

DOI: 10.16781/j.CN31-2187/R.20250578

·人工智能+医学科研·

国内大语言模型与神经内科医生在轻度认知障碍运动干预问答中的表现对比研究

皋文君^{1,2△}, 杜嘉瑞^{3△}, 于龙娟¹, 张玲娟^{1*}

1. 海军军医大学(第二军医大学)第一附属医院, 上海 200433
2. 海军军医大学(第二军医大学)护理系, 上海 200433
3. 复旦大学附属闵行医院神经外科, 上海 201199

[摘要] **目的** 评估国内主流开源大语言模型(LLM)在轻度认知障碍运动干预医学问答中的表现, 并与神经内科医生的回答进行对比, 以探讨LLM在临床决策支持中的潜在应用价值。**方法** 基于多源数据构建问题库, 生成包含25个轻度认知障碍运动相关问题, 涵盖运动类型(1~8题)、运动方案(9~19题)和运动安全(20~25题)3个维度。首先由12名神经内科医生(初级、中级、副高级和高级职称各3名)对各问题进行独立作答, 然后将每个问题向5种LLM提问3次; 邀请3名资深神经内科医生依据循证最佳证据对各回答结果进行评分, 分析LLM和医生回答的符合率及差异。**结果** 5种LLM中Kimi-K2回答与最佳证据建议的完全符合比例最高(84%, 21/25); 临床医生中, 完全符合比例随临床资历升高而递增, 主任医师最高(96%, 24/25), 其次为副主任医师(88%, 22/25)和主治医师(84%, 21/25)。主任医师回答的总体平均得分高于文心一言X1-Turbo、通义千问-max-latest和DeepSeek-V3.1, 差异均有统计学意义(均 $P<0.05$)。运动方案维度中, 不同临床资历医生和各LLM回答得分差异较大。**结论** LLM在轻度认知障碍运动干预问答中的表现接近低年资医生, 但与高年资医生仍有显著差距, 尤其在运动方案制定方面稳定性不足, 目前尚难以替代高年资医生的临床决策。

[关键词] 人工智能; 大语言模型; 认知功能障碍; 运动疗法; 运动方案

[引用本文] 皋文君, 杜嘉瑞, 于龙娟, 等. 国内大语言模型与神经内科医生在轻度认知障碍运动干预问答中的表现对比研究[J]. 海军军医大学学报, 2026, 47(4): 460-465. DOI: 10.16781/j.CN31-2187/R.20250578.

Performance of domestic large language models and neurologists in question-answering regarding exercise interventions for mild cognitive impairment: a comparative study

GAO Wenjun^{1,2△}, DU Jiarui^{3△}, YU Longjuan¹, ZHANG Lingjuan^{1*}

1. The First Affiliated Hospital of Naval Medical University (Second Military Medical University), Shanghai 200433, China
2. Department of Nursing, Naval Medical University (Second Military Medical University), Shanghai 200433, China
3. Department of Neurosurgery, Minhang Hospital, Fudan University, Shanghai 201199, China

[Abstract] **Objective** To evaluate the performance of mainstream open-source large language models (LLMs) in medical question-answering related to exercise interventions for mild cognitive impairment (MCI), and to compare their answers with those of neurologists, so as to explore the potential application value of LLMs in clinical decision support. **Methods** A question bank was constructed based on multi-source data, generating 25 exercise related questions on MCI, covering 3 dimensions: exercise type (questions 1-8), exercise program (questions 9-19), and exercise safety (questions 20-25). First, 12 neurologists with different professional titles (3 each at junior, intermediate, associate senior, and senior levels) independently answered each question. Then, each question was posed 3 times to 5 LLMs. Three senior neurologists scored the answers according to evidence-based best evidence, and the rates of consistency and differences between the answers from LLMs and physicians were analyzed. **Results** Among the LLMs, Kimi-K2 achieved the highest rate of complete consistency with the best evidence (84%, 21/25). Among clinicians, the rate of complete consistency increased with professional title: chief physicians (96%, 24/25) ranked the highest, followed by associate chief physicians (88%, 22/25) and attending physicians (84%, 21/25). The overall mean score of chief physicians was significantly higher than that of Wenxin Yiyan X1-

