CCL25-Eval任务7系统报告：微调与提示协同增强大语言模型的文学语义理解

杨清怡†, 周盼盼†
北京师范大学国际中文教育学院
{yangqingyi, zhppan}@mail.bnu.edu.cn

摘要

本报告基于“第一届中国文学语言理解评测（争鸣）任务”，对Qwen2.5-7B-Instruct模型进行了低秩适配（Low-Rank Adaptation, LoRA）微调实验。任务包括五项主任务：古代文学知识理解、文学阅读完形填空、文学命名实体识别、文学作品风格预测和文学风格转换；另有两项域外任务，涉及现代文学批评倾向与批评挖掘。在有限计算资源条件下，采用LoRA 技术实现了高效参数更新，并结合少量样本提示和高质量指令设计，提升了模型在少样本条件下的鲁棒性与泛化能力。实验结果显示，该方法在五项主任务上取得了良好表现，并在域外任务中展现出显著的跨领域能力。其中，在批评挖掘任务中取得了0.847 的准确率，体现了较强的抽象推理与知识迁移能力。基于本报告方法训练的模型在所有任务的平均指标为0.540，在参赛队伍中排名第三。

关键词： LoRA 微调 ； 中国语言文学理解 ； 提示词工程 ； 跨领域评估

System Report for CCL25-Eval Task 7: Fine-Tuning and Prompting for Enhanced Literary Semantic Understanding in Large Language Models

Yang Qingyi†, Zhou Panpan†
School of International Chinese Language Education, Beijing Normal University
{yangqingyi, zhppan}@mail.bnu.edu.cn

Abstract

This report presents a Low-Rank Adaptation fine-tuning study of the Qwen2.5-7B-Instruct model, conducted as part of the First China Literary Language Understanding Evaluation (ZhengMing) tasks. The evaluation covers five main tasks—classical literary knowledge, cloze reading, named entity recognition, style classification, and style transfer—plus two out-of-domain tasks on modern literary criticism. Under limited resources, LoRA with few-shot prompting and quality instructions improved the model's robustness and generalization in low-data settings. Experimental results demonstrate strong performance across the five primary tasks and notable cross-domain transferability in out-of-domain tasks. Notably, the model achieved an accuracy of 0.847 in the criticism mining task and attained an average score of 0.540 across all tasks, ranking third among participating teams.

Keywords: LoRA Fine-Tuning , Chinese Literary Language Understanding , Prompt Engineering , Cross-Domain Evaluation

©2025 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
† 共同贡献（Equal contribution）

第二十四届中国计算语言学大会论文集，第298页-第303页，济南，中国，2025年8月11日至14日。

1 引言

中国语言文学文本兼具历史性、审美性与语体多样性，涵盖古今语言风格、文体变迁、文化典故与隐喻修辞等多个层面特征。这类文本对语言模型在语义建模、语境推理和风格迁移方面提出了远高于通用语料的挑战。构建系统化、标准化的中文文学语言评测体系，既能从任务设计层面引导模型深入理解人文语义，又有助于提升其在多任务、跨语体场景中的泛化能力，同时也为模型的可解释性、安全性与文化适应性提供了可靠的评估坐标。因此，无论是在语言智能基础能力的刻画方面，还是在中华文化的数字化传承与活化应用中，中文语言文学评测都具有不可替代的战略价值与学术意义。

随着大规模语言模型的发展，其在通用语言理解任务中已表现出强大的能力。然而，中文文学语言理解仍面临诸多独特挑战：一方面，古今汉语在语法结构、词义系统和文化背景方面差异显著，文言文常伴随晦涩语序与典故交织，难以被现代模型有效建模；另一方面，现代文学语言则高度依赖隐喻、风格、情感与作者个体表达，对模型的语境构建与跨文体迁移能力提出更高要求。在外部资源方面，中文文学文本的数据稀缺与标注困难进一步加剧了模型对人文任务的适应难度，使得当前大模型往往仅依赖少量自然语言指令，难以深入挖掘文本的深层含义。此外，大多数现有模型并非专为文学领域设计，缺乏与人文知识相结合的微调机制和评估基准，导致其在此类任务中的性能难以得到系统性验证。上述短板不仅限制了语言模型在人文学科中的深入应用，也成为当前语言智能研究亟需突破的关键瓶颈。

