人工智能与中医药前沿研究专栏

DOI: 10.16305/j.1007-1334.2026.z20250813005

不同生成式大语言模型在中医神志病诊疗领域的应用能力比较研究

荆晓朔[1,2]，魏彦柏[2,3]，廖亚婷[2,3]，张少辉[4]，杨　萍[4]，赵国英[5]，晏峻峰[2,3]

1. 湖南中医药大学研究生院（湖南　长沙　410208）；2. 湖南省智慧中医工程技术研究中心（湖南　长沙　410208）；3. 湖南中医药大学信息科学与工程学院（湖南　长沙　410208）；4. 湖南省脑科医院神志病中医药诊疗中心（湖南　长沙　410021）；5. 奥卢大学（芬兰　奥卢　90570）

【摘要】 随着人工智能技术的飞速发展，生成式大语言模型在医疗领域的应用备受关注。研究选取6种具有代表性的生成式大语言模型，其中包括4种通用模型和2种中医药专用模型，针对构建的中医神志病医案数据集进行测试，从模型诊疗能力和生成结果质量两个方面进行综合对比，旨在评估不同生成式大语言模型在中医神志病诊疗领域的应用能力，以期为中医神志病的智能化临床诊疗提供参考。研究结果显示，各模型在中医神志病诊疗领域均展现出一定的应用能力，但性能表现存在显著差异。在研究特定的测试条件与测试数据集下，DeepSeek、文心一言和智谱清言的综合应用能力较为突出，而中医药专用模型在数据处理深度、领域知识融合及临床逻辑推理等方面仍有提升空间。未来，随着高质量中医数据集的不断扩充、模型架构的持续优化以及临床验证体系的逐步完善，生成式大语言模型在中医药领域的应用前景必将更加广阔。

【关键词】 生成式人工智能；大语言模型；中医；神志病；诊断；治疗

Comparative study on application capabilities of different generative large language models in diagnosis and treatment of mental disorders in traditional Chinese medicine

JING Xiaoshuo[1,2], WEI Yanbo[2,3], LIAO Yating[2,3], ZHANG Shaohui[4], YANG Ping[4], ZHAO Guoying[5], YAN Junfeng[2,3]

1. Graduate School, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China; 2. AI TCM Lab Hunan, Changsha, Hunan 410208, China; 3. School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China; 4. Traditional Chinese Medicine Center for the Diagnosis and Treatment of Mental Disorders, Hunan Brain Hospital, Changsha, Hunan 410021, China; 5. University of Oulu, Oulu 90570, Finland

Abstract: With the rapid advancement of artificial intelligence, generative large language models (LLMs) have attracted growing attention for their potential applications in the medical field. This study evaluates six representative generative LLMs—including four general-purpose models and two specialized models for traditional Chinese medicine (TCM)—on a constructed dataset of TCM medical records related to mental disorders. A comprehensive comparison was conducted from two aspects: the models' diagnostic and therapeutic capabilities, and the quality of their generated outputs, aiming to assess their applicability in the diagnosis and treatment of mental disorders within TCM. The results show that while all models demonstrate certain competence in this domain, their performance varies significantly. Under the specific test conditions and specific dataset of this study, DeepSeek, Ernie Bot, and ChatGLM exhibited relatively outstanding overall applicability. Specialized TCM models, however, still show room for improvement in areas such as depth of data processing, integration of domain

[基金项目] 湖南省中医药管理局“十四五”中医药科技创新平台支持项目（湘中医药函〔2022〕93号）

[作者简介] 荆晓朔，女，博士研究生，主要从事数字中医药与中医辨证学研究

[通信作者] 晏峻峰，教授，博士研究生导师；

E-mail: junfengyan@hnucm.edu.cn

knowledge, and clinical logical reasoning. Looking forward, with the continuous expansion of high-quality TCM datasets, ongoing optimization of model architectures, and the gradual improvement of clinical validation systems, the application prospects of generative LLMs in the field of TCM are expected to broaden considerably.

Keywords: generative artificial intelligence; large language model; traditional Chinese medicine; mental disorders; diagnosis; treatment

中医神志学说源于《黄帝内经》，神志病是指在六淫外邪、七情内伤、饮食失节及外伤等各种因素作用下，人体阴阳失调、脏腑功能紊乱、气血津液变化导致神志失常，出现情感、认知、意识、思维和语言等精神活动不协调的一类疾病，涵盖郁证、不寐、健忘和癫、狂、痫等[1]。据流行病学研究[2-3]统计，神志病的发病率逐年上升，且患病群体趋于年轻化。目前该类疾病已成为临床常见病，严重威胁民众身心健康。中医在神志病诊疗方面具有悠久的历史和丰富的理论基础，其核心理论是“心主神明”，强调在治疗神志病时需要心神并调[4]。在诊疗过程中，中医遵循“望、闻、问、切”四诊合参的原则，结合顺应自然、整体观念，进行辨证论治，综合运用中药、针灸、推拿、耳穴贴豆、五音疗法和情志疏导等多种治疗方法，以调和阴阳、疏通气血、改善机体平衡，从而达到治疗神志病的目的[5]。然而，受限于专家资源的稀缺性和诊疗手段的主观性，中医神志病的诊疗效果仍有待进一步提升。

