CCL25-Eval任务9总结报告：中医辨证辨病及中药处方生成评测

王聪[1,2,a]，赵直倬[1,2,b]，李一硕[1,2,c]，管红娇[1,2,d]，王怡斐[3,e]，李振宇[1,2,f]，鹿文鹏[1,2,g,†]
[1]齐鲁工业大学（山东省科学院），山东省计算中心（国家超级计算济南中心），
算力互联网与信息安全教育部重点实验室，济南，中国
[2]山东省算力互联网与服务计算重点实验室，
山东省基础科学研究中心（计算机科学），济南，中国
[3]山东中医药大学附属医院，济南，中国
[b]zhaozhizhuo@hotmail.com [c]yishuo.li@foxmail.com [e]71000686@sdutcm.edu.cn
[a,f]{cong.wang2024, zhenyu.li2024}@gmail.com [d,g]{hongjiao.guan, wenpeng.lu}@qlu.edu.cn

摘要

中医辨证辨病及中药处方生成评测任务专注于中医“辨证论治”。该任务由齐鲁工业大学（山东省科学院）与山东中医药大学附属医院联合发起，基于真实病历构建了中医“辨证论治”全流程公开数据集**TCM-TBOSD**，覆盖10类中医证型、4类中医疾病及381种常见中药。评测任务设立两个子任务：中医多标签辨证辨病与中药处方推荐，旨在系统评估大模型在中医诊疗全过程中的建模与推理能力。本次评测收到了学术界与产业界的广泛关注，评测共吸引123支队伍参与，35支队伍晋级复赛，最终提交了8份高质量技术报告。评测结果表明，大语言模型在中医任务中展现出良好的适应性与发展潜力，为中医智能化提供了可行路径与技术参考。详细信息可以从网址[1]查看我们的评测任务。

关键词： 中医诊疗；辨证论治；大语言模型；评测

Overview of CCL25-Eval Task 9: TCM Syndrome and Disease Differentiation and Prescription Recommendation

Cong Wang[1,2,a], Zhizhuo Zhao[1,2,b], Yishuo Li[1,2,c], Hongjiao Guan[1,2,d], Yi-Fei Wang[3,e], Zhenyu Li[1,2,f], Wengpeng Lu[1,2,g,*]
[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China
[3]Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, China
[b]zhaozhizhuo@hotmail.com [c]yishuo.li@foxmail.com [e]71000686@sdutcm.edu.cn
[a,f]{cong.wang2024, zhenyu.li2024}@gmail.com [d,g]{hongjiao.guan, wenpeng.lu}@qlu.edu.cn

Abstract

The evaluation task on Traditional Chinese Medicine (TCM) Syndrome and Disease Differentiation and Prescription Recommendation focuses on the TCM principle of ”treatment based on syndrome differentiation” Initiated jointly by Qilu University of Technology (Shandong Academy of Sciences) and the Affiliated Hospital of Shandong University of Traditional Chinese Medicine, this task has established a publicly available dataset named TCM-TBOSD , which captures real-world medical records and supports the full process of TCM syndrome differentiation and treatment. The dataset covers 10 types of TCM syndromes, 4 categories of TCM diseases, and 381 commonly used Chinese herbs. The evaluation includes two subtasks: TCM mulit-label

*通讯作者
[1]https://github.com/QLU-NLP/TCM-Syndrome-and-Disease-Differentiation-and-Prescription-Recommendation

第二十四届中国计算语言学大会论文集，第376页-第386页，济南，中国，2025年8月11日至14日。

syndrome and disease differentiation, and TCM prescription recommendation, aiming to comprehensively assess the modeling and reasoning capabilities of Large Language Models(LLMs) throughout the TCM diagnostic and therapeutic process. This evaluation has garnered widespread attention from both academia and industry. A total of 123 teams registered for the competition, with 35 advancing to the final round, and ultimately eight high-quality technical reports were submitted. Evaluation results demonstrate that LLMs exhibit strong adaptability and promising potential in TCM-related tasks, providing feasible pathways and technical references for intelligent TCM development. More details can be found on the website, where you can view our evaluation tasks.

Keywords: Traditional Chinese Medicine , Treatment Based on Syndrome Differentiation , Large Language Models , Evaluation

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健康做出了重要贡献 (宋勇刚 et al., 2024)。**辨证论治**是中医认识疾病和治疗疾病的核心原则与方法，包含**辨证**和**论治**两个关键过程。其中，辨证是对**证**的识别，**证**是对机体在疾病发展过程中病理变化的综合描述。辨证的基本方法是通过**望**、**闻**、**问**、**切**等手段，收集患者的症状、舌苔、脉象等临床信息，经由分析与综合，辨明疾病的病因和病机，并归纳判断为某种性质的**证**。**论治**则是在辨证基础上，依据辨明的证候确定相应的治疗策略，制定个性化治疗方案，开具合适的中药处方。辨证与论治相辅相成、密不可分，共同构成中医诊疗过程的核心，是中医学理论体系的精髓，也是中国古代科学技术在医学领域的重要体现。

