Overview of CCL25-Eval Task 7: Chinese Literary Language Understanding Evaluation (ZhengMing)

Kang Wang[1], Qing Wang[2], Min Peng[3], Kun Yue[1], Gang Hu[1] †
[1]School of Information Science & Engineering, Yunnan University
[2]School of Literature & Journalism, Sichuan University
[3]School of Artificial Intelligence, Wuhan University
wangkang1@stu.ynu.edu.cn, wqing@stu.scu.edu.cn
pengm@whu.edu.cn, {kyue,hugang}@ynu.edu.cn

Abstract

The 24th Chinese Computational Linguistics Conference (CCL25-Eval) features 12 technical evaluation tasks. Among them, Task 7 is the Chinese Literary Language Understanding Evaluation (ZhengMing). ZhengMing is a universal and scalable evaluation framework designed to assess natural language processing (NLP) tasks in the literary domain, such as text classification, text generation, automated question answering, relation extraction, and machine translation. ZhengMing framework aims to evaluate the performance of large language models (LLMs) in the literary field at a fine-grained level. In this mission, 89 teams signed up for the competition, with 5 teams ultimately submitting results. The highest score achieved is 0.65. This paper presents and discusses the dataset, task descriptions, competition results, and other relevant information for this evaluation task. This paper introduces and presents relevant information about this evaluation task, including the dataset, task description, and competition results. More deatails are available at https://github.com/isShayulajiao/CCL25-Eval-ZhengMing.

Keywords: Chinese Literature , Large Language Model , Dataset&Benchmark

1 Introduction

In recent years, the rapid development of large language models (LLMs) has had a transformative impact on the field of Natural Language Processing (NLP). Leading general-purpose LLMs, such as OpenAI's ChatGPT (Brown et al., 2020), Google's Gemini (Team et al., 2024), and the open-source initiative DeepSeek (Guo et al., 2025), have showcased impressive capabilities across a wide range of tasks. These models have quickly become indispensable tools in the NLP ecosystem, driving advancements in areas such as text generation (Kumichev and others, 2024), named entity recognition (Jung et al., 2024), machine translation (Merx et al., 2024), and more. Their applications have significantly contributed to the development of artificial intelligence research, influencing both academic discourse and real-world technological solutions. Despite these remarkable advancements, large language models still encounter substantial challenges when applied to the study of Chinese language and literature. In particular, their performance on complex literary tasks, which often require deep cultural and linguistic understanding, has not yet reached an ideal level (Cao et al., 2024; Zhong and Yang, 2024). To better serve the needs of Chinese literary research, there is an urgent need to establish specialized evaluation benchmarks and tailored datasets. These efforts would foster a deeper understanding of Chinese literature and facilitate more precise and effective analysis within this domain.

Chinese literature, particularly classical Chinese texts, presents formidable challenges to LLMs due to its linguistic complexity, polysemy, and the rich cultural context embedded in the language (Mair, 2010). In addition to these inherent linguistic difficulties, the high degree of specialization in literary texts necessitates that models not only understand the literary language itself but also grasp the underlying cultural and historical concepts that inform these works. Tasks such as analyzing trends in modern

†Corresponding Author

Proceedings of the 24th China National Conference on Computational Linguistics, pages 288-297, Jinan, China, August 11-14, 2025.

literary criticism, interpreting ancient literary knowledge, predicting the stylistic features of various literary works, and performing tasks like literary style transfer all require access to a rich and diverse corpus that spans centuries of literary history (Cao et al., 2024). Success in these complex tasks relies heavily on the model's ability to capture the subtle nuances of literary language, including its multifaceted meanings and stylistic elements, in order to improve performance in sophisticated literary analysis. The development of LLMs capable of handling such intricate tasks will significantly advance the field of Chinese literary studies and lead to more sophisticated models of textual analysis in the humanities.

To further enhance the capabilities of LLMs in the analysis of Chinese literary language and improve their performance in fine-grained evaluation tasks, we have carefully constructed a benchmark specifically for evaluating Chinese literary language understanding, along with seven high-quality datasets. These datasets not only encompass a variety of literary text types but also involve a range of language comprehension tasks. The Chinese literary language understanding evaluation tasks aim to promote a comprehensive improvement in the models' ability to understand aspects such as language structure, cultural connotations, and literary style.

2 Evaluation Tasks and Datasets

In this section, we present the seven literary tasks (Examples of each Chinese literary task are depicted in Table 1), which include Ancient Literature Knowledge Understanding (LitKU), Literary Reading Cloze (LitRC), Literary Named Entity Recognition (LitLC), Literary Work Style Prediction (LitSP), Literary Language Style Transfer (LitTra), Modern Literary Criticism Orientation (LitCT), and Modern Literary Criticism Mining (LitCM), along with their corresponding datasets. Among them, LitCT and LitCM do not provide a validation set, only a training set, and are used to evaluate the model's generalization ability and robustness, ensuring its adaptability across different texts and tasks.