[收稿日期] 2025-08-26 [接受日期] 2025-10-30

[基金项目] 海军军医大学护理系登峰人才计划项目(2022KYD07)。Supported by Dengfeng Talent Program of Department of Nursing of Naval Medical University (2022KYD07)。

[作者简介] 皋文君, 博士生, 副教授. E-mail: zerowenjun@163.com; 杜嘉瑞, 硕士生, 副主任医师. E-mail: 280609083@qq.com

△共同第一作者(Co-first authors)

*通信作者(Corresponding author). E-mail: lindazhang_cn@126.com

Turbo, Tongyi Qianwen-max-latest, and DeepSeek-V3.1 (all $P < 0.05$). In the exercise program dimension, performance varied considerably among different LLMs and physicians at different professional levels. **Conclusion** The performance of LLMs is comparable to that of junior physicians in question-answering regarding exercise interventions for MCI, but remains significantly inferior to senior physicians, especially in the consistency of developing exercise program. Currently, LLMs cannot yet replace senior physicians in clinical decision-making.

[**Key words**] artificial intelligence; large language models; cognitive dysfunction; motion therapy; exercise protocols

[**Citation**] GAO W, DU J, YU L, et al. Performance of domestic large language models and neurologists in question-answering regarding exercise interventions for mild cognitive impairment: a comparative study[J]. Acad J Naval Med Univ, 2026, 47(4): 460-465. DOI: 10.16781/j.CN31-2187/R.20250578.

轻度认知障碍 (mild cognitive impairment, MCI) 由于起病隐匿、症状轻微, 常被误认为是正常衰老现象, 导致患者忽视就医^[1-2]。研究表明, 提高患者对 MCI 的认识能显著提升认知障碍的预防效果并降低发病率^[3]。在众多 MCI 非药物干预手段中, 运动干预因其疗效确切、安全性高且成本效益良好, 成为 MCI 患者健康管理的核心策略之一。在此背景下, 向公众及患者提供基于循证证据的精准运动指导是健康教育的重中之重。人工智能 (artificial intelligence, AI) 技术的兴起, 为 MCI 的早期识别及大众健康教育提供了全新可能, 其作为临床实践的有力工具, 正逐渐融入医学教育与临床管理流程。研究表明, 生成式 AI 模型在支持临床决策、优化公共卫生信息传播以及提升患者健康素养方面的有效性已获学界认可^[4]。目前, 以 ChatGPT 为代表的国外 AI 模型已在医疗科普、辅助诊断及患者教育等领域广泛应用^[5-6]。与此同时, 国内也相继推出文心一言、通义千问、智谱清言、DeepSeek 等大语言模型 (large language model, LLM), 并逐步拓展至医疗健康场景。然而, LLM 普遍存在的“幻觉”问题, 使得其生成内容在真实性、全面性及专业性方面存在较大差异, 在医学领域尤为明显^[7]。尽管已有研究对国外 LLM 在医学领域中的表现进行了评估^[8], 但针对国内模型的系统性研究仍相对匮乏^[9]。本研究聚焦 MCI 患者运动干预的常见问题, 系统评估 LLM 在 MCI 运动康复指南依从性方面的表现, 并以不同年资临床医生的回答作为参照进行对比分析, 以明确 LLM 在当前发展阶段的适用场景与核心局限。

1 资料和方法

1.1 问题设计 基于以下多源数据构建问题库:

(1) 复旦大学附属闵行医院记忆门诊患者经脱敏

处理的历史咨询记录; (2) 线上健康社区平台 (好大夫在线和春雨医生) 与 MCI 运动干预相关的用户咨询数据; (3) 课题组前期完成的《轻度认知障碍患者运动干预最佳证据总结报告》(待刊出)。通过整合临床实践数据、公众健康咨询需求及循证证据, 初步构建问题库。然后, 邀请 2 名记忆门诊医生 (神经内科副主任医师及以上职称, 且具有至少 5 年记忆门诊或认知障碍专科门诊临床工作经验), 结合临床经验补充易被忽视的关键问题。最终形成的问题库涵盖运动类型 (1~8 题)、运动方案 (9~19 题) 和运动安全 (20~25 题) 3 个维度。