随着人工智能，尤其是自然语言处理(Natural Language Processing)技术的迅速发展，以及电子病历等数字化医疗数据的广泛应用 (Zhao et al., 2024; 刘博 et al., 2017)，中医与人工智能的深度融合已成为中医现代化的重要方向 (张玉洁 et al., 2022; 王欣宇 et al., 2024)。Ren等人 (2022) 构建了首个公开的大规模中医辨证基准数据集（TCM-SD），并提出了一个面向中医领域的预训练语言模型（ZY-BERT），验证了预训练模型在中医辨证任务中的潜力。Yao等人 (2018) 构建了首个中药处方推荐数据集，并提出了一种基于主题模型的中药推荐方法。Zhu等人 (2023)、Wang等人 (2023)、Wei等人 (2024) 的研究则进一步探索了将大语言模型应用于中医问答、知识建模等任务，显著提升了大模型在中医情境下的推理与问答能力，为中医知识的传承与提升我国医疗智能化水平提供了新路径与技术支撑。

尽管先前的研究在中医辨证与处方推荐等任务上取得了一些重要进展 (何圆姣 et al., 2024; Zhao et al., 2024)，但目前仍缺乏一个能够**完整涵盖患者健康信息、辨证信息以及论治信息**的全流程结构化公开数据集。由于**辨证论治**任务本身具有推理链条长、决策过程复杂以及语义模糊等特性 (周学平 et al., 2011; 徐浩, 2024; 杨丽惠 et al., 2024)，现有研究往往仅能聚焦于其中某一环节，例如辨证建模或处方生成 (周开元, 2023; Tang et al., 2021; Zhang et al., 2021)，而难以覆盖从症状采集到最终处方生成的完整诊疗过程。因此，构建一个面向“**辨证-论治**”全过程的高质量基准数据集，已成为推动中医智能化研究向纵深发展的关键步骤。

齐鲁工业大学（山东省科学院）与山东中医药大学附属医院联合组织了本次中医辨证辨病及中药处方生成评测任务。本评测任务构建了一个全新的公开数据集，用于评估大语言模型在中医领域中辨证与论治的能力表现，命名为**T**raditional **C**hinese **M**edicine **T**reatment **B**ased **o**n **S**yndrome **D**ifferentiation (TCM-TBOSD)。该数据集基于脱敏的真实病历构建，涵盖10种中医证型、4种中医疾病、381种中药，总计包含1500条完整诊疗记录。

本次评测任务聚焦于“辨证论治”过程中的关键环节，设计了两个子任务：**中医多标签辨证辨病**任务和**中药处方推荐**任务。与现有的中医相关数据集和评测任务相比，TCM-TBOSD 数据集在辨证与论治两个阶段均具备显著特色，能够更全面地反映真实临床诊疗逻辑。在**辨证阶段**，数据集中详细记录了患者的中医疾病信息与证型信息，其中证型进一步细分为**主证**与**兼证**。主证反映了疾病的核心特征，是判断病机与制定治疗策略的主要依据；兼证则为疾病发展

过程中伴随出现的次要征象，体现病情的多样性和个体差异。主证与兼证的联合表达不仅更全面地揭示了病理状态，还增强了模型对复杂辨证场景的建模能力。在**论治阶段**，数据集提供了覆盖381种常见中药的处方信息，这些处方均来源于真实临床数据，充分体现了中药之间的配伍关系与治疗意图，为模型理解中医处方组合策略提供了重要支持。

本次评测旨在推动人工智能技术在中医领域的应用，助力中医诊疗过程的智能化发展，提出了面向中医“**辨证论治**”全过程的评测任务。评测要求参赛队伍所使用的模型参数规模不超过7B，允许使用各类开源工具与外部数据资源，以促进方法的多样性与实用性探索。

本文主要包含如下内容：第2节主要介绍了中医辨证辨病及中药处方生成评测的相关工作。第3节详细介绍的本次的评测任务的详细信息，例如子任务介绍、数据集、评价指标等。第4节概述了本次评测的参赛情况和参赛队伍的使用方法，并进行了总结分析。最后，第5节对本次评测进行了总结。

2 相关工作

2.1 中医辨病辨证

Ren等人 (2022) 构建了第一个大规模中医辨证基准数据集TCM-SD，并训练了面向中医语料的预训练模型ZY-BERT。该模型集成了中医知识与语言建模能力，在多个辨证分类任务上取得了优于通用模型的表现。该工作首次系统性地引入预训练语言模型用于中医辨证任务，具有开创性意义。

Li等人 (2024) 在TCM-BERT (Yao et al., 2019)的基础上，引入了基于迁移学习的双增强策略，以提升模型在中医辨证任务中的泛化能力。该方法增强了模型对罕见证型的识别能力，在多个不平衡数据场景下展现出良好性能。

随着大语言模型的兴起，其在医疗领域的应用逐渐扩展至中医场景。Wei等人提出了面向中医的专用大语言模型BianCang（扁仓） (Wei et al., 2024)。该模型采用“两阶段训练”策略，首先通过持续预训练引入中医领域知识，随后结合真实医院记录与《中国药典》中提取的ChP-TCM 数据集进行指令微调，以增强其诊断推理与证型识别能力。