Table 2 provides an overview of the basic information of the datasets used in this evaluation, sourced from **(a) Public Dataset** and **(b) Expert-Self-Built Datasets**, and were obtained through the following two methods: (1) Raw Collection: We collaborated with literary domain experts to meticulously clean and filter open-source datasets that were initially released but never used for LLM evaluation. (2) Expert Transcription: Based on the index of the *Chronicle of Modern Chinese Novels* (Chen, 2021), we manually entered, restored, and proofread rare datasets sourced from newspapers during the Republican era.

2.1 Data Collection

We performed a series of operations on public literary datasets, including text filtering, question crafting, and format standardization, to reconstruct Q&A pairs suitable for different literary tasks.

For ***LitCT*** and ***LitCM***, the task content is derived from literary review articles in modern newspapers (with original manuscript sources from valuable archival databases such as the Modern Newspaper Database, National Newspaper Index Database, *Dagongbao*, *Shenbao*, *Dacheng Ancient Paper Archive*, and Republican Era Book Database). They are meticulously annotated over a span of more than 20 years from Sichuan University, starting in 2004. LitCT categorizes the sentiment of literary reviews into three labels: "积极(positive)", "中性(neutral)", or "消极(negative)", offering valuable insights into the emotional tone and evaluative perspectives within critical literary discourse. LitCM focuses on identifying the titles of works being critiqued in literary criticism texts, thus enabling the extraction and in-depth analysis of valuable insights from complex and detailed literary reviews.

For ***LitKU***, It is built upon the Classical Chinese Language Understanding Evaluation Benchmark dataset (Zhang and Li, 2023). This task consists of a series of questions related to classical literature, primarily used to assess the model's ability to understand cultural context, perform sentiment analysis and emotion recognition, as well as integrate knowledge and comprehend cross-domain contexts within the framework of classical Chinese literature. Each question provides a context and four options ("A", "B", "C", "D"), with the model required to make a judgment based on the context and select the correct answer. The questions cover various aspects such as coherence in classical poetry, classical Chinese text analysis, poetry comprehension, and cultural context, making it suitable for Chinese literature education or exam evaluation. This task aims to evaluate the model's understanding of ancient literary knowledge.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 288-297, Jinan, China, August 11-14, 2025.

