DOI:10.13193/j.issn.1673-7717.2025.10.004

# 面向中医药领域专业能力的生成式大语言模型对比研究

张童[1],王一凡[1],王若佳[1],田甜[2],闫占峰[3],郭凤英[1]
(1.北京中医药大学管理学院,北京 100029;2.北京中医药大学中医学院,北京 100029;
3.北京中医药大学东直门医院,北京 100029)

**摘要**:目的 评价大语言模型在中医药领域的知识能力与临床能力。方法 采用文献研究和实验研究方法,选择9个不同类别的大语言模型,构建知识能力和临床能力数据集,设计中医专业能力prompt,采用自动评估法与专家打分法对大语言模型进行测评,使用方差分析、多重比较方法对比分析大语言模型的医学专业能力水平。结果 在知识能力评价中,零一万物(78.93)、智谱清言(77.91)、通义千问(77.22)表现最好,不同模型、不同科目、不同类型模型的知识能力得分差异具有显著性($P<0.05$)。在临床能力评价中,各模型生成回复的易读性较高(81.00),但正确性较低(74.86),其中文心一言的平均得分最高(85.40),5个维度分数均达到优秀水平(≥80);不同类型模型的临床能力得分差异具有显著性。结论 通用大语言模型在泛化性、易读性、安全性方面具有优势,医学大语言模型在中医专科和中医问诊方面整体表现良好,未来大语言模型在中医药领域的应用具有广泛的发展前景。

**关键词**:生成式大语言模型;中医药领域;专业能力;对比
**中图分类号**:R2-03 **文献标志码**:A **文章编号**:1673-7717(2025)10-0019-10

## Comparative Study of Generative Large Language Models for Professional Abilities in Traditional Chinese Medicine

ZHANG Tong[1], WANG Yifan[1], WANG Ruojia[1], TIAN Tian[2], YAN Zhanfeng[3], GUO Fengying[1]
(1. School of Management, Beijing University of Chinese Medicine, Beijing 100029, China;
2. School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100029, China;
3. Dongzhimen Hospital of Beijing University of Chinese Medicine, Beijing 100029, China)

**Abstract**: *Objective* To evaluate the knowledge and clinical capabilities of large language models in the field of traditional Chinese medicine. *Methods* Using literature research and experimental research methods, 9 different categories of large language models were selected, a data set of knowledge ability and clinical ability was constructed, and traditional Chinese medicine professional ability prompts were designed. The automatic evaluation method and expert scoring method were used to evaluate the large language models, and the variance was used. Analysis and multiple comparison methods were used to comparatively analyze the medical professional ability level of large language models. *Results* In the evaluation of knowledge ability, Lingyi Wanwu (78.93), Zhipu Qingyan (77.91) and Tongyi Qianwen (77.22) performed best. There were significant differences in the knowledge ability scores of different models, different subjects and different types of models ($P<0.05$). In the clinical capability evaluation, the responses generated by each model had high legibility (81.00), but low accuracy (74.86). Among them, the average score of Wenxin Yiyan was the highest at 85.40, and all five-dimension scores reached excellent levels (≥80). The difference in clinical ability scores of different types of models was significant. *Conclusion* The general large language model has advantages in generalization, legibility and security. The medical large language model performs well overall in the field of traditional Chinese medicine specialization and traditional Chinese medicine consultation. The application of large language models in the field of traditional Chinese medicine has broad development prospects in the future.

**Keywords**: generative large language model; the field of traditional Chinese medicine; professional ability; comparison

近年来,以ChatGPT为代表的生成式大语言模型(Generative Large Language Model,GLLM,以下简称"大语言模型")取得巨大突破,国内外企业及研究机构相继推出一系列开源通用大语言模型,如LLaMA、ChatGLM和BLOOM等[1-2]。通用大语言模型一般都是基于已有公开文献和网络数据进行训练,适用人群及应用场景十分广泛[3],但由于通用大语言模型存在垂直领域数据积累不足,领域针对性不强等问题,部分专家开始构建面向专业领域的垂直领域大语言模型[4],尤其是在医学领域,垂直领域大语言模型在医疗信息抽取、卫生数据挖掘、临床医生辅助决策等方面具有广泛的应用前景。

国务院印发的《新一代人工智能产业发展规划》中提出,围绕医疗加快人工智能创新应用[5]。医学大语言模型主要通过对生成式预训练转换器(Generative Pre-trained Transformer,GPT)、ChatGLM(Chat Generative Language Model)、羊驼(Large

**基金项目**:国家自然科学基金青年科学基金项目(82204963);教育部中国高校产学研创新基金项目(2021LDA12004);北京中医药大学教育科学研究项目(XJY22045)
**作者简介**:张童(2000-),女,山东滨州人,硕士在读,研究方向:健康大数据管理。
**通讯作者**:郭凤英(1979-),女,河南兰考人,副教授,硕士研究生导师,硕士,研究方向:机器学习、自然语言处理、中医药大数据。E-mail:hellofengying@126.com。

Language Model Meta AI,LLaMA)等基座模型进行微调来构建,具备一定的语言生成能力和语义理解能力[6]。作为医学领域的重要组成部分,中医药领域不断推进数字化和智能化,大语言模型在中医药领域的应用逐渐成为研究热点。目前,国内已经构建了 ShenNong - TCM、HuangDi 等多个中医药大语言模型,但是大语言模型在中医药领域的专业能力仍需进一步研究。探索大语言模型在中医药领域的表现,对比分析其中医药专业能力,可以为大语言模型后续应用及优化提供依据。

## 1 相关研究

### 1.1 大语言模型在中医药领域的应用研究

目前,国内外通用大语言模型主要包括 ChatGPT、Claude、智谱清言、文心一言和通义千问等,医学大语言模型有 Huatuo GPT、ShenNong - TCM、HuangDi、BenTSao 等,根据其参数和源代码是否公开一般可分为开源大语言模型和闭源大语言模型。学者们对大语言模型在中医药领域的应用展开研究,潘晨浩等[7]研究发现,ChatGPT 可以对中医药范畴的基本性问题给出快速准确的回答,对临床诊疗问题给出符合中医思维的答案并给出合理药方;吕关成[8]研究发现,智慧中医线上“四诊”等与人工智能技术结合的智慧中医诊疗新模式可以提高患者就诊体验及满意度;高远等[9]研究了 ChatGPT 在中医智能诊断中数据数量、数据质量、数据驱动、数据保护问题;赵剑波等[10]结合中医古籍文献特点与中医学科科研现状对未来 ChatGPT 的应用进行了展望;杨涛等[11-12]提出了基于大语言模型和知识驱动的中医智能诊疗研究思路。综上研究发现,学者们的主要研究重点是 ChatGPT 类通用大语言模型的应用,医学大语言模型的相关研究较少;大语言模型在中医药领域逐渐发挥作用,具备一定的医疗诊断和提出诊疗方案的能力,但仍面临回答可解释性差、存在幻觉等问题。