为降低模型因上下文理解复杂而导致的答案不一致性, 本研究问题设计主要采用高度结构化的闭合型问题, 具体形式包括二选一的是非题 (是/否) 及固定答案格式的填空题 (如“每周 2 次”), 以确保答案简洁、明确, 便于后续统一归类与定量统计。每个问题均直接对应一个明确的证据总结条目, 以便于验证模型输出结果的循证依据。

1.2 LLM 选择 按照以下标准筛选 LLM: (1) 具备医疗健康领域的相关语料训练; (2) 模型开源, 以确保结果可公开发表; (3) 具有对中文医疗问答的优化能力; (4) 支持患者教育场景的交互式问答。最终确定以下 5 种 LLM: 文心一言 X1-Turbo、通义千问-max-latest、智谱 AI-GLM-4.5V、Kimi-K2 和 DeepSeek-V3.1。2025 年 8 月, 对这 5 种 LLM 进行 3 轮独立测试以系统验证其性能。

1.3 研究设计 首先, 邀请 12 名神经内科医生 (初级、中级、副高级和高级职称各 3 名) 对各问题进行独立作答并记录。依据国家卫生健康委员会标准及工作经验对医生职称进行分类: (1) 初级职称, 即住院医师 (取得医学本科或硕士学历, 通过国家执业医师资格考试并注册后, 经住院医师规范化培训); (2) 中级职称, 即主治医师

(已取得医师资格并在神经内科工作3年以上); (3) 副高级职称,即副主任医师(担任神经内科主治医师满3年以上); (4) 高级职称,即主任医师(担任神经内科副主任医师满3年以上)。其次,通过描述性研究探索上述5类LLM能否根据循证医学提供回答,并将LLM与神经内科医生进行比较。最后,邀请3名未参与本研究、具有神经内科副主任医师及以上职称且拥有5年以上记忆门诊临床工作经验的医生,根据前期已完成的最佳证据总结中的明确建议,对每个问题的LLM/医生回答进行独立评判。评判结果为二分类变量,即“一致”(表示模型/医生回答与证据建议相符)或“不一致”(表示不符)。每个问题的最终得分根据3名医生给出“一致”评价的数量确定,具体评分规则如下:3人均评为“一致”,记3分;2人评为“一致”、1人评为“不一致”,记2分;1人评为“一致”、2人评为“不一致”,记1分;3人均评为“不一致”,记0分。

1.4 数据分析 将LLM输出答案与神经内科医生回答进行比较分析,具体包括以下几个层面:

(1) 将LLM回答、神经内科医生回答与现有最佳

证据建议进行比较;(2) 将LLM回答与神经内科医生回答进行比较;(3) 根据问题维度不同,对各维度的平均得分进行对比。

1.5 统计学处理 采用SPSS 26.0软件进行统计分析。计量资料以 $\bar{x} \pm s$ 描述,采用t检验进行数据分析。检验水准(α)为0.05。

2 结果

2.1 最佳证据建议依从性的总体分布 各LLM及不同临床资历神经内科医生对问题库所有问题的回答与最佳证据建议的符合程度存在一定差异。总体而言,在神经内科医生的回答中,完全符合(评分3分)的比例较高,其中主任医师的表现最佳(96%, 24/25),其次为副主任医师(88%, 22/25)和主治医师(84%, 21/25)。在LLM中,Kimi-K2的完全符合比例最高(84%, 21/25); DeepSeek-V3.1未出现不完全符合(评分1~2分)的情况,但其完全不符合(评分0分)的比例相对较高(24%, 6/25)。住院医师的完全符合比例为72% (18/25),与多数LLM相当。见图1。

