CCL25-Eval 任务5系统报告: 基于千问大模型的古诗词理解与推理研究

王珏
北亚利桑那大学/ 弗拉格斯塔夫, 美国
jw2866@nau.edu

摘要

中国古典诗词语言凝练、意境深远，对自然语言处理系统提出了严峻挑战。本次评测聚焦于古诗词理解与推理，包括词语释义、句子翻译和情感分析三项子任务。本文基于Qwen2.5-14B-Instruct 模型，在LLaMA Factory 框架下采用监督微调（SFT）与LoRA 参数高效微调策略，提升模型在few-shot 条件下的表现。训练数据来自官方发布的多类别JSON 格式语料，经整合与指令格式转换后用于模型训练。实验表明，LoRA 微调显著优于zero-shot 基线。本研究验证了参数高效微调方法在有限数据场景下的有效性。

关键词： 古诗词理解 ； LoRA ； 指令微调 ； few-shot 学习 ； 情感推理

System Report for CCL25-Eval Task 5: Classical Chinese Poetry Understanding and Reasoning based on the Qwen Large Language Model

Jue Wang
Northern Arizona University / Flagstaff, USA
jw2866@nau.edu

Abstract

Classical Chinese poetry is characterized by its concise language and profound imagery, posing significant challenges for natural language processing (NLP) systems. This evaluation focuses on the understanding and reasoning of classical Chinese poetry, including three subtasks: word definition, sentence translation, and sentiment analysis. Based on the Qwen2.5-14B-Instruct model, this study employs supervised fine-tuning (SFT) and the LoRA fine-tuning strategy within the LLaMA Factory framework to enhance model performance under few-shot conditions. The training data, sourced from officially released multi-category JSON-formatted corpora, were integrated and transformed into instruction format for model training. Experimental results demonstrate that LoRA fine-tuning significantly outperforms the zero-shot baseline. This study validates the effectiveness of supervised fine-tuning methods in scenarios with limited data.

Keywords: Classical Chinese poetry understanding , LoRA , Instruction tuning , Few-shot learning , Sentiment reasoning

1 引言

中国古典诗词以其凝练的语言、深邃的意境和丰富的文化内涵，成为中华传统文化的重要组成部分。然而，由于其语言高度浓缩、修辞丰富且依赖历史文化背景，对自然语言处理系统提出了巨大挑战。本次评测任务旨在考察大语言模型在古诗词理解与推理方面的能力，具体包括词语释义、句子翻译以及情感分析三项子任务，涵盖唐诗、宋词等多种形式。

本任务基于Qwen2.5-14B-Instruct，在LLaMA Factory 框架下采用监督微调（SFT）策略，并结合LoRA 参数高效微调方法进行优化。LoRA 通过引入低秩矩阵仅更新少量参数，在保证模型性能的同时显著降低了训练资源消耗，适用于本次few-shot 数据条件下的古诗词理解与推理任务。训练数据来自官方提供的few-shot 数据集，通过遍历多级文件夹结构整合为统一格式，并转换为Alpaca 指令微调格式用于训练。实验结果显示，经过多阶段LoRA 微调后，模型性能显著优于zero-shot 基线。

本文将详细介绍系统的构建过程、训练策略及实验结果，探讨在有限数据条件下如何有效提升模型对古诗词的理解与推理能力。

2 相关工作

近年来，随着大语言模型的发展，其在中文自然语言处理任务中的表现不断提升，尤其在文本理解、生成和推理方面展现出强大能力。Qwen (Yang et al., 2024)、ChatGLM (GLM et al., 2024)、Deepseek (Liu et al., 2024)等系列模型均在多个中文基准测试中取得优异成绩，为包括古典文学在内的复杂语义任务提供了有力支持。

在古诗词理解方面，已有研究尝试将深度学习方法应用于关键词识别、句意翻译及情感分类等任务 (李靓, 2024) (张卫et al., 2021)。然而，由于古诗词语言高度凝练且依赖历史文化背景，传统模型在面对多义词、隐喻或特定修辞时仍存在理解偏差。