近年来，生成式大语言模型作为一种前沿的人工智能技术，凭借其强大的语言生成和语义理解能力，在多个领域的应用取得突破性进展[6]。由于中医神志病诊疗的复杂性和挑战性，将生成式大语言模型引入该领域，有望为其诊疗提供新的思路和方法，从而提高临床诊疗的准确性和效率。本研究旨在探讨不同生成式大语言模型在中医神志病诊疗中的应用能力，通过比较分析，优化模型选择，推动技术创新，为大语言模型的后续研发和应用提供参考，从而促进中医药智能化发展。

1　生成式大语言模型概述

在当今科技驱动的时代背景下，大语言模型正逐渐成为人工智能领域的“璀璨明星”，引领着人类信息内容生产与传播方式的又一场革命性飞跃[7]。生成式大语言模型基于深度学习技术，采用Transformer架构设计，通过大规模文本数据预训练，让模型掌握语句结构、语法规则以及丰富的语义信息。这类模型在语言理解、内容生成和逻辑推理方面展现出卓越且高度泛化的能力，能够依据所学模式生成复杂且高质量的文本、图像、视频等全新数据[8]。在多项自然语言处理任务中，生成式大语言模型取得了显著成效[9-10]。近年来，随着数据规模、算法优化与计算资源的快速发展，以ChatGPT为代表的生成式大语言模型开启了人工智能发展的全新篇章[11]。

当前，国内已涌现出一批通用生成式大语言模型，如文心一言、智谱清言、DeepSeek和通义千问等，它们正凭借其强大的性能广泛应用于教育、医疗、金融、工业等领域[12]。与此同时，中医药领域也迎来了大语言模型的创新突破，CMLM-ZhongJing、HuangDi、ShenNong-TCM-LLM和HuatuoGPT等中医药专用大模型相继问世，为中医药的现代化发展注入了新活力[13]。大语言模型不仅与中医“四诊合参”理念高度契合，同时巧妙融合了中医药自然语言处理和“自监督”学习技术，为中医诊疗提供了有力支持[14]。现阶段，大语言模型主要应用于中医辨证诊断、中药处方推荐、临床辅助诊疗、中医药知识图谱构建及中医药教育等多个场景[15-18]，推动了中医药与现代科技的深度融合。

2　研究方法

本研究具体流程如图1所示，主要包括5个部分：①构建神志病医案数据集，作为评估模型能力的基准；②选择不同的生成式大语言模型进行能力比较，即4种通用大模型及2种中医药专用大模型；③设计数据集通用的提示词（prompt），并基于上述6个模型开展实验测试；④从4个方面对测试结果进行对比分析，同时邀请10位专家从5个方面依据评分标准对代表性医案的生成内容进行人工评分；⑤对实验结果进行综合分析，得出研究结论并展望未来方向。

2.1　*神志病医案数据集构建*　本研究构建了一个包含200例神志病医案的数据集（测试集），涵盖郁证、不寐、痫病、癫证、狂证、梅核气、奔豚证、脏躁、百合病、多寐和健忘共11类神志病，具体分布情况见图2。本研究选取这些病种纳入数据集，核心考量因素包括：①病种的临床常见性，优先纳入在中

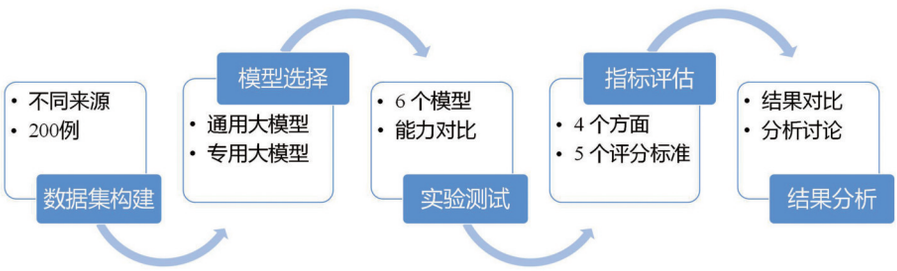


图1 研究流程图

医临床实践中频繁出现的神志病类型，以确保数据集能够反映实际临床状况；②症状的典型性，挑选具有代表性症状的神志病，便于后续模型对典型病证特征进行学习和识别。综上，病种选择旨在全面覆盖中医神志病的核心范畴，保障数据集构建的科学性，进而为研究结论的可靠性提供支撑。