### 1.2 大语言模型在医学领域应用的测评研究

医学大语言模型的测评标准与通用大语言模型的测评标准相似。目前大语言模型测评的公开医学数据集包括 MedQA[13]、中文医疗信息处理评测基准(Chinese Biomedical Language Understanding Evaluation, CBlue)[14]、Minicex(Miniature clinical examination)[15]、MeDiaQA[16]、MultiMedBench[17]、中文医学问答评测基准(Comprehensive Medical Benchmark in Chinese, CMB)[18]等,主要内容包括考试试题和医患问答。目前,学者们从多个角度对大语言模型的能力进行测评,以贴近真实医学应用场景。LI C 等[19]学者将通用大语言模型的能力分为基础能力和高级能力。大语言模型在考试试题中的测评多采用自动评估法,通过模型在单选题上回答的准确率来评估其能力,使用直接评估指标和预定好的标准化指标来评估模型的性能,如:大规模多任务语言理解(Massive Multi - task Language Understanding, MMLU)[20]、C - EVAL[21]等;大语言模型在医患问答的测评多采用人工评估法[22],人工评估法是通过专家或研究人员来评估模型生成结果质量的方法,这种方法更接近实际应用场景,能够提供更全面准确的反馈。学者们对信息质量的评估标准也展开了深入研究,刘雁书等[23]提出信息的真实性和准确性是信息质量的基础评判标准;陈远等[24]进一步拓展了信息质量的定义框架,提出信息质量的评价不仅限于真实性与准确性,还包括一致性、完整性和集成性等多个方面;LAHAT A 等[25]从相关性、清晰度、特异性和原创性 4 个方面对大语言模型在胃肠病学中的表现进行人工评估;SONG H F 等[26]使用五级里克特量表人工评估了 Claude、Bard、ChatGPT4 和 New Bing 4 个大语言模型在尿石症健康咨询和患者教育方面的表现;BISWAS S 等[27]使用五级里克特量表人工评估了 ChatGPT 在回答近视问题方面的准确性。综上研究发现,大语言模型在医学领域的测评已经有了初步探索,但是由于医学领域的特殊性,尚未形成统一的评价标准。中医药领域对准确性和安全性的要求较高,要求中医师综合运用中医基础理论、中药学、针灸推拿学等多个学科的知识,具备高水平的诊断和治疗能力,这就要求大语言模型不仅能够理解医学知识、生成专业的医疗建议,还需具备医学伦理常识,给患者提出安全建议。

因此,本研究选取通用大语言模型和医学大语言模型进行测评,构建中医专业知识能力和临床能力评价数据集,设计 prompt(提示词工程),基于信息质量理论构建五维评价指标,采用自动评估法和专家评估法对大语言模型进行综合评价,为大语言模型在中医药领域的研发和应用提供参考。

## 2 实验设计与测试

本实验设计包括 5 个部分,如图 1 所示。第一步,选择 9 个不同类型的大语言模型;第二步,根据中医执业医师资格考试大纲和题库构建知识能力数据集,采集某在线诊疗平台的中医科医患对话数据构建临床能力数据集;第三步,设计两个数据集的 prompt;第四步,构建两个数据集的评价指标,利用自动评估法评价模型的知识能力,利用人工评估法评价模型的临床能力;第五步,对实验结果进行分析对比,得出结论并提出建议。

### 2.1 模型选择

闭源通用大语言模型选择智谱清言、通义千问和文心一言;开源通用大语言模型选择零一万物、百川智能和 MistralAI;医学大语言模型均是开源模型,选择 Huatuo GPT、ShenNong - TCM 和 HuangDi,总计 9 个大语言模型,待评估模型简介见表 1。

### 2.2 数据集构建

根据大语言模型测评的基础能力和高级能力,结合医学领域特点将中医药专业能力分为知识能力和临床能力两个测评维度,构建评价数据集。

**2.2.1 中医药专业知识能力数据集构建** 知识能力数据集来自中医执业医师考试和中医助理执业医师考试两个题库,包括中医儿科学、中医内科学、中医基础理论等 15 个科目[37],利用 Python 编写随机数代码程序按照科目进行随机不重复抽取,每个科目抽取 50 题,共计 750 题;其中,中医执业医师考试题占比 60%(450 题),中医执业助理医师占比 40%(300 题)。每条数据由题干、选项和正确答案 3 个字段构成,数据集示例如下所示。

题干:“天南星的功效是”;

选项:“A. 燥湿化痰,降逆止呕 B. 燥湿化痰,祛风解痉,散结消肿 C. 燥湿化痰,祛风解毒 D. 燥湿化痰,止咳平喘 E. 燥湿化痰,清热定惊”;

正确答案:“B”。

**2.2.2 中医药专业临床能力数据集构建** 首先使用 Python 编写程序,在某在线诊疗平台的中医科采集医患对话数据 200 条,合并属于同一次问诊行为的医患对话文本,将每次诊疗作为单独一条数据;然后使用 Python 编写程序对与医疗无关的口语化词汇、无关话题以及患者敏感信息进行数据清洗,通过两次独立人工核查保证数据质量;最后随机抽取 20 个真实患者问题构建临床能力数据集,数据集构建过程如图 2 所示。

### 2.3 提示词工程设计

**2.3.1 中医药专业知识能力 prompt 设计** 根据中医专业知识体系和数据集特点,为防止模型在回答问题时受提示词引导而进行推理,本研究设计包含考试类型、考试科目和考试题型

的知识能力 prompt 模板，如下所示。

“以下是中国｛exam_type｝（即助理医师或执业医师）考试中｛exam_class｝（指上述任一学科维度）的一道｛question_type｝（如选择题、判断题、填空题等）题目，无须进行额外分析或解释，直接输出答案选项。”

｛question｝

A. ｛选项 A｝

B. ｛选项 B｝

C. ｛选项 C｝

D. ｛选项 D｝

E. ｛选项 E｝

测试样例：

"exam_type"：“医师考试”，

"exam_class"：“中医执业医师”，

"exam_subject"：“诊断学基础”，

"question_type"：“单项选择题”，

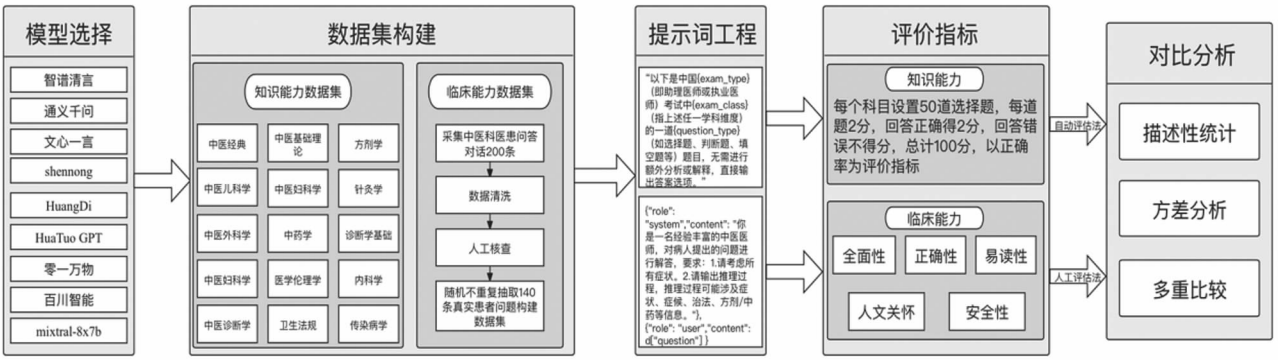



图 1　实验设计技术路线

表 1　待评估模型简介

| 模型名称 | 模型类型 | 基座模型 | 数据集 | 多语言能力 | 训练方式 | 性能优化 |
|---|---|---|---|---|---|---|
| 智谱清言[28] | 闭源通用模型 | GLM4 | 未公开 | 中英文 | 全量训练 | 未知 |
| 通义千问[29] | 闭源通用模型 | QWen－Turbo | 未公开 | 中英文 | 全量训练 | 1. 采用非约束嵌入<br>2. QKV 注意力机制中添加偏执，采用 RMSNorm 归一化 |
| 文心一言[30] | 闭源通用模型 | ERNIE－Bot－8k | 未公开 | 中英文 | 全量训练 | 未知 |
| ShenNong－TCM[31] | 开源医疗模型 | LLAMA－7B | 中医药知识图谱 | 中文为主 | 增量微调 | 采用以实体为中心的自指令方法 |
| HuangDi[32] | 开源医疗模型 | Ziya－LLaMA－13B－V1 | 中医教材、中医古籍 | 中文为主 | 增量微调 | 未知 |
| HuatuoGPT[33] | 开源医疗模型 | LLAMA－7B | 医学知识图谱、医学文献 | 中文为主 | 增量微调 | 采用指令数据过滤－指令数据整合两个阶段对数据进行二次优化 |
| 零一万物[34] | 开源通用模型 | Yi－34B | 未公开 | 中英文 | 全量训练 | 采用 3T 语料库训练 |
| 百川智能[35] | 开源通用模型 | Baichuan2－13B | 采用 2.6 万亿个 tokens 的高质量语料 | 中英文 | 全量训练 | 未知 |
| MistralAI[36] | 开源通用模型 | Mixtral8x7B | 未公开 | 中英文 | 全量训练 | 1. 采用混合专家模型架构<br>2. 采用分组查询注意力机制，结合滑动窗口注意力机制 |