参数高效微调（PEFT）技术的兴起 (Lester et al., 2021)为资源受限场景下的模型适配提供了新思路。LoRA（Low-Rank Adaptation）、Prefix-tuning 和Adapter 等方法通过仅更新少量新增参数，在保持主模型权重冻结的前提下实现了接近全量微调的效果。其中，LoRA (Hu et al., 2021)因其实现简单、训练效率高而被广泛采用，尤其适用于few-shot 或数据稀缺场景。

本评测任务提供的baseline 模型Qwen2.5-7B 在zero-shot 条件下已具备一定古诗词理解能力，但仍有较大提升空间。本文在此基础上，采用LoRA 结合SFT 的方式对Qwen2.5-14B 进行微调，在有限训练数据条件下进一步优化其对古诗词的理解与推理表现。

3 模型构建与训练方法

3.1 模型选择与依据

本次评测任务旨在考察大语言模型对古诗词的理解与推理能力，包括词语释义、句子翻译及情感分析三项子任务。

Qwen2.5-instruct的常用版本包括7B/14B/32B (Team, 2024)。由于比赛规定模型参数量不得超过20B，故选择7B和14B作为候选，通过LLaMA Factory 对两者CMMLU 评分测试 (Li et al., 2023)：7B平均得分80.15，14B 平得均分83.27，14B模型在中文语言性能方面略强于7B模型。综合性能表现与模型规模限制，最终选用`Qwen2.5-14B-Instruct` 作为基础模型。

3.2 数据来源与组织结构

训练数据来自官方发布的`train-data.zip` 文件，包含共计200条古诗词样本。原始数据以多级文件夹形式组织，具体结构如下：

```
├──── 宋词
│     ├──── 北宋词
│     ├──── 南宋词
│     └──── 唐五代词
├──── 唐诗
│     ├──── 七言绝句
│     ├──── 七言律诗
│     ├──── 五言绝句
│     └──── 五言律诗
└──── 杂（诗）
```

每个子目录下均包含一个train.json 文件，记录了对应类别的古诗词训练样本及其标注信息，如题目、作者、内容、关键词解释、白话译文和情感标签等。

3.3 数据预处理流程

为了统一训练格式并适配LLaMA Factory 的监督微调流程 (Zheng et al., 2024)，对所有数据进行了整合与格式转换操作，具体步骤如下：

1. **数据整合**：遍历所有子目录下的train.json 文件，将全部样本合并为一个统一的JSON 文件。

2. **字段提取与映射**：从原始JSON 中提取关键字段，包括：

   title（题目）、author（作者）、content（诗句内容）、keywords（关键词及释义）、trans（白话译文）、emotion（情感表达）

3. **指令格式转换**：将原始数据转换为Alpaca 格式的jsonl 文件，以便进行指令微调。训练样本格式样例如下：

   {"system": "你是一个精通中国古典文学的助手，需要解析古诗词的关键词、翻译句子并分析情感。回答时需严格遵循JSON格式。", "instruction": "请解析以下古诗词，包括关键词解释、句子翻译和情感分析：", "input": "凉州词二首·其一", "output": "{\"idx\": 1, \"ans_qa_words\": {\"凉州词\": \"又名《出塞》\", \"远上\": \"远远向西望去\", \"黄河远上\": \"远望黄河的源头\", \"孤城\": \"指孤零零的戍边的城堡\", \"羌笛\": \"古羌族主要分布在甘、青、川一带\", \"何须\": \"何必\", \"杨柳\": \"指的是《杨柳曲》\", \"度\": \"吹到过\", \"玉门关\": \"汉武帝置，因西域输入玉石取道于此而得名\"}, \"ans_qa_sents\": {\"黄河远上白云间，一片孤城万仞山。羌笛何须怨杨柳，春风不度玉门关。\": \"黄河好像从白云间奔流而来，玉门关孤独地耸峙在高山中。何必用羌笛吹起那哀怨的《杨柳曲》去埋怨春光迟迟不来呢，原来玉门关一带春风是吹不到的啊！\"}, \"choose\": \"苍凉、慷慨、思乡、壮阔\"}"}

