

DOI: 10. 13288/j. 11-2166/r. 2025. 09. 005

思路与方法

基于大语言模型的中医智能预问诊系统的构建与验证

刘奕清¹, 李迎¹, 杨洪军¹, 彭麟景¹, 咸楠星¹, 李坤宁¹,
石启伟¹, 田恒屹¹, 董立峰², 王林³, 赵宇平⁴✉



1. 中国中医科学院医学实验中心, 北京市东城区东直门内南小街 16 号, 100700; 2. 中国科学院自动化研究所; 3. 天津中科闻歌科技有限公司; 4. 中国中医科学院中医基础理论研究所

[摘要] **目的** 基于大语言模型构建一种中医智能预问诊系统, 以提高临床诊疗效率。**方法** 采用 DeepSpeed ZeRO-3 分布式训练策略, 基于 YAYI 2-30B 微调中医大语言模型; 整理中医书籍和临床病案中的关系数据, 设计带权无向图网络, 建立智能体辨证推理模型。利用智能体协作架构实现中医大语言模型与辨证推理模型交互。应用 Loss 函数、BLEU-4 和 ROUGE-L 分别从模型训练、文本生成质量和语言理解能力方面对中医大语言模型进行评估。设计含中医执业医师、执业中药师、中医症状术语识别和经络识别的专业知识测试集, 评估系统的专业知识掌握水平; 通过临床测试对比系统与主治医师在辨证准确率、问诊轮次和问诊时间等指标的差异。**结果** Loss 函数显示经 100 000 次迭代后, 训练损失值逐渐稳定于 0.7 ± 0.08 , 表明中医大语言模型已训练完成, 且具有较好泛化能力。中医大语言模型在 BLEU-4 得分 0.38, 在 ROUGE-L 得分 0.62, 提示其自然语言翻译质量已达标。整理得到 2715 个症状术语、505 种疾病与证型关系、1011 种疾病与主症关系、1 303 600 种不同症状之间关系, 构建了智能体辨证推理模型。在中医执业医师、执业中药师、中医症状术语识别、经络识别模拟测试中的正确率分别达到 94.09%、78.00%、87.50%、68.80%。临床测试中, 该系统辨证准确率达 88.33%, 其问诊轮次、问诊时间均少于主治医师 ($P < 0.01$), 提示该系统具备一定的预问诊能力。**结论** 基于大语言模型的中医智能预问诊系统可以在一定程度上模拟临床中医医生的辨证思维, 提高医生诊疗效率及患者就医体验。

[关键词] 问诊; 大语言模型; 智能预约问诊系统; 智能体

有研究^[1]显示, 三级甲等医院门诊医生平均接诊时间为 2~3 分钟, 在此有限时间内需要完成病史采集等多项临床任务。而与西医相比, 中医更强调个体化辨证论治, 这一特点使得病史采集的时间往往更长。因此, 智能预问诊系统在提升中医诊疗效率方面具有重要价值。2024 年 7 月 19 日, 国家中医药管理局、国家数据局联合发布《关于促进数字中医药发展的若干意见》^[2], 提出发展智能预诊随访等系统, 提升中医药数据智能化采集能力。目前中医问诊相关研究多围绕量表进行^[3-4], 但存在量表条目数量多、中医术语难以理解等不足^[5]。而

中医智能诊断所涉及的分析算法主要集中于症状与证候之间的关系^[6-7], 即较为低层次的数据挖掘与分析, 对传统中医问诊高层次文本语义、认知智能等的研究尚未深入^[8]。

大语言模型 (large language model, LLM) 掌握了人类语言模式, 能生成自然流畅的回复, 创造近乎真实的交流体验。目前 LLM 已在医疗咨询、药物研发等领域取得了巨大突破^[9]。有研究者基于 LLM 设计了针对西医的预问诊系统, 使得医生病历书写时间缩短了 34.44%, 提高了门诊医生工作效率^[1], 但 LLM 在中医领域的应用仍处于研究和探索阶段^[10]。通过对中医领域数据微调, 可让 LLM 更好地理解 and 处理该领域信息。因此, 本研究基于团队自研的雅意大语言模型 (YAYI 2-30B)^[11-12], 微调中医 LLM, 并通过智能体的交互模式^[13], 构