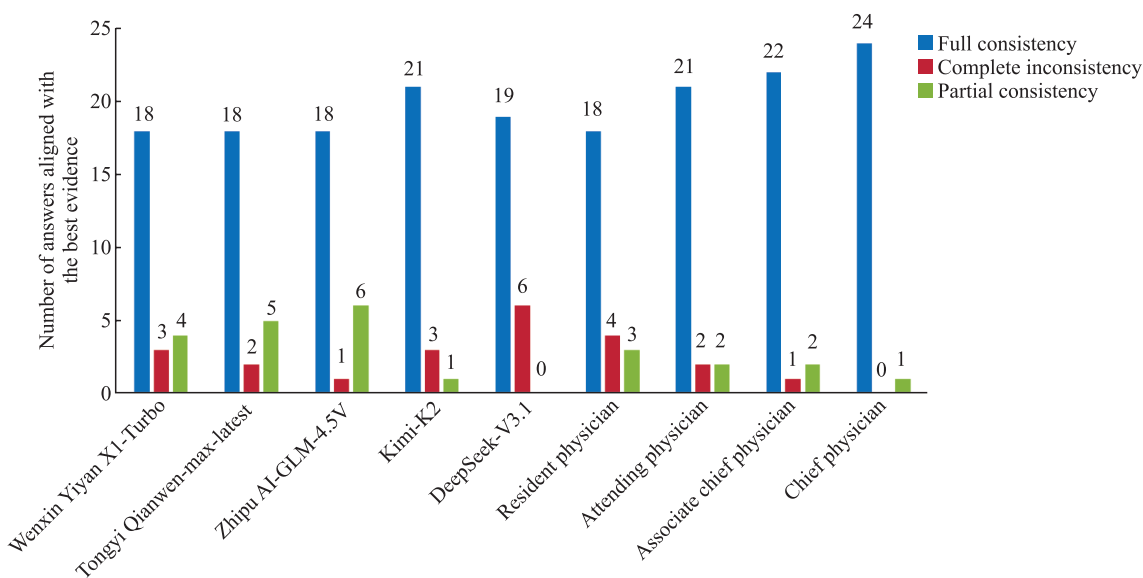


图1 各LLM及不同临床资历神经内科医生对MCI运动干预相关问题的回答与最佳证据建议的符合程度分析
 Fig 1 Analysis of consistency of answers from various LLMs and neurologists at different professional levels to MCI exercise intervention-related questions with best evidence recommendations

LLM: Large language model; MCI: Mild cognitive impairment.

2.2 各LLM与不同临床资历神经内科医生回答的总体得分比较 文心一言X1-Turbo回答的总体平均得分为(2.40±1.08)分,通义千问-max-latest为(2.40±1.04)分,智谱AI-GLM-4.5V为(2.60±0.76)分,

Kimi-K2为(2.60±1.00)分,DeepSeek-V3.1为(2.28±1.31)分;住院医师回答的总体平均得分为(2.36±1.15)分,主治医师为(2.56±1.00)分,副主任医师为(2.72±0.79)分,主任医师为(2.92±0.40)分。

文心一言 X1-Turbo、通义千问-max-latest 和 DeepSeek-V3.1 回答的总体平均得分均低于主任医师,差异均有统计学意义 ($t=-2.32$ 、 -2.32 、 -2.40 , $P=0.03$ 、 0.03 、 0.02),其余 LLM 与不同临床资历医生回答的总体得分比较差异均无统计学意义 (均 $P>0.05$)。

2.3 各 LLM 与不同临床资历神经内科医生对各问题维度回答的平均得分比较 在运动类型 (1~8 题) 维度,不同临床资历神经内科医生和各 LLM 回答

的平均得分两两比较结果均无统计学意义 (均 $P>0.05$)。在运动方案 (9~19 题) 维度,各 LLM 的回答得分范围较广 (2.100~2.600 分),部分模型标准差较大,如 DeepSeek-V3.1 的标准差为 1.449。5 种 LLM 中,对运动方案回答平均得分最高的是 Kimi-K2,其次是智谱 AI-GLM-4.5V。在运动安全 (20~25 题) 维度,LLM 和神经内科医生的回答正确率均达到 100%。见表 1。