Oringial Chinese	CritBias	Context: 一种刊物的能否博得读者的信仰，其大部分的关系当然存于它的内容能否博得读者的同情，但往往同时有几种内容以外的些微缺憾，也会引起读者不良的反感，如出版延期装订拙劣等等... ...其实静的爱少年士官强惟力，而在强连长的身世的叙述中，轻轻补足了一句强连长是一个南洋富商的儿子，可说已嘲尽当今那些半新半旧的女子了。..... Answer: 中性
	CritPred	Context: 沫若：\n 又久没有同你通信了！你前回的信中说：“我们生在这个大砂漠中，我们彼此的信迹... ...《冲积期化石》朋友拿去了，我现在不好如何说。不过这篇小说... ...有暇并续写我的小说。 Answer: 《冲积期化石》
	ACLUE	Context: 静以修身，俭以养德”对我们提出修身养德的基本要求，这句话出自：A,诸葛亮B,欧阳修C,孔子D,孟子 Answer: C
	ReadCom	Context: 鸽子我家养了一群鸽子。它们长着白白的羽毛，尖尖的爪子，圆圆的眼睛，红红的嘴巴。飞的时候，它们张开扇形的尾巴，展开大翅膀，在空中盘旋。写出表示颜色的词有白、XXXXX 。 Answer: 红
	LitNRE	Context: 清明是人们祭扫先人，怀念追思的日子。 Answer: 清明,时间 人们,人物 先人,人物 日子,时间
	AuthIDE	Context: 后来自己看起来，明白了：何尝如此。 Answer: 鲁迅
	ClaTrans	Context: 昌渐骄贵，自言身应符谶，又为妖人王百艺所诳，僭称尊号，乃于越州自称罗平国王，年号大圣，伪命为两浙都将。 Answer: 董昌逐渐骄横显贵，自以为应当与图谶言相应，被妖人王百艺所骗，僭称尊号，就在越州自称为罗平国王，年号叫大圣，任钱为伪两浙都将。
English Tanslation	CritBias	Context: Whether a publication can gain the trust of its readers largely depends on whether its content can win the readers' sympathy. However, there are often minor imperfections beyond the content itself that may evoke negative reactions from readers, such as delays in publication, poor binding, and so on... ... In fact, the love for the young officer is quiet but strong, and within the narrative of the life story of the strong company commander, there is a subtle addition: that the company commander is the son of a wealthy businessman from Southeast Asia. This can be seen as a satire of the semi-new and semi-old women of today... ... Answer: Neutral
	CritPred	Context: It's been a while since I last corresponded with you! In your previous letter, you mentioned: "We are born in this vast desert, and our traces... The friend took away Alluvial Period Fossils, and I'm not sure what to say about it now. However, this novel... I have time and will continue writing my novel. Answer: *Alluvial Period Fossils*
	ACLUE	Context: Quietness cultivates one's character, and frugality nourishes virtue" presents the basic requirements for self-cultivation and moral development. This quote comes from:A. Zhuge LiangB. Ouyang XiuC. ConfuciusD. Mencius Answer: C
	ReadCom	Context: I have a group of pigeons at home. They have white feathers, sharp claws, round eyes, and red beaks. When they fly, they spread their fan-shaped tails, stretch their large wings, and circle in the air. The words that describe colors are white, XXXXX. Answer: Red
	LitNRE	Context: Qingming is a day for people to pay respects to their ancestors, remembering and honoring them. Answer: Qingming, time People, individuals Ancestors, individuals Day, time
	AuthIDE	Context: Later, when I looked at it myself, I understood: It was never like that. Answer: Lu Xun
	ClaTrans	Context: Dong Chang gradually grew arrogant and esteemed, claiming that he was destined to fulfill the prophecies. Deceived by the sorcerer Wang Baiyi, he usurped a royal title, declaring himself the King of Luoping in Yuezhou, adopting the reign title of "Great Sage," and falsely appointing Qian Zhang as the Commander of the Two Zhe regions. Answer: Dong Chang gradually became arrogant and influential, believing that his fate should align with the prophecies. He was deceived by the trickster Wang Baiyi, who convinced him to usurp a royal title. Dong then declared himself the King of Luoping in Yuezhou, adopting the reign title "Great Sage," and appointed Qian Zhang as the false Commander of the Two Zhe regions.

Table 1: Examples of data in the seven Chinese literary datasets.

Task	Dataset	Raw	Instruction	Data Type	Data Source	License
LitCT	CritBias	1,014	141	Modern Chinese	Republic of China Newspapers	MIT License
LitCM	CritPred	1,014	829	Modern Chinese	Republic of China Newspapers	MIT License
LitKU	ACLUE	49,660	49,660	Modern Chinese	Ancient Chinese Corpus	MIT License
LitRC	ReadCom	29,013	29,013	Modern Chinese	Children's Storybooks	CC BY-SA 4.0
LitLC	LitNRE	28,894	27,864	Modern Chinese	Literary Articles	Public
LitSP	AuthIDE	30,324	30,324	Modern Chinese	Modern Chinese literary works	Public
LitTra	ClaTrans	972,467	972,467	Classical Chinese	Ancient Literary Works	Public

Table 2: Detailed information on the raw data used for constructing instruction and evaluation data.

For ***LitRC***, composed of the Modern Chinese Cloze Test (ReadCom), is the first Chinese cloze-style document dataset (Xu et al., 2021), sourced from People's Daily and children's fairy tales. Each document consists of 10 sentences, with the 9th sentence serving as the query. The task requires the model to perform reading comprehension and extract key information from the given literary text to fill in the missing information (e.g., the XXXXX in the input text). The objective is for the model to accurately extract the appropriate words or characters from the text to complete the cloze. This task comprehensively evaluates the model's understanding abilities in complex contexts, including the logical coherence of the surrounding context and the accuracy of information extraction.

For ***LitLC***, the task is composed of the Chinese Literature Named Entity Recognition (NER) dataset (Xu et al., 2017), which is specifically tailored for Chinese literary works. The task requires the model to accurately identify named entities in the literary text, including "时间" (time),"人物" (persons), "人物" (organizations), "事物" (things), "出版商" (publication titles), "数量词" (quantifiers), and "地点" (locations), and output both the identified entities and their corresponding types.

For ***LitSP***, the task is built upon a Chinese author attribution dataset (Zhang, 2025), which requires the model to accurately identify the stylistic features of a given literary text and determine the author of the text. This dataset facilitates a style recognition task across literary works, enabling a more comprehensive evaluation of the model's natural language understanding capabilities as well as its ability to attribute stylistic features to authorship.