针对这些挑战，近年来多个面向中文语言文学的评测任务陆续提出，推动了该领域的研究体系化发展。例如，EvaHan 系列评测专注于古汉语命名实体识别，通过标准化的数据构建与指标体系，推动了古籍信息抽取的自动化与规范化研究(EvaHan, 2025)；CLUE与ChineseGLUE 等通用中文评测平台也逐步纳入成语填空、古文问答、节日文化等具有文学属性的子任务，拓展了模型对传统文化内容的理解能力(Xu et al., 2020; Hu et al., 2020)；CCL25-Eval 的Task 7 “第一届中国文学语言理解评测（争鸣）任务”则首次系统整合了古今文学文本的多类型处理任务，在风格判断、文言文翻译、批评情感识别等方面构建了覆盖多维度的人文语言理解评测框架(isShayulajiao, 2025)。

为系统评估语言模型在中文文学场景下的综合表现，本研究基于“第一届中国文学语言理解评测（争鸣）任务”展开模型训练与迁移能力分析。该评测涵盖七个子任务，横跨古今语言风格、知识理解与语言生成三个维度，能够较为全面地刻画语言模型在中文文学文本处理中的理解深度与风格迁移能力，为未来中文文学智能理解系统的构建与人文语言智能的发展提供了方法参考与评估支撑。

评测任务包括五个设有训练集的主任务，以及两个仅用于推理测试的域外迁移任务：

（1）古代文学知识理解：通过古文选择题评估模型对古汉语语言、语义及典故的理解能力；
（2）文学阅读完形填空：填补文学段落空缺，评估模型的语境理解与风格一致性预测能力；
（3）文学命名实体识别：识别文本中的人物、时间、地点、事物等命名实体，衡量语言结构感知能力；
（4）文学作品风格预测：基于文本内容判断其作者（如鲁迅或莫言），评估模型对文学风格与作者特征的建模能力；
（5）文学语言风格转换：将文言文翻译为现代汉语，测试语言迁移与语义保留能力。

为进一步考察模型的泛化能力，评测还包含两项域外任务，仅提供测试集：

（1）现代文学批评倾向：判断文学批评文本的情感取向（积极、中性或消极）；
（2）现代文学批评挖掘：识别批评文本中的评论对象，评估抽象理解与实体识别能力。

评测任务设计强调模型在语言理解、风格识别、知识迁移等多个维度的表现。本研究侧重探讨在已有通用语言模型基础上，如何通过微调方法、提示词工程与迁移评估设计，提升模型在中文文学理解任务中的能力，并实现跨任务迁移。

2 方法介绍

本研究采用低秩适配（Low-Rank Adaptation, LoRA）技术对预训练语言模型进行高效微调(Hu et al., 2021)，并结合基于任务指令和少量样本提示（few-shot prompting）的提示词构建方式(Brown et al., 2020)，实现对多种中文文学类任务的适应能力提升。

第二十四届中国计算语言学大会论文集，第298页-第303页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

2.1 模型选择

鉴于任务限制模型规模不超过7B，结合现有榜单中各模型在中文语言任务上的表现(Jeinlee1991, 2025; Xu et al., 2023)，本研究选取Qwen2.5-7B-Instruct (Yang et al., 2024)作为主体训练模型，以平衡语义建模能力与文本生成效果。

2.2 LoRA 微调配置

为了在有限计算资源条件下实现高效参数更新，本研究采用LoRA 方法对模型进行轻量微调。微调过程中保持原始模型的偏置项不变，通过在权重路径上引入少量可训练参数，仅更新极小比例的模型权重。在保证模型性能的同时，显著降低了训练成本和显存占用。

同时，本研究采用丢弃法（dropout），将dropout设置为5%，随机将神经网络中5%的神经元的输出设置为零，避免模型对训练数据的过拟合，提升其在小样本条件下的泛化能力。