表 1 5 种 LLM 与不同临床资历神经内科医生对各 MCI 运动干预相关问题维度回答的平均得分比较

Tab 1 Comparison of average scores of answers from 5 LLMs and neurologists with different professional titles in each question dimension related to exercise interventions for MCI

Group	$\bar{x} \pm s$		
	Exercise type $n=8$	Exercise program $n=11$	Exercise safety $n=6$
Chief physician	2.697±0.633	3.000±0.000	3.000±0.000
Associate chief physician	2.397±1.049	2.800±0.632	3.000±0.000
Attending physician	2.297±1.229	2.500±0.972	3.000±0.000
Resident physician	2.297±1.229	2.100±1.287	3.000±0.000
DeepSeek-V3.1	2.297±1.229	2.100±1.449	3.000±0.000
Kimi-K2	2.597±0.937	2.600±0.966	3.000±0.000
Wenxin Yiyao X1-Turbo	2.497±0.667	2.200±1.317	3.000±0.000
Tongyi Qianwen-max-latest	2.597±0.659	2.100±1.197	3.000±0.000
Zhipu AI-GLM-4.5V	2.697±0.423	2.300±1.059	3.000±0.000

LLM: Large language model; MCI: Mild cognitive impairment.

3 讨论

本研究主要评估了 5 种主流 LLM 与不同临床资历神经内科医生在回答 MCI 运动干预相关问题时对最佳证据建议的遵循程度及回答质量。在最佳证据建议整体依从性方面,各 LLM 及不同临床资历医生的表现存在差异。在所有 LLM 中, Kimi-K2 的最佳证据建议依从性最佳,其回答完全符合最佳证据建议的比例 (84%, 21/25) 接近主治医师水平 (84%, 21/25),但仍低于副主任医师水平 (88%, 22/25) 和主任医师水平 (96%, 24/25)。文心一言 X1-Turbo、通义千问-max-latest 和智谱 AI-GLM-4.5V 的回答完全符合率均为 72% (18/25),与住院医师水平相当。DeepSeek-V3.1 的回答完全符合率 (76%, 19/25) 虽高于多数 LLM,但其完全不符合比例高达 24% (6/25),且未出现不完全符合的情况,表明 DeepSeek-V3.1 在回答相关问题时呈现出“全依从”或“全不依从”的两极分化模式。DeepSeek-V3.1 的这种特殊表现可能与其模型设计和训练策略密切相关,该模型

在训练过程中采用了过于严格的知识筛选机制,导致其倾向于生成它认为确定性高的答案^[10]。对于训练数据覆盖良好、最佳证据建议明确的问题, DeepSeek-V3.1 可以高度依从;而对于存在歧义、证据冲突或训练不足的问题,其可能会直接生成偏离最佳证据建议但自身置信度较高的答案,而非尝试给出一个谨慎、模糊的回答。这反映出当前 LLM 在处理医学不确定性问题方面的一种策略选择,其优势在于答案明确,但劣势是输出错误答案时同样坚决,缺乏人类专家的审慎权衡态度^[11]。

本研究结果显示,临床医生的最佳证据建议依从性随职称提升显著增强,主任医师的完全符合率最高 (96%),无完全不符合的情况发生;同时,随着职称的提升,不完全符合的比例并未增加,表明高年资医生丰富的临床经验有助于避免不完全符合最佳证据建议的情况发生。高年资医生 (副主任/主任医师) 通过大量的临床实践,能更加准确地理解最佳证据建议的适用边界和例外情况,从而做出既符合最佳证据建议原则又契合个体情境的判断^[12]。此外, LLM 的整体表现介于住院

医师与主治医师之间,但均未达到高年资医生的水平。这一差异可能源于临床医生实践经验的积累,低年资医生的临床经验相对有限,而高年资医生经过长期实践,临床经验丰富,这种经验是当前数据驱动的LLM难以完全模拟的。

