Overview of CCL25-Eval Task 5: Chinese Classical Poetry Appreciation Evaluation (CCPA) Task

Zhenwu Pei[1], **Yingjie Zhu**[1,2], **Rongbo Chen**[1], **Xuefeng Bai**[1†], **Kehai Chen**[1], **Min Zhang**[1]
[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
{24s151006,zhuyj,23S151077}@stu.hit.edu.cn
{baixuefeng,chenkehai,zhangmin2021}@hit.edu.cn

Abstract

This paper presents a review of CCL2025-Eval Task 5: the First **C**hinese **C**lassical **P**oetry **A**ppreciation Evaluation (CCPA). The primary aim of this task is to evaluate the ability of language models in performing deep semantic understanding and aesthetic appreciation of Chinese classical poetry. The evaluation comprises two tracks: (1) Poetic content understanding, which examines models' ability to interpret both fine-grained and coarse-grained semantics; (2) Poetic emotion recognition, which evaluates models' capacity to identify and analyze emotional expressions. A total of 55 teams registered for the task, among which 7 teams provided valid submissions. The paper provides an in-depth analysis of the submissions and results from all participating teams.

1 Introduction

As a quintessential form of Chinese classical literature, Chinese classical poetry is marked by profound artistic conception, succinct expression, and intricate rhetorical techniques. Compared to modern Chinese, ancient poetry exhibits highly condensed syntax, flexible word-class usage, frequent omission of syntactic components (e.g., subject-object structures), semantic ambiguity, and open-ended forms (Wang et al., 2024; Li et al., 2021). These linguistic and structural features pose significant challenges for natural language processing models, particularly in semantic interpretation and reasoning.

The rapid advancement of large language models (LLMs) in natural language processing (Dong et al., 2024; Koshkin et al., 2024; Li et al., 2025; Fang et al., 2024; Zhang et al., 2025) has spurred growing interest in classical Chinese literature datasets. While existing poetry corpora (Zhang and Li, 2023; Wei et al., 2024; Guo et al., 2023) have enabled preliminary research, such as poetry generation, poetry generation and translation. However, current studies are narrowly designed for isolated tasks like generation (Liu et al., 2020; Agarwal and Kann, 2020; Wang et al., 2016) or style detection (Shao et al., 2021; Ming et al., 2022), lacking comprehensive annotation frameworks that capture the multidimensional nature of classical poetry appreciation;remain at the level of surface feature analysis and have not deeply explored the models' performance in deeper, fine-grained semantic comprehension and reasoning abilities regarding classical poetry, thus falling short of providing a comprehensive evaluation framework.

To tackle these issues, we propose the *CCL-Task5 evaluation: **C**hinese **C**lassical **P**oetry **A**ppreciation Evaluation* (**CCPA**). CCPA systematically evaluates LLMs' comprehensive understanding and appreciation of Chinese classical poetry from two key dimensions — *Poetic content understanding* and *Poetic emotion recognition* — while establishing a scientific, objective, and reproducible evaluation framework. Specifically, the Poetic content understanding task focuses on interpreting key vocabulary and translating sentences within ancient Chinese poems, aiming to evaluate the model's grasp of linguistic and contextual meaning. The Poetic emotion recognition task, on the other hand, is designed to assess the model's ability to identify and infer the emotional expressions conveyed by the poet in the given texts.

