CCL25-Eval任务5系统报告：基于风格改写与投票机制的中文古诗词赏析评测

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

摘要

本研究聚焦于古诗文理解与情感推理任务，面向CCL-EVAL任务5评测中的关键词解释、关键句意译与情感分类三个子任务，以古典诗词为核心语料，通过高质量数据清洗、模型改写和情感推理优化等策略，提升模型对复杂语义和历史情感的建模能力，探索了语言风格适配与生成策略对模型性能的影响。实验表明，经过指令微调的Qwen2.5-14B-Instruct在多项指标上优于7B模型，尤其在情感推理任务中表现突出，准确率达0.714。此外，基于多次生成结果的加权投票机制有效提高了输出稳定性。然而，引入其他古诗文数据训练与模型风格改写未提升任务正确率，暴露出数据一致性与评测机制适配性方面的问题与挑战。本研究验证了大模型在古诗文理解中的能力及提升潜力，未来可从数据质量提升、评测优化与计算效率控制等方面进一步改进。

关键词： 中文古诗词赏析；语言风格改写；加权投票机制；生成稳定性

System Report for CCL25-Eval Task 5: Style Transfer and Weighted Voting for Chinese Ancient Poetry Appreciation

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

Abstract

This work addresses the understanding and emotional reasoning of Chinese Ancient Poetry, focusing on three subtasks from CCL-EVAL Task 5: keyword explanation, paraphrasing, and sentiment classification. We employ strategies including data cleaning, style-based rewriting, and multi-round voting to enhance model performance on complex semantics and historical sentiment. Experiments show that the fine-tuned Qwen2.5-14B-Instruct outperforms the 7B version, achieving 0.714 accuracy in sentiment reasoning. The weighted voting mechanism also improves generation stability. However, incorporating additional classical texts and style rewriting during training did not improve accuracy, highlighting challenges in data consistency and evaluation alignment. Our findings demonstrate the potential of large language models for ancient text understanding and suggest directions for future improvements in data quality and evaluation methods.

† 共同贡献（Equal contribution）

Keywords: Chinese Ancient Poetry Appreciation , Style Transfer , Weighted Voting Mechanism , Generation Stability

1 任务简述

古典诗词作为典型的高文体文本，其翻译任务面临语义抽象、结构压缩与语言风格重建等多重挑战。Chen et al.（2024）在系统评估多种大型语言模型在该任务中的表现基础上，提出将adequacy（语义充分）、fluency（语言通顺）与elegance（风格优雅）作为评估维度，为该任务的建模与评价提供了方法论基础。本次古诗词理解与推理任务旨在测试语言模型在古典文学场景下的语言理解和情感推理能力。本评测任务分为三个子任务：任务A要求对古诗词中的每个关键词进行释义，任务B要求生成诗句对应的现代白话文翻译，任务C需根据整首诗歌内容，推断出诗人所表达的情感倾向，从候选选项中选择最符合的情感类别。本文构建了一个多阶段、模块化的处理框架，分别面向任务A、B和C，采用了不同规模和功能的语言模型，以达到最佳效果。

2 数据处理

2.1 数据清洗

评测官方提供了数据集，分为唐诗、宋词与杂类，进一步唐诗可以细分为五言绝句、五言律诗、七言绝句、七言律诗，宋词可以细分为唐五代词、北宋词、南宋词，包含了Task A、B、C三个子任务的原始训练数据。

为提升模型在Task A中的表现，本研究在官方提供的训练集基础上进行了进一步优化。通过对train.json文件的逐条扫描与验证，发现训练集中存在典故关键词与诗词内容不匹配的情况，影响模型学习效果。本研究对该类样本进行了分类处理：