在所有LLM中,智谱AI-GLM-4.5V和Kimi-K2回答的总体得分较高,接近主治医师水平,且与各职称医师的差异均无统计学意义,提示这2种模型在最佳证据建议遵从方面表现较优,已达到临床医生的中等专业水平。然而,文心一言X1-Turbo、通义千问-max-latest和DeepSeek-V3.1与主任医师的得分差异有统计学意义,说明这3种模型在回答复杂医学问题时仍存在明显局限。各职称医师中,主任医师的回答稳定性最高,进一步说明当前LLM尚难以完全达到高年资医生的可靠水平^[13]。本研究结果提示,在医疗场景中表现较优的模型,如智谱AI-GLM-4.5V和Kimi-K2,或可作为初级职称医师的智能助手,帮助医生快速生成标准化、基于指南的患者教育材料,为低年资医生或全科医生提供符合指南的初步诊疗建议,以辅助其临床决策、缩短学习曲线并减少经验性偏差^[14]。由于LLM在复杂个体化情境中仍与高年资医生存在本质差距,高年资医生应承担最终决策与审核的角色,对LLM或低年资医生提出的建议进行修正与优化^[15]。这也表明未来有必要开发更专业的临床AI工具,以提升AI在真实医疗环境中的实用性和可靠性^[16]。

不同LLM在运动类型和运动方案2个维度的表现差别较大,在运动安全方面,医生和LLM均表现优秀,说明运动完全是临床实践和LLM做出运动康复建议的基础。在运动类型维度中,智谱AI-GLM-4.5V的得分相对最高,其次是通义千问-max-latest和Kimi-K2;运动方案维度中Kimi-K2表现最佳,其次是智谱AI-GLM-4.5V。总体而言,智谱AI-GLM-4.5V和Kimi-K2在各维度中均展现出了较强的理解和应对能力,表明它们在处理MCI患者运动干预问题时均具有较高的适应性和灵活性。研究表明,Kimi-K2的高依从率可能受益于其长文本理解优化,而智谱AI-GLM-4.5V则更专注学术研究辅助^[17]。

综上所述,在制定MCI患者的个体化运动方案时,高年资医生显著优于LLM及低年资医生,突显了临床经验在个体化决策中不可替代的价值。最佳证据建议提供的往往是范围或原则,而如何为

特定患者选择最合适的参数,需要深厚的临床判断力。这正是当前LLM的核心局限,LLM善于处理显性知识,但缺乏隐性的、基于经验的判断力^[18]。本研究结果进一步明确了AI在当前发展阶段中的合理定位:在医疗环境中,LLM可作为辅助工具,为低年资医生提供符合最佳证据建议的初步方案,帮助其缩短学习曲线,减少因经验不足导致的偏差;而高年资医生则应承担最终决策与审核的角色,对LLM或低年资医生提出的建议进行修正与优化^[19-20]。未来研究与应用应致力于构建以“人机协同”为核心的新型诊疗模式,重点探索如何将AI的高效信息处理能力与高年资医生的临床经验深度融合,既充分发挥AI的辅助潜力,又确保医疗决策始终由人类专业判断主导,从而在整体上提升干预质量、诊疗效率及患者满意度。

本研究仍存在一定局限性:(1)结论仅基于MCI运动康复领域最佳证据建议,未涵盖更广泛的临床场景与知识类型,结果外推需谨慎。(2)本研究中设定的评判任务相对结构化,未能充分模拟真实世界中复杂的、个体化的临床情境。因此,尽管部分LLM在本次评测中表现出与住院医师相当的水平,但尚无法全面评估其在复杂临床推理和个体化决策中的表现,也难以量化其与高年资医生在实际诊疗思维上的本质差距。未来研究应聚焦于开发更专业、可靠的临床LLM工具,并重点提升情境化理解与循证决策能力,同时建立标准化评估框架,推动LLM安全、有效地融入临床与医学教育体系^[21]。(3)本研究评估的所有LLM均为公开可用模型,其知识来源于公开的训练数据。然而,在临床实践中,尤其是在神经内科这类专业领域,大量最新的指南、专家共识和医院内部规范可能处于非公开或半公开状态。LLM回答的准确率不仅受其算法影响,更受限于其训练数据中是否包含了回答问题所必需的关键证据。本研究观察到的模型间性能差异,可能部分源于它们训练数据集的覆盖范围和质量不同,而并非单纯的算法能力差异。因此,建议临床专业人员未来可以聚焦于检索增强生成等关键技术的应用,通过构建与实时更新的权威医学知识库的桥梁定向增强LLM的垂直领域专业能力。

