 and bleu_sents at 0.217. It is important to highlight that, despite the lackluster performance in automated evaluation metrics such as bleu scores, from the standpoint of artificial evaluation, the poetry translations generated by Qwen3+DPO have shown commendable content accuracy and expressive fluency.

5 Post-Competition Analysis and Discussion

5.1 Technical Advantages and Successful Experience

Through techniques including prompt engineering, data augmentation, model parameters fine-tuning, and emotional inference strategies, we achieved significant model performance improvements. Precise

Training strategies	score	sim_words	bleu_words	sim_sents	bleu_sents	emo_acc
Qwen3-14B	0.6801	**0.911**	0.218	0.852	0.108	0.838
Qwen3-14B+DPO	0.6816	0.852	0.110	**0.910**	0.217	0.841

Table 5: Qwen3 experiment comparison

temperature coefficient control combined with LoRA fine-tuning effectively balances the quality of the content generated by the model and the resource consumption of its operation(Hu et al., 2021).The scheme of emotion appreciation-based direct judgment of emotional lexicon options significantly enhances emotional inference accuracy. Moreover, fine-grained optimization and full text translation strategies in data augmentation strengthen the model's contextual comprehension capabilities(Chen, 2019). Overall, the high-quality dataset construction in this study provides a solid data foundation for enhanced model performance.

5.2 Challenges and Subsequent Technical Optimization

In this study, even with Task-level Optimal Result Integration, we could only achieve a comprehensive score of 0.755, indicating that LLMs still have much room for improvement in the field of ancient poetry comprehension and inference. This may be related to the model's insufficient capability in integrating cultural background knowledge, which limits its performance in more complex semantic tasks; Furthermore, there exists a prominent contradiction between high generation quality and format stability. For instance, the Qwen3 model demonstrates excellent content quality but poor format stability, while adopting DPO to address format issues would sacrifice some generation quality. Meanwhile, ancient poetry comprehension and inference represent relatively subjective tasks. The current objective evaluation metrics may fail to comprehensively capture the depth and breadth of ancient poetry understanding, proving inadequate for fully assessing poetic imagery and cultural connotations, resulting in incomplete and biased assessments of model performance.

To address these challenges, we propose the following optimization strategies: First, establish a more balanced and comprehensive multi-genre and multi-dynastic ancient poetry dataset to enhance data diversity and representativeness, while collaborating with domain experts and scholars to develop a more sophisticated evaluation framework for ancient poetry comprehension and inference assessment. Second, we explored methods such as contrastive learning, multi-task learning, knowledge distillation, model distillation, and reinforcement learning to enhance model generalization capabilities and multi-task adaptability.Moreover, future work will further explore Qwen3, focusing on its "thinking mode," with support from augmented data and case-based analysis.

6 Conclusion

This paper focuses on the CCL25-Eval task for the comprehension and inference of ancient poetry. It systematically explores the effectiveness of a combined approach using Qwen series LLMs, high-quality dataset construction, data augmentation, model parameter fine-tuning and diverse emotional inference strategies.Experimental results demonstrate that optimized data processing and emotional inference strategies significantly improved model performance in ancient poetry semantic analysis and emotional inference tasks, with some solutions achieving leading performance in official evaluations. However, LLMs still face challenges such as insufficient knowledge integration, content format instability, and subjective evaluation in high-context-dependent tasks. Subsequent work needs to pay more attention to enriching data diversity and improving evaluation systems, while also exploring methods such as multi-task learning, contrastive learning, and knowledge distillation. These efforts are aimed at further enhancing the model's generalization ability and practical application value.

References

Yeh, W. C., Chang, Y. C., Li, Y. H., et al. Rhyming knowledge-aware deep neural network for Chinese poetry generation. *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2019,1-6, doi:

10.1109/ICMLC48188.2019.8949208.

Cui, M. DRIIS: research on automatic recognition of artistic conception of classical poems based on deep learning. *International Journal of Cooperative Information Systems*, 2022. DOI:10.1142/S0218843022500010.

Liu, J. H. Research on the application of natural language processing in the analysis of ancient poetrys and texts// *AIP Conference Proceedings*.AIP Publishing, 2024, 3194(1).

Cao, J. H., Liu, Y., Shi, Y. X., et al. WenMind: A comprehensive benchmark for evaluating large language models in Chinese classical literature and language arts. *Advances in Neural Information Processing Systems*, 2024, 37: 51358–51410.

Yang, A., Yang, B., Zhang, B., et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zheng, X., Li, Y., Chu, H., et al. An Empirical Study of Qwen3 Quantization. *arXiv preprint arXiv:2505.02214*, 2025.

Kamalloo, E., Rezagholizadeh, M., Ghodsi, A. When chosen wisely, more data is what you need: A universal sample-efficient strategy for data augmentation. *arXiv preprint arXiv:2203.09391*, 2022.

Chen, K. H. Research on Context Representation Methods for Machine Translation. *Harbin Institute of Technology*, 2019. DOI:10.27061/d.cnki.ghgdu.2019.000172.(in Chinese)

Rafailov, R., Sharma, A., Mitchell, E. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023, 36: 53728–53741.

Zheng, Y., Zhang, R., Zhang,J., et al. Llamafactory: Unified efficient fine-tuning of 100+ language models[J]. *arXiv preprint arXiv:2403.13372*, 2024.

Shi, M. W., Lin, M., Sun, Y. R. et al. Research on the Similarity Matching of Ancient Texts Based on Multi-Stage Knowledge Distillation. *Computer Engineering and Applications*, 1–14 [2025-05-11].(in Chinese)

Li, T., Yang, H. G., Liu, K. et al. Knowledge distillation methods for large models: a research review. *Journal of the Hebei Academy of Sciences*, 2025, 42(02):94–96. DOI:10.16191/j.cnki.hbkx.2025.02.004.(in Chinese)

Fu, C. S., Li, P., Wu, Y. D. et al. Research on The Emotion Analysis of Poetry Based on Imagery. *SmartTech Innovations*,2022,28(12):9-16.(in Chinese)

Hu, E. J., Shen, Y., Wallis, P., et al. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv, abs/2106.09685.*,2021