[†] Corresponding Author

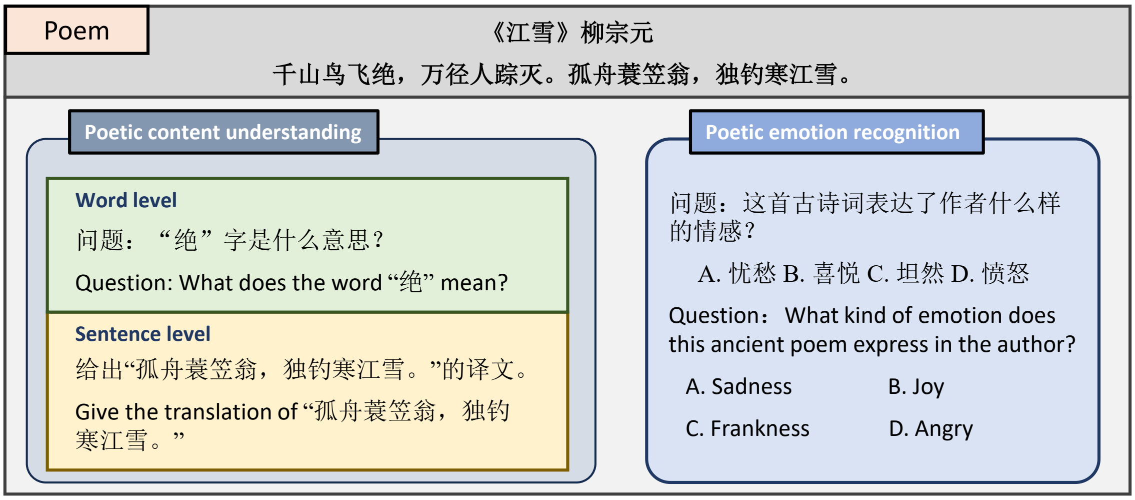

Poem
《江雪》柳宗元
千山鸟飞绝，万径人踪灭。孤舟蓑笠翁，独钓寒江雪。
Poetic content understanding
Word level
问题：“绝”字是什么意思？
Question: What does the word “绝” mean?
Sentence level
给出“孤舟蓑笠翁，独钓寒江雪。”的译文。
Give the translation of “孤舟蓑笠翁，独钓寒江雪。”
Poetic emotion recognition
问题：这首古诗词表达了作者什么样的情感？
A. 忧愁 B. 喜悦 C. 坦然 D. 愤怒
Question：What kind of emotion does this ancient poem express in the author?
A. Sadness
B. Joy
C. Frankness
D. Angry

Figure 1: Tasks related to Chinese Classical Poetry Appreciation, including Poetic content understanding and Poetic emotion recognition.

A total of 55 teams registered for the competition, and ultimately, we received 7 valid submissions from 7 teams. The details of the task are outlined in Section 2. Section 3 describes the data collection and preprocessing procedures employed for this task. In Section 4, we define the evaluation metrics for each subtask and present the methodology for computing the overall task score. Section 5 introduces the baseline models used for each track. Section 6 provides a comprehensive overview of the participating teams, summarizes their submitted results, and offers an in-depth analysis and discussion.

2 Task Description

The evaluation is divided into two different parts, each of which tests the specific ability to understand and appreciate Chinese classical poetry. These tasks are designed to evaluate the performance of the tested system in understanding the semantics intrinsic emotions of classical poetry, thereby providing a basis for further improving the model's ability to understand and generate classical literary works.

2.1 Track 1: Poetic Content Understanding

Understanding the content of poetry forms the foundation of poetry appreciation, with its essence lying in the systematic comprehension of the content's semantic information. Thus, we first propose a hierarchical analytical framework starting from two fundamental dimensions—vocabulary and syntax—and build a quantifiable evaluation system for content understanding. Specifically, this task primarily encompasses the following two subtasks.

Word-level Understanding (WU) The language of classical Chinese poetry is often more nuanced than modern Chinese, with word meanings shaped by historical, cultural, and contextual factors. For example, in most poetry, “liu” (willow) may signify both a tree and the sentiment of farewell, or contain implicit literary allusions. Therefore, we design this task to evaluate the model's ability to accurately understand such deeper semantic meanings within the poetic context.

Sentence-level Understanding (SU) This task aims to assess the model's ability to translate Chinese classical poetry into modern Chinese. It requires the model not only to accurately convey the semantic content of Chinese classical poetry, but also to capture its linguistic elegance and cultural depth. Compared to general translation, translating classical poetry is more challenging due to its highly condensed language, loosely structured syntax, and frequent use of metaphors and cultural allusions.

Category	FQ	SQ	FR	SR	SC
#Poems	29	51	80	53	278

Table 1: Distribution of poetry types in the dataset, where FQ denotes *Five-character quatrains*, SQ denotes *Seven-character quatrains*, FR denotes *Five-character regulated verses*, SR denotes *Seven-character regulated verses*, and SC denotes *the poetry from the Song dynasty*.

2.2 Track 2: Poetic Emotion Recognition

The Poetic Emotion Recognition (ER) task aims to leverage natural language processing techniques to analyze the intrinsic features of Chinese classical poetry, such as imagery and rhetorical devices. By incorporating historical and cultural context, this task requires the model to infer the poet's emotional attitudes and internal states conveyed through the poetry, thereby enabling a deeper assessment of the model's ability to understand and interpret subtle emotional shifts in classical literary works.