2.3 提示词设计

为提升模型在多任务场景下的理解能力与输出质量，本研究采用“任务指令+固定few-shot示例”的方式构建对话格式输入，帮助模型在上下文语境中准确理解任务意图并生成输出。如图 1展示了文言文翻译任务的Few-shot 对话格式。本研究针对模型推理过程中的典型错误（bad cases）进行了系统性分析，结合误判现象对提示语进行了精细化优化，如图 2所示。

文言文翻译任务示例

[{’role’: ’user’, ’content’: ’你精通古代文学文化，请将下列文言文翻译为现代汉语，确保语言通顺、意思准确。你只需要输出翻译后的现代汉语，不需要输出任何解释或标点修饰。文言文：臣本布衣，躬耕于南阳，苟全性命于乱世，不求闻达于诸侯。’},
{’role’: ’assistant’, ’content’: ’翻译：我本是平民，在南阳种地，在乱世中苟且保全性命，不求在诸侯中出名。’},
{’role’: ’user’, ’content’: ’你精通古代文学文化，请将下列文言文翻译为现代汉语，确保语言通顺、意思准确。你只需要输出翻译后的现代汉语，不需要输出任何解释或标点修饰。文言文：桃花潭水深千尺，不及汪伦送我情。’},
{’role’: ’assistant’, ’content’: ’翻译：桃花潭的水即使有千尺深，也比不上汪伦送别我的情谊深厚。’},
{’role’: ’user’, ’content’: ’你精通古代文学文化，请将下列文言文翻译为现代汉语，确保语言通顺、意思准确。你只需要输出翻译后的现代汉语，不需要输出任何解释或标点修饰。文言文：后坐事将刑，御谨身殿亲鞫之，获免。’},
{’role’: ’assistant’, ’content’: ’翻译：后犯罪行将受刑，太祖亲临谨身殿讯问，得以释免。’}]

Figure 1: 文言文翻译任务的提示词示例

命名实体识别任务提示词

你是一位命名实体识别专家。
任务：请识别下列句子中的时间、人物、机构、事物、出版物名、数量词或地点等命名实体，输出格式为“实体名称,实体类型”，多个实体之间以空格分隔，不要换行，不需要输出其他内容。
输出示例：一排,数量词白杨树,事物万千,数量词白杨树,事物农家乐,机构
请记住：我们、你们、他们、你、我、她、他，这类人称代词都是人物。
句子：”清明是人们祭扫先人，怀念追思的日子。”
答:

Figure 2: 命名实体识别任务的优化后提示词示例

(c) 2025 中国中文信息学会计算语言学专业委员会

通过明确的任务指令与多轮示例交互，模型能更有效地识别当前输入的任务类型，并在上下文中生成符合预期的高质量输出。通过上述优化策略，本研究显著提升了模型在少样本场景下的表现稳定性与输出准确性，为模型在多个中文语言理解任务中的迁移泛化能力提供了基础支持。

2.4 微调参数配置

本研究基于单块NVIDIA RTX 4090（24GB显存）GPU 环境进行模型微调，采用全监督微调策略，并借助监督微调训练器（Supervised Fine-tuning Trainer, SFTTrainer）统一管理训练流程。主要超参数配置如下：

- 学习率设为1×10^{-4}，引入权重衰减系数0.05以减少过拟合风险。
- 训练时单卡批次大小（per device batch size）为1，采用梯度累计策略（gradient accumulation steps=4），等效于批次大小4，以平衡显存使用和训练效率。
- 训练总轮数（epochs）设为1，适用于固定的少样本训练场景。
- 训练采用bfloat16 混合精度模式，以提升显存利用率并兼顾训练稳定性和速度。
- 使用线性学习率调度（linear scheduler），配合以上超参数实现平滑学习率变化。
- 训练中启用随机种子（seed=1126）以保证实验可复现。
- 其他配置包括按序列长度分组（group_by_length=True）、保留全部输入字段（remove_unused_columns=False）等，以优化数据加载和训练效率。