2.2 中药处方推荐

Yao等人 (2018) 基于包含33,765条中医处方的数据集，提出了一种结合中医理论的主题建模方法，用于揭示中药配伍背后的潜在结构和生成逻辑。该方法通过引入中医领域知识，有效增强了模型在中药推荐、症状建议及处方模式发现等方面的能力。

PresRecST (Dong et al., 2024) 首次提出了使用渐进式中医辨证和论治过程来进行中医处方推荐。研究团队构建了TCM-Lung 数据集，并整合中医知识图谱，帮助模型更好地捕捉中医实体和诊疗顺序的关系。PresRecST 证明了将结构化中医知识嵌入模型流程中，能够显著提高模型对真实临床任务的适应能力。

随着大语言模型的发展，研究者开始探索其在中医处方推荐任务中的应用。TCMLLM-PR (Tian et al., 2024) 整合了来自教材、《中国药典》和真实临床记录的多源数据，构建了包含68,654条样本的高质量训练集，并实现了模型的高效微调。在真实病历的还原程度上，该模型生成的处方最接近真实医生的决策。

3 评测任务

3.1 数据集介绍

评测数据基于医院脱敏病历构建，总计1500条记录。数据被分为训练集、验证集和测试集，分别包含800、200和500条记录。数据均来源于医院真实病历，经过医师标注和核验以确保质量。如表1所示，数据集涵盖16个字段，包括患者的基本信息、健康状况、中医证型与疾病，以及治疗处方。

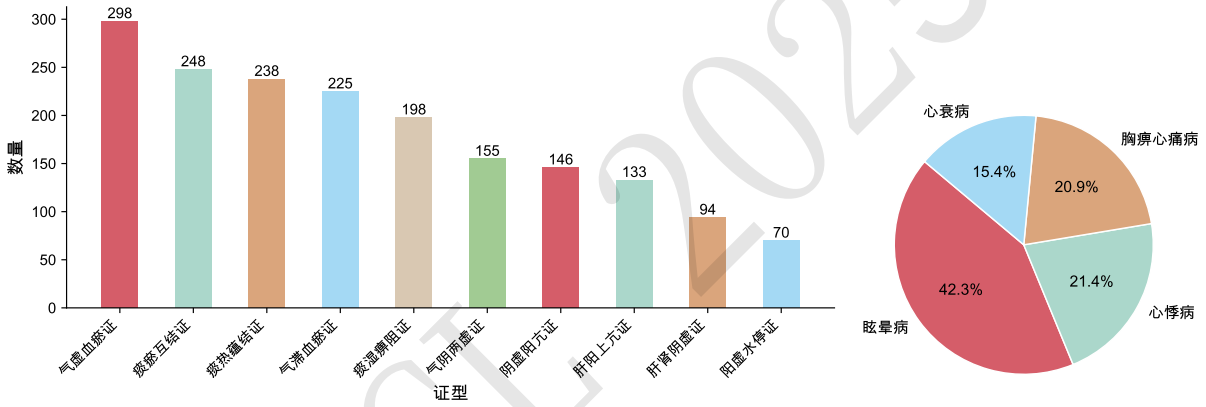
3.2 数据集分析

本数据集基于脱敏后的真实中医病历构建，具有较高的专业性与可信度。各字段涵盖患者的基础信息（如性别、年龄、职业等）、主诉与症状表现、诊疗过程中的主观与客观观察（如中医四诊、体格检查、辅助检查）、中医辨证结果（证型与疾病）以及最终生成的治疗处方，完整地反映了中医“辨证-论治”的实际流程。

第二十四届中国计算语言学大会论文集，第376页–第386页，济南，中国，2025年8月11日至14日。

字段	说明
ID	患者入院的唯一id
性别	男或女
职业	患者的职业信息，如职员、退（离）休人员等
年龄	患者的年龄
婚姻	描述婚姻状况，如已婚、未婚等
病史陈述者	入院时描述患者身体状况的人员与患者本人的关系，如患者本人
发病节气	患者出现病情时所处于的节气，如清明、小雪等
主诉	患者在就诊时向医生描述的最主要、最直接的不适或症状，用一句简短的文本概括描述，通常是患者就医的主要原因
症状	患者入院时所表现出的主要症状和体征的概述
中医望问切诊	医师对患者进行“望”、“闻”、“切”后，对患者状态的描述
病史	包括现病史、既往史、个人史、婚育史、家族史
体格检查	患者的体格检查
辅助检查	患者的其他检查项目，如CT、心电图报告等
证型	患者对应的中医证型，如气虚血瘀证、痰热蕴结证等
疾病	患者对应中医的疾病，如心悸病、胸痹心痛病等
处方	患者中药处方，如黄芪、白芷等（不包括剂量）

表 1: 数据字段说明



(a) 中医证型分布　　(b) 中医疾病分布

图 1: (a) 数据集中10种中医证型的分布情况。(b) 数据集中4种中医疾病的分布情况。

在证型分布方面，数据集包含10种临床常见中医证型，构建了较为完整的辨证体系。统计分析显示，单证型病历占主体，占比高达79.7%；多证型病历则进一步细分为主证与兼证，真实反映了临床实践中证候兼夹的复杂情况。从图 1(a)可以看出，证型分布较为均衡，其中“气虚血瘀证”出现频率最高，是最具代表性的证型类型。