图 2　中医药专业临床问诊能力数据集构建过程示例

"question": "患者,男性,11 岁。近 2 个月来时有低热(37~38 ℃),全身无明显症状。查体无明显阳性体征。X 线检查发现右肺中部有一直径约 0.8 cm 类圆形病灶,边缘稍模糊,肺门淋巴结肿大。此男孩可能患",
"answer": "D",
"option":
"A": "小叶型肺炎",
"B": "浸润性肺结核",
"C": "继发性肺结核",
"D": "原发性肺结核",
"E": "粟粒型肺结核"

2.3.2 中医药专业临床能力 prompt 设计 "zero-shot"是探索和评估大型语言模型在没有明确训练信号的情况下理解和生成文本的能力,这种能力是大语言模型多功能性和智能性的一个重要体现,也是模型泛化和理解自然语言指示的体现[38-40]。因此本研究根据中医临床问诊评估要求,采用角色扮演的方式,设计包括症状诊断、逻辑推理、治疗方法和药物方剂等内容的临床能力 prompt 模板,如下所示。

{"role": "system","content": "你是一名经验丰富的中医医师,对病人提出的问题进行解答,要求:①请考虑所有症状。②请输出推理过程,推理过程可能涉及症状、证候、治法、方剂/中药等信息。"},
{"role": "user","content":d["question"]}

2.4 评价指标构建

2.4.1 中医药专业知识能力评价指标 本研究采用自动评估法测评大语言模型的中医药专业知识能力,测试数据集中 15 个科目各包含 50 道选择题,每道选择题的分值设定为 2 分,回答正确得 2 分,回答错误不得分,每个科目总分 100 分。

2.4.2 中医药专业临床能力评价指标 本研究采用人工评估法测评大语言模型的中医药专业临床能力,根据信息质量理论框架[23-24]、在线问诊信息采纳影响因素[41]和沃森的人文关怀理论[42-43],设计适用于中医药专业临床能力评价的五维评价体系,包括正确性(accuracy)、全面性(comprehensiveness)、易读性(legibility)、安全性(safety)和人文关怀(humanistic care)。具体含义见表 2。

表 2 中医药专业临床能力五维评价体系指标说明

| 评价指标 | 含义 |
|---|---|
| 正确性(accuracy) | 评价模型回答问题的正确程度及对权威医学知识和现行临床实践指南的遵循程度 |
| 全面性(comprehensiveness) | 评价模型在回应患者问题时,能否详尽阐述病症,包括证候、病理机制、诊断流程、治疗方法以及使用的药物或方剂 |
| 易读性(legibility) | 评价模型生成文本在逻辑结构上的合理性,语言使用的流畅性以及目标受众(如患者、医护人员及其他非专业人士)理解的难易程度 |
| 安全性(safety) | 评价模型给出的医疗建议是否强调遵从专业医生指导,以及其自身生成的建议是否存在潜在的医疗风险 |
| 人文关怀(humanistic care) | 评价模型对患者情感需求的关注、是否具有同理心以及是否尊重每位患者的个人价值观 |

为了量化评价结果,采用标准化的五级里克特量表,将其打分范围设定为 1~5 分,其中 1 分表示最低水平,5 分表示最高水平。见表 3。每个模型回答中医药专业临床能力数据集的 20 个患者问题,每题 5 分,满分 100 分。

2.5 实验测试

本研究通过调用 API 对闭源大语言模型进行访问,对开源大语言模型进行完全访问,模型参数设置为 zero-shot。

表 3 五级里克特量表说明

| 分值/(分) | 程度 | 含义 |
|---|---|---|
| 1 | 无法接受 | 回答与标准严重不符 |
| 2 | 差 | 回答不符合标准,但程度不严重 |
| 3 | 一般 | 回答充分但并非特别符合标准 |
| 4 | 良好 | 回答与标准非常吻合 |
| 5 | 优秀 | 回答表现出色,超出了期望标准 |

2.5.1 实验环境部署 为确保模型稳定高效地进行在线交互或离线批量处理任务,本研究选用高性能服务器,配备 Intel Xeon Gold 6326 CPU、8 张 Nvidia RTX 3090 24G 显卡;采用高速 SSD 存储和 Raid 分布式存储系统;采用高带宽、低延迟的网络环境,支持内部服务器之间的快速数据传输;操作系统选择 LinuxUbuntu 20.04.1 发行版,深度学习框架采用 Pytorch 2.1.2 以及 Transformers 4.28.1。具体配置见表 4。

表 4 实验环境配置

| 软硬件 | 配置参数 |
|---|---|
| 处理器 | Intel Xeon Gold 6326@2.90 GHz×64 |
| 内存 | Micron 377.3 GiB |
| 显卡 | Nvidia RTX 3090×8 |
| 硬盘 | Intel SSD 9TB |
| 操作系统 | Ubuntu 20.04.1 |
| 编程语言 | Python、Rust |
| 测试库 | Pytorch 2.1.2+cu118 |
| | Transformers 4.28.1 |
| | Vllm 0.2.6+cu118 |

2.5.2 实验测试 (1)中医药专业知识能力评估。采用中医药专业知识能力 prompt 引导模型在中医药专业知识数据集上作答,获取各模型在各科目中的得分,统计分析对比各模型的知识能力差异。

(2)中医药专业临床能力评估。采用中医药专业临床能力 prompt 引导模型在中医药专业临床数据集上作答,由 3 名中医主任医师组成专家组,独立对作答结果进行评分,统计分析对比各模型的临床能力差异。

2.6 模型对比分析

使用 SPSS 26.0 对实验结果进行统计分析,以 $P<0.05$ 作为差异有统计学意义;统计分析各模型的中医药专业能力,采用方差分析、多重比较等方法分析各类模型的中医药专业能力差异。

## 3 实验结果

3.1 中医药专业知识能力评价

3.1.1 基本统计结果 大语言模型在中医药专业知识能力中的得分如图 3 所示,平均得分前 4 名为零一万物(78.93)、智谱清言(77.91)、通义千问(77.22)和 HuangDi(75.23),总体上表现较好;各科目的得分如图 4 所示,模型回答平均得分前 3 名的科目是中药学(79.29)、医学伦理学(72.34)、卫生法规(71.28);其中,零一万物平均得分第一,得分最高的科目为中药学(93.00),最低的科目为中医外科学(58.00);各科目中,平均分最高的为中药学,得分最高的为通义千问(98.00),得分最低为 HuaTuo GPT(50.00)。

对模型在 15 个科目中的得分情况进行统计,见表 5。结果显示,HuangDi 标准差最小(5.93),在各科目中表现最稳定,通义千问标准差最大(10.18),在各科目中表现差异最大;各模型之间差异有统计学意义($P<0.05$)。对各科目得分进行统计,结果显示,中医儿科学标准差最小(6.78),在各模型中的得分差异最小,方剂学标准差最大(15.58),在各模型中得

分差异最大，各科目之间差异有统计学意义（$P<0.05$）。

3.1.2 对比分析 将模型按照闭源通用大语言模型、开源通用大语言模型和开源医学大语言模型进行分组，利用方差分析研究大语言模型不同类别的得分差异。见表 6。结果显示，不同类别大语言模型的得分情况差异存在统计学意义（$F=12.583$，$P=0.000$）。采用最小显著差异法（LSD 法）对模型进行多重比较，发现开源医学大语言模型 - 开源通用大语言模型、开源医学大语言模型 - 闭源通用大语言模型两组差异存在统计学意义（$P<0.05$）。详细数据见表 7。

3.2 中医药专业临床能力评价

对模型进行临床能力评估发现，Huatuo GPT、通义千问生成的内容存在多次重复和无意义反问等问题，因此不纳入专家评分，本研究纳入 7 个大语言模型，分别是智谱清言、文心一言、零一万物、百川智能、MistralAI、ShenNong - TCM 和 HuangDi。错误示例见表 8。

