

System Report for CCL25-Eval Task 5: Hierarchical Multi-Task Prompt Fine-Tuning and PPO Reinforcement for Classical Chinese Poetry Comprehension and Sentiment Reasoning

Jingjun Tang, Zhiwen Tang[†]

School of Information Science & Engineering, Yunnan University, Kunming 650091
12024215193@stu.ynu.edu.cn, zhiwen.tang@ynu.edu.cn

Abstract

We present a hierarchical multi-task framework to enhance classical Chinese poetry understanding and sentiment reasoning using large language models. Centered on Qwen2.5-14B-Instruction or Xunzi-Qwen-14B, we construct a 1,225-sample corpus of Tang and Song poems with parallel translations and multi-label sentiment annotations (e.g., nostalgia, patriotism, contemplation). The task is divided into comprehension, translation, and sentiment inference, each guided by dynamic prompting and task-specific templates. We employ mixed supervised fine-tuning to better capture syntactic and metaphorical patterns. For sentiment reasoning, we apply proximal policy optimization (PPO) with a custom reward function, boosting accuracy from 0.771 to 0.807 ($p < 0.01$). Our model achieves a 0.714 comprehensive score, outperforming single-task baselines by 12.6%. Ablation studies further confirm the benefits of multi-task learning in promoting cross-task knowledge transfer.

Keywords: Classical Chinese Poetry, Multi-Task Fine-Tuning, Data Augmentation, Proximal Policy Optimization

1 Introduction

This task poses significant challenges. First, the condensed syntax and extensive use of literary allusions in classical Chinese poetry hinder accurate semantic parsing. Second, the layered emotions and inherent ambiguity of poetic language demand deep reasoning capabilities that often exceed the limitations of large language models trained predominantly on modern vernacular texts (Sun et al., 2019; Dong and Lapata, 2016; Zhang et al., 2023; Shen et al., 2024). Third, the strong cultural specificity of classical poetry necessitates a nuanced understanding of historical and contextual background, further complicating interpretation (Hu, 2024; Li et al., 2018).

To address these challenges, we propose a hierarchical multi-task framework. For syntactic compactness and allusion resolution, we employ prompt-based fine-tuning and introduce data augmentation techniques (Lester et al., 2021). To enhance reasoning over complex emotional structures, we design a hierarchical prompting strategy tailored to each subtask. To incorporate cultural knowledge, we integrate a PPO-based reinforcement learning module (Schulman et al., 2017; Ouyang et al., 2022; Havrilla et al., 2024). By leveraging multi-task learning (Ruder, 2017; Samant et al., 2022), our method facilitates comprehensive understanding and consistently surpasses strong baselines.

2 Related Work

Early research on classical Chinese poetry primarily addressed tasks such as word segmentation, named entity recognition, and syntactic analysis (Sun et al., 2019; Dong and Lapata, 2016). In recent years, the focus has shifted toward poem generation and translation, with models aiming to preserve both semantic meaning and poetic form (Hu, 2024; Li et al., 2018). Notably, ChineseBERT introduced glyph and pinyin

[†] Corresponding author

embeddings to enhance Chinese language modeling capabilities (Sun et al., 2021). However, few studies have systematically approached poetry understanding through a unified multi-task reasoning framework.

Prompt tuning has emerged as a parameter-efficient method for adapting large language models in low-resource scenarios (Lester et al., 2021). Multi-task learning (MTL) has demonstrated effectiveness in leveraging shared representations across related tasks (Ruder, 2017). Additionally, reinforcement learning with human feedback—particularly using proximal policy optimization (PPO)—has been shown to improve model alignment and output quality (Ouyang et al., 2022; Schulman et al., 2017). In this work, we integrate these three methodologies within the domain of classical Chinese literature and empirically validate the effectiveness of this unified framework.

3 Methodology

3.1 Framework Overview

Understanding classical Chinese poetry requires nuanced modeling of its intricate linguistic structures, layered emotional expressions, and rich cultural references. To address these challenges holistically, we propose a unified three-stage framework that integrates data augmentation, prompt-based supervised fine-tuning, and reinforcement learning with composite rewards. An overview of the framework is illustrated in Figure 1.