For ***LitTra***, the task is built upon the Classical-Modern dataset (NiuTrans, 2025), which contains pairs of Classical Chinese sentences and their modern Chinese translations. These examples encompass the linguistic diversity of ancient China, covering areas such as history, philosophy, and literature, while highlighting the concise nature and cultural context of Classical Chinese. The sentences are standardized to retain typical Classical Chinese expressions while ensuring the modern translations are clear and accessible. Given the significant differences in grammar between Classical and Modern Chinese, this translation task challenges the model's ability to understand and parse complex linguistic forms.

2.2 Instruction Construction

After integrating the raw data, the dataset for this evaluation competition consists of 85,840 samples, requiring further careful prompt selection, design, and refinement. Given the expertise needed for effective prompt design, we assemble a team of experts to annotate and review the prompts for each task. However, due to the adaptive nature of LLMs, some results may exhibit errors or anomalies. To ensure the reliability and generalizability of the prompts, we manually evaluate and retain 15-20 high-quality prompts per dataset. Specifically, a team of five graduate students and four undergraduates, all with backgrounds in literature and linguistics, conducted a comprehensive review. Each participant assessed five prompts based on three criteria—Accuracy (ACC), Naturalness (NAT), and Informativeness (INF)—using a Likert scale (DeVellis and Thorpe, 2021) ("Strongly Disagree (0)", "Disagree (1)", "Agree (2)", "Strongly Agree (3)"). Only prompts with an average score (AVG) above 2 were retained. These prompts were then iteratively tested on various large language model platforms, including OpenAI's ChatGPT (Brown et al., 2020), Baidu's ERNIE Bot (Sun et al., 2020), and Deepseek (liu, 2024). After five rounds of cross-validation, only prompts with scores above 2 are kept.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 288-297, Jinan, China, August 11-14, 2025.

