CCL25-Eval 任务9系统报告：基于大模型及指令微调方法的中医辨证辨病及中药处方生成研究

李南书
赣西肿瘤医院/ 江西省萍乡市
879075435@qq.com

摘要

辨证论治是中医认识疾病和治疗疾病的核心原则和方法，其基本思想是通过望、闻、问、切的方法，收集患者症状、舌苔、脉象等临床信息，通过分析、综合，辨清疾病的病因、病机，概括、判断为某种性质的证，进而制定个性化的治疗方案，开具合适的中药处方予以治疗。本研究探究如何增强大模型根据格式化，标准化的中医病例自动生成相对应的辨证辨病及中药处方的能力。本研究将任务拆分为辩证辨病与中药处方生成两个任务，使用的训练框架是LLamafactory，使用的大模型是开源模型（qwen2.5-7B-Instruct(Qwen Team, 2024)，qwen3-4B）。首先设置lora参数为LLamafactory默认参数，修改参数中验证集比例为0.2，epoch为5，进行lora监督微调，获得验证集相对最佳的epoch。然后，设置lora参数为默认，修改其中的epoch参数为验证集最佳epoch+1，同时对模型进行全数据lora调参优化，择其中相对最优者。最后对全数据进行full微调，与lora调参最优模型比较，择其更优者。最终在B榜中获得score1：0.648，score：0.4259，总score：0.5369，综合排名第一的成绩。

关键词： 大模型 ； 中医辩证辨病 ； 中药处方生成 ； Lora监督微调 ； Full监督微调 ； LLamafactory

System Report for CCL25-Eval Task 12: End-to-End Model and Instruction Fine-Tuning-Based Chinese Speech-Oriented Entity-Relation Triplet Extraction Research

Nanshu LI
Ganxi Cancer Hospital /Pingxiang City, Jiangxi Province
879075435@qq.com

Abstract

Syndrome differentiation and treatment is the core principle and method of traditional Chinese medicine (TCM) for understanding and treating diseases. Its basic idea is to collect clinical information such as patients' symptoms, tongue signs, and pulse conditions through the four diagnostic methods of inspection, auscultation and olfaction, inquiry, and palpation. By analyzing and synthesizing this information, the cause and pathogenesis of the disease are identified, and the disease is generalized and judged as a syndrome of a certain nature. Subsequently, a personalized treatment plan is formulated, and appropriate Chinese herbal prescriptions are prescribed for treatment.

第二十四届中国计算语言学大会论文集，第363页-第368页，济南，中国，2025年8月11日至14日。

This study explores how to enhance the ability of large models to automatically generate corresponding syndrome differentiation and Chinese herbal prescriptions based on formatted and standardized TCM medical records. The task is divided into two subtasks: syndrome differentiation and disease identification, and Chinese herbal prescription generation. The training framework used in this study is LLamafactory, and the large models employed are open-source models (qwen2.5-7B-Instruct(Qwen Team, 2024), qwen3-4B). First, the LoRA parameters were set to the default values of LLamafactory, the validation set ratio in the parameters was modified to 0.2, and the number of epochs was set to 5 for LoRA supervised fine-tuning to obtain the relatively optimal epoch on the validation set. Then, with the LoRA parameters kept as default, the epoch parameter was adjusted to the optimal validation epoch + 1, and LoRA parameter tuning and optimization were performed on the models using the full dataset to select the relatively optimal model. Finally, full fine-tuning was carried out on the full dataset, and the results were compared with those of the optimal LoRA-tuned model to select the better one. Eventually, the study achieved scores of 0.648, 0.4259, and an overall score of 0.5369 on the(B Benchmark), ranking first comprehensively.