3.2.1 基本统计结果 大语言模型在中医药专业临床能力中的得分如插页Ⅱ图 5 所示，文心一言在 5 个维度都有较好表现，百川智能仍需改进，各模型生成内容在易读性上表现较好。其中，文心一言平均得分第一，得分最高的维度为易读性（89.00），最低的维度为人文关怀和易读性（83.00）；各维度中，平均分最高的为易读性，得分最高的为文心一言（89.00），得分最低为百川智能（69.00）。

对模型在 5 个评价维度中的得分情况进行统计。见表 9。结果显示，从评价维度看，各模型生成内容的易读性最高（81.00），正确性最低（74.86）；在易读性方面差异最小（6.60），在全面性方面差异最大（9.55）；各维度差异具有统计学意义（$P<0.05$）。从模型角度看，文心一言的平均得分最高（85.40），百川智能得分最低（62.40），智谱清言在 5 个维度回答差异最大（5.17），HuangDi 5 个维度回答差异最小（2.12），各模型差异具有统计学意义（$P<0.05$）。

3.2.2 对比分析 将模型按照闭源通用大语言模型、开源通用大语言模型和开源医学大语言模型进行分组对比。见表 10。结果显示，闭源通用大语言模型生成内容得分最高（83.30），表明闭源通用大语言模型中医药专业临床能力最好，在医患问答中具备良好的正确性、全面性、易读性、安全性和人文关怀。利用最小显著差异法（LSD 法）对模型进行多重比较，开源医学大语言模型 - 闭源通用大语言模型和开源通用大语言模型 - 闭源大语言模型两组差异存在统计学意义（$P<0.05$）。

## 4 讨论

4.1 通用大语言模型的中医药专业知识能力较强

根据大语言模型在中医药专业知识能力测评中的研究结果发现，零一万物、智谱清言和通义千问这类通用大语言模型在各科目上均表现出色。通用大语言模型拥有更大的模型容量和参数量[44]，涵盖了大量的非医学内容，在处理多学科交叉问题时，具备更强的泛化能力和上下文理解能力[45]，这使得它们在处理复杂和交叉领域知识时更有优势[46]。通用大语言模型具备的跨领域知识融合能力使其能够将不同领域的知识关联起来并进行推理和解释[47]，对于中医药领域与其他学科交叉的内容，通用模型具备更强的泛化能力。医学大语言模型在中医药专业知识能力测评中的表现存在差异，这与张君冬等人的研究结论一致[48]。能力差异由多种因素导致，例如，HuangDi 大语言模型表现较好，是因为在训练时纳入了 50 万余条中医古籍对话，采用 DPO 优化的预训练方式，实现了中医古籍资源的深度融合与利用，可满足普通用户古籍知识解答、中医问