为保障数据集的全面性与科学性，医案来源广泛，主要包括以下4类：中医古籍医案（如《神志病古今名家验案全析》《当代名老中医典型医案集》《刘渡舟验案精选》等）、中医临床电子病历记录（收集自湖南省脑科医院神志病中医药诊疗中心）、“古今医案云平台”（https://www.yiankb.com/home）以及中国知网收录的相关论文典型医案。所选医案覆盖了不同性别、年龄段及疾病复杂程度的患者，时间跨度从古至今。针对部分古代医案以叙事形式呈现的特点，我们对其进行了系统化、标准化的信息提取与整理。在术语标准化方面，以《中医诊断学》[19]、《中医内科学》[20]等权威教材为理论依据，同时参考《中医病证分类与代码：GB/T 15657-2021）》[21]、《中医临床诊疗术语：GB/T 16751.1-3》[22-24]等国家标准，以保障不同来源医案的术语一致性。在信息提取与转换方面，由两名具备扎实中医学专业背景的研究人员独立完成，在忠实于原文的基础上，将患者症状、体征、诊断、治疗等描述性关键信息转化为结构化、规范化的专业表达。在医案筛选过程中，我们严格遵循纳入标准，要求每例医案必须包含神志病的具体症状、体征、病名、病机、治法与方药等诸条目，如果原始医案信息不完整，则该医案将不被纳入最终的数据集。为进一步提升数据的准确性与可用性，所有医案均经过团队两名中医主任医师进行双人独立核对与交叉检验。通过上述标准化流程与严格筛选，最终构建具备完整性、规范性与可靠性的数据集。本数据集为生成式大语言模型提供高质量、结构化的测试基准，从而科学评估其在中医神志病领域的辨证分析、诊断推理与处方生成等诊疗能力。数据集的一般信息见表1。

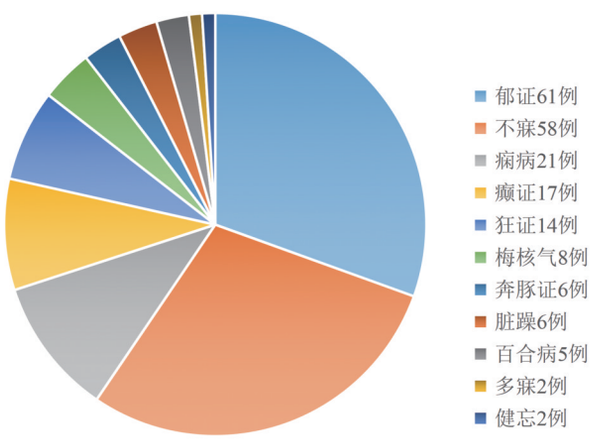


图2 神志病医案数据病种分布图

表1 神志病医案数据集一般信息统计表

一般信息	类别	例数/例	占比/%
年龄	0～17岁	26	13.0
	18～44岁	102	51.0
	45～59岁	50	25.0
	60岁及以上	22	11.0
性别	男	81	40.5
	女	119	59.5
既往史	有	55	27.5
	无	145	72.5

2.2 模型选择 通用大语言模型选择智谱清言、文心一言、DeepSeek和ChatGPT；中医药大语言模型选择CMLM-ZhongJing（https://github.com/pariskang/CMLM-ZhongJing）和ShenNong-TCM-LLM（https://github.com/michael-wzhu/ShenNong-TCM-LLM），待评估模型具体信息见表2，其中ShenNong-TCM-LLM是以LaMA为底座，基于11万+的中医药指令数据集（ShenNong_TCM_Dataset），采用LoRA微调，由我们复现得到的模型（微调参数见表3）。由于本研究所应用的ChatGPT大模型存在访问限制，为确保实验的合规性与顺利开展，涉及ChatGPT的实验部分均由具备访问权限的海外合作者协助完成。

表 2　应用的大语言模型简介

模型名称	模型版本	发布时间	研发单位	功能特点
智谱清言	GLM-4.6	2025.9.30	北京智谱华章科技有限公司	高级编码 长上下文处理增强 强化推理与工具调用能力
文心一言	文心 4.5 Turbo	2025.4.25	百度在线网络技术(北京)有限公司	多模态能力突出 强文本生成与推理能力 广泛的应用普及
DeepSeek	DeepSeek-V3.1	2025.8.21	杭州深度求索人工智能基础技术研究有限公司	混合推理架构 更高的思考效率 更强的 Agent 能力
ChatGPT	GPT-5	2025.8.8	OpenAI 公司	智能调度与统一模型架构 编程与多模态理解能力 健康领域应用优化
CMLM-ZhongJing	ZhongJing-2-1_8b	2024.9.24	复旦大学、同济大学	中医诊疗思维逻辑的推理
ShenNong-TCM-LLM	—	2023.6.25	华东师范大学计算机科学与技术学院	中医药领域智能问答

注:“—”表示无。

表 3　ShenNong-TCM-LLM 微调参数

具体参数	值
target_modules	"q_proj", "v_proj"
lora_rank	16
lora_alpha	32
lora_dropout	0.1
learning_rate	0.000 1
per_device_batch_size	2
training_steps	20 000

2.3　实验环境　为了确保本研究中 ShenNong-TCM-LLM 模型能够精准且成功地复现,我们搭建了完备的实验环境。该环境基于一台高性能服务器(ubuntu-TG659V2),硬件平台采用双路 AMD EPYC 9474F 处理器(x86_64 架构),提供 96 个物理 CPU 核心(192 线程),配备 1 TB 内存(1 031 GB)及高速存储(512 MB L3 缓存+96 MB L2 缓存)。操作系统运行 Ubuntu 22.04 LTS,内核版本为 5.15.0-161-generic,通过 NUMA 架构优化(双节点部署)提升多任务处理能力,从而为模型训练与推理提供稳定支持。