Keywords: large language models , TCM syndrome and disease differentiation , Chinese herbal prescription generation , LoRA supervised fine-tuning , Full supervised fine-tuning , LLamafactory

1 引言

中医作为中国传统医学的重要组成部分，历经数千年的发展，已形成独具特色的理论体系和诊疗方法，对中国乃至全球人民的医疗健康做出了重要贡献。

对于辩证辨病任务（以下简称任务1）本研究主要运用的是qwen2.5-7B-Instruct大模型，分别进行了lora微调和full微调，选择两种微调方式中最佳的模型参数，最终在B榜score1获得0.648分；对于中药处方生成任务（以下简称任务2）本研究开始运用的也是qwen2.5-7B-Instruct，因lora微调与full微调成绩均较差，故而在后期临时选用了qwen3-4B模型，在截止日期前勉强在B榜score2获得了一个0.4259的成绩。

2 方法

2.1 监督微调

预训练的大模型擅长通用任务（文本生成，文本理解等），但在具体任务中表现可能不足。通过监督微调，模型可学习任务特定的模式，本研究通过构建提示词模版与数据融合，使模型加强了处理理解中医病例数据以及根据病例生成对应的辩证辨病和中药处方的能力。

2.2 Lora

LoRA（低秩适应）是Meta 于2021 年提出的参数高效微调技术(Hu et al., 2022)，旨在减少大语言模型（LLMs）微调时的参数量，同时保持接近全量微调的性能。其核心思想是通过低秩分解来近似参数更新，从而大幅降低可训练参数的数量。（本研究lora微调使用的是一张RTX4090，显存24G）

2.3 Full

“Full 微调”（Full Fine-Tuning）是指对深度学习模型的所有参数进行调整和优化的过程，与仅调整部分参数的微调方式（如LoRA、QLoRA 等）不同。在自然语言处理（NLP）领域，特别是大语言模型（LLM）的训练中，Full 微调通常用于使预训练模型更好地适应特定任务或领域的数据分布，但由于其更新全部的参数，因此需要消耗更多的计算资源。（本研究full微调使用的是两张A800显卡，显存共160G）

3 研究设置

3.1 数据集介绍

数据集基于医院脱敏病历数据而构建，共涉及10种中医证型（下称证型）、4种中医疾病、381种中药，共计1500条数据。任务旨在评估辨证论治的算法性能，包括两个子任务。子任务1：中医多标签辨证辨病基于给定的患者临床文档，判断患者所患的证型和疾病。子任务2：中药处方推荐基于给定的患者临床文档，为患者推荐合适的中药处方。

3.2 评价指标

中医多标签辨证辨病任务采用以下评测指标:

1.辨证任务准确率（acc of syndrome differentiation）

$$syndrome_{acc} = \frac{NUM(y \cap \hat{y})}{NUM(y)}$$

其中，y是数据集样本中真实证型的列表和$\hat{y}$是模型预测的数据集样本中的证型列表；$NUM(x)$代表数量函数，用来计算x 的数量。

2.辨病任务准确率（acc of disease differentiation）

$$disease_{acc} = \frac{NUM(y \cap \hat{y})}{NUM(y)}$$

其中，y是数据集样本中真实疾病的列表和$\hat{y}$是模型预测的数据集样本中的疾病列表；$NUM(x)$代表数量函数，用来计算x 的数量。

3.评价总指标

$$task1_{acc} = \frac{1}{2}(syndrome_{acc} + disease_{acc})$$

中药处方推荐任务：给定患者的基本信息和健康情况，给患者推荐一组中草药用作中药处方。输入：诊疗记录中各类临床信息构成的文本输出：中药处方，一组中草药的集合。

1.Jaccard相似系数（Jaccard Similarity Coefficient）Jaccard相似系数用于衡量两个集合的相似度，Jaccard相似系数的取值范围为[0, 1]，值越大表示预测结果与真实标签的相似度越高。计算公式如下：

$$Jaccard(y, \hat{y}) = \frac{NUM(y \cap \hat{y})}{NUM(y \cup \hat{y})}$$

其中，y是真实处方，$\hat{y}$ 是模型预测的处方，NUM(x)代表数量函数，用来计算x的数量。

2.Recall Recall 用于衡量预测结果中，与真实标签匹配的数量占所有真实标签总数的比例。计算公式如下：

$$Recall(y, \hat{y}) = \frac{NUM(y \cap \hat{y})}{NUM(y)}$$

其中，y是真实处方，$\hat{y}$ 是模型预测的处方，NUM(x)代表数量函数，用来计算x的数量。

3.Precision

Precision 用于衡量预测结果中，与真实标签匹配的数量占预测标签总数的比例。计算公式如下：

$$Precision(y, \hat{y}) = \frac{NUM(y \cap \hat{y})}{NUM(y)}$$

其中，y是真实处方，$\hat{y}$ 是模型预测的处方，NUM(x)代表数量函数，用来计算x的数量。

4.F1分数F1分数是Precision和Recall的调和平均数，用于综合衡量模型的准确性和召回率。计算公式如下：

$$F1(y, \hat{y}) = 2 \cdot \frac{Precision(y, \hat{y}) \cdot Recall(y, \hat{y})}{Precision(y, \hat{y}) + Recall(y, \hat{y})}$$