疾病标签方面如图 1(b)所示，数据集中共涉及4种中医疾病，其中“眩晕病”最为常见，占比达42.3%，这种疾病分布特点为深入研究特定病种的辨证规律提供了数据基础。

在处方数据上，本任务将中药推荐视作一个多标签分类问题。统计结果如图 2所示，处方标签的种类繁多且分布极不均衡，呈现出显著的长尾分布特征。我们选取了出现频次最高的前100种药物进行分析，结果发现少数药物频繁出现，而大部分药物仅在极少数样本中出现，为模型的训练带来了挑战。

总体而言，该数据集具有三个显著优势：首先，完整保留了中医诊疗的临床思维链条，在内容结构上具有较强的医学完整性和中医特征；其次，证型标注体系兼顾标准规范与临床实际；最后，药物分布特征真实反映了中医处方规律。这些特点使得该数据集既适合用于辨证论治全流程的端到端建模，也可为长尾分布处理、小样本学习等机器学习前沿问题提供具有中医特色的研究场景。

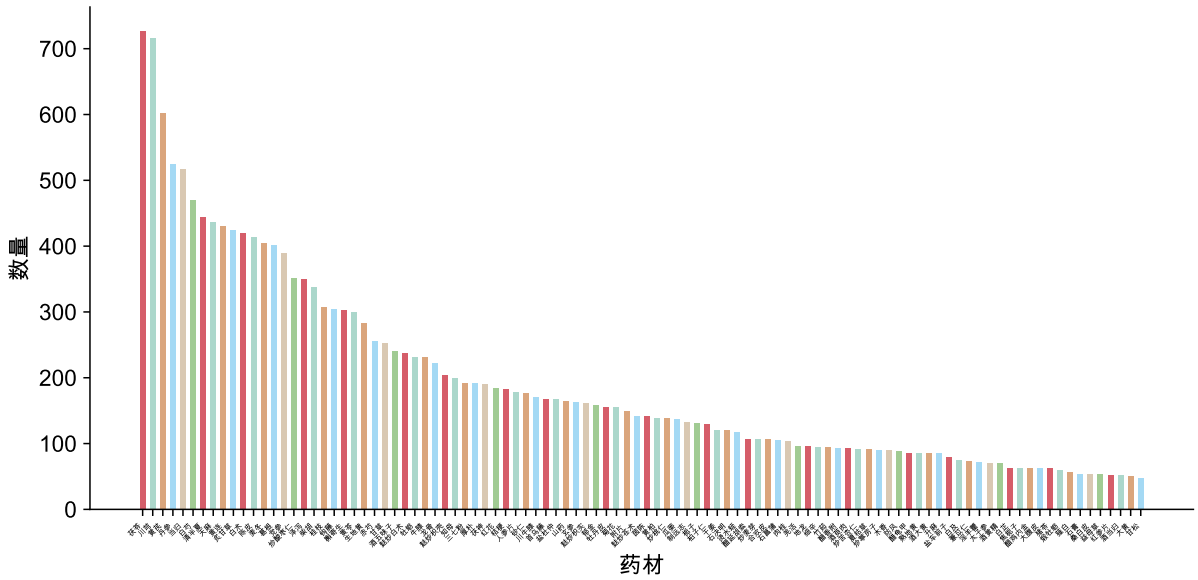


图 2: 数据集中前100种中药的分布情况

3.3 子任务一：中医辨病辨证

3.3.1 任务简介

辨病是对患者的病因病机、病情发展等进行整体性理解，而辨证则更加注重根据病情某一发展阶段的病理特点做出阶段性判断。辨证是中医学的核心任务之一，也是实施个性化诊疗的关键环节。中医辨病辨证的目标是借助自然语言处理技术，从病历和症状描述中提取信息，快速分析主证、兼证与疾病的关系，从而提高辨病辨证的效率和准确性，辅助医生更精准地完成诊断。

3.3.2 评测指标

辨证任务准确率：

该指标用于评估模型在中医证型预测上的性能。计算公式如下：

$$syndrome_{acc} = \frac{NUM(y \cap \hat{y})}{NUM(y)} \tag{1}$$

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。

该指标用于评估模型预测证型与真实证型的匹配程度，能够有效衡量单证型及多证型（含主证与兼证）的预测准确性。对于多证型预测这一更具挑战性的任务，模型不仅需要准确识别所有相关证型，还需正确区分主证与兼证，才能被视为完全正确。这一要求充分体现了中医临床辨证中主次分明的特点，也使得该指标能够更精准地反映模型的实际辨证能力。

辨病任务准确率：

该指标评估模型在中医疾病诊断上的准确率，计算方式与辨证任务类似：

$$disease_{acc} = \frac{NUM(y \cap \hat{y})}{NUM(y)} \tag{2}$$

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。

在中医临床实践中，疾病诊断通常采用单标签形式，即每个病例对应一个主要疾病诊断。该指标重点评估模型对患者核心疾病的辨识能力，要求模型能够从复杂的临床症状中准确提取关键特征，做出符合中医理论体系的疾病判断。