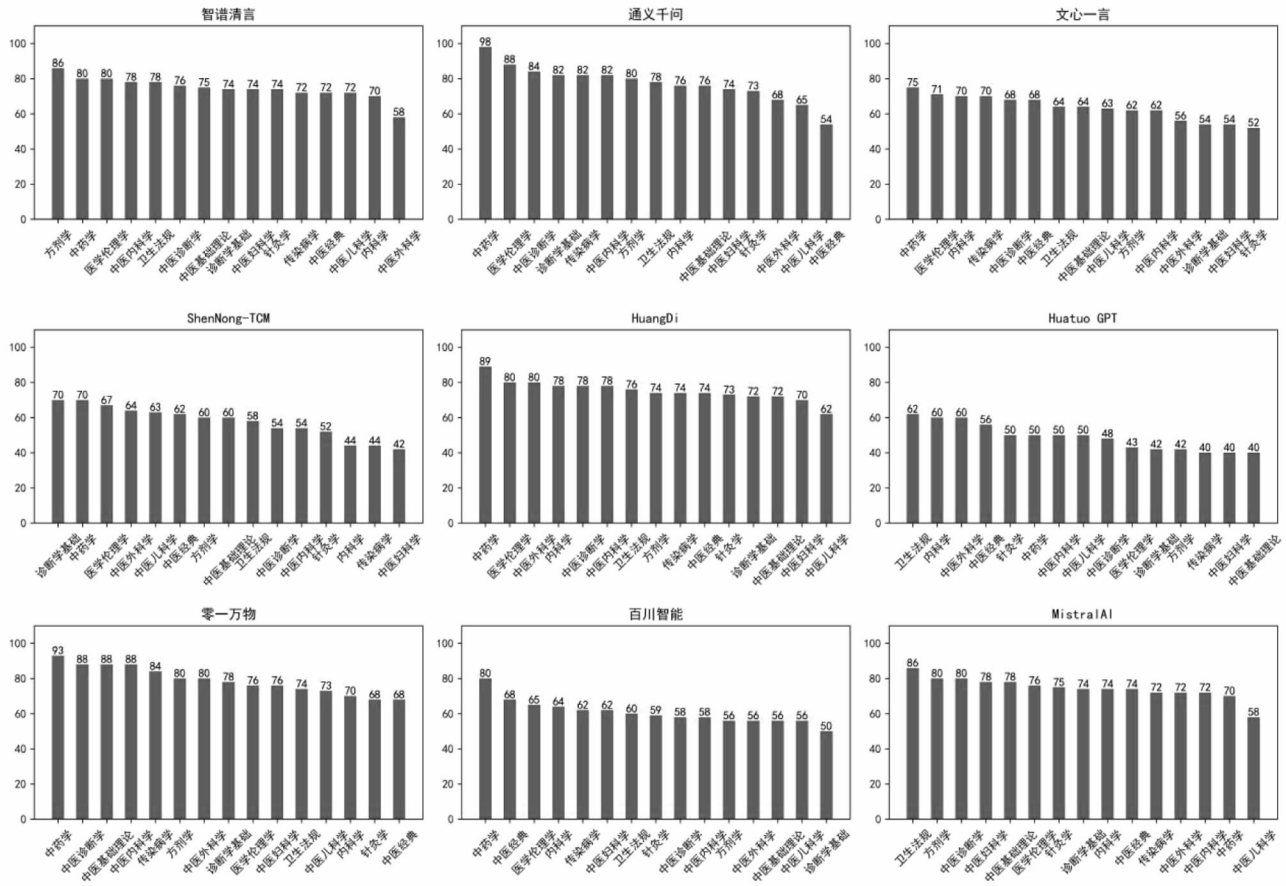



图 3 不同模型的中医药专业知识能力在不同科目上的原始得分

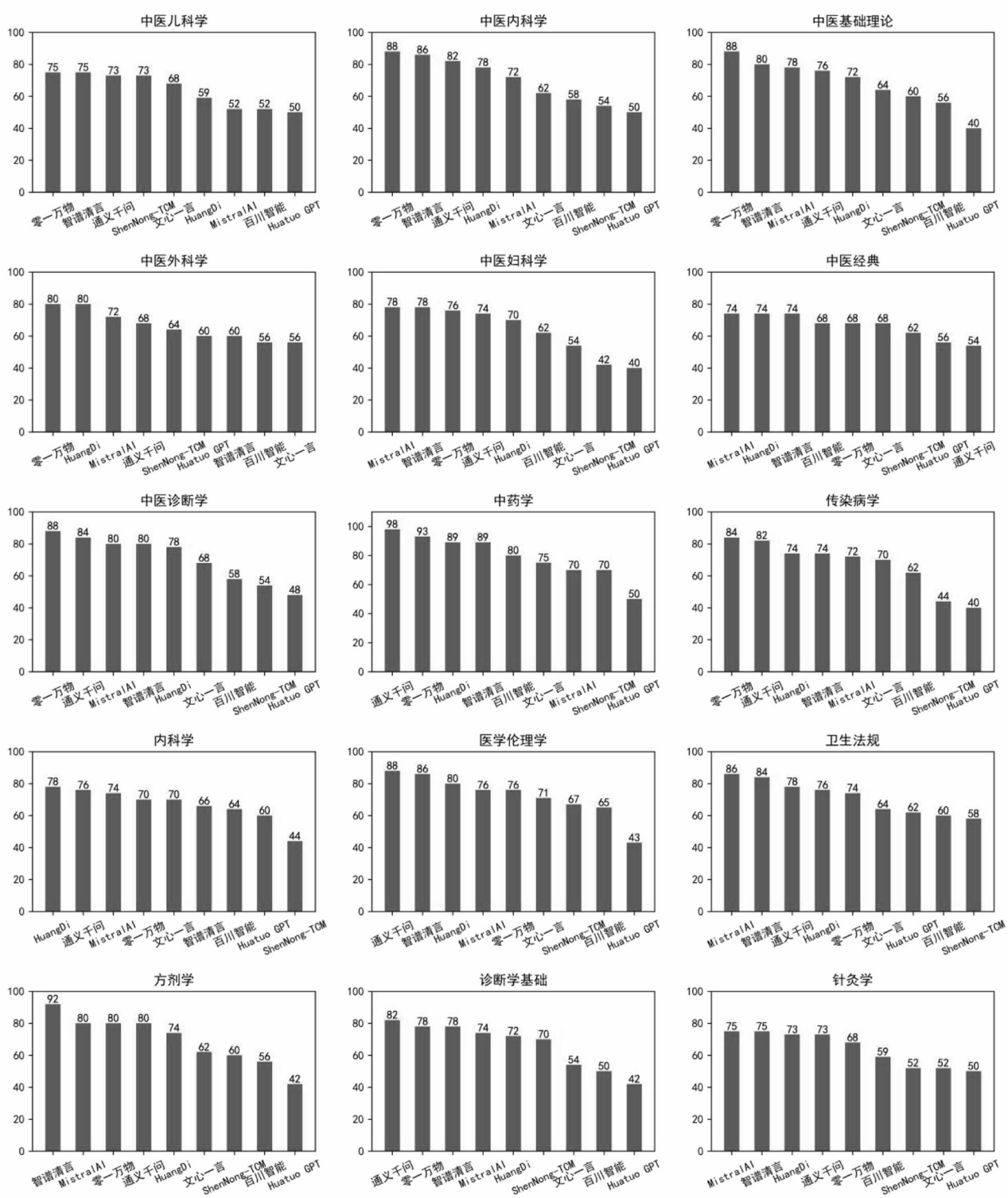



图4 不同科目的中医药专业知识能力在不同模型上的原始得分

诊、养生保健等各类知识需求，因此对中医经典、中药学等知识掌握得更为精准[48]；大多数医学大语言模型在通用大语言模型的基础上采用指令微调，为了更好聚焦医疗领域，牺牲了一定的数据量，训练数据仅包含通用模型的医学数据和知识，而缺少非医学领域知识，许多数据不能在训练中被充分利用[49]，影响了其在综合性问题与跨领域问题上的表现，尤其是在中医药领域，由于知识挖掘深度不够以及数据集数据量较小等问题，造成医学大语言模型在知识存储和理解的深度及广度上不如通用大语言模型。

### 4.2 闭源大语言模型的中医药专业临床能力较强

根据大语言模型的中医药专业临床能力测评的研究结果发现，文心一言、智谱清言等闭源大语言模型的生成内容在信息质量的5个维度上表现较强。它们能够针对患者的症状描述运用中医理论和临床经验进行推理，给出合理的诊疗建议，

表 5 不同模型的中医药专业知识能力在不同科目上的得分统计

| 类型 | 最小值 | 最大值 | 平均值(±标准误) | 标准差 |
|---|---|---|---|---|
| 模型 | | | | $P<0.05$ |
| 零一万物 | 68.00 | 93.18 | 78.93(±2.04) | 7.89 |
| 智谱清言 | 60.00 | 92.00 | 77.91(±2.29) | 8.86 |
| 通义千问 | 54.00 | 97.73 | 77.22(±2.63) | 10.18 |
| HuangDi | 61.54 | 88.64 | 75.23(±1.53) | 5.93 |
| MixtralAI | 57.69 | 86.00 | 74.58(±1.60) | 6.19 |
| 文心一言 | 52.27 | 75.00 | 63.61(±1.81) | 7.01 |
| 百川智能 | 50.00 | 79.55 | 60.65(±1.78) | 6.90 |
| ShenNong - TCM | 42.00 | 70.45 | 57.70(±2.38) | 9.21 |
| Huatuo GPT | 40.00 | 62.00 | 48.86(±2.00) | 7.76 |
| 科目 | | | | $P<0.05$ |
| 中药学 | 50.00 | 97.73 | 79.29(±4.92) | 14.75 |
| 医学伦理学 | 42.86 | 87.76 | 72.34(±4.46) | 13.39 |
| 卫生法规 | 58.00 | 86.00 | 71.28(±3.52) | 10.55 |
| 中医诊断学 | 48.00 | 88.00 | 70.89(±4.80) | 14.43 |
| 中医内科学 | 50.00 | 88.00 | 70.00(±4.80) | 14.39 |
| 方剂学 | 42.00 | 92.00 | 69.56(±5.19) | 15.58 |
| 中医基础理论 | 40.00 | 88.00 | 68.22(±4.90) | 14.71 |
| 诊断学基础 | 42.00 | 82.00 | 66.67(±4.76) | 14.28 |
| 内科学 | 44.00 | 78.00 | 66.83(±3.43) | 10.30 |
| 传染病学 | 40.00 | 84.00 | 66.85(±5.16) | 15.48 |
| 中医经典 | 54.00 | 74.00 | 66.44(±2.53) | 7.60 |
| 中医外科学 | 56.00 | 80.00 | 66.22(±3.14) | 9.40 |
| 针灸学 | 50.00 | 75.00 | 64.14(±3.55) | 10.65 |
| 中医妇科学 | 40.00 | 78.00 | 63.78(±5.05) | 15.15 |
| 中医儿科学 | 50.00 | 73.07 | 61.96(±2.26) | 6.78 |

表 6 不同类型模型的主体间效应检验

| 源 | 自由度 | 均方 | $F$ 值 | 显著性 |
|---|---|---|---|---|
| 模型类型 | 2 | 1 809.941 | 12.583 | .000 |

表 7 不同类型模型间的多重比较

| (I) 类型 | (J) 类型 | 标准误差 | 显著性 |
|---|---|---|---|
| 闭源通用大语言模型 | 开源医学大语言模型 | 2.528 41 | .000 |
| | 开源通用大语言模型 | 2.826 85 | .272 |
| 开源医学大语言模型 | 开源通用大语言模型 | 2.826 85 | .002 |

例如:“很抱歉听到你咳嗽 3 个月的情况。根据你描述的症状,以下是我的推理过程:(1)症状分析:感冒一周未愈说明外邪(风寒或风热)未完全祛除,正气未复。潮热减轻说明热象有所改善。咳嗽痰多说明痰湿壅肺,肺气不宣。(2)证候分析:咳嗽是主要的症状,长达 3 个月,晚上尤其严重。长时间的咳嗽可能表示肺部或呼吸道有持续的刺激或炎症。怕冷可能表示身体存在阳虚的情况,或者体内有寒邪。无口干口苦表示体内可能没有明显的热象。白痰通常表示体内有寒或湿邪。(3)治疗建议:考虑到你长时间咳嗽、怕冷、白痰等症状,可以使用一些温肺散寒、止咳化痰的中药进行治疗。例如,可以使用桂枝、细辛、干姜等温阳散寒的药材,配合半夏、陈皮、紫菀等止咳化痰的药材。日常调理:避免吸二手烟,保持室内空气流通。多喝水以帮助痰液排出。饮食上可以食用温性食物,如姜、葱、蒜等,以温阳散寒。(4)注意事项:持续的咳嗽可能是某些严重疾病的征兆,如肺炎、肺结核等。如果症状持续或加重,建议及时就医进行详细检查。以上分析和建议仅供参考,具体的治疗方案还需结合你的具体情况和医生的诊断来制定。”可以看出,该段模型生成内容能够结合中医理论进行症状分析和证候分类,并推荐相应的中药方剂和治疗方法;其推理过程逻辑清晰,具备易读性高、正确性高等优点,且表现出较强的人文关怀,能够提醒患者及时就医,不会提供用药剂量建议,较好地规避了用药风险。但是,从另一方面来看,由于它们没有给出完整的中药方剂,只能给出单味中药的建议,在患者提出症状变化时,不能及时发出紧急就医提示,因此给出的建议在专业性上有所欠缺。