基金项目: 中国中医科学院科技创新工程 (ZN2023A02); 国家中医药管理局国家中医药多学科交叉创新团队项目 (ZYYCXTD-D-202205); 中国中医科学院自然科学基金培育项目 (JJPY2023005)

✉ 通讯作者: 18810084632@163.com

建基于中医 LLM 的智能预问诊系统,通过引导式问诊,模拟中医医生的辨证思维,实现患者症状信息智能化采集,以突破传统量表的线性问诊模式,提高医生诊疗效率并优化就医体验。

1 预问诊系统总体设计

总体设计方面,预问诊系统由中医 LLM 与智能体辨证推理模型构成(见附图 1,请扫描本文二维码获取),协同实现预问诊功能。首先患者向中医 LLM 描述当前最难受的症状即主诉,中医 LLM 会将主诉处理为症状术语,并在智能体辨证推理模型中得到与主诉相关的最潜在的症状术语,再由中医 LLM 将症状术语转化为容易理解的自然语言,向患者进行提问。这样引导交互式的一问一答,直至系统收集到患者当前支持中医诊疗的所有症状。

1.1 中医 LLM 构建

1.1.1 数据来源与处理 本研究数据整合自多源异构渠道,具体包括:1)古籍文献数据:基于《中华医典》《爱如生》等电子数据库,参照《中国中医古籍总目》^[14]、《中国医籍通考》^[15]、《中国医籍考》^[16]目录框架,结合课题组积累的古籍资料,系统整理中医古籍 1500 本;2)其他结构化数据:整合行业标准、可获取的现代公开信息及开源数据集。所有数据经清洗后统一以 JSONL 格式存储,文本编码采用 UTF-8 标准。

采用人工智能与人工审核相结合的混合方法对数据预处理。基于规则匹配算法进行文本初步清洗与格式标准化,利用 TF-IDF 算法计算文本特征向量,对数据集进行初步去重,引入基于 BERT 的文本表示模型计算语义相似度,并设定阈值筛除高度相似的冗余数据,保证数据安全性。长、短问题-答案对(question-answer pair, QA 对)在 LLM 微调中具有重要作用,主要用于知识注入、行为调整、错误修正^[17]。将已分类的数据作为先验知识导入 YAYI 2-30B,生成多个维度的 QA 对。对 LLM 生成的 QA 对进行预处理,删除重复、缺少主语、生成无关内容的数据。最终构建了一个高质量中医数据集,包含 100 万条经严格质量控制的数据条目,兼具高纯净度与临床适用性,并采用了等量的通用数据进行联合训练,以增强模型的泛化能力。

1.1.2 中医 LLM 微调训练 本研究采用 DeepSpeed ZeRO-3^[18]分布式训练框架进行模型微调,采用 AdamW 优化器($\beta_1=0.9$, $\beta_2=0.999$)进行模型微调,设置初始学习率为 5×10^{-6} 、批大小 4、最

大序列长度 4096 词元及 2 个训练轮次。训练策略包含 0.1 随机失活率、10% 验证集划分、线性学习率调度器(10% 预热阶段步长)结合余弦退火策略(最小学习率为初始值 10%),并采用 FP16 混合精度训练、梯度裁剪(阈值 1.0)及 4 步梯度累积优化训练过程。硬件配置为 8 块英伟达 A100 图形处理器(80GB/卡)集群,基于英伟达集体通信库(NVIDIA Collective Communications Library, NCCL)后端实现分布式训练,单轮耗时 108 小时。通过动态权重衰减(系数 0.01)、验证集持续监控及最优模型保存点保存机制防止过拟合,同时利用可视化张量面板实时追踪损失曲线、学习率动态、GPU 资源利用率等关键指标以优化训练稳定性。