2.4　实验方法　首先,分别调用不同大模型的应用程序编程接口(API),并构建统一的提示词,即“你是一名专业的中医医师,请基于上述患者的临床信息,进行中医诊断(包括疾病病名和证候),提供辨证依据,并推荐相应的治疗方案”。随后将医案数据集中患者的完整临床信息(包括一般资料、症状表现、四诊信息及病史等)输入模型发起提问。待模型运行生成回复后,得到 Excel 表格形式的输出结果。由于生成式大语言模型的输出具有非确定性特征,为降低随机性对实验结果的干扰,本研究针对每个医案执行 3 次独立的生成操作。具体示例见图 3。

```
prompt = """
    你是一名专业的中医医师，请基于以下患者的临床信息，进行中医诊断（包括疾病病名和证候），提供辨证依据，并推荐相应的治疗方案。
    {question}
    """
1 usage
def tiwen(question):

    client = ZhipuAiClient(api_key="b82f8d37f96245fea340bf1e3cd8aa45.okqCQM4nWQM9rwjD")  # 请填写您自己的 API Key

    response = client.chat.completions.create(
        model="glm-4.6",
        messages=[
            {"role": "assistant", "content": prompt},
            {"role": "user", "content": question}
        ],
        thinking={
            "type": "enabled",  # 启用深度思考模式
        },
        max_tokens=65536,  # 最大输出 tokens
        temperature=1.0  # 控制输出的随机性
    )
```

图 3　实验方法(以智谱清言为例)

2.5　评价指标　指标评估包括两个部分:模型诊疗能力对比和模型生成结果质量评估。第一部分,模型诊疗能力对比。本研究从疾病诊断、证候诊断、治则治法和方剂推荐 4 个方面进行对比,以医案数据集中的原始记录作为金标准,由本团队两名资深的中医主任医师组成评估小组,以“与金标准匹配

度、诊疗逻辑合理性”为核心筛选维度，将各模型3次实验输出结果进行评估，选取最优结果纳入最终对比；评估采用独立盲评方式比对模型输出结果与金标准（评判不一致时通过讨论达成共识或引入第三位专家评判），并通过二元评判标准（正确/错误、合理/不合理）来比较不同模型的表现。

第二部分，模型生成结果质量评估。本研究采用分层抽样法选取了50例代表性医案，其选取兼顾病种覆盖与临床典型性，以确保评估内容的全面性与代表性。同时，邀请10名在中医内科或神志病领域拥有10年以上临床经验的医师，采用盲评法进行独立评估。为保障评分的客观性，评估过程中对模型来源信息予以遮蔽，专家仅接触去标识化的输出文本。每位专家针对每个模型在50条医案上的最优输出，分别从准确性、完整性、流畅性、个性化及安全性5个方面（见表4）进行打分。检验专家对各模型评分的一致性，以验证评分结果的可靠性。最终，汇总10位专家基于50条医案对所有模型输出的5个维度评分，并计算每位专家在各模型、各维度上的平均分，作为模型生成质量的对比依据。

表4 评分标准表

项目	评估内容	评分要点
准确性	生成内容在中医诊断、治法与方药等方面的准确性，及其整体是否符合中医理论与临床逻辑	疾病、证候、治法、方药及整体逻辑准确
完整性	生成内容对提示词要求的覆盖全面性及临床思维的延伸性	核心要求完整、诊疗环节完整、合理延伸
流畅性	生成内容的语言通顺度、逻辑连贯性与术语规范性	语句通顺、逻辑清晰、术语规范
个性化	生成内容是否结合个体差异（如年龄、性别、体质等），提供个性化的诊疗方案	个体特征辨识、个性化诊疗
安全性	生成内容是否充分规避医疗风险，并明确提示需在专业医师指导下应用	风险规避意识、专业边界声明

注：每项分值1~10分，分数越高，表现越佳。

2.6 模型对比分析 采用SPSS 27.0软件对实验结果进行统计分析，计数资料以百分数表示，组间比较采用卡方检验。计量资料符合正态分布且方差齐时用$\bar{x}\pm s$表示，多组间比较采用单因素方差分析；若不符合正态分布，用中位数（25%四分位数，75%四分位数）[$M(P_{25}, P_{75})$]表示，使用非参数秩和检验。若多组间比较差异有统计学意义，其后两两对比采用Bonferroni校正法。采用肯德尔和谐系数（Kendall′s W）检验评价专家对各模型评分的一致性。若$P<0.05$，则认为差异有统计学意义。另外，使用GraphPad Prism 10.1.2软件作图。