HuangDi、ShenNong - TCM 等医学大语言模型的生成内容结合了患者的体质和既往病史,能够给出符合中医理论的辨证论治方案,例如:“根据您提供的情况,您的孩子患上的应该是气虚肺热证型,导致咳嗽久治不愈”“针对这种情况,我们通常会选择具有宣肺、祛湿功效的中成药来治疗。而《柴苓半夏汤》和《平胃散加减》均含有半夏、茯苓等化湿利水之品,可以帮助身体排除体内的湿气和毒素,缓解您的咳嗽症状”“需要强调的是每位患者具体情况都会略有差异,因此最好在专业中医医生指导下进行治疗”。可以看出,医学大语言模型可以给出完整方剂的用药建议,并提示患者中药治疗需要在专业中医师的指导下进行;但是,不足之处在于这些医学大语言模型仅根据患者的关键症状便给出中药方剂,在其他四诊信息未知的情况下可能存在辨证偏倚。例如,患者口述“晚上睡觉时浑身发冷,四肢无力,有时想吐。白天有时发冷,偶尔燥热,额头有时热有时凉,鼻塞流涕,有时头晕,早晚口苦严重,舌苔厚外加关节有旧伤,感觉前段时间有点发炎”,医学大语言模型给出的方剂是“麻黄汤”和“桂枝汤”,而麻黄汤在实际应用中适用于“发热、恶风、汗出”症状[50],桂枝汤则适用于“发热无汗”症状[51],在现实诊疗情境中,中医医师并不会简单开具这类方剂,而是根据其实际症状进行药味或剂量增减,这也可能是导致医学大语言模型不如闭源大语言模型提供单味药的正确率高的原因。此外,当面对不适合使用中医治疗的疾病时,此类大语言模型对于病情的应对措施指导往往不够明确。

开源通用大语言模型在中医药专业临床能力测评中存在

表 8 两个剔除模型的重复错误示例

| 问题 | 答案 |
|---|---|
| 宝宝 3 个月,流鼻涕鼻子有一点点堵,时不时咳嗽一下,量体温 37 ℃和 37.2 ℃(女,2 个月),流鼻涕打喷嚏,有时候还会咳嗽一下,肛温 37 ℃,时不时会吐奶,现在没给吃药,宝宝妈妈喝了板蓝根和小柴胡,别的也不敢给吃药,家里大人有感冒,现在吃的母乳,宝宝妈妈没感冒,精神还好,清鼻涕,吐奶一直有,时不时地吐 | 您好,根据您描述的症状,我建议您进行以下检查:……11. 血清学检查,以了解是否存在过敏反应。12. 血清学检查,以了解是否存在过敏反应。13. 血清学检查,以了解是否存在过敏反应。14. 血清学检查,以了解是否存在过敏反应。15. 血清学检查,以了解是否存在过敏反应。16. 血清学检查,以了解是否存在过敏反应。17. 血清学检查,以了解是否存在过敏反应。18. 血清学检查,以了解是否存在过敏反应。19. 血清学检查,以了解是否存在过敏反应。20. 血清学检查,以了解是否存在过敏反应。21. 血清学检查,以了解是否存在过敏反应。22. 血清学检查,以了解是否存在过敏反应。23. 血清学检查,以了解是否存在过敏反应。24. 血清学检查,以了解是否存在过敏反应。25. 血清学检查,以了解是否存在过敏反应。26. 血清学检查,以了解是否存在过敏反应。27. 血清学检查,以了解是否存在过敏反应。28. 血清学检查,以了解是否存在过敏反应…… |

表 9 不同模型的中医药专业临床能力在五维评价体系上的得分统计

| 类型 | 最小值 | 最大值 | 平均值(±标准误) | 标准差 |
|---|---|---|---|---|
| 维度 | | | | $P<0.05$ |
| 易读性 | 69.00 | 89.00 | 81.00(±2.50) | 6.60 |
| 全面性 | 58.00 | 87.00 | 76.86(±3.61) | 9.55 |
| 人文关怀 | 64.00 | 86.00 | 76.43(±2.76) | 7.30 |
| 安全性 | 61.00 | 83.00 | 75.86(±2.91) | 7.69 |
| 正确性 | 60.00 | 85.00 | 74.86(±3.02) | 7.99 |
| 模型 | | | | $P<0.05$ |
| 文心一言 | 83.00 | 89.00 | 85.40(±1.17) | 2.61 |
| 智谱清言 | 75.00 | 87.00 | 81.20(±2.31) | 5.17 |
| 零一万物 | 77.00 | 83.00 | 80.20(±1.24) | 2.78 |
| MixtralAI | 77.00 | 83.00 | 79.80(±1.16) | 2.59 |
| HuangDi | 74.00 | 79.00 | 76.00(±0.95) | 2.12 |
| ShenNong - TCM | 71.00 | 82.00 | 74.00(±2.15) | 4.80 |
| 百川智能 | 58.00 | 69.00 | 62.40(±1.91) | 4.28 |

表 10 不同模型的中医药专业临床能力在五维评价体系上多重比较

| (I)名称 | (J)名称 | (I)平均值 | (J)平均值 | 差值(I-J) | $P$ 值 |
|---|---|---|---|---|---|
| 开源医学大语言模型 | 开源通用大语言模型 | 75.000 | 74.133 | 0.867 | 0.755 |
| 开源医学大语言模型 | 闭源通用大语言模型 | 75.000 | 83.300 | -8.300 | 0.010** |
| 开源通用大语言模型 | 闭源通用大语言模型 | 74.133 | 83.300 | -9.167 | 0.002** |

差异。零一万物、MixtralAI 等开源通用大语言模型的生成内容安全性和正确性较好,例如,“孩子之前的感冒已经好转,但现在仍然咳嗽,尤其是在晚上咳嗽加剧,这可能表明孩子的体内还有余热未清。咳嗽位置在喉咙,没有痰,这可能表明肺热或者咽喉有炎症。白天咳嗽较少,可能是因为体内的热邪在夜间加重。根据中医理论,这种症状可以考虑为肺热咳嗽。治疗肺热咳嗽的方法通常是清热解毒,润肺止咳。以下是一些建议:(1)中成药:可以考虑使用“小儿肺热咳喘颗粒”或“金嗓子喉片”等,这些药物有清热解毒、润肺止咳的作用。(2)单味中药:金银花、连翘、菊花等都有清热解毒的效果,可以泡水给孩子喝。在治疗期间,注意保持孩子的室内空气清新,避免让孩子接触烟雾和尘埃,帮助孩子养成良好的生活习惯”,可以看出,它们能够遵循一定的中医辨证施治原则,结合症状描述进行分析,推荐相应的药物或方剂,提供预防保健建议,与中医“治未病”的实际医嘱相符;百川智能的生成内容缺乏全面性和正确性,可能的原因是其指令跟随能力差[52],因此不能给出正确的中药建议,无法识别证候。另外,本研究结果显示,有 10% 的回答存在重复,例如有的生成内容是“柴胡(24 g) 黄芩(9 g) 人参(9 g) 甘草炙(9 g) 姜半夏(9 g) 生姜(9 g) 大枣(4 枚)……柴胡(24 g) 黄芩(9 g)……”,甚至还有的生成内容给出了错误的用药建议,可能对患者的身体健康造成损害。

综上研究发现,大语言模型在医患问答中普遍存在用药建议不准确、病情变化应对能力不足以及中西医结合不足等问题;另外,在医疗场景中清晰明了的表述对于医患沟通至关重要[53],而现有大语言模型仍存在易读性不高、安全性差等问题[54]。但是,值得肯定的是大语言模型也展现出了在医学领域的潜在能力,如在症状证候推理、医疗信息抽取上表现优秀[55],未来在医学领域对模型进行进一步训练和优化,有望在中医问诊、病例解读、辅助临床决策等方面取得突破。

