System Report for CCL25-Eval Task 7: A Two-stage Framework for Aligning LLM to Chinese Literature via Fine-Tuning and Prompting

Fan Su, Yiming Qin, Aijia Zhao, Zhenxu Wang, Zekang Huang
School of Information Science & Engineering, Yunnan University, Yunnan, CN
{sufan, qinyiming, zhaoaijia}@stu.ynu.edu.cn
wangzhenxu@hos.ynu.edu.cn, huangzekang@stu.ynu.edu.cn

Abstract

This system report presents our approach and results for the First Chinese Literary Language Understanding Evaluation (ZhengMing) task at CCL25-Eval. The ZhengMing evaluation benchmark consists of seven subtasks: Biases in Modern Literary Criticism, Modern Literary Criticism Mining, Classical Chinese Literature Comprehension, Literary Reading Comprehension, Literary Named Entity Recognition, Literary Language Style Transfer, and Literary Work Style Prediction. To address these tasks, we propose a two-stage framework named StageAli to align large language models (LLMs) to the Chinese literature domain. In the first stage, we employ Low-Rank Adaptation (LoRA) to fine-tune an LLM on Chinese literary datasets, aiming to adapt the model to Chinese literature domain. In the second stage, we utilize a combination of prompting strategies to further unleash the potential of the fine-tuned model in addressing the Chinese Literary Language Understanding task. Our proposed StageAli framework achieves second place in the overall evaluation, demonstrating the effectiveness of our method.

Keywords: Chinese Literary Language Understanding , LLM , Prompting Strategy

1 Introduction

Chinese literature embodies a profound cultural heritage through its rich history, vast literary corpus, and diverse expressive styles. A deeper understanding and mastery of Chinese literature not only contributes to the preservation and transmission of traditional Chinese culture, but also is instrumental in promoting its development and revitalization in contemporary contexts. Nevertheless, it typically requires long-term learning and extensive reading to achieve a deep understanding of literary texts, thereby limiting its accessibility to non-experts.

Recently, a great number of general-purpose LLMs–such as GPT-4 (Achiam et al., 2023), DeepSeek (Guo et al., 2025)–have demonstrated extraordinary capabilities and significant potential across various natural language processing (NLP) tasks (Han et al., 2021; Wang et al., 2023; Kojima et al., 2022; Wei et al., 2022). Various studies have explored leveraging the power of LLMs in specific domains (Cui et al., 2023; Yang et al., 2024; Cao et al., 2024), highlighting their potential to revolutionize domain-specific applications. To adapt general-purpose LLMs to specific tasks, fine-tuning has proven to be an effective transfer learning technique. It involves further training of a pre-trained model on smaller, domain-specific datasets, enabling it to achieve improved performance on specialized applications. However, in-domain tasks often encompass a variety of subtasks, making it difficult for fine-tuning to achieve consistently strong performance across all of them. Recent studies in prompt engineering have highlighted the underestimated potential of LLMs, suggesting that these models inherently possess the capabilities necessary to perform diverse tasks (Chen et al., 2023; Xie et al., 2021; Min et al., 2022). Notably, techniques such as in-context learning (ICL), chain-of-thought (CoT), and carefully crafted prompts have been shown to effectively enhance model performance in target domains without necessitating parameter updates (Cahyawijaya et al., 2024; Wei et al., 2022; Gao et al., 2020).

Proceedings of the 24th China National Conference on Computational Linguistics, pages 278-287, Jinan, China, August 11-14, 2025.

Building on the above insights, we see LLM as a powerful bridge between non-experts and Chinese literature. To this end, we propose StageAli, a two-stage framework, designed to address the ZhengMing task. Rather than relying solely on either fine-tuning or prompt engineering, our approach integrates both paradigms to effectively handle the challenges posed by ZhengMing. In phase one, we select a base LLM based on its average performance of all subtasks, and then use fine-tuning to adapt it to Chinese literature domain. In phase two, we apply a set of prompting strategies to further explore the inherent potential of the fine-tuned model.