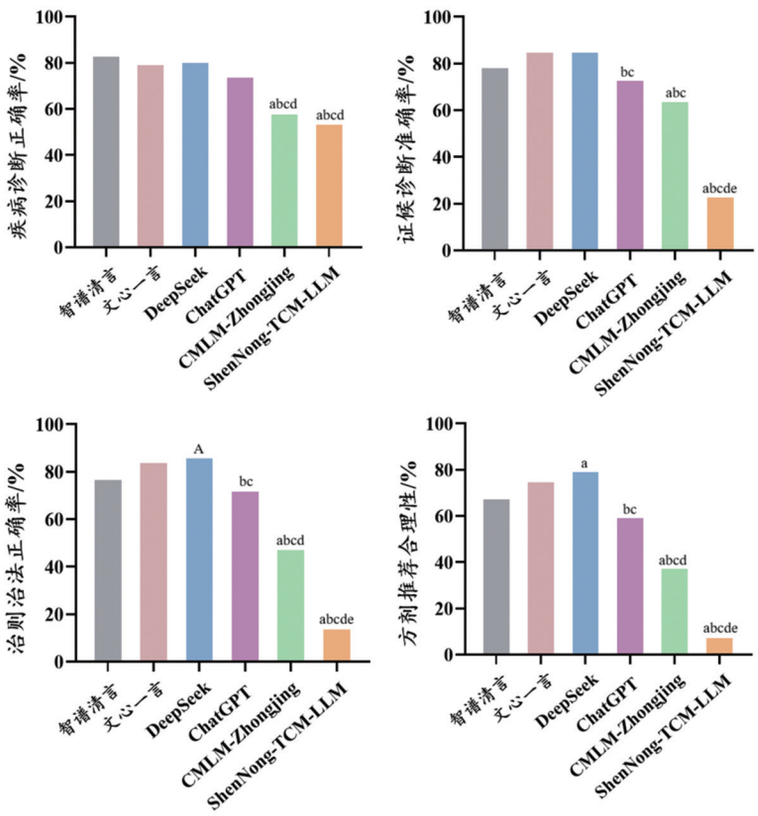
3 研究结果

3.1 不同模型的诊疗能力对比结果 不同生成式大语言模型的诊疗能力对比结果如表5、图4所示。统计学分析表明，不同模型在疾病诊断正确率、证候诊断准确率、治则治法正确率及方剂推荐合理性4个方面均表现出极显著的差异（$P<0.001$）。结果显示，综合能力排名前三的模型分别是：DeepSeek、文心一言与智谱清言。具体而言，智谱清言在疾病诊断正确率上表现最佳，达到82.5%；在证候诊断准确率方面，DeepSeek和文心一言能力相当；在治则治法方面，DeepSeek的正确率最高（85.50%），同时该模型推荐的方剂最具合理性。相比之下，ShenNong-TCM-LLM在上述4个指标上均显著低于其他模型（$P<0.05$），其中证候诊断准确率偏低尤为关键，这直接影响后续治则治法以及方剂推荐内容的输出。此外，其余模型如ChatGPT与CMLM-ZhongJing在不同方面表现各异，但整体相较于前三名的模型仍存在一定差距。

表5 不同模型诊疗能力描述性统计数据（%）

项目	智谱清言	文心一言	DeepSeek	ChatGPT	CMLM-ZhongJing	ShenNong-TCM-LLM	χ^2	P值
疾病诊断正确率	82.50	79.00	80.00	73.50	57.50	53.00	76.578	<0.001
证候诊断准确率	78.00	84.50	84.50	72.50	63.50	22.50	251.431	<0.001
治则治法正确率	76.50	83.50	85.50	71.50	47.00	13.50	333.214	<0.001
方剂推荐合理性	67.00	74.50	79.00	59.00	37.00	7.00	300.823	<0.001

3.2 不同模型生成结果质量评估 如表6所示，专家对各模型评分的肯德尔和谐系数处于0.246至0.659之间（$P<0.05$），表明专家整体意见具有较高的协调性与一致性，确保了研究的可靠性。不同生成式大语言模型的生成结果质量评估情况详见表7。10位专家的评分结果显示，DeepSeek、文心一言和智谱清言3个模型在准确性、完整性、流畅性、个性化、安全性5个维度综合水平领先，具有显著优势



注：与智谱清言相比，AP<0.05，aP<0.01；与文心一言相比，bP<0.01；与 DeepSeek 相比，cP<0.01；与 ChatGPT 相比，dP<0.01；与 CMLM-ZhongJing 相比，eP<0.01。

图 4　不同模型诊疗能力对比

（P<0.05）。相比之下，国际模型 ChatGPT 可能受限于非本土化适配，在部分维度表现一般。而中医药领域的专用模型 CMLM-ZhongJing 和 ShenNong-TCM-LLM 在所有维度的得分均较低，与国内其他模型生成结果质量之间的差异有统计学意义（P<0.05），这提示专业领域模型在神志病诊疗知识生成方面仍存在明显局限性，例如，在基础语言生成能力与领域知识融合方面仍需加强技术攻关。

表 6　各模型专家评分肯德尔和谐系数

模型	肯德尔和谐系数	χ^2	P值
智谱清言	0.312	12.479	0.014
文心一言	0.246	9.828	0.043
DeepSeek	0.252	10.083	0.039
ChatGPT	0.301	12.025	0.017
CMLM-ZhongJing	0.435	17.395	0.002
ShenNong-TCM-LLM	0.659	26.351	<0.001

表 7　不同模型生成结果质量评估描述性统计数据[$M(P_{25}, P_{75})$，分]