3.4 模型训练设置

训练采用LLaMA Factory 工具链进行监督微调（SFT）和LoRA 参数高效微调。主要配置如下：

训练模式：SFT + LoRA
- 实验1：zero-shot，未进行微调
- 实验2：LoRA 微调3轮
- 实验3：LoRA 微调3+3轮
- 实验4：LoRA 微调3+3+3轮
- 实验5：LoRA 微调3+3+3+3轮

训练过程中冻结主模型权重，仅更新LoRA 引入的低秩矩阵参数，有效降低了显存占用和训练成本。在趋动云平台采用单卡40G显存即可完成本训练任务。

3.5 推理与预测策略

针对测试集数据，模型需完成三项任务：

1. 对qa_words 列表中的关键词进行释义；

2. 对qa_sents 列表中的句子进行白话翻译；

3. 在choose 提供的情感选项中选出最符合的一项，并返回其索引值。

第二十四届中国计算语言学大会论文集，第206页-第211页，济南，中国，2025年8月11日至14日。

为提升生成结果的准确性与稳定性，推理阶段采用基于采样的解码方式，并设置以下参数：

- `top_p` = 0.8：使用nucleus sampling，保留概率累积达到80%的词候选集合，避免低概率错误词汇干扰；

- `temperature` = 0.5：降低softmax 温度，使分布更集中，减少随机性，提高输出一致性；

此外，在提示工程中加入明确的格式引导语句和示例输出结构，强制模型按照JSON 格式输出结果。推理后还设计了轻量级后处理模块，用于校验输出结构是否合法、修正格式错误并提取最终答案字段。

综上所述，该推理策略在保证输出质量的前提下提升了模型预测的鲁棒性，有助于获得更高BLEU 和BERTScore 分数。

4 实验结果与分析

本节展示了基于Qwen2.5-14B-Instruct 模型在不同微调策略下的实验结果，并对各项评价指标进行了系统性分析。所有实验均在LLaMA Factory 框架下完成，使用LoRA 参数高效微调方法进行优化。

4.1 实验设置与评价指标

本次评测任务采用多维度指标评估模型性能，主要包括：

理解任务（Task A）：

- BLEU 值：衡量生成释义和译文与参考答案之间的n-gram 匹配程度 (Papineni et al., 2002)；

- 中文BERTScore：基于中文预训练语言模型的相似度评分，更贴合语义一致性 (Zhang et al., 2020)。

推理任务（Task B）：

- 准确率（Accuracy）：判断模型在情感选项中选择正确答案的比例。

最终得分由两项任务加权平均得出：

$$\text{task\_score} = 0.5 \times \text{Task A 得分} + 0.5 \times \text{Task B 准确率}$$

4.2 实验结果汇总

表 1 展示了各实验设置下的详细得分情况，并与官方Baseline（Qwen2.5-7B zeroshot）进行了对比。

实验编号	score	taskA	emo_acc	taskB	bleu_words	bleu_sents	sim_words	sim_sents
Qwen2.5-7B (Zero-shot)	0.6670	0.7710	0.7710	0.5640	0.2300	0.2410	0.8730	0.9110
实验1 (Zero-shot)	0.6846	0.8100	0.8100	0.5590	0.1980	0.2530	0.8720	0.9120
实验2 (LoRA 3轮)	0.7119	0.7920	0.7920	0.6320	0.3690	0.3460	0.8940	0.9180
实验3 (LoRA 3+3轮)	0.7241	0.8130	0.8130	0.6350	0.3800	0.3450	0.8970	0.9170
实验4 (LoRA 3+3+3轮)	0.7241	0.8100	0.8100	0.6380	0.3930	0.3450	0.8970	0.9160
实验5 (LoRA 3+3+3+3轮)	0.7180	0.8200	0.8200	0.6160	0.3510	0.3120	0.8910	0.9120