2 Task Description

This task encompasses seven tracks, ranging from classical to modern Chinese literature, and comprehensively covers all essential aspects of Chinese literary understanding. Each track corresponds to a specific dataset, namely CritBias, CritPred, ACLUE, ReadCom, LitNRE, ClaTrans and AuthIDE, respectively.

2.1 Track 1: Biases in Modern Literary Criticism

It centers on sentiment analysis of modern literary criticism, with sentiment labels classified into three categories: positive, negative, and neutral. The dataset is constructed from literary review articles sourced from modern Chinese periodicals.

2.2 Track 2: Modern Literary Criticism Mining

This track is also an out-of-domain evaluation subtask, focusing on the extraction of reviewed publication names from literary criticism texts. It serves as an important resource for the automatic extraction and analysis of information from Chinese literary review texts.

2.3 Track 3: Classical Chinese Literature Comprehension

This track involves classical Chinese literary questions, aimed at evaluating models' abilities in cultural context understanding, sentiment analysis, knowledge integration and cross-domain comprehension.

2.4 Track 4: Literary Reading Comprehension

This track is a cloze-style reading comprehension task. It provides a thorough assessment of models' comprehension skills in complex contexts and can be applied in advancing the intelligent capabilities of models for educational use.

2.5 Track 5: Literary Named Entity Recognition

This track is a typical name entity recognition task, involving the extraction of seven entity types: time, person, organization, object, publication name, quantifier and location.

2.6 Track 6: Literary Language Style Transfer

This track focuses on classical Chinese translation and involves domains like history, philosophy, literature, and others. It is designed to assess models' capabilities in cross-lingual transfer and generalization.

2.7 Track 7: Literary Work Style Prediction

This track involves identify the author of the given literary text by analyzing its stylistic features, including linguistic expression, syntactic selection and contextual background.

3 Methodology

Figure 1 presents the overview of StageAli framework. This framework consists of two stages: model fine-tuning and prompt construction. In stage one, we apply LoRA to fine-tune a base LLM using four literary datasets provided by ZhengMing: ACLUE, ReadCom, LitNRE, and AuthIDE. In stage two, we construct customized prompts for each track through two procedures. First, we obtain Ti_1 by manually optimizing the original prompt. Then, the second procedure varies across tracks. For tracks 4, 5, and 6, we apply embedding-based demonstration retrieval to identify the most semantically similar instances,