2.5 模型评估与验证设置

实验对5个任务进行了标准化推理与性能评估。评估过程中使用了vLLM 推理后端，并结合HuggingFace 接口加载基于LoRA 微调后的因果语言模型。模型推理精度设置为`float16`，每轮推理的最大批量大小为16，以提升推理效率。

针对不同任务的输出特点，我们设置了任务专属的最大生成长度（`max_length`），例如在文学作品风格预测和现代文学批评倾向等短文本任务中，将最大长度限制为5，以防止模型生成冗余内容；而对于文学风格转换等生成型任务，则将最大生成长度调整至128 或更高，以保证生成文本的完整性和翻译质量。具体长度的设置基于任务输出的平均文本长度和实验调优结果确定，以兼顾生成质量与计算效率。

所有任务执行时均设置`--no_cache`参数以禁用缓存机制，以确保每条样本独立推理，避免状态残留对结果产生干扰。模型输出结果被写入标准路径，供后续评估指标的统计与分析使用。

2.6 迁移评估

在模型训练完成后，我们进一步评估了所训练的五个模型在两个“域外任务”上的迁移能力，具体任务为现代文学批评倾向与现代文学批评挖掘。每个模型均基于五类主任务中的单一任务进行训练，再用于在这两个未见任务上的直接推理评估。

3 评测结果与分析

为系统评估模型在各任务上的表现，表1汇总了各微调模型在相应子任务中的主要指标结果。对于现代文学批评倾向与现代文学批评挖掘两个域外任务，表中展示的是由古代文学知识理解任务微调模型所取得的最优性能，以反映其跨任务迁移能力。

3.1 跨任务迁移能力评估

在两个域外迁移任务——现代文学批评倾向与现代文学批评挖掘中，模型在准确率（Accuracy）和完全匹配率（Exact Match, EM）上分别实现了7.8% 与11.2% 的性能提升，充分体现了通过古代文学知识理解任务训练所得模型在抽象概念推理与文化语境融合方面的能力具备良好的跨任务迁移性。这表明古代文学知识理解任务所涉及的语义理解、文化背景整合

(c) 2025 中国中文信息学会计算语言学专业委员会

任务	指标	基线	本文结果	提升
现代文学批评倾向	Accuracy	0.390	0.468	+0.078
现代文学批评挖掘	EM	0.735	0.847	+0.112
古代文学知识理解	Accuracy	0.475	0.509	+0.034
	Macro-F1	0.468	0.540	+0.072
	MCC	0.317	0.372	+0.055
文学阅读完形填空	EM	0.019	0.705	+0.686
文学命名实体识别	Entity F1	0.028	0.551	+0.523
文学作品风格预测	Accuracy	0.794	0.771	—
	Macro-F1	0.802	0.762	—
	MCC	0.612	0.591	—
文学风格转换	BERTScore-F1	0.700	0.677	—
	BARTScore	-5.588	-3.606	+1.982
	Average	-2.444	-1.464	+0.980

Table 1: 实验结果

与抽象推理能力可能促成了更强的知识迁移与风格适应效果，模型不仅学会了对历史语境下的复杂意象进行解读，也能将此类知识用于分析近现代评论，从而提升对文本主体和情感倾向的捕捉精度，也为后续在多源风格文本上的泛化应用提供了有力支持。

然而，基于古代文学知识理解任务微调所得模型在批评倾向判断任务中的整体表现仍显有限，在准确率（Accuracy）和马修斯相关系数（Matthews Correlation Coefficient, MCC）两个指标上分别为0.468 和0.310，表明该模型在主观情感倾向识别方面仍存在一定挑战。宏平均F1（Macro-F1）分数为0.461，反映各类别标签分类精度均不理想，可能由标注主观性、语义模糊等因素导致，标签边界不清晰也降低了MCC 表现。反映模型对情感细粒度识别能力仍需增强，后续可引入更多主观性语言与批评风格标注样本以提升鲁棒性。