CritBias	query: 我提供的这段文本是对茅盾的文学作品的评论，这段文本是从近现代报刊中提取而来，你需要对这段评论文本进行情感分析，你仅需从[“积极”,“中性”,“消极”]中选择一个作为你的答案且你的答案不需要添加额外的解释说明。Text:站在无尽头人生沙漠的旅途上，你们要些什么，悲观的人们... ...一种满足的刺激，道德家伪善者于我们有什么需要呢? answer: 积极 choices: [“消极”, “中性”, “积极”], gold: 2 text: 站在无尽头人生沙漠的旅途上，你们要些什么... ...一种满足的刺激，道德家伪善者于我们有什么需要呢?
CritPred	query: 我给出的这段评论文本是从近现代报刊中摘取的，请你判断出这段文本所评论的出版物名，你给出的答案只需提供书名，例如：《洋泾浜奇侠》。\n Context:一种刊物的能否博得读者的信仰... ...其实静的爱少年士官强惟力，而在强连长的身世的叙述中，轻轻补足了一句强连长是一个南洋富商的儿子，可说已嘲尽当今那些半新半旧的女子了。.....\n Answer: answer: 《幻灭》 text: 一种刊物的能否博得读者的信仰... ...其实静的爱少年士官强惟力，而在强连长的身世的叙述中，轻轻补足了一句强连长是一个南洋富商的儿子，可说已嘲尽当今那些半新半旧的女子了。.....
ACLUE	query: 根据所给的古文文学问题和ABCD四个选项，选择你认为最合适的一个。你应该给出“A”、“B”、“C”或“D”，你只需输出字母。Context:下列古诗词前后文连贯性最高的是A,阴雨难侵牖—春虫足哺儿—年年秋报喜—牛女有佳期B,敛眉语芳草—何许太无情—正见离人别—春心相向生C,神爽窗前月—风流洞口烟—山源久湮没—重此渊泉D,家有良医病转多—无栖泊处最诳讹—水如蓝也花如锦—依旧檐声滴旧窠Answer: answer: B choice: [“A”, “B”, ‘C”, “D”] gold: 1 text: 下列古诗词前后文连贯性最高的是A,阴雨难侵牖—春虫足哺儿—年年秋报喜—牛女有佳期B,敛眉语芳草—何许太无情—正见离人别—春心相向生C,神爽窗前月—风流洞口烟—山源久湮没—重此渊泉D,家有良医病转多—无栖泊处最诳—水如蓝也花如锦—依旧檐声滴旧窠
ReadCom	query: 对于我现在展示的这段从《儿童文学》中截取的文本，请告知我一个最适合填写在文中XXXXX处的词语或汉字，你的答案仅需要用一个词语或一个字来表示而无需给出多余的解释说明以及标点符号，你不需要在你的答案中出现“答案：”、“Answer:”等部分。\n Text:刺猬是一种有趣的小动物... ...它浑身长满了又短又密的硬刺。这段话在写刺猬外表时先写了头、眼睛和耳朵、牙齿、XXXXX 、四肢，又写了爪子, 最后写了硬刺。 answer: 门牙 text: 刺猬是一种有趣的小动物... ...它浑身长满了又短又密的硬刺。这段话在写刺猬外表时先写了头、眼睛和耳朵、牙齿、XXXXX 、四肢，又写了爪子, 最后写了硬刺。
LitNRE	instruction: 现在我呈上一句话，这句话来自于一篇文学文章，你需要确定这句话中代表时间、人物、机构、事物、出版物名、数量词或地点的命名实体并将实体名称以及实体类型按照指定的格式输出。你的答案格式应为：“实体名称，实体类型”，例如：“这时,时间母亲,人物国家,机构风雪,事物《甲申三百年祭》,出版物名几个,数量词南北山头,地点”，若有多对命名实体以及类型，你需要用空格隔开，不需要进行换行。需要注意的是你不需要输出除答案之外的其他内容，你仅需要输出实体以及其对应的类型而不需要给出相关的解释说明且你给出的答案必须在同一行中输出。\n Context:俗话说，穷人的孩子早当家，母亲11岁时姥姥去世了，16岁出嫁后，没有几年，奶奶也去世了。\n Answer: answer: 穷人,人物孩子,人物母亲,人物11岁时,时间姥姥,人物16岁出嫁后,时间几年,时间奶奶,人物 label: [“O”, “O”, “O”, “O”, “B_人物”, “I_人物”, “O”, “B_人物”, “I_人物”, “O”, “O”, “O”, “O”, “B_人物”, “I_人物”, “B_时间”, “I_时间”, “I_时间”, “I_时间”, “B_人物”, “I_人物”, “O”, “O”, “O”, “O”, “B_时间”, “I_时间”, “I_时间”, “I_时间”, “I_时间”, “I_时间”, “O”, “O”, “O”, “B_时间”, “I_时间”, “O”, “B_人物”, “I_人物”, “O”, “O”, “O”, “O”, “O”] text: 俗话说，穷人的孩子早当家，母亲11岁时姥姥去世了，16岁出嫁后，没有几年，奶奶也去世了。
AuthIDE	query: 这一段文字选自鲁迅的《呐喊》、《华盖集》、《彷徨》、《朝花夕拾》、《而已集》、《南腔北调》、《二心集》、《花边文学》，莫言的《红高粱家族》、《丰乳肥臀》，请判断出这句话所对应的作者，你的答案应该是“鲁迅”或“莫言”，你只需要输出作者名字。\n Text: 但其实，除了极少数的第一流作品以外，一切全没有什么现实底的申诉的。\n Answer: answer: 但也许批评界有时也是“只许州官放火不准百姓点灯”，正如天才之在文坛一样的。 choices: [“鲁迅”, “莫言”] gold: 0 answer: 但其实，除了极少数的第一流作品以外，一切全没有什么现实底的申诉的。
ClaTrans	query: 现在你需要完成翻译任务，给你的输入是中国的古代文言文，你需要把给出的句子翻译为现代汉语。Context:昌渐骄贵，自言身应符谶，又为妖人王百艺所诳，僭称尊号，乃于越州自称罗平国王，年号大圣，伪命为两浙都将。 answer: 董昌逐渐骄横显贵，自以为应当与图谶言相应，被妖人王百艺所骗，僭称尊号，就在越州自称为罗平国王，年号叫大圣，任钱为伪两浙都将。 text: 昌渐骄贵，自言身应符谶，又为妖人王百艺所诳，僭称尊号，乃于越州自称罗平国王，年号大圣，伪命?为两浙都将。

Table 3: Examples of data in the seven Chinese literary instruction datasets.

Finally, we combined the constructed prompts with the cleaned and manually filtered original datasets to create the instruction-tuning datasets: CritBias, CritPred, ACLUE, ReadCom, LitNRE, AuthIDE and ClaTrans. In this evaluation competition, CritBias and CritPred, as out-of-domain tasks, only provided the test set. Examples of each Chinese literary task's instruction dataset are shown in Table 3.

3 Evaluation Setup

3.1 Metrics

For LitCT, LitKU, and LitSP, we use Accuracy (ACC), Weighted F1, Macro F1, and atthews correlation coefficient (MCC) (Chicco and Jurman, 2020) as evaluation metrics. For LitCM and LitRC, Exact Match (EM) is used as the evaluation metric. For LitTra, we use BERTScore-F1 and BARTScore (Yuan et al., 2021) as the evaluation metrics. For LitLC, Entity F1 is used as the evaluation metric. For the overall performance of each task, we take the average of all the metrics (All.Avg) for that task. The average value across all tasks is used as the final score for the participants' submissions.