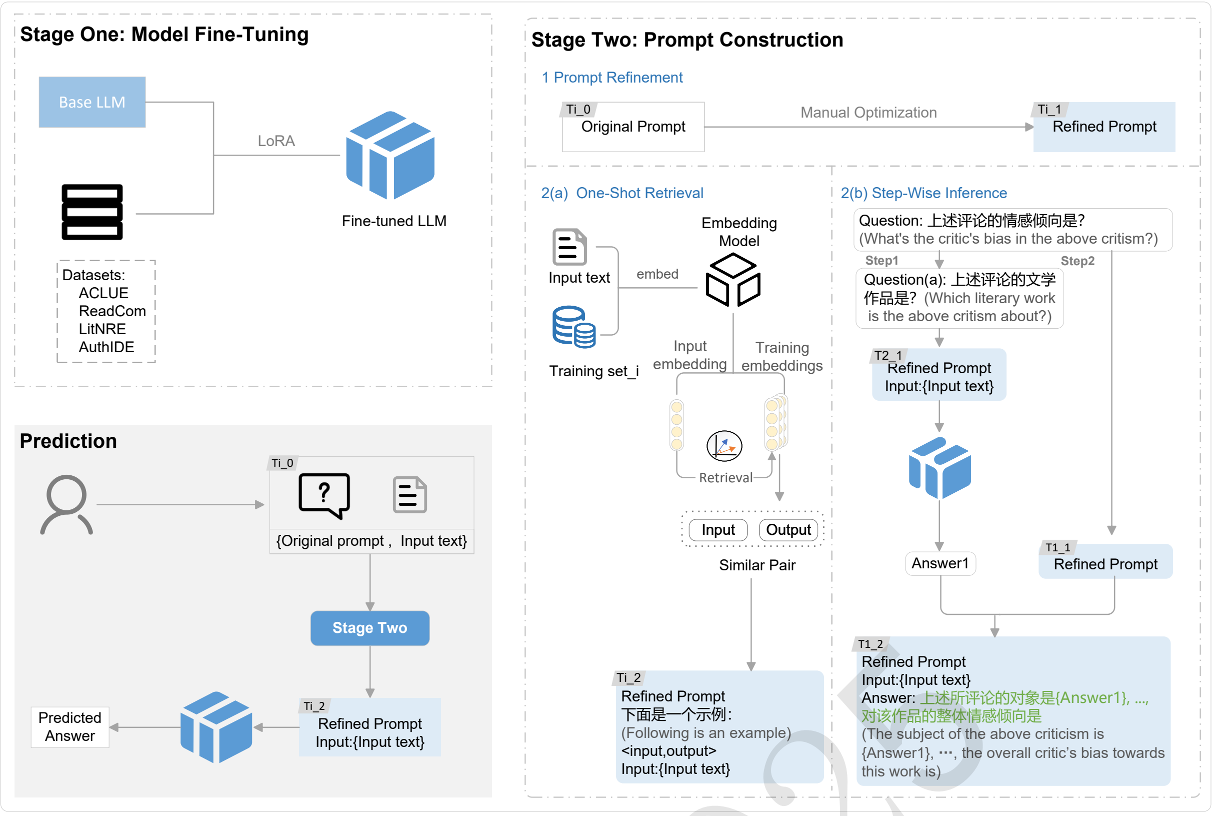

Figure 1: The overall architecture of StageAli consists of two stages: model fine-tuning and prompt construction. In stage two: prompt refinement is the first procedure of prompt construction. 2(a) and 2(b) are the second procedures, applied for track 4, 5, 6 and track 1, respectively. Prompt labeled with T_i_j denotes the prompt constructed for $\mathrm{track_i}$ after $\mathrm{procedure_j}$. Training set_i denotes the corresponding training dataset of $\mathrm{track_i}$. The words highlighted in green are reasoning hints.

which are subsequently used to construct Ti_2. For track 1, we adopt a step-wise inference strategy to decompose the one-step inference into two steps. The predicted answer from step one is used as a reasoning hint in constructing T1_2, aiming to guide the model toward more accurate inference.

3.1 Optimal Base Model Fine-Tuning

Low-Rank Adaptation (LoRA) (Hu et al., 2022), a widely used parameter-efficient fine-tuning algorithm for LLM, introduces trainable low-rank matrices into the weight update process. It allows effective adaptation with minimal additional parameters, thereby achieving high efficiency and wide applicability. In this work, we leverage LoRA to fine-tune an LLM for the target task. To identify a suitable base LLM, we first investigate the performance of five general-purpose LLMs with parameter sizes under 7 billion (7B) on the official test datasets, including Baichuan-7B (Inc., 2023), Qwen1.5-7B (Bai et al., 2023), Qwen2-7B (Team, 2024), LLama-2-7b-hf (Touvron et al., 2023) and InternLM2-7B (Cai et al., 2024). The evaluation metric is set as the average score of seven subtasks. According to the experimental results shown in Table 3, we choose InternLM2-7B as our base model. Subsequently, to strengthen the model's proficiency in the Chinese literature domain, we apply LoRA to fine-tune InternLM2-7B using the four official datasets mentioned above.

3.2 Prompting Strategy

In this work, we adopt three different prompting strategies to further harness the potential of the fine-tuned model, including prompt refinement, one-shot demonstration and step-wise inference.

Prompt Refinement. Prompt engineering has been a promising approach for adapting LLM to specific tasks. An informative prompt can be essential for steering LLMs toward more accurate and contextually appropriate responses. Building upon the original prompts, we manually refine a more specific version for each track by incorporating a detailed task description and explicit task requirements. The task requirements specify both the expected output format and the required output content.