1.2 智能体辨证推理模型构建

考虑医疗场景问答的严肃性,避免生成式人工智能输出的幻觉问题,LLM 输出结果必须经过严谨的中医辨证推理。为此,本研究构建了智能体辨证推理模型,包括数据来源与处理、算法构建两个部分(见附图 2,请扫描本文二维码获取),以保证问诊结果的可解释性。

1.2.1 数据来源与处理 临床中医问诊主观性很强,不同医生有不同的问诊习惯。为确保问诊的科学性、客观性,本研究使用的症状、证候术语数据来源于全国中医药行业高等教育“十四五”规划教材的《中医内科学》^[19]、《中医外科学》^[20]、《中医妇科学》^[21]、《中医儿科学》^[22]及课题组 2020—2024 年积累的 2000 份真实世界临床医案。症状具有唯一性,对于同一人体异常在同一粒度下仅有一个标准化表达^[23],因此对症状进行人工拆分处理,以获取同一粒度下的词汇表达,拆分过程参考《病状术语规范化基础》^[24]、《中医症状鉴别诊断学》^[25]。

对电子数据进行拆分,将复合的症状拆分成最小的词汇,如将“高热烦躁”拆分为“高热”“烦躁”。解决中医症状一义多词是症状标准化工作的一项重要内容,因此得到拆分词汇后,本研究参考《中医临床名词术语 第 1 部分·内科学》^[26]对词汇进行同义词合并,如将“情志抑郁”“情怀抑郁”“精神抑郁”词汇合并为“抑郁”。操作环节由 3 位中医执业医师执行,2 位具有高级职称的中医专家进行质量审核。

本研究依据临床实践指南、专家共识,梳理了疾病与证型的关系(如感冒有风寒感冒、风热感冒等证型),整理了疾病与主症的关系(如感冒的主

症为鼻塞、流涕、喷嚏、头痛、恶寒、发热、全身不适)。其次,基于《黄帝内经》理论框架,本研究构建症状互斥、相交关系判定标准如下:互斥关系整理规则是病性互斥如:寒热(如“恶寒”与“恶热”)、虚实(如“声高气粗”与“声低气怯”)等对立属性;病机互斥特指阴阳偏胜相关病机(如“喜冷”与“喜暖”);性别互斥指生殖系统特异性症状(如“月事不调”与“睾丸坠痛”)。相交关系整理规则是使用apriori算法,挖掘症状与症状的交叉关系,并通过人工核验以验证合理性,如“神疲倦怠”与“便质溏稀”。数据处理结果包含四类,分别是症状术语(2715条)、疾病与证型的关系(505条)、疾病与主症的关系(1011条)、症状与症状的关系(1303600条)。

1.2.2 算法构建 构建疾病与证型、疾病与主症、证型与症状、症状与症状的智能体辨证推理模型(局部示例见附图3,请扫描本文二维码获取),构造一个带权的无向图。定义带权无向图 G :

$$G = (V, E, w)$$

其中 $V = \{V_d, V_s\}$, V_d 是疾病顶点, V_s 是症状顶点; $E = \{E_{ds}, E_{ss}\}$, E_{ds} 是疾病与症状构成的边, E_{ss} 是症状与症状构成的边; w 是2个顶点之间边的权重,分别是 V_d 与 V_s 之间边的权重、 V_s 与 V_s 之间边的权重。

假设用户自然语言输入的症状经中医LLM提取为标准化症状 V_{s_i} ,将此顶点的状态设为1,计算该用户此次所患疾病的概率 P_i ,计算公式如下:

$$P_i = \frac{\sum_{i=1}^k w_{ds_{i-1}} - \sum_{i=1}^m w_{ds_{i-0}}}{\sum w_{ds}}$$

计算所有症状的优先 $bias_i$,计算公式如下: $bias_i = \log_{10}(Sequence_{value_i})$ 取优先值最高的症状,作为下一轮问询的症状。

2 预问诊系统的评价

2.1 机器评估

Loss曲线可以反映LLM训练效果,如附图4(请扫描本文二维码获取)所示,在训练过程中,经过100000次迭代,训练损失从初始较高值下降,最后训练损失逐渐下降最终收敛至相对稳定状态,表明中医LLM(13B参数)已训练完成,且有较好泛化能力。