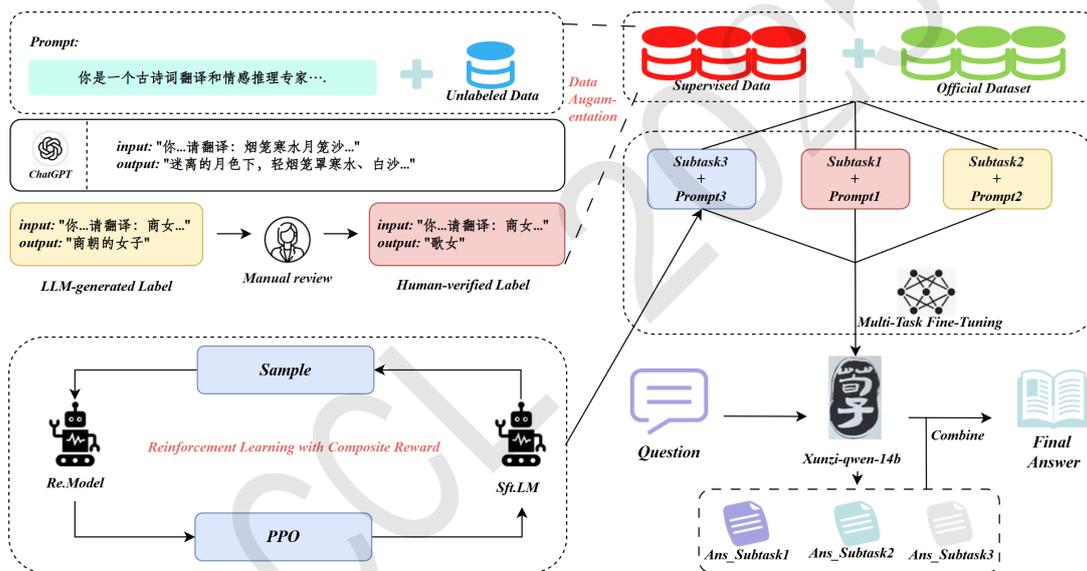


Figure 1: Overall framework for classical Chinese poetry comprehension and sentiment reasoning

Data Augmentation. We constructed a high-quality dataset comprising 1,225 classical poems drawn from authoritative sources such as *Complete Tang Poems* and *Complete Song Lyrics*. To enrich the annotation quality, we leveraged ChatGPT to automatically label word meanings, sentence translations, and emotional tags, followed by meticulous manual correction to ensure accuracy and consistency.

Multi-Task Supervised Fine-Tuning. We decompose the overall task into three subtasks: word interpretation, sentence translation, and emotion reasoning. Each subtask is initially fine-tuned independently using prompt-based templates tailored to its objective. Subsequently, a joint multi-task fine-tuning stage is conducted to enable cross-task knowledge sharing and enhance overall model generalization.

Reinforcement Learning with Composite Rewards. To further improve emotional inference, we apply reinforcement learning using the proximal policy optimization (PPO) algorithm. The reward function integrates both emotion classification accuracy and cultural consistency, encouraging the model to generate sentimentally accurate and culturally appropriate responses.

3.2 Data Augmentation

Due to the scarcity of annotated data for classical Chinese poetry, we adopt a ChatGPT-assisted data augmentation strategy to enhance training diversity and improve the model’s ability to handle nuanced linguistic and emotional patterns. This augmented data is integrated into the supervised fine-tuning stage to support multi-task learning.

To construct the initial dataset, we curated 1,225 representative poems from authoritative sources, including the *Complete Tang Poems* and *Complete Song Lyrics*. Selection criteria emphasized genre variety, thematic richness, and historical representativeness to ensure wide coverage of poetic forms and sentiments.

For data augmentation, we designed structured prompt templates tailored to each subtask (e.g., word explanation, verse translation, emotion annotation), and applied the ChatGPT-4 API to generate augmented samples. The outputs included paraphrased interpretations, sentiment labels, and sentence-level translations, simulating human annotations.

To ensure quality, all generated annotations underwent a manual verification process, including initial screening to eliminate incorrect or stylistically inconsistent outputs and corrections to address fine-grained semantic or cultural inaccuracies. This process ensured both reliability and domain fidelity in the final training corpus.

3.3 Multi-Task Fine-Tuning

Multitask fine-tuning plays a central role in adapting large language models to the complex demands of classical Chinese poetry analysis. We formulate the problem as a combination of three interrelated tasks: comprehension, translation, and sentiment reasoning. While these subtasks share a common foundation in poetic understanding, they differ significantly in output structure, interpretive focus, and reasoning depth.