One-shot Demonstration. In-context demonstrations have proven highly effective in improving the performance of LLMs on a wide range of tasks. Label space, distribution of the input text, the format of the input-label pairs and the number of demonstrations are all keys factors to its effectiveness (Min et al., 2022). However, demonstrations may also lead to excessively long inputs and the selection of demonstrations serves as a critical factor that significantly impacts the model's final performance. Therefore, in this work, we adopt a one-shot demonstration strategy to avoid overly long inputs and focus on the method of selecting examples with high-quality. Prior work by (Liu et al., 2021) have demonstrated that semantically similar examples are better at unleashing the capabilities of LLMs such as GPT-3. Building on this insight, we employ two similarity measures: BM25 and cosine similarity derived from semantic embeddings. As for semantic embeddings, we evaluate three pre-trained models for generating these embeddings, including BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), Llama2-7B (Touvron et al., 2023). To extract semantic representations, we use four different methods, as illustrated in Equations (1)–(5), where d_i denotes that the embeddings are extracted from the last i layers of the model, avg refers to average pooling, and cls indicates the use of the [CLS] token embedding. $E_i^{bert} \in \mathbb{R}^{1\times1024}$, $E_i^{roberta} \in \mathbb{R}^{1\times1024}$, $E_i^{llama2:d_1} \in \mathbb{R}^{1\times4096}$ and $E_i^{llama2:d_5} \in \mathbb{R}^{5\times4096}$ are the final semantic embeddings of the input text x_i obtained using the four methods described above, respectively.

$$E_i^{bert} = BERT(x_i|cls, d_1) \tag{1}$$

$$E_i^{roberta} = RoBERTa(x_i|cls, d_1) \tag{2}$$

$$E_i^{llama2:d_1} = Llama2(x_i|avg, d_1) \tag{3}$$

$$P_i^{llama2} = Llama2(x_i|avg, d_5) \tag{4}$$

$$E_i^{llama2:d_5} = Avg((Softmax(P_i^{llama2} P_i^{llama2^T}))P_i^{llama2}) \tag{5}$$

Step-wise Inference. For track 1, its criticism texts often involve multiple literary works and some of the sentiments are expressed implicitly, which can easily cause LLMs to lose focus or misidentify the target of the critique. Inspired by the previous work (Xie et al., 2023), we reformulate this subtask as a two-step question. In the first step, we extract the main subject of the critique by prompting the model with the instruction designed for track 2, which is specifically aimed at identifying the target of criticism. In the second step, we integrate the predicted answer from the first step as a reasoning hint into the refined prompt for track 1 to guide the model towards a more accurate answer.

For this task, we apply the three prompting strategies described above to address the relatively low performance observed in tracks 1–2 and 4–6 after fine-tuning, as shown in Table 4. We construct the final optimized prompt for above 5 tracks mainly through two procedures. First, we manually refine the original prompt into a clearer and more task-specific version for each track. Next, we additionally apply one-shot demonstration for track 4-6 and step-wise inference for track 1 to construct their final version of prompt. The final prompts used for each track are shown in Table 1.

4 Experimental Setup

Datasets. For this task, ZhengMing (isShayulajiao, 2025) provides seven datasets: CritBias, CritPred, ACLUE, ReadCom, LitNRE, ClaTrans and AuthIDE, corresponding to track 1 to 7, respectively. Except for CritBias and CritPred, which contain only test set, the remaining five datasets are all divided into training, validation, and test sets. Details of each dataset are shown in Table 2. During fine-tuning, we use four datasets–ACLUE, ReadCom, LitNRE and AuthIDE–excluding all the test sets and ClaTrans, since incorporating the latter during fine-tuning was found to degrade the model's overall performance. For demonstration retrieval, we use the training dataset corresponding to each track as the retrieval base to