Table 1: 多轮实验得分对比

4.3 LoRA 微调的有效性

从表中可以看出，引入LoRA 微调后，模型表现显著提升。与Zero-shot 实验相比，实验2（LoRA 微调3轮）的总得分从0.6846 提升至0.7119，说明即使在few-shot 场景下，LoRA 也能有效增强模型对古诗词的理解能力。

此外，BLEU得分也有明显改善，特别是在词语释义方面`bleu_words` 从0.198 提升至0.369，表明LoRA 微调有助于模型学习更精确的词汇解释。

实验2至实验5尝试了多阶段迭代训练策略，每次增加3轮训练。结果显示：实验3与实验4均取得了0.7241 的最高总分；，这种渐进式学习有助于观察模型的知识结构；

实验5进一步延长训练周期（LoRA 3+3+3+3轮），但模型性能并未持续提升，反而略有下降（总得分降至0.7180）。说明在few-shot 条件下，过度迭代会削弱模型泛化能力；特别是`bleu_sents` 明显降低（从0.345 下降至0.312），提示模型在长句生成上出现了偏差；因此，建议在后续工作中控制训练轮次，或引入早停机制防止过拟合。

4.4 与Baseline 的对比

将实验组与官方Baseline（Qwen2.5-7B zeroshot）进行对比可发现：模型在`taskA` 上表现优于Baseline（0.813 vs 0.771）；在`taskB` 上也有明显提升（0.638 vs 0.520）；BLEU和BERTScore 等具体指标也全面领先，说明所采用的LoRA 微调策略具有较强优势。

综上所述，本任务在有限的训练数据条件下，通过合理利用LoRA 微调和多阶段训练策略，成功提升了模型对古诗词内容与情感的理解与推理能力。

5 结论与展望

本文基于Qwen2.5-14B-Instruct 模型，在LLaMA Factory 框架下采用监督微调（SFT）与参数高效微调（PEFT）策略，参与了古诗词理解与推理评测任务。通过多轮LoRA 微调优化模型表现，在few-shot 条件下取得了总得分0.7241 的成绩，排名第9位。

实验结果表明，LoRA 微调能有效提升模型对古诗词的理解能力，尤其在词语释义与白话译文生成方面效果显著。同时，多阶段训练策略有助于模型逐步建立知识结构，避免一次性学习多个复杂任务带来的干扰。此外，研究发现过度迭代可能导致过拟合，影响模型泛化能力，因此建议控制训练轮次或引入早停机制。

本评估任务严格遵守赛制要求，未使用任何检索增强生成（RAG）技术，所有预测结果均由模型自身理解和生成完成。未来可尝试以下改进方向：

- 探索最新发布的Qwen3 或其它系列模型，进一步提升语言理解与生成能力；
- 引入外部词典或文化背景知识辅助关键词解释 (李佳斌et al., 2025)；
- 尝试课程学习 (Bengio et al., 2009)或任务融合策略，优化训练流程。

综上所述，本系统在有限数据条件下，通过合理选择模型与优化训练策略，成功提升了古诗词理解与推理性能。

参考文献

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Tim Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *ArXiv*, abs/2306.09212.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models, September.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

张卫, 王昊, 邓三鸿, and 张宝隆. 2021. 面向数字人文的古诗文本情感术语抽取与应用研究. 中国图书馆学报, 47(4):113–131.

李佳斌, 魏庭新, 曲维光, 李斌, 冯敏萱, and 王东波. 2025. 大语言模型下古诗笺注知识库的构建与应用. 图书馆论坛, 45(3):99–109.

李靓. 2024. 基于深度学习的古诗词情感分析. 电脑编程技巧与维护, (5):124–126.