To address these differences, we design tailored prompt templates for each subtask, guiding the model to produce task-specific outputs. For example, comprehension prompts encourage lexical or syntactic interpretation, while translation prompts elicit modern equivalents, and sentiment prompts focus on emotional inference. The detailed structure of each template is shown in Table 1.

We adopt a cross-task training strategy, allowing the model to learn shared representations while maintaining sensitivity to the distinct objectives of each task. This joint training approach facilitates knowledge transfer across subtasks, improving overall model generalization and robustness.

Subtask1-Prompt	instruction:”你是一个古诗词专家，擅长于将古诗词翻译为白话文...” input:”请阅读下列古诗，回答相关问题...”
Subtask2-Prompt	instruction:”你是一个古诗词专家，擅长于将古诗句翻译为白话文...” input:”请阅读下列古诗，回答相关问题...”
Subtask3-Prompt	instruction:”你是一个古诗词专家，擅长于分析出古诗词所表达的情感...” input:”请阅读下列古诗，回答相关问题...”

Table 1: Prompt Templates for Different Subtasks

4 Reinforcement Learning with Composite Reward

For the sentiment reasoning subtask (Subtask 3), we employ Proximal Policy Optimization (PPO) (Schulman et al., 2017) to fine-tune the model via reinforcement learning, guided by a composite reward function that balances prediction accuracy and contextual coherence (Ziegler et al., 2019; Lyu et al., 2019). The full process is illustrated in Figure 2 and consists of the following key components:

Candidate Generation. The SFT-trained model takes the input poem and associated prompt to generate a set of candidate sentiment explanations (E_{cand}).

Sentiment Selection. A reinforcement learning policy model (Re.Model) selects one candidate sentiment e_{selected} from E_{cand} .

Reward Calculation. The reward $R_t(s_t, a_t)$ for the selected sentiment is computed using a compound function:

$$R_t(s_t, a_t) = \lambda_{\text{acc}} R_{\text{acc}}(s_t, a_t) + \lambda_{\text{cons}} R_{\text{cons}}(s_t, a_t) \quad (1)$$

where λ_{acc} and λ_{cons} are non-negative weights that control the influence of the two reward components. The prediction accuracy reward $R_{\text{acc}}(s_t, a_t)$ is defined as:

$$R_{\text{acc}}(s_t, a_t) = \begin{cases} \omega_{\text{match}} & \text{if } e_{\text{selected}} = e_{\text{gold}} \\ -\omega_{\text{mismatch}} & \text{otherwise} \end{cases} \quad (2)$$

Here, e_{gold} denotes the ground-truth sentiment, $\omega_{\text{match}} \geq 0$ is the reward for correct prediction, and $\omega_{\text{mismatch}} \geq 0$ is the penalty for incorrect prediction.

The sentiment consistency reward $R_{\text{cons}}(s_t, a_t)$ evaluates how well the selected sentiment aligns with the poem’s content and stylistic context. This is computed using a semantic similarity score between the generated sentiment expression and the reference text, based on embeddings from the text2vec. The raw similarity score is then scaled by a predefined function f_{scale} (Chaudhari et al., 2024).

Policy Update. The PPO algorithm then updates the parameters of the policy model using the calculated reward signals, encouraging the model to favor sentiment selections that are both accurate and contextually grounded (Ouyang et al., 2022).