（1）标注典故关键词未出现在标题/正文中：出现关键词与文本无关，如“芙蓉塘”“轻雷”等未在原诗中体现，已剔除或替换为正确版本。

（2）文本错误导致匹配失败：原诗错误（如《逢入京使》文本实为《黄鹤楼送孟浩然之广陵》），如训练集中已存在正确版本则予以替换；无正确版本则补充添加。

（3）表述方式差异：如“XX一句”“XX两句”等表述在文本中不可直接查找，统一调整为原句内容以便模型识别。

（4）典故拆分或跨句现象：如“浮云蔽日”被拆为“浮云”“蔽日”，导致匹配失败，已调整匹配逻辑为支持连续短语。

（5）繁简体与异体字问题：如“馀辉”“拚一醉”等造成关键词漏检，统一处理为简体写法。

在完成数据清理与优化后，本研究为Task A设计了“关键词+ 上下文补充”的提示词框架，为每个关键词从原诗中抽取相邻诗句构成局部上下文作为训练集，增强模型对词义的理解能力。如”messages”: [”role”: ”user”, ”content”: ”你是一个古诗词专家，请解释“中书舍人”的含义。上下文：标题：奉和中书舍人贾至早朝大明宫。”, ”role”: ”assistant”, ”content”: ”官名，时贾至任此职。”]。

2.2 数据增强

2.2.1 译文风格改写

近年来的研究表明，大语言模型（LLM）在微调阶段对训练数据语言风格的“熟悉度”显著影响其下游表现。Ren 等（2024）通过对比使用LLM 自生成语料与人工标注数据进行微调的实验，发现前者在推理类任务中表现更优，尤其在多步推理与风格依赖性较强的任务中优势明显，进而推断模型对自身生成内容具备更高的风格适应性和语言亲和性，表现为更低的困惑度（perplexity）与更强的收敛性，从而提升泛化性能。该结论为本研究采用风格统一、结构清晰的伪语料构建策略提供了有力理论支持。Wu 等（2025）进一步从token 层面分析了影响微调稳健性的关键因素，提出困惑度较低的token 分布有助于模型训练的稳定性与鲁棒性，其理论基

础与Ren 等（2024）关于“语言熟悉度”的假设形成互补，为本研究从数据层面优化模型理解能力提供了另一视角的支撑。

在上述研究启发下，本文提出基于Qwen2.5-72B 的译文改写增强方法，对训练集中原始古诗词翻译进行再生成，以构建符合模型表达习惯的伪标注数据，从而提升模型对古典语体的风格适应能力与表达一致性。

2.2.2 辅助训练集构建

此外，鉴于原始训练语料主要涵盖唐诗与宋词，为进一步提升模型对古诗文的覆盖广度与理解深度，本文引入“古诗文网”数据构建辅助训练集。首先，以“题目+作者”为匹配，剔除辅助训练集中与测试集重合的部分；然后，在仅保留唐宋诗词的基础上，随机抽取999首诗词的原文、注释和译文作为辅助训练语料。经格式标准化处理，该扩展语料被统一纳入Task B 的训练数据体系中。该补充语料在体裁风格、情感类型与表达复杂度方面相较原始数据具有更高多样性，有助于模型在训练阶段形成更稳健的语义建模能力与情感辨识能力。一方面，它有效弥补了原始训练集中样本来源相对单一的问题；另一方面，其标准化译文风格亦可在一定程度上缓解微调过程中语言风格漂移所带来的性能波动。通过丰富的跨诗人、跨主题样本输入，模型得以进一步学习如何从多样化的语言表达中提取出抽象的语义特征，并理解关键词与情感之间的关系，为后续在复杂古诗词上的推理任务打下基础。

2.3 加权投票

在Task C 中，模型需从预设选项中进行情感类别判定，任务形式为单项选择题。为提升预测鲁棒性与输出一致性，本文引入多轮生成（multi-pass decoding）与加权投票机制。在推理阶段，本研究使用Qwen2.5-14B-Instruct 模型，结合多组经过人工设计的提示词（prompt），对每首古诗词执行5 次独立生成，并选取其中成绩最好的3 次。每次生成的情感预测结果视作一次“投票”，根据训练阶段不同提示词下生成结果的质量表现，经验性地为不同预测选项分配权重（分别为0.5、0.75 和1），最终通过加权多数投票方式确定该诗词的情感分类输出。