本研究使用BLEU-4、ROUGE-L两个指标评估大模型生成结果。BLEU-4是一种广泛用于评估机器翻译质量的指标,通过比较机器生成的翻译文本

与一个或多个参考翻译之间的相似度来计算分数;ROUGE-L被用来评估自动生成文本的质量^[27]。课题组在前期调研中,评估得到2360亿参数的开源模型DeepSeek-V2^[28]在中医领域表现最佳,故在本研究中,将中医LLM与DeepSeek-V2进行能力评估比较。结果显示,中医LLM(BLEU-4=0.38, ROUGE-L=0.62), DeepSeek-V2(BLEU-4=0.43, ROUGE-L=0.65),提示DeepSeek-V2在通用文本生成中略优于中医LLM。

2.2 中医专业知识评估

为评估中医LLM的中医专业能力,本研究构建一系列验证集,覆盖中医执业医师、执业中药师、症状术语识别、经络识别方面的选择题、填空题、简答题、论述题等多类题型,过程均基于预设标准答案进行自动化评估。测试结果见表1,中医LLM在中医执业医师、执业中药师、症状术语识别方面正确率高于DeepSeek-V2,在经络识别方面正确率低于DeepSeek-V2。

表1 中医大语言模型与DeepSeek-V2在不同验证集中的专业能力测试正确率比较

Table 1 Correctness rate comparison of the TCM large language model and DeepSeek-V2 in professional competency tests (%)

组别	中医执业医师	执业中药师	症状术语识别	经络识别
中医大语言模型	94.09	78.00	87.50	68.80
DeepSeek-V2	75.60	69.00	85.50	86.40

2.3 临床测试评估

2.3.1 临床测试方案 以2024年6-7月就诊于中国中医科学院广安门医院门诊的慢性非萎缩性胃炎患者360例为研究对象(伦理批件号:2023-163-KY-01)。诊断标准参照中华医学会消化病学分会制定的《中国慢性胃炎共识意见(2017年,上海)》^[29],中医辨证标准参照中国中西医结合学会消化系统疾病专业委员会拟定的《慢性非萎缩性胃炎中西医结合诊疗共识意见(2017年)》^[30];选取年龄18~75岁,符合慢性非萎缩性胃炎诊断标准,且签署知情同意书自愿参加者,排除患有严重心、脑、肾或血液系统疾病者。

研究人员使用症状客观化量表收集患者所有症状,作为后续可参考的数据。患者向预问诊系统输入主诉,系统对患者进行提问后患者回复,直到系统停止提问,研究人员记录预问诊系统问诊轮次、问诊时间、证型结果。中医主治医师再次提问患

者,实验人员记录医生问诊轮次、问诊时间、证型结果。系统与主治医师分别得到的证型结果由2位具有高级职称的中医专家参考《慢性非萎缩性胃炎中西医结合诊疗共识意见(2017年)》^[30]进行最终判定,两位专家评估结果经Kappa检验,一致性系数为0.79,表明专家间具有较好的一致性。对于评估不一致的病例,两位专家通过讨论达成共识,形成最终判定结果作为评估标准。数据使用GraphPad Prism 10软件进行分析,计量资料均符合正态分布,以均值±标准差($\bar{x}\pm s$)表示,使用配对 t 检验分析;计数资料以例数(%)表示,采用卡方检验;以 $P<0.01$ 表示有显著统计学意义。

2.3.2 临床测试结果 相较主治医师,基于中医LLM的预问诊系统在问诊轮次、问诊时间方面减少,差异有统计学意义($P<0.01$),见表2。预问诊系统判定错误的有42例,该42例患者主治医师也均判定错误。在辨证准确率上,系统正确率为88.33%,略高于主治医师的85.28%,但差异无统计学意义($P>0.05$),见表3。