3.2 Baselines

To evaluate the performance of Chinese-specific LLMs, we select several representative models with 6B-8B parameters as baselines for a fair comparison. The detailed evaluation results of these LLMs are presented in Table 4. Among them, InternLM2-7B, the best-performing model, is chosen as the baseline for this evaluation. Table 5 displays the baseline for this evaluation. In addition, for fairness, this evaluation discourages parameter stacking. LLMs submitted by teams must have fewer than 7B parameters, and the use of commercial LLM APIs is prohibited to maintain a competitive environment.

Dataset	Metric	Baichuan2-7B (Yang et al., 2023)	Qwen2-7B (Yang et al., 2024)	LLaMA-3-8B (Dubey et al., 2024)	InternLM2-7B (Cai et al., 2024)	DeepSeek-7B (Bi et al., 2024)	Yi-6B (Young et al., 2024)	Xunzi-7B (lclcji, 2025)
CritBias	ACC	0.007	0.156	0.021	0.390	0.050	0.064	**0.404**
	Weighted F1	0.014	0.231	0.036	**0.475**	0.048	0.086	0.451
	Macro F1	0.008	0.194	0.051	**0.397**	0.076	0.11	0.344
	MCC	-0.032	0.062	0.035	**0.216**	-0.002	0.055	0.182
CritPred	EM	0.186	0.164	0.130	**0.735**	0.028	0.015	0.000
ACLUE	ACC	0.310	**0.492**	0.274	0.475	0.213	0.072	0.275
	Weighted F1	0.284	**0.479**	0.192	0.467	0.118	0.166	0.175
	Macro F1	0.280	**0.478**	0.189	0.468	0.115	0.163	0.173
	MCC	0.087	**0.333**	0.037	0.317	0.000	0.062	0.049
ReadCom	EM	0.001	**0.038**	0.000	0.019	0.000	0.003	0.000
LitNER	EM	0.003	0.006	0.130	0.028	0.013	**0.041**	0.005
AuthIDE	ACC	0.405	0.500	0.460	**0.794**	0.525	0.418	0.656
	Weighted F1	0.384	0.473	0.485	**0.802**	0.386	0.648	0.400
	Macro F1	0.384	0.473	0.485	**0.802**	0.371	0.386	0.648
	MCC	0.154	0.202	0.063	**0.612**	0.105	0.200	0.349
ClaTrans	BERTScore-F1	0.655	0.629	0.546	0.700	**0.707**	0.627	0.591
	BERTScore	-0.505	-5.256	-5.795	-5.588	-4.930	-5.826	**-4.571**
Best Result Count		0	5	0	8	1	1	2

Table 4: The detailed evaluation results of various 6B-8B general LLMs on the seven Chinese literary instruction datasets (CritBias, CritPred, ACLUE, ReadCom, LitNRE, AuthIDE and ClaTrans). Results in bold indicate the best performance across all models.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 288-297, Jinan, China, August 11-14, 2025.

Task	Dataset	Metrics		All.Avg
LitCT	CritBias	ACC Weighted F1 Macro F1 MCC	0.390 0.475 0.397 0.216	0.370
LitCM	CritPred	EM	0.735	0.735
LitKU	ACLUE	ACC Weighted F1 Macro F1 MCC	0.475 0.467 0.468 0.317	0.432
LitRC	ReadCom	EM	0.019	0.019
LitLC	LitNRE	Entity F1	0.028	0.028
LitSP	AuthIDE	ACC Weighted F1 Macro F1 MCC	0.794 0.802 0.802 0.612	0.753
LitTra	ClaTrans	BERTScore-F1 BARTScore	0.700 -5.588	-2.444

Table 5: Evaluation metrics and baselines for the seven literary tasks.

4 Submissions and Results

The competition attracted 26 teams, including universities such as Beijing Normal University, Beijing Institute of Technology, East China Normal University, and Yunnan University, as well as companies like Pacific Insurance and iFlytek. Before the final test result submission deadline, five teams submitted evaluation results that met the requirements. The official ranking of this evaluation is shown in Table 6.

Team	Score	Rank
East China Normal University	0.65	1
Yunnan University	0.60	2
Beijing Normal University	0.54	3
Beijing Institute of Technology	0.50	4
Shanghai University of Electric Power	-0.20	5

Table 6: Final score and ranking of outstanding teams.