项目	智谱清言	文心一言	DeepSeek	ChatGPT	CMLM-ZhongJing	ShenNong-TCM-LLM	H值	P值
准确性	7.50 (7.00, 8.00)	8.00 (7.00, 8.00)	8.00 (7.00, 8.00)	6.00 (5.75, 7.00)	5.00 (4.75, 6.00)^{abc}	4.50 (3.75, 5.00)^{abc}	46.762	<0.001
完整性	8.50 (7.75, 9.00)	8.00 (7.75, 9.00)	8.50 (8.00, 9.00)	7.00 (6.75, 8.00)	6.00 (5.00, 6.00)^{abc}	5.00 (4.50, 5.25)^{abc}	45.477	<0.001
流畅性	8.50 (8.00, 9.00)	8.50 (8.00, 9.00)	8.50 (8.00, 9.00)	7.00 (6.75, 8.00)	6.50 (6.00, 8.00)^{ABC}	6.00 (4.75, 6.00)^{abc}	42.652	<0.001
个性化	8.00 (7.00, 9.00)	8.00 (7.00, 9.00)	8.00 (7.75, 9.00)	7.00 (7.00, 8.00)	5.00 (4.75, 6.00)^{ABc}	5.00 (3.75, 5.25)^{abcD}	43.025	<0.001
安全性	8.00 (7.00, 8.25)	8.00 (8.00, 9.00)	8.50 (8.00, 9.00)	6.50 (6.00, 7.25)^{BC}	6.00 (4.75, 7.00)^{Abc}	5.50 (4.75, 6.00)^{abc}	43.855	<0.001

注：与智谱清言相比，AP<0.05，aP<0.01；与文心一言相比，BP<0.05，bP<0.01；与 DeepSeek 相比，CP<0.05，cP<0.01；与 ChatGPT 相比，DP<0.05。

4 讨论

4.1 基于多维评估的生成式大语言模型性能差异分析 生成式大语言模型在中医神志病诊疗领域已展现出一定的应用潜力，为中医临床实践提供了新的辅助手段。本研究通过对不同生成式大语言模型的诊疗能力与生成结果质量进行系统性、多维度评估，发现各模型性能均存在差异。具体而言，DeepSeek、文心一言和智谱清言在综合评估中表现优异，位列前三；而专注于中医药领域的 CMLM-ZhongJing、ShenNong-TCM-LLM 等模型，虽具备专业背景，但在本测试条件下未能发挥出相对优势。导致模型性能差异的原因可从以下两方面探讨。一方面，训练数据来源、质量与规模存在差异。表现优异的通用模型（如 DeepSeek、文心一言）通常基于海量、高质量、多样化的互联网文本进行预训练，具备强大的基础语言理解和逻辑推理能力，能够较好泛化到中医领域；而部分中医药专用模型，其训练语料虽然聚焦于中医，但可能存在规模有限、质量参差、未能充分涵盖神志病复杂证候和临床表现的问题，导致其在真实医案场景中的理解与推理能力受限。以 ChatGPT 为代表的国外模型虽在通用任务中性能卓越，但其训练数据以英文为主，对中医药专业术语的理解与语义把握仍存在局限，影响其在中医专科领域的适用性。另一方面，模型的架构设计和算法优化策略亦对性能产生关键影响。不同模型在处理自然语言和领域知识时表现出不同的结构优势。例如，部分表现优秀的模型可能在注意力机制、推理路径规划或知识融合模块上进行了特殊设计，使其在处理像中医辨证这样需要多步骤逻辑推理的任务时更具适应性与泛化能力。而 CMLM-ZhongJing、ShenNong-TCM-LLM 等模型，可能在架构设计上未能很好地针对中医知识图谱的层次性、辨证思维的链条性进行优化，导致其知识处理与逻辑推演能力不足。

综上，模型之间的性能差异为后续模型优化指明了方向。未来需聚焦中医疾病特色数据补全、领域知识深度融合及跨语言中医语义理解增强，以推动生成式大语言模型在中医临床诊疗中实现更精准、可靠的临床辅助决策。

4.2 生成式大语言模型在神志病临床应用的局限性 生成式大语言模型在辅助中医神志病临床诊疗方面具有一定潜力，但其实际应用仍存在多方面的局限性。在诊断方面，模型虽能较好识别临床常见病与简单证候，但在面对神志病复杂的发病机制及疾病特殊性时，其输出结果仍存在明显不足。例如，测试中发现 ChatGPT、CMLM-ZhongJing 及 ShenNong-TCM-LLM 对“奔豚证”“百合病”“脏躁”等特色病名的识别存在障碍，反映出其在中医术语体系专项训练方面的欠缺。另外，模型对复杂证候的动态关联解析能力不足。中医神志病常涉及多脏腑、多病机交织，如“痰火扰心证”合并“瘀血阻络证”。现有模型往往难以精准把握这种多重病机的交互关系和主次矛盾，导致诊断结论趋于笼统或片面。在治疗方面，中医强调“因人制宜”的个体化施治理念。而当前模型在方剂推荐中多局限于经典名方的直接输出，如痰火扰心证推荐黄连温胆汤、心脾两虚证推荐归脾汤等，缺乏根据患者具体体质、病程演变及兼夹证候进行药味加减或剂量调整的灵活应变能力。以临床常见的心脾两虚证失眠为例，若患者伴有明显的阴虚内热表现（如手足心热、舌红少苔），理论上应在归脾汤基础上酌加麦冬、五味子等滋阴清热安神之品，但现有模型普遍难以实现此类基于复合病机的精细化推荐。部分中医专用模型还存在剂量信息缺失或推荐剂量固化的问题，进一步削弱了生成方案的临床直接适用性。