Track	Dataset	Prompt Template / Prompt Template in English
1	CritBias	以下是一段对茅盾作品的文学评论,...,判断这段评论的整体情感倾向是积极，中性还是消极。 **评价方法**：首先确定评价对象，...， **评价标准**：如果整体积极情感超过消极，则整体为积极，...， **回答要求**：需要注意的是你只需要从[“积极”,“中性”,“消极”]中选择,…, Text:{input text} **Answer:上述评论的对象是：{Answer1}**，通过综合分析对{Answer1}的情感评论语句，判断整体的情感倾向是
		The following is a literary review of a work by Mao Dun. Your task is to determine whether the overall sentiment expressed in the review is Positive, Neutral, or Negative. **Evaluation Method**: First, identify the subject of the Review, ... **Evaluation Criteria**: If the overall positive sentiment outweighs the negative, classify the review as Positive. ... **Answer Requirements**: You only need to choose one from the following options: ["Positive", "Neutral", "Negative"]. Text: {input text} **Answer: The subject of the criticism is: {Answer1}.** Based on a comprehensive analysis of the sentiment expressions regarding {Answer1}, the overall critic's bias towards this work is:
2	CritPred	你将阅读一段来自中国近现代报刊的文学评论文本,…, 判断该段文字评论的中心作品名称。 **要求:** 1.统一使用《》符号标注中文作品名称,…, 2.给出的需要是真实存在的作品名称。 3.若文本涉及多部作品，请提取评论重点聚焦的作品名称。 请根据上述要求，准确指出被评论出版物名称： Context:{input text} \nAnswer:
		You will be given a passage of literary criticism from modern or late modern Chinese newspapers or journals. Your task is to identify the main literary work being discussed in the passage. **Requirements:** 1.Use 《》 to denote the name of the work. 2.The identified title must correspond to a real, published literary work. 3.If the text mentions multiple works, extract the one that is the primary focus of the criticism. Please follow the instructions above and accurately identify the name of the work under criticism. Context: {input text} \nAnswer:
3	ACLUE	Original prompt
4	ReadCom	**任务**：阅读给定儿童文学文本，理解其内容，并判断文中 “XXXXX”最适合填入的词语。 **要求**：1.需要充分理解给定文本内容，判断语境中最合适的词语,…. **下面是一个示例：**\n**输入：**{example input text} \n**输出**：{example ouput} 输入：{input text} \n输出：
		Task: Read the given children's literature text, understand its content, and determine the most appropriate word to fill in the blank marked “XXXXX” in the passage. **Requirements:**1.Fully comprehend the content of the given text and choose the word that best fits the context, … **The following is an example:** \n**Input:** {example input text} \n**Output:** {example output} Input: {input text} \nOutput:
5	LitNRE	任务：请从以下中文文学短句中识别出所有的命名实体。… 要求：1. 输出格式为“实体名称，实体类型”，多个实体之间用空格隔开，不换行,…, 下面是一个示例：\n输入：{example input text}\n输出：{example output} 请按照要求对下面的中国文学短句进行命名实体识别。 输入：{input text} \n输出：
		Task: Please identify all named entities in the following Chinese literary phrase. … **Requirements**: 1.The output format should be: “Entity Name, Entity Type”, with multiple entities separated by spaces, and no line breaks. … **The following is an example:** \n**Input**：{example input text} \n**Output**：{example ouput} Please perform named entity recognition on the following Chinese literary phrase according to the requirements.. Input: {input text} \nOutput:
6	ClaTrans	文言文是一种古老的书面语言，具有浓厚的历史文化背景。将文言文转换为白话文的过程需要充分理解原文的意思，…, **任务：**给你一段文言文，请你将这段文言文翻译成现代文。 **要求：**1.保证翻译内容准确且符合现代汉语的表达习惯,... **下面是一个示例:** \n**输入：**{example input text} \n**输出：**{example output} 请按照要求对下面句子进行翻译。 输入：{input text} \n输出：
		Classical Chinese is an ancient written language rich in historical and cultural significance. Converting Classical Chinese into modern vernacular Chinese requires a thorough understanding of the original text,…, **Task:** You will be given a passage in Classical Chinese. Please translate it into modern Chinese. **Requirements:** 1.Ensure the translation is accurate and conforms to the norms of modern Chinese expression. … **The following is an example:** \n**Input:** {example input text} \n**Output:** {example output} Please translate the following sentence according to the requirements. Input: {input text} \nOutput:
7	AuthIDE	Original prompt

Table 1: The prompt templates used for each track. Original prompt refers to the prompt provided in the dataset without any refinement.