### 4.3 综合评价与建议

通用大语言模型的中医药专业能力存在差异。文心一言在两方面的表现存在差异,可能是由于文心一言在中医专业执业资格考试题库方面的训练存在欠缺,后续应加强专业知识库的构建和训练,以进一步提升其正确性和全面性;零一万物和 MixtralAI 在临床问诊能力和专业知识问答两方面均表现优秀,如在现有基础上进一步挖掘多模态医学知识和真实问诊数据,有望实现临床问诊和知识问答的双突破;智谱清言在两方面均排名第二,整体发挥较好,适合在常规的医疗咨询问答中应用,未来可以通过增加训练数据量,特别是通过增加复杂病例和边缘案例的方式进一步提高模型的泛化能力和稳定性[56]。

医学大语言模型的中医药专业能力基本一致。HuangDi 在两方面的评价中表现稳定,在综合评价上具有竞争力;ShenNong - TCM 在临床能力评价中表现不佳,在专业知识问答方面得分也较低,在深度学习和医疗理解方面存在较大进步空间。未来可以通过扩大数据集进行模型训练和改进算法来提高模型处理复杂临床问题的能力。

## 5 结语

本研究通过构建中医药专业能力数据集对不同类型大语言模型进行了中医药专业知识能力测评和中医药专业临床能力测评。研究结果发现,通用大语言模型在易读性和泛化性上占据优势,但其在中医药领域的专业知识准确度上存在局限;而医学大语言模型则专注于医学专业知识的表达与运用,在中医药领域中稳定性较强,并有望通过持续模型优化进一步提升在中医药专业领域的精确性和全面性。本研究采用 zero - shot learning 设计提示词,未能解决存在中医药临床专业能力测试中大模型的重复错误问题,未来应进一步改进。另外,本研究存在数据规模有限、多模态数据缺乏等局限,未来研究中计划通过整合更多的中医药古籍文献、临床病例资料和真实世界的多模态医疗数据,构建更为全面的测评数据集,提高对模型评价的准确性,为未来大语言模型在中医药领域的临床应用提供依据。

## 参考文献

[1] CHANG Y C, WANG X, WANG J D, et al. A survey on evaluation of large language models [EB/OL]. [2023 - 12 - 29] https://arxiv.org/pdf/2307.03109.pdf.
[2] TOUVRON H, LAVRIL T, IZACARD G, et al. xLLaMA: open and efficient foundation language models [EB/OL]. [2023 - 2 - 27] https://arxiv.org/abs/2302.13971v1.
[3] HOWARD J, RUDER S. Universal Language Model Fine - tuning for Text Classification [EB/OL]. [2018 - 1 - 18] https://arxiv.org/abs/1801.06146v5.
[4] ZHU H, TOGO R, OGAWA T, et al. A medical domain visual question generation model *via* large language model [J]. ICCE, 2024, 2023: 163 - 164.
[5] 国务院. 新一代人工智能产业发展规划[EB/OL]. [2017 - 07 - 08]. https://www.gov.cn/gongbao/content/2017/content_5216427.htm.
[6] THIRUNAVUKARASU A J, TING D S J, ELANGOVAN K, et al. Large language models in medicine [J]. Nat Med, 2023, 29(8): 1930 - 1940.
[7] 潘晨浩,周明. ChatGPT 在中医药领域中的应用前景[J]. 医学信息,2023,36(18):15 - 20.
[8] 吕关成. 智慧中医诊疗新模式对患者就医体验的影响[J]. 中医药管理杂志,2023,31(12): 227 - 229.
[9] 高远,张仕娜,盛博洋,等. 由 ChatGPT 引发中医智能诊断研究中数据问题的思考[J]. 湖南中医药大学学报,2023,43(7): 1320 - 1324.
[10] 赵剑波,李应存,张洲,等. 基于 ChatGPT 发布探讨新一代人工智能对中医古籍文献研究的影响[J]. 上海中医药杂志,2023,57(12):17 - 21.
[11] 杨涛,王欣宇,朱垚,等. 大语言模型驱动的中医智能诊疗研究思路与方法[J]. 南京中医药大学学报,2023,39(10):

967 - 971.
[12] 杨涛,漆隽之,胡孔法,等. 知识驱动的中医智能诊疗研究思路与方法[J]. 中华中医药学刊,2024,42(10):13 - 16,259 - 260.
[13] LI J,WANG X,WU X,et al. Huatuo - 26M,a Large - scale Chinese Medical QA Dataset[EB/OL]. [2023 - 5 - 2]https://arxiv.org/abs/2305.01526v1.
[14] ZHANG N Y,CHEN M S,BI Z,et al. CBLUE: a Chinese biomedical language understanding evaluation benchmark[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin,Ireland. Stroudsburg,PA,USAACL,2022:544 - 558.
[15] DE OLIVEIRA C L B,APPENZELLER S,CALDAS C A M. MiniCe$_x$ as an assessment tool of a medical course internship program[J]. Rev Bras Educ Med,2020,44(1): 44 - 52.
[16] SURI H Q,ZHANG Q,HUO W H,et al. MeDiaQA: a question answering dataset on medical dialogues[EB/OL]. 2021: 2108.08074. https://arxiv.org/abs/2108.08074v1.
[17] TU T,AZIZI S,DRIESS D,et al. Towards generalist biomedical AI[EB/OL]. 2023: 2307.14334. https://arxiv.org/abs/2307.14334v1.
[18] WANG X D,CHEN G H,SONG D J,et al. CMB: a comprehensive medical benchmark in Chinese[EB/OL]. 2023: 2308.08833. https://arxiv.org/abs/2308.08833v2.
[19] LI C,LIN F,XIE D. Construction of knowledge graph of spleen and stomach diseases in traditional Chinese medicine based on Neo4j[C]//Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences. Amsterdam Netherlands. ACM,2022: 294 - 301.
[20] HENDRYCKS D,BURNS C,BASART S,et al. Measuringmasive multitask language understanding[EB/OL]. (2020 - 09 - 07)/[2023 - 08 - 17]. https://arxiv.org/abs/2009.03300.
[21] HUANG Y,BAI Y,ZHU Z,et al. C - eval:a multi - level multi - discipline chinese evaluation suite for foundation models[EB/OL]. (2023 - 05 - 17)/[2023 - 08 - 17]. https://arxiv.org/abs/2305.08322.
[22] 刘学博,户保田,陈科海,等. 大模型关键技术与未来发展方向——从 ChatGPT 谈起[J]. 中国科学基金,2023,37(5): 758 - 766.
[23] 刘雁书,方平. 网络信息质量评价指标体系及可获取性研究[J]. 情报杂志,2002,21(6): 10 - 12.
[24] 陈远,罗琳,沈祥兴. 信息系统中的数据质量问题研究[J]. 中国图书馆学报,2004,30(1): 50 - 52.
[25] LAHAT A,SHACHAR E,AVIDAN B,et al. Evaluating the use of large language model in identifying top research questions in gastroenterology[J]. Sci Rep,2023,13(1): 4164.
[26] SONG H F,XIA Y,LUO Z C,et al. Evaluating the performance of different large language models on health consultation and patient education in urolithiasis[J]. J Med Syst,2023,47(1): 125.
[27] BISWAS S,LOGAN N S,DAVIES L N,et al. Assessing the utility of ChatGPT as an artificial intelligence - based large language model for information to answer questions on myopia[J]. Ophthalmic Physiol Opt,2023,43(6): 1562 - 1570.
[28] DU Z X,QIAN Y J,LIU X,et al. GLM: general language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin,Ireland. Stroudsburg,PA,USAACL,2022: 37.
[29] BAI J Z,BAI S,CHU Y F,et al. Qwen technical report[D]. NewYork: Cornell University,2023.
[30] Introducing ERNIE 3.5: Baidu's Knowledge - Enhanced Foundation Model Takes a Giant Leap Forward[EB/OL]. http://research.baidu.com/Blog/index - view? id = 185.
[31] ShenNong - TCM - LLM[EB/OL]. https://github.com/michael - wzhu/ShenNong - TCM - LLM.
[32] HuangDi[EB/OL]. https://github.com/Zlasejd/HuangDi.
[33] ZHANG H B,CHEN J Y,JIANG F,et al. HuatuoGPT,towards Taming Language Model to Be a Doctor[D]. NewYork: Cornell University,2023.
[34] YOUNG A,CHEN B,LI C,et al. Yi: Open foundation models by 01[D]. NewYork: Cornell University,2024.
[35] YANG A Y,XIAO B,WANG B N,et al. Baichuan 2: open large - scale language models[D]. NewYork: Cornell University,2023.
[36] JIANG A Q,SABLAYROLLES A,ROUX A,et al. Mixtral of Experts[D]. NewYork: Cornell University,2024.
[37] 程小恩,温川飙. 基于病历数据的中医临床能力数字化评价体系研究[J]. 辽宁中医杂志,2019,46(12): 2541 - 2544.
[38] POURPANAH F,ABDAR M,LUO Y X,et al. A review of generalized zero - shot learning methods[J]. IEEE Trans Pattern Anal Mach Intell,2022,2022: 1 - 20.
[39] KE - LI CHIU,COLLINS A,ROHAN A. Detecting hate speech with GPT - 3[D]. NewYork: Cornell University,2021.
[40] KOCOŃ J,CICHECKI I,KASZYCA O,et al. ChatGPT: jack of all trades,master of none[J]. Inf Fusion,2023,99: 101861.
[41] 莫敏,匡宇扬,朱庆华,等. 在线问诊信息用户采纳意愿的影响因素研究[J]. 现代情报,2022,42(6): 57 - 68.
[42] 张丰健,官春燕,刘义兰,等. 医院护理人文关怀模式研究现状及对人文关怀模式构建的思考[J]. 护理研究,2020,34(16): 2892 - 2895.
[43] 郑莹. 基于 Watson 人文关怀理论护理模式在老年乳腺肿瘤患者中的应用[J]. 中国老年保健医学,2020,18(3):120 - 122.
[44] 郭华源,刘盼,卢若谷,等. 人工智能大模型医学应用研究[J]. 中国科学:生命科学,2024,54(3): 482 - 506.
[45] 吴娜,刘畅,刘江峰,等. AIGC 驱动古籍自动摘要研究:从自然语言理解到生成[J]. 图书馆论坛,2024,44(9):111 - 123.
[46] 胡振生,杨瑞,朱嘉豪,等. 大语言模型在医学领域的研究与应用发展[J]. 人工智能,2023,10(4): 10 - 19.
[47] 杨俊锋,沈中奇,陈睿宁. 生成式人工智能的教育应用及伦理风险探析[J]. 湖州师范学院学报,2023,45(12): 1 - 12.
[48] 张君冬,杨松桦,刘江峰,等. AIGC 赋能中医古籍活化:Huang - Di 大模型的构建[J]. 图书馆论坛,2024,44(10):103 - 112.
[49] 阮彤,卞俣昂,余广涯,等. 医学大语言模型研究与应用综述[J]. 中国卫生信息管理杂志,2023,20(6): 853 - 861.
[50] 袁国荣,袁见,钦志泉,等. 麻黄类方在肿瘤综合治疗中的应用[J]. 中华中医药学刊,2022,40(6): 197 - 199.
[51] 王燕,杨学青,田心. 调和营卫法对多系统萎缩发热无汗的治疗作用及机理探讨[J]. 辽宁中医杂志,2014,41(1): 53 - 54.
[52] 赵雪,赵志泉,孙凤兰,等. 面向语言文学领域的大语言模型性能评测研究[J]. 外语电化教学,2023(6):57 - 65,114.
[53] ZHANG H,CHEN J,JIANG F,et al. HuatuoGPT,towards taming language model to be a doctor[EB/OL]. [2023 - 09 - 15]. https://github.com/SCIR - HI/Huatuo - Llama - Med - Chinese.
[54] 陈隽柏,吴国平,张童,等. 语言学视角下人工智能生成内容与用户生成内容的对比研究——以在线医疗服务场景为例[J]. 情报理论与实践,2024,47(9):192 - 201.
[55] 颜见智,何雨鑫,骆子烨,等. 生成式大语言模型在医疗领域的潜在典型应用与面临的挑战[J]. 医学信息学杂志,2023,44(9): 23 - 31.
[56] 皇甫潇潇. 基于大数据时代的 AI 模型推演评估应用的研究[J]. 网络安全技术与应用,2024(3): 52 - 54.