表2 基于中医大语言模型的智能预问诊系统与中医主治医师问诊轮次与时间比较

Table 2 Comparison of consultation rounds and duration between the TCM large language model-based intelligent pre-consultation system and attending physicians ($\bar{x}\pm s$)

组别	问诊轮次/次	问诊时间/min
主治医师	19.63±3.36	4.52±0.21
预问诊系统	15.08±1.09 ^{a)}	1.53±0.02 ^{a)}

注:a)与主治医师相比, $P<0.01$ 。

表3 基于中医大语言模型的智能预问诊系统与中医主治医师辨证准确率比较

Table 3 Comparison of syndrome identification accuracy between the TCM large language model-based intelligent pre-consultation system and attending physicians

组别	总数	正确数	错误数	准确率/%
主治医师	360	307	53	85.28
预问诊系统	360	318	42	88.33

3 讨论

问诊是中医师诊察疾病的方法,在四诊中占有重要地位,受到历代医家的高度重视。智能化预问诊系统能提前收集患者症状信息、提高医生诊疗效率,但与智能脉诊和舌诊相比,其发展相对滞后^[31]。滞后的原因主要由于中医问诊数据标准化程度较低,用于训练问诊系统的数据质量、数量、疾病范围存在较大差异,同时机器的语义理解深度性差、

语言模型泛化能力不足。LLM技术为基于自然语言处理的人机交互智能预问诊系统提供了可能^[32],然而具有中医辨证推理思维的预问诊系统尚待开发。基于此,本研究提出了一种基于LLM的中医智能预问诊系统的构建方法,并对系统进行临床测试。

本研究基于雅意大语言模型,采用DeepSpeed ZeRO-3策略,在8块英伟达A100图形处理器(80 GB/卡)上进行微调训练,收集古籍文献结构化数据、真实世界临床病历,基于《黄帝内经》理论框架,整理不同症状之间互斥、相交关系,训练了智能体辨证推理模型,将中医辨证思维客观化、公式化。Loss函数表明中医LLM已训练完成,且有较好泛化能力。在机器评估方面,通过BLEU-4, ROUGE-L自然语言理解指标证明,小参数量13B的中医LLM表现能力与超大参数DeepSeek-V2相当;在专业知识评估方面,中医LLM在中医执业医师、执业中药师、症状术语识别等测试中准确率远高于DeepSeek-V2;临床测试方面,基于中医LLM的预问诊系统在问诊轮次、问诊用时方面均优于主治医师,辨证准确率方面也显示出一定优势,证明中医预问诊模型构建初步成功,且一定程度上具备主治医师在辨证方面的推理能力。

本研究打破了传统中医问诊客观化研究中基于量表询问患者的模式,创新性构建了人机交互引导式的中医智能预问诊系统,提升了问诊效率。本研究过程中,存在系统询问症状不够精细、中医LLM缺少评价标准、辨证逻辑难以客观化评估等问题,未来可通过构建专科专病中医智能问诊系统、共建中医LLM标准、依据辨证方法分别编写专科专病辨证逻辑专家共识等方案逐步解决。智能中医问诊研究在未来将迎来更多创新和变革,在跨学科融合与综合治疗、实时检测与预警、智能化辅助诊疗、教育与培训等领域发挥优势。

参考文献

[1]侯佳音,徐博,何萍.基于AI的诊前预问诊系统的设计与应用[J].中国数字医学,2024,19(7):17-22.

[2]国家中医药管理局,国家数据局.国家中医药管理局 国家数据局关于印发《关于促进数字中医药发展的若干意见》的通知[EB/OL].(2024-07-19)[2024-11-30].
https://www.gov.cn/zhengce/zhengceku/202408/content_6968519.htm.