Datasets	Literary Task	Instruction Set	Test Set	Average Length	Length Range
CritBias	Biases in Modern Literary Criticism	-	141	2,572	227-17,432
CritPred	Modern Literary Criticism Mining	-	829	1970	16-16,275
ACLUE	Classical Chinese Literature Comprehension	49,660	2,000	136	27-2,146
ReadCom	Literary Reading Comprehension	29,013	2,000	315	65-955
LiNRE	Literary Named Entity Recognition	27,864	2,750	45	2-2,007
ClaTrans	Literary Language Style Transfer	972,467	2,000	20	1-1,452
AuthIDE	Literary Work Style Prediction	30,324	2,000	32	2-621

Table 2: Details of Datasets

find the most similar example. For evaluation, we use test sets paired with their corresponding optimized prompts to assess the model's performance.

Metrics. Each track is evaluated using different metrics, in accordance with the official guidelines. Specifically, Accuracy (ACC), Weighted F1-score, Macro F1, and MCC are used for tracks 1, 3 and 7. Exact Match (EM) is adopted for tracks 2 and 4, while Entity Weighted F1-score is used for track 5. For track 6, we use BERTScore-F1 (Zhang et al., 2019) and BARTScore (Yuan et al., 2021).

Parameters. For fine-tuning, we utilize the LLaMA Factory framework (Zheng et al., 2024) and set the learning rate to 5e-5, the cutoff length to 1024, the batch size to 2 and train for one epoch. For evaluation, we follow the official guideline to set the max_gen_toks to 1024, 200, 1024, 200, 200, 20 and 1024 for track 1-7, respectively.

5 Results and Analysis

5.1 Comparison of different general LLMs

Table 3 shows the performance of five general-purpose LLMs on the test sets. InternLM2-7B outperforms the other four LLMs on four datasets and achieves the highest average score across seven tracks. Therefore, we select InternLM2-7B as our base model for the subsequent experiments.

5.2 Ablation Study of different components

Table 4 presents us with the comparison results of different components. According to those results, we can observe that:

(1) Fine-tuning substantially improves the performance of InternLM2-7B on several subtasks. Specifically, it achieves a 10% increase on CritPred, a remarkable 131% improvement on ACLUE, a nearly 29-fold enhancement on ReadCom, a 13% gain on ClaTrans, and a 27% increase on AuthIDE. However, these gains are accompanied by performance decreases on CritBias and LitNRE, by 32% and 11%, respectively. These results indicate that while fine-tuning effectively improves the model's capabilities in the Chinese literature domain, its impact across subtasks within the domain remains uneven.

(2) Integrating prompt refinement and one-shot demonstration further enhances the fine-tuned model's performance on ReadCom, ClaTrans and LitNRE by 75%, 17% and 40%, respectively. Among the retrieval methods for demonstration, embedding-based methods consistently outperform the traditional