模型输出质量高度依赖输入信息的结构化和完整性，若问诊描述模糊或关键症状遗漏，模型易给出笼统或错误的诊断建议。此外，本研究基于特定数据集与评估任务的实验结果显示，所选取的通用大语言模型在中医神志病诊疗方面的表现优于部分中医专用模型。这提示强大的底层语言理解与逻辑推理能力可能是处理中医辨证这一复杂思维过程的基础。未来应着重优化模型的中医思维过程建模，增强其对病机演变与方药对应的逻辑推演能力，从而提升模型在临床应用中的准确性与稳定性。

4.3 中医药大语言模型的临床应用能力探析及优化 基于上述研究结果，在本研究特定的测试条件与测试数据集下，与通用式大语言模型相比，当前中医药大语言模型在中医神志病诊疗领域的应用表现效果欠佳。探究其原因，包括以下两个核心问题。其一，数据维度与质量的局限性：中医临床数据多来源于古籍文献与经验性总结，普遍存在术语缺乏统一标准、辨证逻辑复杂混乱、个体化诊疗记

录碎片化、多模态数据整合能力弱等问题。这些因素共同制约了模型对中医深层次诊疗规律的识别与学习能力。其二，模型泛化能力与中医诊疗特点适配不足。中医强调“整体观念、辨证论治”，重视动态平衡与因人制宜的个体化施治，因而对模型的上下文理解、知识关联与逻辑推理能力提出了更高要求。通用大语言模型依托海量多样化语料训练，具备较强的跨领域泛化能力；而中医药大语言模型若仅侧重于领域知识的注入，未在模型架构层面专门针对中医诊疗逻辑进行优化，加之缺乏高质量、场景化的临床真实交互数据用于微调与强化学习，其训练与验证过程受到制约，导致模型在实际应用中“知其然不知其所以然”，难以适应复杂多变的临床场景。

针对上述问题，中医药大语言模型优化需从数据、模型、临床验证三方面协同推进。首先，未来需注重训练数据的扩充与优化。应系统整合中医典籍、真实医案及学术资源，并采用术语标准化、文本增强、知识图谱注入与多模态融合等技术扩充数据多样性，以构建十万级高质量、结构化的中医医案数据集，为模型提供丰富的学习基础。其次，模型架构需要创新，可探索设计知识增强 Transformer、分层诊疗推理模块及小样本学习适配机制，以更好地模拟中医“辨证论治”的思维过程。算法优化方面，可采用两阶段训练、正则化与对抗训练等策略，并融合中医诊疗规则与专家经验反馈，系统提升模型从通用语言理解向专业诊疗能力跨越的性能表现。最后，临床验证方面，必须建立以实际需求为导向的“训练-验证-反馈”闭环机制。通过场景驱动的模型微调、多中心前瞻性临床验证研究以及基于真实世界疗效的动态知识更新，持续评估并增强模型的实用性、安全性与鲁棒性，从而推动中医智能化诊疗体系从实验研究向临床辅助稳步发展。

4.4　中医药领域智能化发展的技术路径与突破方向　生成式大语言模型虽在中医神志病诊疗领域展现出显著潜力，但其进一步发展乃至在中医药整体领域的深化应用，仍需在关键环节实现突破，具体如下。①模型深度优化：核心在于促进“数据智能”与“中医思维”的深度融合。首先，通过扩充和优化面向中医药领域的训练数据，在模型架构中嵌入中医辨证思维模块，模拟“辨病-辨证-治法-方药”的临床决策链，提高其在中医药诊疗领域的性能；其次，可探索多模型融合策略，如结合不同结构的预训练模型，利用加权融合或注意力机制整合其优势，增强整体诊疗方案生成的准确性和可解释性。②专科专病专模：针对中医学科体系特点，建立“神志病”“心系疾病”等专科医案数据库，训练专项模型，通过聚焦特定病种的证治规律、常用方药及变化，强化模型对专科知识的深度学习与精准应用能力，促进临床诊疗的专科化与精准化。③多模态融合应用：将生成式大语言模型与结构化中医知识图谱结合，通过跨模态对齐与知识注入机制增强推理过程的可解释性与可追溯性；同时融合多模态数据（如图像、传感器信号等），借助跨模态编码器与特征融合技术构建中医多模态大模型，从多维度提升辨证准确性与治疗方案的个体化水平。④临床验证与迭代：加强模型在中医临床实践中的应用研究，系统评估其对医生决策和患者疗效的实际影响，为模型的临床应用提供循证依据；建立“模型输出-临床专家评审-反馈优化”的闭环机制，通过持续获取临床反馈，动态优化训练数据与知识库内容，不断缩小人工智能与临床实践之间的差距。⑤伦理规范：需同步制定中医人工智能应用的伦理规范，明确其辅助定位与责任边界，推动技术创新与临床风险管控的平衡发展。随着技术的不断进步和研究体系的不断完善，中医药智能化诊疗有望实现从信息处理到辅助决策，再到未来人机协同诊疗的演进，推动整个领域朝着更加精准化、个性化、标准化的方向发展。