[3]侯春蕾,许颖,崔延婕,等.2型糖尿病中医问诊量表的研制及临床应用概述[J].世界科学技术-中医药现

- 代化, 2021, 23(2): 396-401.
- [4]刘国萍, 王忆勤, 董英, 等. 中医心系问诊量表的研制及评价[J]. 中西医结合学报, 2009, 7(1): 20-24.
- [5]王芬, 刘铜华, 丁雷, 等. 基于概率图的中医动态交互问诊与智能辨证的数学模型的探索[J]. 世界科学技术-中医药现代化, 2023, 25(10): 3370-3376.
- [6]LIU GP, YAN JJ, WANG YQ, et al. Application of multi-label learning using the relevant feature for each label in chronic gastritis syndrome diagnosis [J]. Evid Based Complement Alternat Med, 2012, 135387. doi: 10.1155/2012/135387.
- [7]黄仲羽, 刘凤斌. 基于多维计算机自适应测试的中医证候量化方法学探索[J]. 时珍国医国药, 2018, 29(1): 234-236.
- [8]余楷杰, 袁芳君, 马庆宇, 等. 机器学习驱动中医诊断智能化的发展现状、问题及解决路径[J]. 中国中医基础医学杂志, 2024, 30(3): 398-406.
- [9]TAN S, XIN X, WU D. ChatGPT in medicine: prospects and challenges: a review article [J]. Int J Surg, 2024, 110(6): 3701-3706.
- [10]柏灿, 王洁. 人工智能大语言模型在中医药领域的应用[J]. 西昌学院学报(自然科学版), 2024, 38(2): 62-69.
- [11]XIAO X, WANG Y, XU N, et al. YAYI-UIE: A chat-enhanced instruction tuning framework for universal information extraction [EB/OL]. (2023-12-24) [2024-10-16]. <http://arxiv.org/abs/2312.15548>.
- [12]LUO Y, KONG Q, XU N, et al. YAYI 2: Multilingual open-source large language models [EB/OL]. (2023-12-22) [2024-10-16]. <http://arxiv.org/abs/2312.14862>.
- [13]刘阳. 智能体交互模式大模型时代的新型交互范式[J]. 通信世界, 2024(13): 26-29.
- [14]薛清录. 中国中医古籍总目[M]. 上海: 上海辞书出版社, 2007.
- [15]严世芸. 中国医籍通考[M]. 上海: 上海中医学院出版社, 1990.
- [16]丹波元胤. 中国医籍考[M]. 北京: 人民卫生出版社, 1956.
- [17]KUMAR A, KHARADI A, SINGH D, et al. Automatic question-answer pair generation using Deep Learning [EB/OL]. (2021-09-02) [2025-02-01]. <https://ieeexplore.ieee.org/xpl/conhome/9544088/proceeding>.
- [18]HANINDHITO B, PATEL B, JOHN LK. Bandwidth characterization of DeepSpeed on distributed large language model training [EB/OL]. (2024-05-05) [2025-02-01]. <https://ieeexplore.ieee.org/xpl/conhome/10589923/proceeding>.
- [19]吴勉华, 石岩. 中医内科学[M]. 5版. 北京: 中国中医药出版社, 2021.
- [20]陈红风. 中医外科学[M]. 5版. 北京: 中国中医药出版社, 2021.
- [21]冯晓玲, 张婷婷. 中医妇科学[M]. 5版. 北京: 中国中医药出版社, 2021.
- [22]赵霞, 李新民. 中医儿科学[M]. 5版. 北京: 中国中医药出版社, 2021.
- [23]赵宗耀, 何安宇, 张宇初, 等. 论症状的性质及其对症状标准化研究的意义[J]. 中医杂志, 2022, 63(19): 1801-1805.
- [24]王志国. 病状术语规范化基础[M]. 北京: 人民卫生出版社, 2015.
- [25]姚乃礼. 中医症状鉴别诊断学[M]. 2版. 北京: 人民卫生出版社, 2005.
- [26]全国中医标准化技术委员会. 中医临床名词术语 第1部分: 内科学[S]. 北京: 中国标准出版社, 2023.
- [27]苑中旭, 李理, 何凡, 等. 融合思维链与知识图谱的中医问答模型[J/OL]. 计算机工程与应用. (2024-10-15) [2025-02-12]. <https://link.cnki.net/urlid/11.2127.TP.20241015.1111.006>.
- [28]DEEPSEEK-AI, LIU A, FENG B, et al. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model [EB/OL]. (2024-05-07) [2024-10-17]. <https://arxiv.org/abs/2405.04434>.
- [29]中华医学会消化病学分会. 中国慢性胃炎共识意见 (2017年, 上海) [J]. 胃肠病学, 2017, 22(11): 670-687.
- [30]中国中西医结合学会消化系统疾病专业委员会. 慢性非萎缩性胃炎中西医结合诊疗共识意见 (2017年) [J]. 中国中西医结合消化杂志, 2018, 26(1): 1-8.
- [31]中华中医药学会. 2023年度中医药重大科学问题、工程技术难题和产业技术问题[J]. 中医杂志, 2023, 64(14): 1405-1421.
- [32]丁勇, 蔡秀军, 薛翀, 等. 基于大模型的智能预问诊系统创新与实践[J]. 中国信息化, 2024(6): 68-69.