3.2 上下文理解与生成质量

在文学阅读完形填空任务中，EM从0.019升至0.705，提升高达68.6%。这一显著提升体现了模型对长文本上下文的整体把握和细节预测能力——通过Few-Shot Prompting，模型能够有效利用示例中的上下文逻辑模式，快速学会填词策略，从而在少量样本下仍具备强大的生成准确性与连贯性。

3.3 多粒度实体识别与信息抽取

在文学命名实体识别任务中，实体F1分数（Entity F1）提升至0.551，较基线增加0.523。该改进说明模型在处理文言和现代汉语中多类型实体（时间、人物、地点等）的边界识别和类别判定上均有所增强。LoRA微调使模型参数高效聚焦于实体识别子空间，而Few-Shot Prompting则提供了对特定命名规则的示例指导，有助于纠正模型对隐喻或多义词的误判。

3.4 文学风格敏感性与生成平衡

文学作品风格预测与文学风格转换两项任务分别检验了模型对篇章风格特征与语言迁移流畅性的把控能力。尽管文学作品风格预测任务中准确率略有下降，但其MCC维持在0.591，表明在二分类情形下模型仍具较高的判别稳定性；文学风格转换任务中，BERTScore-F1微降而BARTScore显著改善（+1.982），反映出生成文本在句法衔接和可读性方面的实质性增强。

4 结语

本研究展示了基于Qwen2.5-7B-Instruct 模型，采用LoRA与少样本学习策略的训练方法，在中文文学语言理解任务中展现出良好的泛化能力与较强的综合表现。通过任务定制提示词、

(c) 2025 中国中文信息学会计算语言学专业委员会

知识注入式训练和高效推理部署，模型在多个任务中实现稳定、优质的表现，尤其在任务迁移与风格适应上表现突出，为多类型中文语言处理任务提供了有力支撑。

致谢

本报告所使用的训练与评估算力由趋动云平台赞助支持。

参考文献

李斌, 冯敏萱, 许超, 等. 2025. EvaHan2025 古汉语命名实体识别评测总结报告[R/OL]. 中国人工智能学会语言智能专委会. [2025-04-30]. https://github.com/SCIR-HI/EvaHan2025.

T. Brown, B. Mann, N. Ryder, et al. 2020. Language Models are Few-Shot Learners [C/OL]//Advances in Neural Information Processing Systems: Vol. 33. Curran Associates, Inc.: 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

E. J. Hu, Y. Shen, P. Wallis, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models [A/OL]. *arXiv*. http://arxiv.org/abs/2106.09685 DOI:10.48550/arXiv.2106.09685.

M. Hu, Z. Li, Z. Zhang, et al. 2020. ChineseGLUE: A Benchmark for Natural Language Understanding in Chinese [EB/OL]. *arXiv*. http://arxiv.org/abs/2004.05986 DOI:10.48550/arXiv.2004.05986.

isShayulajiao. 2025. isShayulajiao/CCL25-Eval-ZhengMing [CP/OL]. [2025-04-30]. https://github.com/isShayulajiao/CCL25-Eval-ZhengMing.

Jeinlee1991. 2025. jeinlee1991/chinese-llm-benchmark [CP/OL]. [2025-04-30]. https://github.com/jeinlee1991/chinese-llm-benchmark.

L. Xu, J. Hou, Y. Zhang, et al. 2020. CLUE: A Chinese Language Understanding Evaluation Benchmark [C/OL]//Proceedings of the 28th International Conference on Computational Linguistics (COLING). https://www.cluebenchmarks.com/.

L. Xu, A. Li, L. Zhu, et al. 2023. SuperCLUE: A Comprehensive Chinese Large Language Model Benchmark [A/OL]. *arXiv*. http://arxiv.org/abs/2307.15020 DOI:10.48550/arXiv.2307.15020.

A. Yang, B. Yang, B. Zhang, et al. 2025. Qwen2.5 Technical Report [A/OL]. *arXiv*. http://arxiv.org/abs/2412.15115 DOI:10.48550/arXiv.2412.15115.

CCL 2025

(c) 2025 中国中文信息学会计算语言学专业委员会