3 Datasets

3.1 Data Collection

The proposed datasets are drawn from authentic classical poetry texts collected from online platforms[1], including annotations, translations, emotion recognition, and artistic appreciation. To ensure genre diversity, we collect poetry from a broad range of classical types—such as *five-character quatrains* and *Seven-character quatrains*—as detailed in Table 1. These real-world poetry exhibits diverse styles, expressive nuances, and interpretative variations, offering a comprehensive and realistic foundation for evaluating models' understanding of Chinese classical poetry.

3.2 Data Processing

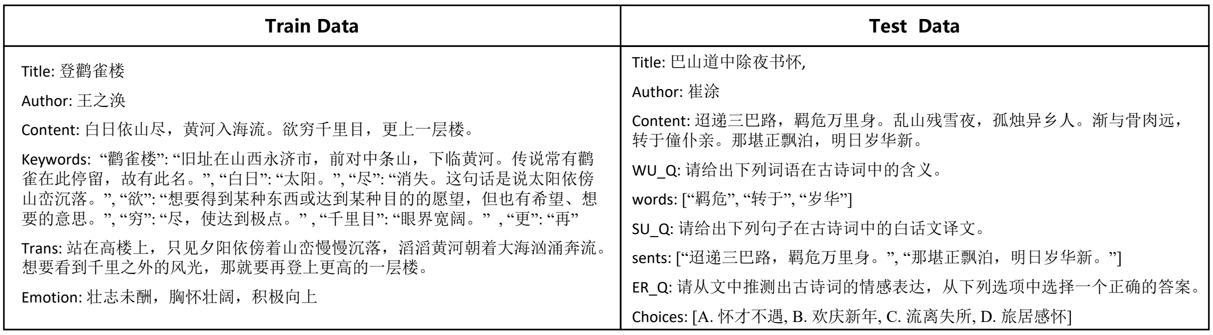
To improve dataset quality, we designed a comprehensive preprocessing pipeline combining semantic filtering and noise reduction techniques. We first employed a text similarity algorithm alongside GPT-4o-mini (Hurst et al., 2024) for semantic-level deduplication, effectively removing around 20% of redundant samples. To further enhance textual integrity, we applied regular expression-based rules and character-level filters to eliminate extraneous symbols, irrelevant content, and nonsensical entries, thereby improving the dataset's linguistic precision and semantic consistency. This process ensures a clean and reliable foundation for subsequent model training and evaluation.

The final dataset features a rich training set of original classical poetry paired with corresponding appreciation texts, while the test set targets key dimensions of understanding—vocabulary, syntax, and emotion—offering a comprehensive benchmark for evaluating models' language comprehension and reasoning abilities. A representative sample is illustrated in Figure 2 and the statistics of the dataset are presented in Table 2.

4 Evaluation Metrics

Considering the distinct characteristics of each task, we select appropriate evaluation metrics accordingly. For the Poetic content understanding, we employ two automatic metrics: BLEU (Papineni et al., 2002) and BERTScore (BS), to thoroughly assess both the surface-level text overlap and deeper semantic similarity between the model-generated answers and reference responses. Specifically, BLEU measures the n-gram matching degree, while BERTScore evaluates semantic closeness by leveraging contextual embeddings. In the Poetic emotion recognition task, accuracy (ACC) is used as the main evaluation criterion. The model receives a score of 1 if its predicted label exactly matches the ground truth; otherwise, it is scored 0. This binary scoring system is designed to quantify the model's accuracy in inferring emotional states.

[1] https://www.gushiwen.cn/



Train Data	Test Data
Title: 登鹳雀楼 Author: 王之涣 Content: 白日依山尽，黄河入海流。欲穷千里目，更上一层楼。 Keywords: “鹳雀楼”: “旧址在山西永济市，前对中条山，下临黄河。传说常有鹳雀在此停留，故有此名。”, “白日”: “太阳。”, “尽”: “消失。这句话是说太阳依傍山峦沉落。”, “欲”: “想要得到某种东西或达到某种目的的愿望，但也有希望、想要的意思。”, “穷”: “尽，使达到极点。” , “千里目”: “眼界宽阔。” , “更”: “再” Trans: 站在高楼上，只见夕阳依傍着山峦慢慢沉落，滔滔黄河朝着大海汹涌奔流。想要看到千里之外的风光，那就要再登上更高的一层楼。 Emotion: 壮志未酬，胸怀壮阔，积极向上	Title: 巴山道中除夜书怀, Author: 崔涂 Content: 迢递三巴路，羁危万里身。乱山残雪夜，孤烛异乡人。渐与骨肉远，转于僮仆亲。那堪正飘泊，明日岁华新。 WU_Q: 请给出下列词语在古诗词中的含义。 words: [“羁危”, “转于”, “岁华”] SU_Q: 请给出下列句子在古诗词中的白话文译文。 sents: [“迢递三巴路，羁危万里身。”, “那堪正飘泊，明日岁华新。”] ER_Q: 请从文中推测出古诗词的情感表达，从下列选项中选择一个正确的答案。 Choices: [A. 怀才不遇, B. 欢庆新年, C. 流离失所, D. 旅居感怀]