其中，y是真实处方，$\hat{y}$ 是模型预测的处方。

5.药物平均数量(Avg Herb)

药物平均数量用于衡量模型推荐的中药方剂数量与真实标签数量的接近程度。计算方法是通过比较模型推荐的中药数量和真实标签的中药数量，并计算它们的匹配度。匹配度越高，表示模型推荐的中药数量越接近真实标签的数量。计算公式如下：

$$AVG(y, \hat{y}) = 1 - \frac{|NUM(y) - NUM(\hat{y})|}{\max(NUM(y), NUM(\hat{y}))} \tag{1}$$

其中，y是真实处方，$\hat{y}$ 是模型预测的处方，NUM(x)代表数量函数，用来计算x的数量。max(a,b)代表取a,b 中的最大值，—x—代表计算x的绝对值。

6.评价总指标

$$task2\_score = \frac{1}{3} \cdot \frac{1}{N} \sum_{i=1}^{N} [Jaccard(y_i, \hat{y}_i) + F1(y_i, \hat{y}_i) + AVG(y_i, \hat{y}_i)] \tag{2}$$

其中y_i是第i条样本的真实处方，$\hat{y}_i$是模型预测的第i条样本的处方，N表示样本总数。

3.3 数据预处理

原病例数据有id，性别，职业，年龄，婚姻，病史陈述者，发病气节，主诉，症状，中医望闻问切，病史，体格检查，辅助检查，疾病（测试集为空），症型（测试集为空），处方（测试集为空），16项特征，均予以保留。

对于任务1，将疾病，症型作为output，同时删除处方特征。

然后以prompt1：“你作为三甲医院主任中医师，请严格按照以下规则分析中医病例： 诊断规则1. 疾病判断：-必须且只能从以下选项选择唯一疾病：[胸痹心痛病, 心衰病, 眩晕病, 心悸病] - 根据主诉、病史和核心症状确定疾病类型2.证型判断：- 必须从以下证型选择1-2个：[气虚血瘀证, 痰瘀互结证, 气阴两虚证, 气滞血瘀证, 肝阳上亢证, 阴虚阳亢证, 痰热蕴结证, 痰湿痹阻证, 阳虚水停证, 肝肾阴虚证] - 选择依据优先级：(1) 典型证候组合优先(2) 病程发展阶段判断(3) 核心症状的病理关联性 输出规范1. 格式要求：证型1—证型2,疾病名（注意：证型最多两个，用—分隔；疾病名单独一个）2. 严格禁止：- 添加解释性文字- 使用非列表中的名称- 证型超过两个或疾病超过一个请按以下逻辑分析：1. 解析病例中的核心症状和体征2. 匹配对应的疾病诊断标准3. 分析证候的病理组合4. 生成最终诊断”。为提示词，构建数据集。

对于任务2，将处方特征作为output。

然后以prompt2：“你是一名专业的中医师，请从提供的中医病例，给出推荐的中药处方，基于以下规则规则：1.所有的药材必须是中药材，2.最终输出格式为：[药物名1, 药物名2, 药物名3,...] 3.同一种药材只能出现一次”。为提示词，构建数据集。

最后将两组数据构造成符合LLamafactory要求的格式分别加入其data数据库并在数据库中注册。

4 研究流程（任务1和任务2基本公用同一套流程）

4.1 基本参数设置

cutoff_len: 2048
learning_rate: 0.0001
num_train_epochs: 5
lr_scheduler_type: cosine
warmup_ratio: 0.1
bf16: true
optim: adamw_torch

表 1: 模型训练基本参数设置

lora_rank: 8
lora_alpha: 16
lora_dropout: 0
lora_target: all
val_size: 0.2

表 2: LoRA参数配置表

4.2 获取相对最佳epcoh

选择qwen2.5-7B-Instruct模型，采用默认的lora参数设置以及设置验证集比例为0.2即：最终结果推荐val最佳位置在epcoh为4（即step200）附近，因此最终选择之后的全数据训练epoch为5。