评价总指标：

为了综合评估模型在辨证辨病上的整体表现，采用辨证和辨病准确率的算术平均值作为最终评价指标：

$$task1\_acc = \frac{1}{2}(syndrome_{acc} + disease_{acc}) \tag{3}$$

3.4 子任务二：中药处方推荐

3.4.1 任务简介

在实际诊疗过程中，医生首先根据患者的四诊信息判断出“疾病”和“证型”，随后基于辨证审因和治法确定的结果，按照组方原则选择合适的药物进行合理配伍，组成中药处方。自然语言处理技术能够从海量临床数据中提取有价值的信息，帮助医生快速、准确地进行辨证论治，并推荐个性化的中药处方。中药处方推荐的目标是根据患者的详细描述，自动生成一组中草药作为处方（不包括中药剂量）。

3.4.2 评测指标

Jaccard相似系数：Jaccard指标用于衡量预测处方与真实处方之间的药物重叠程度，反映整体配伍相似性，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：

$$Jaccard(y, \hat{y}) = \frac{NUM(y \cap \hat{y})}{NUM(y \cup \hat{y})} \tag{4}$$

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。

该指标特别适用于评估中药复方配伍的整体相似度，能够有效反映模型对中医“君臣佐使”组方原则的把握程度。通过计算预测处方与标准处方的药物重合度，不仅可以衡量核心药物的匹配情况，还能评估辅助药物的配伍合理性，从而全面考察模型对中医复方整体架构的还原能力。

Recall、Precision、F1分数：

这三项指标从不同角度评估预测处方的准确性：

$$\begin{aligned} F1(y, \hat{y}) &= 2 \cdot \frac{\text{Precision}(y, \hat{y}) \cdot \text{Recall}(y, \hat{y})}{\text{Precision}(y, \hat{y}) + \text{Recall}(y, \hat{y})} \\ Recall(y, \hat{y}) &= \frac{NUM(y \cap \hat{y})}{NUM(y)} \\ Precision(y, \hat{y}) &= \frac{NUM(y \cap \hat{y})}{NUM(\hat{y})} \end{aligned} \tag{5}$$

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。

召回率(Recall)：衡量模型对真实处方中关键药物的覆盖能力，避免漏荐重要药物。精确率(Precision)：评估预测药物的准确性，避免推荐无关或错误药物。F1分数：综合召回率与精确率的调和平均数，平衡模型的全面性与精准性。

药物平均数量：

该指标评估预测处方与真实处方的药物数量一致性：

$$AVG(y, \hat{y}) = 1 - \frac{|NUM(y) - NUM(\hat{y})|}{max(NUM(y), NUM(\hat{y}))} \tag{6}$$

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，$maxa(a, b)$ 代表取a , b 中的最大值，$|x|$ 代表计算x 的绝对值。

中医处方通常包含6–15味中药，药物数量过少可能遗漏关键配伍，过多则可能降低疗效或增加副作用。该指标在于引导模型生成符合临床实际规模的处方。

评价总指标：为全面反映模型性能，最终得分整合上述三项指标：

$$task2\_score = \frac{1}{3} \cdot \frac{1}{N} \sum_{i=1}^{N} [Jaccard(y_i, \hat{y_i}) + F1(y_i, \hat{y_i}) + AVG(y_i, \hat{y_i})] \tag{7}$$

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。该指标要求模型同时兼顾药物配伍、关键药物覆盖和处方规模。

4 评测结果

4.1 评测情况

参赛单位	队伍编号	队长	子任务1	子任务2	最终成绩
赣西肿瘤医院	Team.1	李南书	**0.6480**	0.4259	**0.5369**
个人	Team.2	李彦	0.6020	0.4348	0.5184
大连理工大学	Team.3	康益扬	0.5710	**0.4632**	0.5171
大连理工大学	Team.4	刘泓宇	0.5770	0.4300	0.5035
新疆大学	Team.5	左梓呈	0.5530	0.4515	0.5022
北京智览医疗科技有限公司	Team.6	张欣欣	0.5840	0.4188	0.5014
山东职业学院	Team.7	李晗	0.5690	0.4306	0.4998
云南大学	Team.8	张坚	0.5390	0.4189	0.4789

表 2: 参赛队伍的B榜成绩

本次评测自2025年2月10日开启报名，吸引了来自国内高校与科研机构、医院及科技企业的123支队伍参与A榜初赛，其中35支队伍成功晋级B榜复赛阶段。A榜于2025年5月4日截止，B榜于2025年5月6日开放，并于2025年5月15日正式结束比赛。参赛队伍的来源分布反映了本任务在学术界与产业界的广泛关注。

赛后，我们收到了8支B榜参赛队伍提交的技术报告，并在统一环境下对这些方法进行了复现与评估。表 2 展示了这8支队伍在各子任务及最终得分上的表现，其中最终得分是两个子任务得分的平均值。为便于叙述，我们在文中使用队伍编号来指代对应的参赛队伍。