Figure 2: Examples from the training and test sets

Tasks	Training Set		Test Set	
	#Poems	#QAs	#Poems	#QAs
WU	164	1304	327	960
SU	164	164	327	665
ER	164	164	327	327

Table 2: Statistics of our dataset across different task categories. “#Poems” denotes the number of unique Chinese classical poems, and “#QAs” refers to the number of corresponding question–answer pairs.

4.1 Track 1: Poetic Content Understanding

The total score of Track 1 consists of two subtasks: Word-level Understanding and Sentence-level Understanding. We calculate evaluation indicators such as BLEU or BERTScore for these two subtasks separately and assign equal weights to them. The specific calculation formula is as follows:

$$\text{Score}_{\text{Track1}} = 0.5 \times F(\text{WU}) + 0.5 \times F(\text{SU}) \tag{1}$$

Where $F(\cdot)$ represents the scoring function for each subtask.

4.2 Track 2: Poetic Emotion Recognition

Track 2 focuses on the task of Poetic emotion recognition for Chinese classical poetry. We use accuracy (ACC) as the only evaluation metric to measure the model's judgment ability in the emotion classification task. In each test question, if the model's prediction result is consistent with the reference answer, it is scored as 1 point; otherwise, it is scored as 0 points. The final score is the average accuracy of all test samples, and its calculation formula is as follows:

$$\text{Score}_{\text{Track2}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i) \tag{2}$$

Where N is the total number of test samples, $\hat{y}_i$ represents the model's prediction result for the ith sample, y_i is its corresponding reference answer, and $\mathbb{I}(\cdot)$ is an indicator function, which takes a value of 1 when the prediction is correct and 0 otherwise.

4.3 Overall Score

The final score is the weighted sum of the scores of Track 1 and Track 2 according to their importance weights. Assuming that the weights of the two tasks are α and β, the total score is calculated as follows:

$$\text{score}_{\text{overall}} = \alpha \cdot \text{score}_{\text{Track1}} + \beta \cdot \text{score}_{\text{Track2}} \tag{3}$$

Under the default setting, we set $\alpha = \beta = 0.5$, which means that the two tasks are equally important.

Team Name	Organization	Overall Score	WU		SU		ER
			BLEU	BS	BLEU	BS	ACC
LittleStudent	Tsinghua University	**0.758**	**0.486**	**0.920**	0.419	**0.925**	0.829
AI4S	Beihang University	0.757	0.405	0.909	**0.436**	0.914	0.847
Suoyiran	Qilu Normal University	0.749	0.426	0.907	0.304	0.909	**0.862**
JueJue	Northern Arizona University	0.724	0.380	0.897	0.345	0.917	0.813
Potted plants	Individual	0.725	0.389	0.903	0.269	0.914	0.832
Gogogo	Individual	0.714	0.343	0.886	0.288	0.905	0.823
Ketang	Yunnan University	0.714	0.323	0.885	0.269	0.921	0.829
Baseline	Harbin Institute of Technology (Shenzhen)	0.667	0.230	0.873	0.241	0.911	0.771

Table 3: Poetic content understanding and Poetic emotion recognition results.

5 Results and Analysis

In this section, we first introduce the selection and configuration of the baseline system, followed by a presentation and analysis of the submitted results from participating teams. We then summarize their solution strategies from three key perspectives: model training, data construction and enhancement, and task-specific optimization and reasoning.

5.1 Baseline

We select Qwen2.5-7B-Instruct (Yang et al., 2025) as the baseline model. The default parameter settings of VLLM[2] are used in the experiment, where the maximum generation length is 2048 tokens, temperature is 1, top-p is 1, and top-k is set to 50.