5 Overview of Methods

The team from East China Normal University utilized their academic expertise in classical Chinese literature and paleography to its fullest extent, focusing on the LitKU task. They carried out continual pre-training (CPT) using a self-built, domain-specific corpus of classical Chinese, comprising 11,216,638 tokens. This corpus included annotated texts from the Classical Chinese Literature Website, with lecture notes from courses such as the History of Ancient Chinese Literature and Chinese Phonology, as well as canonical reference materials such as Common Knowledge of Ancient Chinese Culture(Wang, 2008), Outline of Paleography(Qiu, 1988), Shuowen Jiezi (written during the Eastern Han Dynasty and completed during the reign of Emperor Han An, it is the first Chinese dictionary and one of the earliest in the world), and Comprehensive Dictionary of Chinese Characters (Xu, 1990). In the second stage, the team conducted supervised fine-tuning (SFT) using the instruction-tuning dataset from the organizers. They also explored integrating Chain-of-Thought (CoT) prompting to enhance instruction tuning, but ultimately found that the combination of CPT and SFT was most effective. The team achieved a final score of 0.65, ranking first among all participating teams.

The team ranked second, from Yunnan University, employed the Low-Rank Adaptation (LoRA) method to efficiently fine-tune the InternLM2-7B (Cai et al., 2024) model by introducing trainable low-rank matrices. This optimization facilitate more effective adaptation to the unique linguistic structures of Chinese literature. Additionally, they enhanced the prompts in the instruction-tuning dataset provided by the task organizers by supplementing them with rich task background information and more comprehensive guiding prompts. This approach ensured that the model had a clearer understanding of the context

Proceedings of the 24th China National Conference on Computational Linguistics, pages 288-297, Jinan, China, August 11-14, 2025.

and objectives of each task. Furthermore, the team adopted a one-shot learning strategy in conjunction with a stepwise inference approach (Xie et al., 2023), guiding the model to perform systematic, step-by-step reasoning (Liu et al., 2021). This holistic approach led to significant performance improvements across various subtasks, demonstrating the efficacy of both the fine-tuning strategy and the reasoning framework.

The team ranked third, from Beijing Normal University, designed instructions specifically for the task of Chinese literary understanding in order to simulate real-world interactive scenarios. They adopted a few-shot approach to reconstruct the input format in a dialog-based manner, with each data instance preset with 2 to 5 examples. Additionally, they performed low-rank adaptation fine-tuning on the Qwen2.5-7B-Instruct (Qwen2.5, 2024) model.

The team ranked fourth, from Beijing Institute of Technology, first constructed an unsupervised dataset and designed a dual-instruction unsupervised training mechanism. They applied the LoRA (Hu et al., 2022) algorithm for the initial fine-tuning of Baichuan2-7B-Base (Yang et al., 2023). For the second fine-tuning, they employed a sparse fine-tuning approach integrated with Adapter technology, enabling the model to better adapt to the linguistic characteristics of Chinese literature.

The team ranked fifth employed the Qwen2-7B (Yang et al., 2024) model along with a zero-shot strategy to evaluate performance across seven literary tasks. They directly applied the pre-trained model to assess its ability to transfer and adapt to complex literary tasks without any further fine-tuning. This approach aimed to examine the model's generalization capabilities and its effectiveness in handling a range of nuanced literary challenges, providing insights into its inherent potential for tackling diverse language understanding tasks.

6 Results Analysis

This competition showcased a wide range of advanced techniques and strategies aimed at enhancing LLMs for the task of classical Chinese literary understanding. Across submissions, participants demonstrated a strong emphasis on domain adaptation, instruction tuning, and efficient fine-tuning, reflecting current trends in the development of specialized NLP systems for complex literary tasks. A common and effective approach was the use of Continual Pre-training (CPT) on curated classical Chinese corpora. These domain-specific datasets—often including annotated texts, lecture notes, and canonical reference works—provided valuable linguistic and cultural grounding for the models. This stage was frequently followed by Supervised Fine-tuning (SFT) on task-specific instruction datasets, allowing models to better align with downstream objectives.

Several teams adopted LoRA to enable parameter-efficient fine-tuning of large models. This technique was often combined with additional strategies such as prompt engineering, including the enrichment of prompts with contextual or guiding information to enhance model comprehension and output quality. Others explored few-shot learning and dialog-based input reconstructions, simulating interactive scenarios to improve task relevance and model robustness. In terms of inference, techniques such as step-wise reasoning and chain-of-thought prompting were investigated to guide models through complex reasoning steps. Notably, some teams experimented with zero-shot settings, directly applying pre-trained models to evaluate their inherent generalization capabilities without further task-specific adaptation. While these methods show promising results, they each have limitations. The continual CPT approach heavily relies on corpus quality and diversity, which may restrict the model's generalization to unseen literary styles. The LoRA method, though efficient, may fail to capture the necessary details for complex tasks. Additionally, zero-shot strategies and sparse fine-tuning methods perform poorly on nuanced literary challenges without further task-specific adjustments. Future research can build on these methods to improve their ability to handle more complex Chinese literary comprehension tasks.