该策略的理论动因可追溯至集成学习。集成多模型或多结果可以显著提升模型在复杂任务中的表现（Dong 等, 2020）。加权投票作为集成学习中的经典方法，能根据模型输出的置信度或其他启发式指标，动态调整不同候选结果的影响力，从而提升最终预测的鲁棒性与泛化能力。

此外，Zeng 等（2025）也指出，对于语言建模任务而言，单次生成存在显著的波动性与非稳定性，而采用“多次采样+ 策略集成”的方法（如majority voting）可以在无需重新训练模型的情况下，有效提升输出的正确率与一致性。此外，Zeng 等人指出，模型在测试阶段通过prompt扰动和候选答案重排序等手段进行推理集成，能够在zero-shot 或few-shot 设定下获得更稳健的表现。

因此，本文采用的多次生成机制不仅是一种工程上的性能增强手段，也符合当前关于LLM推理能力扩展的一致性策略研究结论。在古典诗词情感判断这一语义高度浓缩的任务背景下，多轮投票机制可有效缓解模型对提示词敏感、答案边界模糊等带来的噪声影响，从而在不牺牲计算效率的前提下提升整体推理可靠性。

3 模型与训练设置

3.1 模型选型

本研究首先选用Qwen2.5-7B模型进行训练与测试，模型展现出一定的古文理解能力，但在Task B意译与Task C情感推理上误判数量较多。由于大模型的规模效应，参数规模更大的模型，通常在处理复杂语义与历史文化语境方面展现更优性能。评测要求仅能使用20B以下的开源模型，参考各项榜单，本研究综合对比Meta-Llama-3.1-8B-Instruct、Baichuan-M1-14B-Instruct、Qwen2.5-14B-Instruct，最终选择Qwen2.5-14B-Instruct作为基座模型进行微调。实验结果显示其在所有任务上的得分均有显著提升，说明更强的语言建模能力在古文任务中具备优势。

3.2 超参数配置

针对本任务的特殊性，本研究对模型参数进行了适当的调整。以下是一些关键的训练超参数设置:

第二十四届中国计算语言学大会论文集，第187页-第191页，济南，中国，2025年8月11日至14日。

- 学习率：3e-4，选择此值是为了在训练过程中平衡收敛速度与稳定性。
- 循环次数：5，保证模型在有限的训练周期内充分学习。
- 批次大小：32，为提高训练效率与稳定性，选择了较为适中的批次大小。
- 学习率调整策略：采用余弦退火（cosine）策略，帮助学习率平滑衰减，避免训练过程中的震荡。
- 序列长度：设置为16384，这是为了处理较长的输入序列，尤其是古诗词中长句的建模。
- LoRA配置：LoRA的秩值设置为8，阿尔法值设置为32，丢弃率为0.1，目标模块为ALL。通过LoRA注入机制，实现了在大模型下的高效微调。

4 结果分析与讨论

4.1 总体表现

最终模型（Qwen2.5-14B-Instruct）在任务C情感推理上取得82.3%的准确率，整体得分达到71.4%，显著优于Qwen2.5-7B版本。

4.2 各任务表现指标

从下表可以看到，模型在句子相似度（sim_sents）和情感判断准确率（emo_acc）方面得分较高，表明其对古文语义结构和整体情绪的把握能力较强。关键词BLEU得分相对偏低，说明在局部词义表达上尚有改进空间。

任务类别	得分
词相似度（sim_words）	0.886
词BLEU（bleu_words）	0.343
句相似度（sim_sents）	0.905
句BLEU（bleu_sents）	0.288
Task A: 关键词解释	0.823
Task B：句子意译	0.606
Task C：情感推理（emo_acc）	0.823
多任务平均（score）	0.714

Table 1: 各子任务得分情况

4.3 关键发现

模型规模对性能具有显著影响。Qwen2.5-14B-Instruct 模型在古诗文理解任务中整体优于7B 版本，尤其在Task C（情感分类）任务中表现突出，反映出大模型在捕捉复杂语义关联与抽象情感表达方面具有更强建模能力，验证了语言建模能力扩展对高文体任务的性能提升效应。