Construction and Validation of a Large Language Model-Based Intelligent Pre-Consultation System for Traditional Chinese Medicine

LIU Yiqing¹, LI Ying¹, YANG Hongjun¹, PENG Linjing¹, XIAN Nanxing¹, LI Kunning¹, SHI Qiwei¹, TIAN Hengyi¹, DONG Lifeng², WANG Lin³, ZHAO Yuping⁴

1. Experimental Research Center, China Academy of Chinese Medical Sciences, Beijing, 100700; 2. Institute of Automation, Chinese Academy of Sciences; 3. Tianjin Wenge Technology Co., Ltd; 4. Institute of Basic Theory for Chinese Medicine, China Academy of Chinese Medical Sciences

ABSTRACT **Objective** To construct a large language model (LLM)-based intelligent pre-consultation system for traditional Chinese medicine (TCM) to improve efficacy of clinical practice. **Methods** A TCM large language model was fine-tuned using DeepSpeed ZeRO-3 distributed training strategy based on YAYI 2-30B. A weighted undirected graph network was designed and an agent-based syndrome differentiation model was established based on relationship data extracted from TCM literature and clinical records. An agent collaboration framework was developed to integrate the TCM LLM with the syndrome differentiation model. Model performance was comprehensively evaluated by Loss function, BLEU-4, and ROUGE-L metrics, through which training convergence, text generation quality, and language understanding capability were assessed. Professional knowledge test sets were developed to evaluate system proficiency in TCM physician licensure content, TCM pharmacist licensure content, TCM symptom terminology recognition, and meridian identification. Clinical tests were conducted to compare the system with attending physicians in terms of diagnostic accuracy, consultation rounds, and consultation duration. **Results** After 100 000 iterations, the training loss value was gradually stabilized at about 0.7 ± 0.08 , indicating that the TCM-LLM has been trained and has good generalization ability. The TCM-LLM scored 0.38 in BLEU-4 and 0.62 in ROUGE-L, suggesting that its natural language processing ability meets the standard. We obtained 2715 symptom terms, 505 relationships between diseases and syndromes, 1011 relationships between diseases and main symptoms, and 1 303 600 relationships among different symptoms, and constructed the Agent of syndrome differentiation model. The accuracy rates in the simulated tests for TCM practitioners, licensed pharmacists of Chinese materia medica, recognition of TCM symptom terminology, and meridian recognition were 94.09%, 78.00%, 87.50%, and 68.80%, respectively. In clinical tests, the syndrome differentiation accuracy of the system reached 88.33%, with fewer consultation rounds and shorter consultation time compared to the attending physicians ($P < 0.01$), suggesting that the system has a certain pre-consultation ability.

Conclusion The LLM-based intelligent TCM pre-diagnosis system could simulate diagnostic thinking of TCM physicians to a certain extent. After understanding the patients' natural language, it collects all the patient's symptom through guided questioning, thereby enhancing the diagnostic and treatment efficiency of physicians as well as the consultation experience of the patients.

Keywords consultation; large language models; intelligent pre-consultation system; agent

(收稿日期: 2024-11-18; 修回日期: 2025-03-04)

[编辑: 焦 爽]