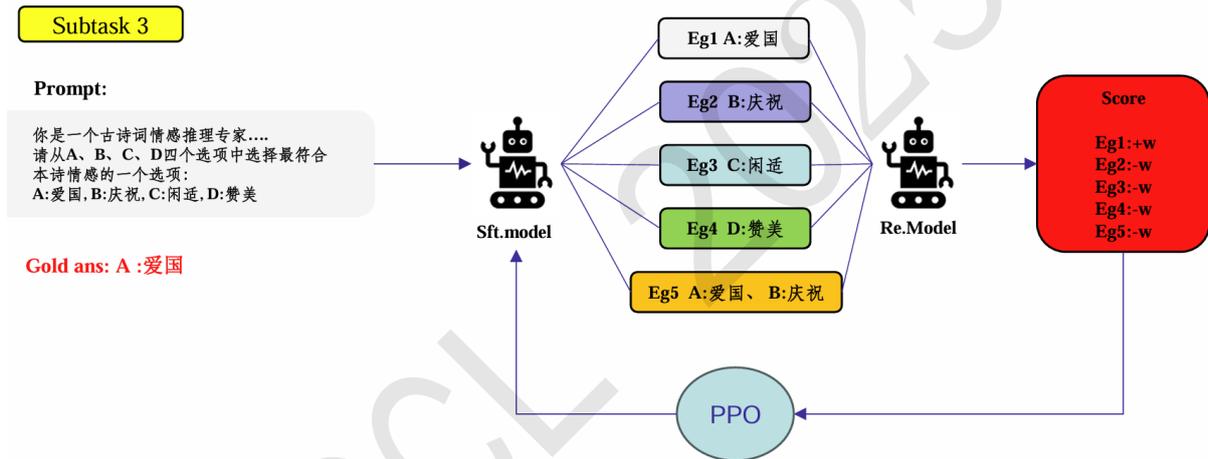


Figure 2: Diagram illustrating the PPO-based sentiment reasoning process

5 Experiments

5.1 Setting

During SFT, we used the AdamW optimizer (learning rate: $5e-5$, batch size: 2, max sequence length: 256) for up to 25 epochs, employing a 100-step learning rate warm-up. For sentiment reasoning, we fine-tuned the policy network using Proximal Policy Optimization (PPO) with the Adam optimizer (learning rate: 2×10^{-5} , weight decay: 1×10^{-4}). Training ran for 10 epochs with a total batch size of 2048 and a mini-batch size of 256. PPO configuration included $\gamma = 0.1$, $\epsilon = 0.2$, GAE $\lambda = 0.95$, and Beta=0.1. Gradient clipping (0.5) and reward normalization were applied. A random seed of 42 was used, and the loss function was sigmoid.

We set the reward weights to $\lambda_{\text{acc}} = 0.9$ and $\lambda_{\text{cons}} = 0.1$ in the composite reward function. This choice reflects the nature of the subtask, where sentiment prediction is treated as a multiple-choice question. Since identical predictions are likely to be semantically correct, prediction accuracy plays a dominant role and is thus assigned a higher weight.

Model	score	sim_sents	emo_acc	bleu_sents	taskA	sim_words	taskB	bleu_words
Team1	0.759	0.945	0.823	0.605	0.823	0.902	0.695	0.328
Team2	0.758	0.925	0.829	0.419	0.829	0.920	0.688	0.486
Team3	0.757	0.914	0.847	0.436	0.847	0.909	0.666	0.405
Ours	0.714	0.921	0.829	0.269	0.829	0.885	0.599	0.323
Baseline	0.667	0.911	0.771	0.241	0.771	0.873	0.564	0.230
Team4	0.706	0.916	0.829	0.294	0.829	0.876	0.584	0.248
Team5	0.700	0.913	0.853	0.244	0.853	0.865	0.546	0.163

Table 2: Performance comparison of different models on classical Chinese poetry tasks

5.2 Results and Analysis

Table 2 compares our model (**Ours**) with baselines (Teams 1-5, Baseline). While our overall score (0.714) is slightly lower than the top teams, we demonstrate competitive or superior performance on key subtasks, notably `emo_acc` (0.829) and Task A (0.829). Our model underperforms on translation metrics (`bleu_sents`, `sim_sents`, `bleu_words`, `sim_words`) and Task B. This highlights our framework’s strength in sentiment and comprehension tasks.

5.3 Ablation Study

To thoroughly evaluate the contributions of different training components, we conducted an ablation study, with results summarized in Table 3. We systematically compared the full configuration of our model against variants where either interleaved training, PPO, or both were removed. Our analysis reveals that the full configuration consistently delivers the best performance on both Task A (accuracy) and Task B (accuracy) across all backbone models. The progressive removal of these components leads to a noticeable decline in performance, demonstrating the significant positive impact of both interleaved training and PPO within our framework. Notably, the XunZi-Qwen-14b model achieves its highest scores under the full setting, further validating the combined effectiveness of these training strategies.