4.2 方法分析

多数队伍均基于当前主流开源大语言模型进行任务微调，表明具备较强中文理解能力的大模型在中医领域的迁移能力已得到广泛认可。其次，**低秩适配微调(LoRA)**成为普遍采用的训练范式，尤其适用于本任务对参数规模和计算资源的限制。此外，多个团队通过引入**知识增强**（如中医知识图谱、数据集扩展等）以及向量检索技术，提升了模型在中医专业任务中的表现稳定性与准确性。值得注意的是，一些队伍还尝试融合**强化学习**方法，如Group Relative Policy Optimization (GRPO) (Shao et al., 2024)，以进一步提升系统的生成质量。

Team 1 提出了一种多阶段LoRA (Hu et al., 2022) 微调策略。他们首先在训练集中以默认参数进行监督微调，以确定最优epoch；随后使用该epoch 在全量数据上进行精细化调整，并将最终结果与全参数微调模型进行性能对比，从而选取最优版本。整体过程充分挖掘了模型的性能上限。

Team 2 构建了一个训练数据集，包含临床特征清洗、中医知识增强及思维链 (Wei et al., 2022)推理内容。他们结合GRPO 强化学习策略，提升了模型对中医任务的适应能力。同时，该队还引入了Self-Consistency (Wang et al., 2022) 推理机制，显著提高了生成结果的稳定性和准确性。

Team 3 在子任务一中分别对Qwen2.5-7B (Qwen Team, 2024)、Mistral-7B (Jiang et al., 2023) 和Baichuan2-7B (Baichuan, 2023) 进行基于QLoRA (Dettmers et al., 2023) 的微调，并引入多模型集成投票策略。在子任务二中，该队构建了一个融合向量检索、监督学习与强化学习的中药推荐系统。通过候选处方生成和GRPO 优化，显著提升了推荐精度。

Team 4 专注于中医知识融合与诊疗推理，在子任务一中构建了知识增强提示模板，并通过基于QLoRA 的微调增强了模型的辨证能力。在子任务二中，在构建中药疗效知识提示模版的基础上，该队采用硬投票策略进行模型推理集成，进一步提升了推荐结果的正确性和有效性。

Team 5 针对中医辨证辨病与处方推荐中数据标注不足的问题，提出融合大模型与可控文本生成的数据增强方法，利用少量标注样本构建高质量扩展数据，结合LoRA微调实现高效适

配。实验验证了方法的有效性，在不引入外部数据下取得良好性能，为中医药智能化诊疗提供了可行方案。

Team 6 采用了多模型融合策略，在辨病任务中集成了Qwen2.5-7B-Instruct、DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI, 2025) 和ZY-BERT 三种模型，通过硬投票方法生成最终预测结果。在辨证任务中，该队采用了基于ZY-BERT 的多标签分类模型；在处方推荐方面，则聚焦于prompt 优化与参数调优。

Team 7 构建了基于原始病历文本的高质量标注数据集，并结合中医诊疗理论框架，设计了融合多标签分类与序列生成的双任务协同模型架构。该队在多个Qwen 模型间进行了系统性微调与性能评估，并结合多个子任务的评价指标进行了全面分析。

Team 8 构建了一个基于大语言模型的中医智能诊断系统，采用分阶段的多轮对话机制模拟临床推理流程，包括辨证、诊病与处方生成三个环节。通过引入思维链（CoT）技术和多任务微调策略，提升了模型的推理连贯性与诊断准确性。实验表明，系统在各项任务中表现良好，且辨证与诊病阶段的准确率直接影响最终处方质量，为中医智能化诊断提供了可行的技术路径与理论支持。

综上所述，各参赛队伍围绕中医“**辨证-论治**”一体化任务进行了多样化的建模探索，充分展示了当前大语言模型在中医任务中的强大表达能力与可塑性。从方法角度来看，多数队伍采用LoRA、QLoRA 等高效参数微调策略，以适配任务资源限制与数据分布特征。同时，提示工程设计、中医知识融合以及多阶段训练机制成为提升模型对复杂任务适应能力的关键手段。

值得关注的是，**强化学习如GRPO在本次评测任务中表现出显著优势**。Team 2 在其“全参数微调加强化学习”双阶段优化框架中，通过GRPO 策略有效缓解了辨证辨病任务中的多标签不平衡问题，同时结合Self-Consistency 推理技术，显著提升了模型在推理任务中的稳定性和准确性。Team 3 在中药处方推荐任务中，引入了强化学习阶段，并结合任务评估指标设计专属奖励函数，成功优化了模型在生成式推荐过程中的策略选择能力，从而显著提升了最终处方的准确率与临床合理性。

从本次评测结果可以看出，**强化学习为大语言模型在复杂医学任务中的表现提升提供了新思路**。本次评测的成果不仅验证了大模型技术在中医诊疗场景下的适应能力，也为后续中医智能系统的研发提供了方法参考与技术路径探索。

4.3 案例分析

在中医人工智能辅助诊疗研究中，辨证论治始终是核心挑战。尽管近年来深度学习技术取得了显著进展，但由于中医语言特有的复杂性、证候的动态组合特性以及药物配伍的灵活性，现有模型在实际临床应用中仍存在明显的误判和系统性偏差。本文通过深入分析典型临床案例如表 3所示，旨在揭示当前模型在理解中医语言时面临的关键技术瓶颈。