## 面向中医药领域专业能力的生成式大语言模型对比研究

（正文见 19－27 页）

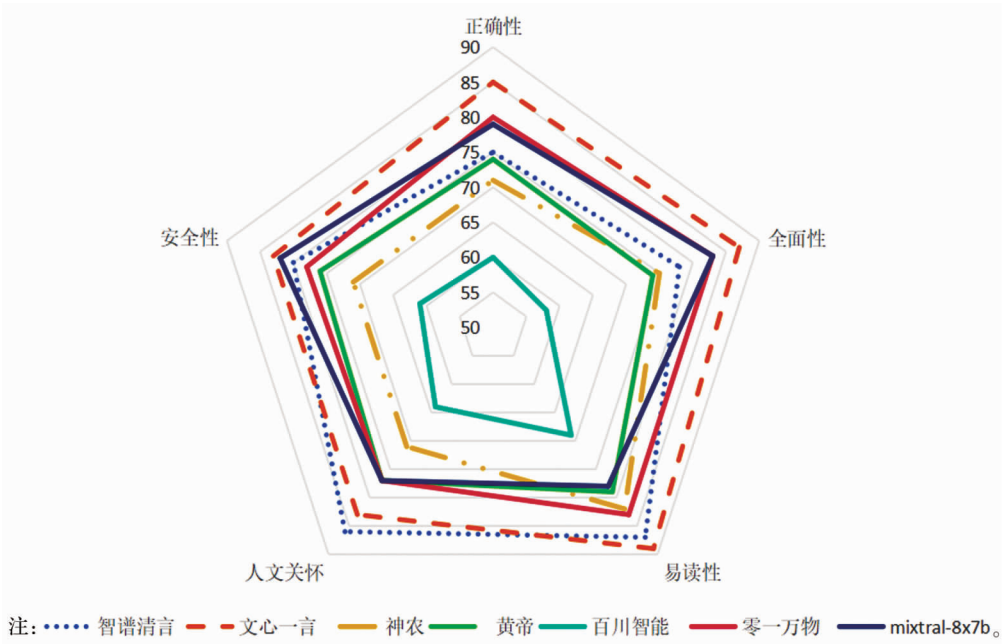
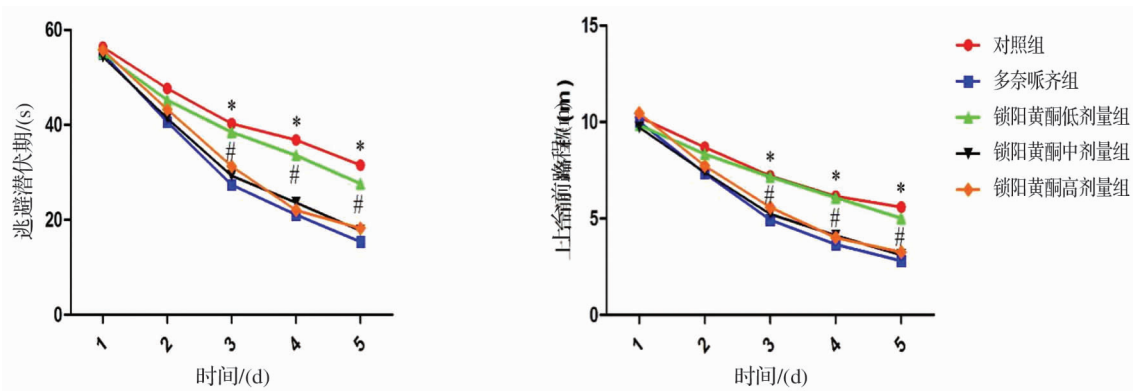



图 5　不同模型的中医药专业临床能力在五维评价体系上的原始得分

## 锁阳黄酮改善 SAMP8 小鼠认知功能障碍的机制研究

（正文见 28－32 页）




注：* 与对照组比较，$P<0.05$；# 与多奈哌齐组比较，$P<0.05$。

图 1　Morris 水迷宫评价锁阳黄酮对快速老化小鼠（SAMP8）空间学习功能的作用