高质量训练语料是优化模型表现的关键基础。在Task A（关键词释义）中，针对原始训练数据的系统性清洗与规范化处理带来了显著的性能提升。包括异体字统一、典故匹配修正与上下文提示优化等措施，有效增强了模型对关键词语义的辨析能力，进一步强调了数据一致性与语义准确性在词义建模任务中的基础性作用。

语料扩展与改写效果受限于评测机制的适配性。在Task B（句子意译）任务中，尽管改写后的译文在语义传达上更为自然，但由于BLEU 等自动评估指标对词序与表层形式极为敏感，导致风格偏离参考译文时反而得分下降。同时，古诗文网引入的外部数据存在主题与表达风格的不一致性，未能有效提升整体任务表现，暴露出语料扩展策略在缺乏风格对齐与评测一致性的情况下存在“适配性失配”的问题。

多次生成与加权投票机制有效提升情感推理任务稳定性。在Task C 中，采用多prompt多次生成结合加权投票策略，显著缓解了模型对提示敏感、输出波动性强的问题。该策略不仅

提升了情感分类的准确率，也验证了Zeng 等（2025）所提出的测试阶段集成机制对提升模型稳健性与一致性的有效性，展示了在高语义压缩文本场景下投票机制的实用价值。

5 结论

本研究系统探讨了大语言模型在古典诗文理解与情感推理任务中的应用潜力与优化路径，围绕CCL-EVAL 任务5中的关键词解释、关键句意译与情感分类三项子任务，构建了涵盖数据清洗、风格改写与生成策略的模块化处理框架。实验结果表明，Qwen2.5-14B-Instruct 在各项任务上均优于7B 版本，尤其在情感分类任务中展现出显著优势，验证了模型规模扩展对高文体语义建模的正向效应。

在数据策略方面，关键词匹配逻辑优化与清洗规范处理显著提升了模型在词义判断任务中的准确性；风格改写策略虽未显著提升BLEU 分数，但从语言稳定性与表达一致性角度增强了模型的泛化能力；而辅助语料引入过程中暴露出评测机制与风格一致性之间的适配挑战，提示未来需更精细化地控制数据来源与表达规范。

在推理策略方面，结合多轮生成与加权投票的集成机制有效缓解了模型对prompt 敏感与输出不确定性问题，显著提升了情感分类任务的鲁棒性与一致性，验证了测试阶段策略融合的实用性。

综上，本研究不仅验证了Qwen 系列模型在古典文本建模任务中的可行性，也表明在面向古文类任务进行大模型优化时，应在数据一致性、风格适配与推理稳定性之间实现多维协同。未来可进一步从高质量语料构建、评测指标体系优化以及生成成本控制等方向持续改进，为中文古典语料智能处理提供可推广的范式参考。

参考文献

A. Chen, L. Lou, K. Chen, et al. 2024. Large Language Models for Classical Chinese Poetry Translation: Benchmarking, Evaluating, and Improving. *arXiv:2408.09945*.

X. Dong, Z. Yu, W. Cao, et al. 2020. A Survey on Ensemble Learning. *Frontiers of Computer Science*, 14(2):241–258. doi:10.1007/s11704-019-8208-z.

X. Ren, B. Wu, and L. Liu. 2024. I Learn Better If You Speak My Language: Understanding the Superior Performance of Fine-Tuning Large Language Models with LLM-Generated Responses. *arXiv:2402.11192*.

C. C. Wu, Z. R. Tam, C. Y. Lin, et al. 2025. Clear Minds Think Alike: What Makes LLM Fine-Tuning Robust? A Study of Token Perplexity. *arXiv:2501.14315*.

Z. Zeng, Q. Cheng, Z. Yin, et al. 2025. Revisiting the Test-Time Scaling of o1-like Models: Do They Truly Possess Test-Time Scaling Capabilities? *arXiv:2502.12215*.