5.2 Main Results

The evaluation task received registrations from 55 teams, with 7 teams ultimately submitting their results. Participants adopted a range of methods to tackle the challenges of classical Chinese poetry understanding, spanning word-level interpretation, sentence-level semantic comprehension, emotion recognition, and aesthetic judgment. The final results are shown in Table 3. We can observe that the team from Tsinghua University achieves the most consistent and competitive results, ranking first overall. It obtained the highest BLEU and BERTScore in the Word-level Understanding (WU) task, and also led in BERTScore for the Sentence-level Understanding (SU) task, indicating strong capabilities in both semantic comprehension and generation quality. The team from Beihang University achieves the top BLEU score in the SU task, indicating a particular strength in generating fluent and faithful modern Chinese translations of classical poetry. In the Emotion Recognition (ER) task, the team from Qilu Normal University outperformed others with the highest classification accuracy (0.862), showing a solid grasp of the emotional and cultural nuances embedded in classical texts. In comparison, the baseline system consistently underperformed in all metrics, particularly in BLEU scores for WU and SU, underscoring the difficulty of the tasks and the importance of task-specific optimization for classical poetry understanding.

5.3 Model Training Strategy

To accomplish this task, participating teams adopted a range of fine-tuning strategies and optimization techniques, reflecting diverse technical perspectives. AI4S, Suoyiran, and primarily employed parameter-efficient LoRA fine-tuning (Hu et al., 2022). AI4S achieved a 9.7% performance improvement by updating only low-rank matrix parameters. Suoyiran optimized the temperature coefficient during inference to balance output quality and stability, while Juejue combined supervised fine-tuning (SFT) with LoRA, leveraging a unified instruction format to enhance multi-task generalization.

In contrast, Pengzai and Little Student adopted full-parameter fine-tuning. Pengzai focused on improving word- and sentence-level comprehension, whereas Little Student employed a two-stage training strategy: first conducting full-parameter fine-tuning on external classical poetry corpora to establish a strong semantic foundation, followed by LoRA-based fine-tuning on our training data to boost task-specific performance.

[2] https://github.com/vllm-project/vllm

Additionally, KETANG proposed a staged SFT approach combined with an attention gating mechanism to dynamically regulate the information flow between shared and task-specific representations. Meanwhile, Proximal Policy Optimization (PPO)-based reinforcement learning (Schulman et al., 2017) was applied to refine the emotion reasoning subtask, yielding further gains in performance.

5.4 Data Construction and Enhancement Strategy

All teams prioritize data quality, structural standardization, and task adaptation, employing diverse yet complementary methods. AI4S and KETANG constructed multi-level annotated corpora with task-aligned formats to enhance semantic and emotional modeling. JUEJUE and SUOYIRAN focused on unifying data formats and ensuring input quality through folder-based extraction and multi-stage processing workflows.

To mitigate data scarcity, PENGZAI leveraged LLMs for automatic data augmentation and quality assessment, thereby increasing both volume and diversity. GOGOGO combined rigorous data cleaning with pseudo-labeling using Qwen2.5-72B and incorporated external sources such as "Gushiwen" to enrich coverage of classical poetry styles.

LITTLE STUDENT curated datasets with Alpaca-style instruction formats, applied staged training on external poetry corpora, and leveraged translation results as enhanced prompts to maintain semantic consistency.

Overall, these approaches reflect a balance between standardized data processing and task-aware augmentation, effectively supporting improved model generalization and performance across subtasks.

5.5 Task Optimization and Reasoning Strategy

For task optimization and reasoning, participating teams adopted a range of strategies to enhance model performance across subtasks. AI4S employed explicit subtask decomposition, aligning terminology explanation, semantic translation, and sentiment reasoning within a unified framework—demonstrating strong contextual understanding across multiple levels. JUEJUE and GOGOGO focused on generation optimization, where the former adopted sampling-based generation with post-processing, and the latter introduced multi-round generation combined with weighted voting to improve output stability and judgment reliability. Meanwhile, PENGZAI showed that while full-parameter fine-tuning improved translation accuracy, it required additional techniques to perform well in emotion-related tasks. And LITTLE STUDENT employed a multi-model collaborative approach, using prior translation outputs as prompts for downstream tasks, thereby improving coherence and consistency across subtasks.

Furthermore, for poetic emotion recognition task, KETANG and SUOYIRAN explored targeted enhancements. KETANG incorporated PPO-based reinforcement learning with composite rewards (e.g., accuracy and contextual consistency) to improve classification performance. SUOYIRAN experimented with sentiment dictionary matching and emotion-based appreciation, with the latter achieving the highest accuracy (0.862) in ER task, highlighting the effectiveness of hybrid strategies.

These approaches form a broad technical spectrum from architectural integration to specialized optimization. Hybrid strategies, especially in sentiment analysis, proved advantageous, offering valuable insights for multi-task collaborative optimization in complex scenarios.