Track	Dataset	Metrics	Baichuan-7B	Qwen1.5-7B	Qwen2-7B	LLama-2-7b-hf	InternLM2-7B
1	CritBias	ACC	0.021	0.227	0.156	0.043	**0.390**
		weighted F1-score	0.035	0.337	0.231	0.074	**0.475**
		Macro F1	0.050	0.278	0.194	0.088	**0.397**
		MCC	0.029	0.142	0.062	0.078	**0.216**
2	CritPred	EM	0.045	0.160	0.164	0	**0.735**
3	ACLUE	ACC	0.267	0.395	**0.492**	0.263	0.475
		weighted F1-score	0.226	0.388	**0.479**	0.175	0.467
		Macro F1	0.228	0.388	**0.478**	0.171	0.468
		MCC	0.029	0.205	**0.333**	0.011	0.317
4	ReadCom	EM	0	0.013	**0.038**	0	0.019
5	LitNRE	Entity F1	0.006	0.014	0.006	0.007	**0.028**
6	ClaTrans	BERTScore-F1	0.584	0.620	0.629	0.561	**0.700**
		BARTScore	-5.518	-5.290	**-5.256**	-5.712	-5.588
7	AuthIDE	ACC	0.515	0.532	0.500	0.444	**0.794**
		weighted F1-score	0.357	0.482	0.473	0.349	**0.802**
		Macro F1	0.340	0.482	0.473	0.349	**0.802**
		MCC	-0.048	0.171	0.202	0.004	**0.612**
Average Score			-0.272	-0.163	-0.155	-0.294	**-0.016**

Table 3: Performance comparison of five general LLMs based on their popularity and strong performance across seven tracks. The Average Score value is calculated by averaging the scores of all tasks, where the score of each task is defined as the mean value of all its associated metrics.

BM25 algorithm, underscoring the superior capability of pre-trained models in capturing semantic information. In particular, the $\mathrm{Llama2}^{d5}$ method proves to be the most effective in providing high-quality examples. Although its performance on LitNRE is slightly lower than that of BERT, it achieves the best overall results across three tasks. This can be attributed to its weighted integration of information from the last five layers, which enables a more precise semantic representation of literary texts.

(3) For two out-of-domain subtasks, our approach also achieves competitive performances. Through prompt refinement, the fine-tuned model's performance increases by 75%, 2% for CritBias and CritPred, respectively. Moreover, incorporating step-wise inference further enhances the model's performance on CritBias by 38%. These results suggest that our prompt refinement method and step-wise inference are effective in unleashing the potential of InternLM2 in the Chinese literature domain. By providing clearer instructions and decomposing the reasoning process into intermediate steps, these methods help guide the model toward more accurate answers and make complex sentiment analysis more tractable.

5.3 Overall performance of StageAli

Our team submits the final results of tracks 1–5 and 7 for official evaluation. The overall score for ZhengMing is 0.632 and ranks second among all participating teams. Table 5 compares our approach's final results with the official baseline. It is evident that StageAli substantially outperforms the baseline,

Track	Dataset	FT	PR	One-Shot	SW	Metrics		Average
1	CritBias	✓				ACC	0.262	0.252
						weighted F1-score	0.321	
						Macro F1	0.307	
						MCC	0.119	
		✓	✓			ACC	0.482	0.440
						weighted F1-score	0.480	
						Macro F1	0.432	
						MCC	0.367	
		✓	✓		✓	ACC	**0.702**	**0.607**
						weighted F1-score	**0.682**	
						Macro F1	**0.566**	
						MCC	**0.477**	
2	CritPred	✓				EM	0.811	0.811
		✓	✓			EM	**0.830**	**0.830**
3	ACLUE	✓				ACC	**0.998**	**0.998**
						weighted F1-score	**0.998**	
						Macro F1	**0.998**	
						MCC	**0.997**	
4	ReadCom	✓				EM	0.569	0.569
		✓		BM25		EM	0.700	0.700
		✓		BERT		EM	0.779	0.779
		✓		RoBERTa		EM	**0.995**	**0.995**
		✓		Llama2^{d1}		EM	0.989	0.989
		✓		Llama2^{d5}		EM	**0.995**	**0.995**
5	LitNRE	✓				Entity F1	0.025	0.025
		✓	✓	BM25		Entity F1	0.033	0.033
		✓	✓	BERT		Entity F1	**0.040**	**0.040**
		✓	✓	RoBERTa		Entity F1	0.039	0.039
		✓	✓	Llama2^{d1}		Entity F1	0.034	0.034
		✓	✓	Llama2^{d5}		Entity F1	0.035	0.035
6	ClaTrans	✓				BERTScore-F1	0.730	-2.126
						BARTScore	-4.982	
		✓	✓	BM25		BERTScore-F1	0.808	-1.871
						BARTScore	-4.550	
		✓	✓	BERT		BERTScore-F1	0.793	-1.960
						BARTScore	-4.712	
		✓	✓	RoBERTa		BERTScore-F1	0.806	-1.879
						BARTScore	-4.564	
		✓	✓	Llama2^{d1}		BERTScore-F1	0.828	-1.808
						BARTScore	-4.443	
		✓	✓	Llama2^{d5}		BERTScore-F1	**0.832**	**-1.773**
						BARTScore	**-4.378**	
7	AuthIDE	✓				ACC	**0.968**	**0.960**
						weighted F1-score	**0.968**	
						Macro F1	**0.968**	
						MCC	**0.936**	