在证型预测方面，模型对多证型组合的识别能力尤为薄弱。研究发现，模型表现出明显的“高频证型偏好”，即倾向于选择数据集中出现频率较高的单一证型，而忽略临床实际中常见的次要证型。以本案例为例，患者实际证型为“阴虚阳亢证（主证）合并气虚血瘀证（兼证）”，但模型仅预测出“气虚血瘀证”，遗漏了更为关键的主证“阴虚阳亢证”。这一现象充分暴露出当前模型在理解中医病历文本时，尚缺乏对症状-证型复杂关联的深层语义理解能力。相比之下，在疾病预测任务中，由于采用单标签分类且疾病类别数量有限，模型能够达到相对理想的预测效果。

处方推荐任务的表现凸显出中医智能诊疗面临的深层挑战。中医组方不仅严格遵循“君臣佐使”的配伍原则，更因学术流派和诊治学说的多样性而呈现出“同病异治”的特点。当前模型虽能捕捉主流治疗方案的用药规律，却难以兼顾各具特色的小众流派用药思路。典型案例显示，模型更倾向于推荐数据集中占主导地位的主流流派处方，而对一些虽临床有效但数据量较少的小众或者独居特色的流派的独特配伍则往往难以准确捕捉。这种数据分布的不均衡性，使得模型难以全面反映中医诊疗的丰富多样性，为处方推荐的临床适配带来显著挑战。

通过深入案例分析发现，当前中医智能诊疗系统在辨证辨病和处方生成方面仍存在显著局限：首要挑战在于模型对中医“辨证论治”复杂关联体系的深层语义理解不足；其次面临药物配伍规律的形式化表达困境；加之学术流派和诊治学说的多样性导致的系统性偏差，这些因素共同制约了模型的临床应用价值。这充分表明单纯依赖数据驱动的方法存在局限。

字段	描述
性别	男
职业	退（离）休人员
年龄	79岁
婚姻	已婚
病史陈述者	本人
发病节气	立夏
主诉	主诉：阵发性头晕21年，加重伴胸闷2天。
症状	阵发性头晕伴胸闷，咳嗽，痰少，无心慌胸痛，双下肢不自觉震颤，怕热，乏力，无口干口苦，纳差，眠差，大便干，小便调。
中医望问切诊	中医望闻切诊：表情自然，面色少华，形体正常，动静姿态，语气清，气息平，无异常气味，舌暗红，苔薄白，脉沉弦。
病史	现病史***否认家族性遗传病史。
体格检查	体温***病理反射未引出。
辅助检查	颅脑MRI+MRA：***建议进一步检查。
证型	阴虚阳亢证|气虚血瘀证
疾病	眩晕病
处方	麦冬, 北沙参, 桑叶, 玉竹, 炙甘草, 炒白扁豆, 天花粉

表 3: 案例数据，其中病史等字段的文本过长已省略

5 评测总结

本次中医辨证辨病及中药处方生成评测任务以“**辨证论治**”全过程为核心，旨在推动人工智能技术在中医诊疗中的应用与能力提升。评测任务由齐鲁工业大学（山东省科学院）与山东中医药大学附属医院联合发起，依托脱敏的真实病历数据构建了高质量的结构化数据集TCM-TBOSD，并公开发布。该数据集填补了当前中医领域中缺乏涵盖患者信息、辨证诊断与处方推荐的一体化数据资源的空白。

本次评测共吸引了123支队伍报名，最终35支队伍进入B榜复赛，展现了广泛的社会关注和技术热度。参赛队伍覆盖高校、研究机构、医院与企业，方法包括微调策略、提示工程、知识注入、数据增强和强化学习等，充分体现了大语言模型在中医任务中的应用潜力。从模型方法角度来看，LoRA 和QLoRA 等轻量级微调技术成为主流选择，显著降低了模型适配成本。多数队伍通过设计微调策略并引入中医知识，提升了诊疗推理能力。值得注意的是，部分队伍在辨病辨证与中药推荐任务中引入了GRPO 等强化学习策略，取得了显著成效，体现了强化学习在复杂医学推理场景下的广阔应用前景。

本次评测任务验证了大模型技术在中医诊疗任务中的可行性和有效性，为中医智能化研究奠定了坚实基础，同时为中医现代化发展提供了新的技术路径和实践范式。

致谢

本论文受到国家自然科学基金项目(No.62376130)、济南市“新高校20条”项目(No.202333008)、齐鲁工业大学(山东省科学院)科教产融合试点工程重大创新类项目(No.2024ZDZX08)、山东省科技型中小企业创新能力提升工程项目(No.2024TSGC0094)的资助。

参考文献

Zhizhuo Zhao, Xuping Peng, Yong Li, Hao Wu, Weiyu Zhang, and Wenpeng Lu. 2024. *Thinking the Importance of Patient's Chief Complaint in TCM Syndrome Differentiation.* In *Proceedings of the 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2024)*, pages 1534–1539, Tianjin, China.

Zhizhuo Zhao, Wenpeng Lu, Xueping Peng, Lumin Xing, Weiyu Zhang, and Chaoqun Zheng. 2024.

Automated ICD Coding via Contrastive Learning with Back-Reference and Synonym Knowledge for Smart Self-Diagnosis Applications. IEEE Transactions on Consumer Electronics, 70(3):6042–6053.