6 Conclusions and Future Work

This paper presents an overview of CCL25-Eval Task 5: the first Chinese Classical Poetry Appreciation Evaluation (CCPA). This task aims to assess models' capabilities in two core aspects of Chinese classical poetry understanding: Track 1:Poetic Content Understanding, which evaluates a system's ability to accurately interpret phrases and verses in ancient poetry. Track 2: Poetic Emotion Recognition, which tests the system's capacity to infer the poet's emotions based on a deep comprehension of the poem.

A total of 55 teams registered for the evaluation, among which 7 teams successfully submitted their systems. We conducted a comprehensive analysis of the participating methods and summarized their characteristics and performance. These insights not only shed light on the current state of Chinese classical poetry processing but also offer valuable guidance for future research in natural language understanding and cultural heritage-oriented AI.

Despite the progress achieved, several challenges remain for future exploration:

- **Diachronic Semantics and Ambiguity**: Classical Chinese poetry often contains words whose meanings have evolved over time or vary with context. Disambiguating these requires deeper linguistic modeling and historical knowledge integration.
- **Metaphor and Allusion Recognition**: Poets frequently employ literary devices such as metaphors, allusions, and symbolic imagery. Accurately identifying and interpreting these expressions remains a major hurdle for current models.
- **Emotion Reasoning Beyond Sentiment Analysis**: Poetic emotion is often implicit, multi-layered, and context-dependent. Future models must move beyond basic sentiment classification and develop deeper reasoning abilities grounded in literary context.
- **Data Scarcity and Annotation Difficulty**: High-quality annotated datasets for classical Chinese texts are limited, and annotation itself demands domain expertise. Constructing richer, large-scale, and multi-level datasets is an urgent need.
- **Cross-modal and Human-in-the-loop Integration**: Incorporating visual and audio elements of poetry (e.g., calligraphy, recitation) and leveraging human feedback could further enhance appreciation and understanding capabilities.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (62406091, 62276077, U23B2055) and Guangdong Basic and Applied Basic Research Foundation (2024A1515011205).

References

Rajat Agarwal and Katharina Kann. 2020. Acrostic poem generation. *arXiv preprint arXiv:2010.02239*.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge? *arXiv preprint arXiv:2406.11657*.

Jiangnan Fang, Cheng-Tse Liu, Jieun Kim, Yash Bhedaru, Ethan Liu, Nikhil Singh, Nedim Lipka, Puneet Mathur, Nesreen K Ahmed, Franck Dernoncourt, et al. 2024. Multi-llm text summarization. *arXiv preprint arXiv:2412.15487*.

Geyang Guo, Jiarong Yang, Fengyuan Lu, Jiaxin Qin, Tianyi Tang, and Wayne Xin Zhao. 2023. Towards effective ancient chinese translation: dataset, model, and evaluation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 416–427. Springer.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Transllama: Llm-based simultaneous translation system. *arXiv preprint arXiv:2402.04636*.

Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. Ccpm: A chinese classical poetry matching dataset. *arXiv preprint arXiv:2106.01979*.

Xingzuo Li, Kehai Chen, Yunfei Long, Xuefeng Bai, Yong Xu, and Min Zhang. 2025. Generator-assistant stepwise rollback framework for large language model agent. *arXiv preprint arXiv:2503.02519*.

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

Yifei Ming, Ying Fan, and Yixuan Li. 2022. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pages 15650–15665. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Yizhan Shao, Tong Shao, Minghao Wang, Peng Wang, and Jie Gao. 2021. A sentiment and style controllable approach for chinese poetry generation. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4784–4788.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. *arXiv preprint arXiv:1610.09889*.

Shanshan Wang, Derek F Wong, Jingming Yao, and Lidia S Chao. 2024. What is the best way for chatgpt to translate poetry? *arXiv preprint arXiv:2406.03450*.

Yuting Wei, Yuanxing Xu, Xinru Wei, Simin Yang, Yangfu Zhu, Yuqing Li, Di Liu, and Bin Wu. 2024. Ac-eval: Evaluating ancient chinese language understanding in large language models. *arXiv preprint arXiv:2403.06574*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yixuan Zhang and Haonan Li. 2023. Can large langauge model comprehend ancient chinese? a preliminary test on aclue. In *Proceedings of the Ancient Language Processing Workshop*, pages 80–87.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025. Exploring translation mechanism of large language models. *arXiv preprint arXiv:2502.11806*.