Table 4: Performance comparison of different components. FT: model fine-tuning, PR: prompt refinement, One-Shot: one-shot demonstration, SW: step-wise inference. ✓ indicates whether the strategy is used. Average refers to the mean value of all evaluation metrics. Bold text represents the best result.

particularly on tracks 1–4 and 7. Although its performance on track 5 remains limited, StageAli delivers consistent improvements across the remaining six tracks and demonstrates strong generalization ability. These results underscores its efficacy in adapting and aligning the LLM to the Chinese literature domain.

Approach	Track 1	Track 2	Track 3	Track 4	Track 5	Track 7	Average
Baseline	0.370	0.735	0.432	0.019	0.028	0.753	0.334
StageAli (Our)	**0.607**	**0.830**	**0.998**	**0.995**	**0.035**	**0.960**	**0.632**

Table 5: Comparison of final results. Scores for each track are the average of all associated evaluation metrics. The average is calculated by dividing the total sum of the scores from six submitted tracks by 7.

6 Conclusion

In this paper, we propose a two-stage framework named StageAli to address the challenged posed by Chinese Literary Language Understanding Task (ZhengMing) to LLMs. In addition to fine-tuning the optimal backbone LLM, InternLM2-7B, to adapt it to Chinese literary task, we also utilize prompting strategies to further harness the capabilities of the fine-tuned model. Providing the LLM with more informative prompts and a semantically similar demonstration facilitates more accurate and contextually relevant responses. Additionally, decomposing the reasoning process into multi steps helps address complex sentiment analysis and reduces confusion to some extent. However, there remains substantial room for improvement in certain subtasks, like Literary Named Entity Recognition. Moreover, step-wise inference could also be applied to other subtasks, including Literary Name Entity Recognition and Modern Literary Criticism Mining. For future work, we plan to extend the step-wise inference to additional literary tasks and optimize the prompting strategies to improve overall performance.

Acknowledgments

This work is supported by the General Program of Applied Basic Research of Yunnan Province (No. 202301AT070184), the Open Project Program of Yunnan Key Laboratory of Intelligent Systems and Computing (No. ISC22Y08), and the Open Research Project of the Yunnan University Resilience and Excellence Children's Character Development Platform (No. K207003250006). We also express our gratitude to the VirtaiCloud platform for providing computational resources.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024. Tonggu: Mastering classical chinese understanding with knowledge-grounded large language models. *arXiv preprint arXiv:2407.03937*.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023. Self-icl: Zero-shot in-context learning with self-generated demonstrations. *arXiv preprint arXiv:2305.15035*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

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.

Baichuan Inc. 2023. Baichuan-7b: An open-source 7b-parameter language model. https://github.com/baichuan-inc/Baichuan-7B. Accessed: 2025-06-05.

isShayulajiao. 2025. Ccl25-eval-zhengming. https://github.com/isShayulajiao/CCL25-Eval-ZhengMing. Accessed: 2025-06-04.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot ner with chatgpt. *arXiv preprint arXiv:2310.10035*.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. 2024. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.