Backbone	Module	Task A (Acc)	Task B
Qwen2.5-14b-Instruction	Full Setting	0.777	0.577
	- Interleaved Training	0.765	0.567
	- PPO	0.749	0.577
	- Interleaved Training & PPO	0.737	0.497
TongGu-7b-Instruction	Full Setting	0.787	0.556
	- Interleaved Training	0.755	0.533
	- PPO	0.750	0.550
	- Interleaved Training & PPO	0.739	0.507
XunZi-Qwen-14b	Full Setting	0.829	0.599
	- Interleaved Training	0.805	0.548
	- PPO	0.798	0.562
	- Interleaved Training & PPO	0.767	0.527

Table 3: Ablation study results on Task A, Task B, and Emotion Accuracy (Emo_acc)

6 Conclusion and Future Work

Classical Chinese poetry presents significant NLP challenges. Our high-quality corpus and hierarchical SFT proved effective, with mixed SFT enhancing knowledge transfer. Ablation studies confirmed PPO significantly improved emotion accuracy. Our framework achieved promising results, and future work will focus on corpus expansion, further RL exploration, application to other classical texts, and enhancing cultural reasoning.

Acknowledgements

This work is supported by the Open Research Project of the Yunnan University Resilience and Excellence Children’s Character Development Platform (K207003250007), and Yunnan Fundamental Research Project (202501AT070231).

References

- Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *arXiv preprint arXiv:2404.08555*.
- Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. *arXiv preprint arXiv:1601.01280*.
- Alex Havrilla, Yuqing Du, Sharath Chandra Rapparthi, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskiy, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. 2024. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*.
- Yongchen Hu. 2024. An automatic generation of poetry using generative pre-trained transformer with fine tuning. In *2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC)*, pages 1–5. IEEE.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.
- Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. 2018. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3890–3900.
- He Lyu, Ningyu Sha, Shuyang Qin, Ming Yan, Yuying Xie, and Rongrong Wang. 2019. Advances in neural information processing systems. *Advances in neural information processing systems*, 32.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Rahul Manohar Samant, Mrinal R Bachute, Shilpa Gite, and Ketan Kotecha. 2022. Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions. *IEEE Access*, 10:17078–17097.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. *arXiv preprint arXiv:2405.04655*.
- Weiwei Sun, Yufei Chen, Xiaojun Wan, and Meichun Liu. 2019. Parsing chinese sentences with grammatical relations. *Computational Linguistics*, 45(1):95–136.
- Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv preprint arXiv:2106.16038*.
- Wei Zhang, Hao Wang, Min Song, and Sanhong Deng. 2023. A method of constructing a fine-grained sentiment lexicon for the humanities computing of classical chinese poetry. *Neural Computing and Applications*, 35(3):2325–2346.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Appendix

A Examples of the fine-tuning data

A.1 Subtask1:

{

"instruction": "你是一个古诗词专家，擅长于将古诗词翻译为白话文\n\n示例：\n请阅读下列古诗，回答相关问题（注意：只需要回答问题，其他的都不要输出，请严格按照示例中的答案格式来回答问题）：\n古诗词题目：泊秦淮\n古诗词内容：烟笼寒水月笼沙，夜泊秦淮近酒家。商女不知亡国恨，隔江犹唱后庭花。 \n\n问题：\n将下列内容翻译为白话文，要求翻译的白话文语意连贯，读起来句子通顺，朗朗上口（注意：白话文写在冒号后面，和问题同一行，不要换行，并且白话文）：\n1、泊：\n2、商女：\n3、后庭花：\n\n答案：\n1、泊：停泊\n2、商女：歌女\n3、后庭花：歌曲《玉树后庭花》的简称\n"

"input": "\n请阅读下列古诗，回答相关问题（注意：只需要回答问题，其他的都不要输出，请严格按照示例中的答案格式来回答问题）：\n古诗词题目：奉和中书舍人贾至早朝大明宫\n古诗词内容：鸡鸣紫陌曙光寒，鸂鶒皇州春色阑。金阙晓钟开万户，玉阶仙仗拥千官。花迎剑佩星初落，柳拂旌旗露未干。独有凤凰池上客，阳春一曲和皆难。 \n\n问题：\n将下列内容翻译为白话文，要求翻译的白话文语意连贯，读起来句子通顺，朗朗上口（注意：白话文写在冒号后面，和问题同一行，不要换行）：\n1、和：\n2、中书舍人：\n3、大明宫：\n4、紫陌：\n5、曙光：\n6、鸂鶒：\n7、皇州：\n8、阑：\n9、金阙：\n10、万户：\n11、玉阶：\n12、仙仗：\n13、剑佩：\n14、旌旗：\n15、凤凰池上客：\n16、阳春\n"