Overall, the competition illustrated multiple promising directions for adapting large language models to classical Chinese literary tasks. The integration of domain knowledge, efficient fine-tuning methods, and thoughtful prompt design collectively contributed to improved model performance, offering valuable insights for future research in domain-specific NLP and literary text understanding.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 288–297, Jinan, China, August 11–14, 2025.

7 Conclusion

In conclusion, the CCL25-Eval Task 7 on Chinese Literary Language Understanding Evaluation (ZhengMing) has effectively attracted a diverse set of participants from prominent academic institutions and industry leaders. However, due to the high difficulty of the tasks and the large number of tasks involved, the final number of submissions was relatively limited. The meticulously constructed benchmark, along with seven high-quality datasets, provides a rigorous framework for systematically evaluating large language models' ability to process complex Chinese literary texts. The competition results provide valuable insights into the current state of LLMs in specialized domains, revealing both the strengths and limitations of existing methods for adapting LLMs to the literary field. Overall, this evaluation marks a significant contribution to advancing natural language processing research in Chinese literature and sets the foundation for future developments in fine-grained literary language understanding.

8 Limitations

This evaluation task, while comprehensive, has several limitations. The datasets focus on a limited range of literary genres and may not capture the full diversity of Chinese literature. Certain complex aspects of literary understanding, such as deep cultural reasoning and metaphor interpretation, are not fully addressed. Additionally, the relatively small number of final submissions may affect the robustness of comparative analysis. Future work is needed to broaden dataset coverage and improve evaluation methodologies for more nuanced literary tasks.

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), the Open Research Project of the Yunnan University Resilience and Excellence Children's Character Development Platform (No.K207003250006), and the computational resource sponsorship from VirtualCloud Platform.

References

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Tom Brown, Benjamin Mann, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zheng Cai, Maosong Cao, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Jiahuan Cao, Yang Liu, Yongxin Shi, Kai Ding, and Lianwen Jin. 2024. Wenmind: A comprehensive benchmark for evaluating large language models in chinese classical literature and language arts. *Advances in Neural Information Processing Systems*, 37:51358–51410.

Siguang Chen. 2021. *Chronicle of Modern Chinese Novels (1922-1949)*. Wuhan Press.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13.

Robert F DeVellis and Carolyn T Thorpe. 2021. *Scale development: Theory and applications*. Sage publications.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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.

Sung Jae Jung, Hajung Kim, and Kyoung Sang Jang. 2024. Llm based biological named entity recognition from scientific literature. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 433–435. IEEE.

Gleb Kumichev et al. 2024. Medsyn: Llm-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer.

lclcjj. 2025. Xunziallm. Accessed: 2025-04-26. https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM.

Jiachang Liu, Dinghan Shen, et al. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Victor H Mair. 2010. *The Columbia history of Chinese literature*. Columbia University Press.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented llm prompting: a study on the mambai language. *arXiv preprint arXiv:2404.04809*.

NiuTrans. 2025. Classical chinese (ancient chinese) - modern chinese parallel corpus. https://github.com/NiuTrans/Classical-Modern. Accessed: 2025-01-19.

Xiguei Qiu. 1988. *Wenxuexue Gaikuang (Outline of the Study of Literature)*. Commercial Press, Beijing.

Qwen2.5. 2024. Qwen2.5: A party of foundation models. https://qwenlm.github.io/blog/qwen2.5/.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Li Wang. 2008. *Chinese Cultural Knowledge*. World Publishing Corporation, Beijing.

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*.

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *arXiv preprint arXiv:1711.07010*.

Shusheng Xu, Yichen Liu, Xiaoyu Yi, Siyuan Zhou, Huizi Li, and Yi Wu. 2021. Native chinese reader: a dataset towards native-level chinese machine reading comprehension. *arXiv preprint arXiv:2112.06494*.

Zhongshu Xu. 1990. *Hanyu Dazidian (The Comprehensive Dictionary of Chinese Characters)*. Sichuan Publishing Group, Sichuan Dictionary Publishing House, Hubei Changjiang Publishing Group, Chongwen Press, Chengdu, first edition edition.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

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

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.

Chen Peng Zhang. 2025. Chinese authorship identification dataset. https://gitee.com/zhang-chen-peng/Chinese-Authorship-Identification-Dataset. Accessed: 2025-01-19.

Tianyang Zhong and Zhenyuan others Yang. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.