4.3 最佳lora参数确定

选择不同的lora参数值以及不同步数保存的lora模型，确定两个任务的lora最优模型：

LoRA配置	Checkpoint步数				
	160	180	200	220	250
rank=8, α=16	0.5435	0.5400	0.5500	0.5453	0.5425
rank=16, α=32	0.5450	0.5500	0.5600	0.5550	0.5525
rank=32, α=64	0.5455	0.5525	0.5550	0.5610	0.5650
rank=64, α=128	0.5500	0.5620	0.5650	0.5650	0.5620
rank=128, α=256	0.5450	0.5510	0.5600	0.5550	0.5580

表 3: 不同LoRA配置的任务1结果

LoRA配置	Checkpoint步数				
	160	180	200	220	250
rank=64, α=128	0.3738	0.3842	0.3912	0.4102	0.4025

表 4: 不同LoRA配置的任务2结果（参照任务1经验，仅选一组LoRA参数）

4.4 full微调

在LLamafactory默认full微调参数下，修改部分参数为4.1中的基本参数作为整体参数设置，选取不同步数保存的full模型，确定两个任务的full最优模型：

Checkpoint步数				
160	180	200	220	250
0.6450	0.6455	0.6470	0.6480	0.6480

表 5: full微调后，任务1的结果

Checkpoint步数				
160	180	200	220	250
0.3748	0.3842	0.3915	0.4105	0.3925

表 6: full微调后，任务2的结果

4.5 qwen3-4B模型lora微调

因qwen2.5-7B-Instruct模型在任务2中不管是lora微调还是full微调结果都不太理想，因此尝试用qwen3-4B模型lora微调任务2，结果如下：

LoRA配置	Checkpoint步数				
	160	180	200	220	250
rank=64, α=128	0.4120	0.4184	0.4237	0.4259	0.4232

表 7: qwen3-4B模型lora微调后，任务2的结果

5 结论

1.根据上述实验结果，任务1选取full微调的qwen2.5-7B-Instruct模型，其最好成绩为0.6480是第220steps保存的模型，任务2选取lora微调的qwen3-4B模型，其最好成绩为0.4259是参数$rank = 64, \alpha = 128$下保存的第220steps模型。

2.本次比赛后期由于时间和资源的问题，采用的研究方法比较粗糙，因此不管在任务1还是任务2，在更加精细的参数设置下，应该可以取得更好的成绩。

3.对任务1，full微调取得了有显著优势的成绩，可能是任务相对简单，只需要从10个症型里面选择一个或两个，以及从四个病症里面选择1个，据此推测对于相对简单的类似任务（如从几个数据里面选固定数量的数据的任务），7B量级的大模型full微调比lora微调可能会有显著优势。

4.对于任务2，full微调结果和lora微调结果差不多，且分数均较低，可能是任务相对复杂，需要从381种中药中选取药物，且对于选取药物的数量没有限制，以及其中没有任何逻辑性的描述（比如由于存在什么病例数据，因此要选用某种药物即把选药用药的逻辑与病例里面的描述结合起来）。据此推测对于7B量级的模型，在任务复杂，且lora粗调分数也比较低的情况下，full微调相对lora微调并不占据优势。

5.一个衍生的不成熟想法：中医治疗患者疾病，是否可以与患者生辰八字（是一个相对单一解释性的数据）结合起来，根据患者不同的八字，定性患者的体质，如火体，水体，依此配合药的药性，对于同一类疾病，进行相对特异性的用药。（如在药量上予以考虑，如在用药种类上予以考虑）

参考文献

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and others. 2022. *Lora: Low-rank adaptation of large language models. ICLR*, 1(2):3.

Qwen Team. 2024. *Qwen2.5: A Party of Foundation Models*. https://qwenlm.github.io/blog/qwen2.5/.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. *Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems*, 36:10088–10115.

Sibo Wei, Xueping Peng, Yi-fei Wang, Jiasheng Si, Weiyu Zhang, Wenpeng Lu, Xiaoming Wu, and Yinglong Wang. 2024. *BianCang: A Traditional Chinese Medicine Large Language Model*. arXiv preprint arXiv:2411.11027.