Mucheng Ren, Heyan Huang, Yuxiang Zhou, Qianwen Cao, Yuan Bu, and Yang Gao. 2022. *TCM-SD: A benchmark for probing syndrome differentiation via natural language processing.* In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 908–920.

Liang Yao, Yin Zhang, Baogang Wei, Wenjin Zhang, and Zhe Jin. 2018. *A topic modeling approach for traditional Chinese medicine prescriptions. IEEE Transactions on Knowledge and Data Engineering*, 30(6):1007–1021.

Wei Zhu, Wenjing Yue, and Xiaoling Wang. 2023. *ShenNong-TCM: A Traditional Chinese Medicine Large Language Model.* GitHub repository. https://github.com/michael-wzhu/ShenNong-TCM-LLM.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. *HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge.* arXiv preprint arXiv:2304.06975.

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.

Xiaochen Li, Kui Chen, Jiaxi Yang, Cheng Wang, Tao Yang, Changyong Luo, Nan Li, and Zhi Liu. 2024. *TLDA: A transfer learning based dual-augmentation strategy for traditional Chinese Medicine syndrome differentiation in rare disease. Computers in Biology and Medicine*, 169:107808.

Liang Yao, Zhe Jin, Chengsheng Mao, Yin Zhang, and Yuan Luo. 2019. *Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. Journal of the American Medical Informatics Association*, 26(12):1632–1636.

Xin Dong, Chenxi Zhao, Xinpeng Song, Lei Zhang, Yu Liu, Jun Wu, Yiran Xu, Ning Xu, Jialing Liu, Haibin Yu, and others. 2024. *PresRecST: A novel herbal prescription recommendation algorithm for real-world patients with integration of syndrome differentiation and treatment planning. Journal of the American Medical Informatics Association*, 31(6):1268–1279. Publisher: Oxford University Press.

Haoyu Tian, Kuo Yang, Xin Dong, Chenxi Zhao, Mingwei Ye, Hongyan Wang, Yiming Liu, Minjie Hu, Qiang Zhu, Jian Yu, and others. 2024. *TCMLLM-PR: Evaluation of large language models for prescription recommendation in traditional Chinese medicine. Digital Chinese Medicine*, 7(4):343–355. Publisher: Elsevier.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and others. 2024. *DeepseekMath: Pushing the limits of mathematical reasoning in open language models.* arXiv preprint arXiv:2402.03300.

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.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. *Self-consistency improves chain of thought reasoning in language models.* arXiv preprint arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and others. 2022. *Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems*, 35:24824–24837.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. *Mistral 7B.* arXiv preprint arXiv:2310.06825. https://arxiv.org/abs/2310.06825.

Baichuan. 2023. *Baichuan 2: Open Large-scale Language Models*. arXiv preprint arXiv:2309.10305. https://arxiv.org/abs/2309.10305.

DeepSeek-AI. 2025. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv preprint arXiv:2501.12948. https://arxiv.org/abs/2501.12948.

Yuqi Tang, Zechen Li, Dongdong Yang, Yu Fang, Shanshan Gao, Shan Liang, and Tao Liu. 2021. *Research of insomnia on traditional Chinese medicine diagnosis and treatment based on machine learning. Chinese Medicine*, 16:1–21. Publisher: Springer.

Qi Zhang, Jianhang Zhou, and Bob Zhang. 2021. *Computational traditional Chinese medicine diagnosis: a literature survey. Computers in Biology and Medicine*, 133:104358. Publisher: Elsevier.

宋勇刚and 邹小伟. 2024. 基于数据与模型驱动的数智中医药发展研究. 时珍国医国药, 35(11):2639–2642.

刘博, 杜建强, 聂斌, 刘蕾, 张鑫, and 郝竹林. 2017. 基于二阶*HMM*的中医诊断古文词性标注. 计算机工程, 43(07):211–216.

张玉洁, 白如江, 许海云, 韩靖, and 赵梦梦. 2022. 融合多自然语言处理任务的中医辅助诊疗方案研究——以糖尿病为例. 数据分析与知识发现, 6(01):122–133.

王欣宇, 杨涛, and 胡孔法. 2024. 基于大语言预训练模型的中医个性化处方推荐研究. 中华中医药学刊, 42(04):15–18+264.

何圆姣and 刘国华. 2024. 融合*BERT*和*GCN*的中医问诊辅助诊断算法研究. 南开大学学报*(*自然科学版*)*, 57(02):70–78.

周学平, 叶放, 郭立中, 过伟峰, 吴勉华, and 周仲瑛. 2011. 以病机为核心构建中医辨证论治新体系——国医大师周仲瑛教授学术思想探讨. 中医杂志, 52(18):1531–1534.

徐浩. 2024. 构建“方—病—证”中医辨证论治新体系. 中国中西医结合杂志, 44(12):1503–1506.

杨丽惠, 李靖华, 周天, and 胡凯文. 2024. 中医辨证论治肺结节研究现状与思考. 中华中医药杂志, 39(03):1431–1436.

周开元. 2023. 基于图神经网络的中医处方对话推荐. 广州大学.