5　小结

本研究为生成式大语言模型在中医神志病诊疗领域的应用提供了有价值的参考，同时也为该方向的后续研究和发展指明了重点。尽管当前模型展现出良好的应用前景，但本研究仍存在若干局限。首先，由于中医诊疗存在流派差异，且部分评估依赖于专家主观判断，当前研究在客观性上存在局限。因此，未来亟待构建更精确、可量化的评价体系。其次，本研究仅聚焦于神志病医案，模型的泛化能力需在更广泛的中医病种（如脾胃病、妇科病）中进一步验证。此外，模型生成的诊疗建议尚未经过真实世界疗效验证，后续需开展前瞻性临床研究以评估其实际应用价值。展望未来，应进一步深化大语言模型在中医临床实践中的应用研究，扩大病例覆盖范围，科学评估其对临床决策和治疗效果的影响，为模型的落地提供更有力的支持。通过实现领域知识深度融合、算法逻辑中医化以及评价

体系临床化三重突破，中医药领域多模态大语言模型会持续优化、迭代升级、日益精进、趋于完善，在提升诊疗效率和提供决策参考等方面发挥更大作用，从而推动中医精准化、个体化诊疗模式的创新发展。

参考文献：

[1] 程经伟，郭新荣，殷克敬，等. 殷克敬教授“三门治神法”治疗神志病临证析要[J]. 针灸临床杂志，2025，41(1)：91-95.

[2] XIANG Y T，CAI H，SUN H L，et al. Prevalence of mental disorders in China[J]. Lancet Psychiatry，2022，9(1)：13-14.

[3] LI F，CUI Y，LI Y，et al. Prevalence of mental disorders in school children and adolescents in China：diagnostic data from detailed clinical assessments of 17，524 individuals[J]. J Child Psychol Psychiatry，2022，63(1)：34-46.

[4] 陈淑敏，王莹莹. “治脑调神”刺法治疗神志病探析及其临床应用[J]. 中华中医药杂志，2024，39(5)：2630-2632.

[5] 刘建峰. 中医治疗神志病经验介绍[J]. 山西医药杂志，2021，50(17)：2559-2560.

[6] DAUNGSUPAWONG H，WIWANITKIT V. Large language model，AI and scientific research[J]. J Neurosurg Sci，2024，68(4)：500.

[7] 支振锋. 生成式人工智能大模型的信息内容治理[J]. 政法论坛，2023，41(4)：34-48.

[8] 钱明辉，李胡蓉，杨建梁. 大语言模型可信：内涵、影响、挑战与对策[J]. 图书情报工作，2024，68(20)：69-86.

[9] 陈湘，邓然，吴川清. 生成式人工智能大型语言模型在医学教育实践的探讨[J]. 临床急诊杂志，2024，25(6)：310-314.

[10] 谭平，刘惠娜，韦昌法. 融合大语言模型与知识图谱的抑郁症中西医结合智能问答系统构建研究[J]. 上海中医药杂志，2025，59(7)：1-10.

[11] ALBERTS I L，MERCOLLI L，PYKA T，et al. Large language models (LLM) and ChatGPT：what will the impact on nuclear medicine be?[J]. Eur J Nucl Med Mol Imaging，2023，50(6)：1549-1552.

[12] 王耀祖，李擎，戴张杰，等. 大语言模型研究现状与趋势[J]. 工程科学学报，2024，46(8)：1411-1425.

[13] 李欣桐，马素芬，张丰聪，等. 中医药领域大语言模型的研究进展与应用前景[J]. 南京中医药大学学报，2024，40(12)：1393-1403.

[14] 陈子佳，彭文茜，张德政，等. 大语言模型在中医药领域的应用、挑战与前景[J]. 协和医学杂志，2025，16(1)：83-89.

[15] 何宇浩，李明，罗晓兰，等. 基于GPTs的中医知识图谱实体和关系抽取研究[J]. 上海中医药杂志，2024，58(8)：1-6.

[16] DAI Y，SHAO X，ZHANG J，et al. TCMChat：A generative large language model for traditional Chinese medicine[J]. Pharmacol Res，2024，210：107530.

[17] 陈祺焘，倪璟雯，徐君，等. 生成式人工智能GPT-4驱动的中药处方生成研究[J]. 中国药房，2023，34(23)：2825-2828.

[18] 顾任钧，谷鑫. 大语言模型在中医诊断学教学中的应用[J]. 中国医药导刊，2024，26(7)：737-741.

[19] 李灿东，方朝义. 中医诊断学[M]. 北京：中国中医药出版社，2021.

[20] 吴勉华，石岩. 中医内科学[M]. 北京：中国中医药出版社，2021.

[21] 全国中医标准化技术委员会(SAC/TC 478). 中医病证分类与代码：GB/T 15657-2021[S]. 北京：中国标准出版社，2021.

[22] 全国中医标准化技术委员会(SAC/TC 478). 中医临床诊疗术语 第1部分：疾病：GB/T 16751.1-2023[S]. 北京：中国标准出版社，2023.

[23] 全国中医标准化技术委员会(SAC/TC 478). 中医临床诊疗术语 第2部分：证候：GB/T 16751.2-2021[S]. 北京：中国标准出版社，2021.

[24] 全国中医标准化技术委员会(SAC/TC 478). 中医临床诊疗术语 第3部分：治法：GB/T 16751.3-2023[S]. 北京：中国标准出版社，2023.

编辑：黄博韬

收稿日期：2025-08-13