"output": "\n答案：\n1、和：即和诗，是用来和答他人诗作的诗，依照别人诗词的格律或内容作诗词。可和韵，可不和韵。 \n2、中书舍人：官名，时贾至任此职。 \n3、大明宫：宫殿名，在长安禁苑南。 \n4、紫陌：指京师的街道。 \n5、曙光：破晓时的阳光。 \n6、鸂鶒：婉转的叫声。 \n7、皇州：京都。 \n8、阑：尽。 \n9、金阙：皇宫金殿。 \n10、万户：指皇宫中宫门。 \n11、玉阶：指皇宫中大明宫的台阶。 \n12、仙仗：天子的仪仗。 \n13、剑佩：带剑、垂佩绶，都为高官之饰物，此指禁卫军的武装。 \n14、旌旗：旗帜的总称。 \n15、凤凰池上客：指贾至。凤凰池，也称凤池，这里指中书省。 \n16、阳春：古曲名，即宋玉《对楚王问》中提到的《阳春》《白雪》，后以之比喻作品高妙而懂得的人很少。 \n"

}

A.2 Subtask2:

{

"instruction": "你是一个古诗词专家，擅长于将古诗句翻译为白话文\n\n示例：\n请阅读下列古诗，回答相关问题（注意：只需要回答问题，其他的都不要输出，请严格按照示例中的答案格式来回答问题）：\n古诗词题目：泊秦淮\n古诗词内容：烟笼寒水月笼沙，夜泊秦淮近酒家。商女不知亡国恨，隔江犹唱后庭花。 \n\n问题：\n将下列内容翻译为白话文，要求翻译的白话文语意连贯，读起来句子通顺，朗朗上口（注意：白话文写在冒号后面，和问题同一行，不要换行）：\n1、烟笼寒水月笼沙：\n2、夜泊秦淮近酒家：\n\n答案：\n1、烟笼寒水月笼沙：迷离的月色下，轻烟笼罩寒水、白沙\n2、夜泊秦淮近酒家：夜晚船只停泊在秦淮边靠近岸上的酒家\n"

"prompt": "\n请阅读下列古诗，回答相关问题（注意：只需要回答问题，其他的都不要输出，请严格按照示例中的答案格式来回答问题）：\n古诗词题目：水调歌头·明月几时有\n古诗词内容：明月几时有？把酒问青天。不知天上宫阙，今夕是何年。我欲乘风归去，又恐琼楼玉宇，高处不胜寒。起舞弄清影，何似在人间？转朱阁，低绮户，照无眠。不应有恨，何事长向别时圆？人有悲欢离合，月有阴晴圆缺，此事古难全。但愿人长久，千里共婵娟。 \n\n问题：\n将下列内容翻译为白话文，要求翻译的白话文语意连贯，读起来句子通顺，朗朗上口（注意：白话文写在冒号后面，和问题同一行，不要换行）：\n1、明月几时有？\n2、把酒问青天。：\n"

"output": "\n答案：\n1、明月几时有？：明月从什么时候才开始出现的？\n2、把酒问青天。：我端起酒杯遥问苍天。 \n"

}

A.3 Subtask3:

{

```
"instruction": "你是一个古诗词专家，擅长于分析出古诗词所表达的情感。\\n\\n示例：\\n请阅读下列古诗，回答相关问题（注意：只需要回答问题，其他的都不要输出，请严格按照示例中的答案格式来回答问题）：\\n古诗词题目：泊秦淮 \\n古诗词内容：烟笼寒水月笼沙，夜泊秦淮近酒家。商女不知亡国恨，隔江犹唱后庭花。\\n\\n问题：\\n最符合本诗情感是什么？：\\n\\n答案：\\n爱国\\n",  
"input": "\\n请阅读下列古诗，回答相关问题（注意：只需要回答问题，其他的都不要输出，请严格按照示例中的答案格式来回答问题）：\\n古诗词题目：奉和中书舍人贾至早朝大明宫\\n内容：鸡鸣紫陌曙光寒，莺啭皇州春色阑。金阙晓钟开万户，玉阶仙仗拥千官。花迎剑佩星初落，柳拂旌旗露未干。独有凤凰池上客，阳春一曲和皆难。\\n\\n问题：\\n最符合本诗情感是什么？：\\n",  
"output": "\\n答案：\\n渴望繁荣与美好生活的心情"  
}
```

CCL 2025