[参考文献]

- [1] 中华医学会神经病学分会痴呆与认知障碍学组.阿

- 尔茨海默病源性轻度认知障碍诊疗中国专家共识2024[J]. 中华神经科杂志, 2024, 57(7): 715-737. DOI: 10.3760/cma.j.cn113694-20240320-00172.
- [2] 田萌, 宋玉磊, 张薛晴, 等. 轻度认知障碍病人睡眠特征的潜在剖面分析及影响因素[J]. 护理研究, 2025, 39(7): 1068-1075. DOI: 10.12102/j.issn.1009-6493.2025.07.003.
- [3] EISER A R. Mild cognitive impairment and the missed opportunity to prevent dementia[J]. *Am J Med*, 2024, 137(8): 689-691. DOI: 10.1016/j.amjmed.2024.03.031.
- [4] THEODOSIOU A A, READ R C. Artificial intelligence, machine learning and deep learning: potential resources for the infection clinician[J]. *J Infect*, 2023, 87(4): 287-294. DOI: 10.1016/j.jinf.2023.07.006.
- [5] SHOOL S, ADIMI S, SABOORI AMLESHI R, et al. A systematic review of large language model (LLM) evaluations in clinical medicine[J]. *BMC Med Inform Decis Mak*, 2025, 25(1): 117. DOI: 10.1186/s12911-025-02954-4.
- [6] KAMMINGA N C W, KIEVITS J E C, PLAISIER P W, et al. Do large language model chatbots perform better than established patient information resources in answering patient questions? A comparative study on melanoma[J]. *Br J Dermatol*, 2025, 192(2): 306-315. DOI: 10.1093/bjd/ljae377.
- [7] 刘泽垣, 王鹏江, 宋晓斌, 等. 大语言模型的幻觉问题研究综述[J]. 软件学报, 2025, 36(3): 1152-1185. DOI: 10.13328/j.cnki.jos.007242.
- [8] ZHAO C Y, SONG C, JIANG C Y, et al. Large language models (LLMs) in tuberculosis patients' health consultation and patient education: a comprehensive performance analysis study[J]. *J Infect*, 2025, 91(1): 106527. DOI: 10.1016/j.jinf.2025.106527.
- [9] 邢倩, 何达. 医疗大语言模型的评价现状及思考[J]. 健康发展与政策研究, 2025, 28(1): 65-72, 79. DOI: 10.12458/HDPR.202407099.
- [10] 掌上张寨. 国产大语言模型深度对比: 豆包、通义千文、DeepSeek、智谱清言、Kimi 与文心一言[EB/OL]. (2025-05-19) [2025-08-18]. <https://mp.weixin.qq.com/s/aNPi9zj67yiDfwalN35QQ>.
- [11] KOGA S. Exploring the pitfalls of large language models: inconsistency and inaccuracy in answering pathology board examination-style questions[J]. *Pathol Int*, 2023, 73(12): 618-620. DOI: 10.1111/pin.13382.
- [12] BALDUCCI L, COHEN H J, ENGSTROM P F, et al. Senior adult oncology clinical practice guidelines in oncology[J]. *J Natl Compr Canc Netw*, 2005, 3(4): 572-590. DOI: 10.6004/jnccn.2005.0032.
- [13] HAGER P, JUNGMANN F, HOLLAND R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making[J]. *Nat Med*, 2024, 30(9): 2613-2622. DOI: 10.1038/s41591-024-03097-1.
- [14] VRDOLJAK J, BOBAN Z, VILOVIĆ M, et al. A review of large language models in medical education, clinical decision support, and healthcare administration[J]. *Healthcare*, 2025, 13(6): 603. DOI: 10.3390/healthcare13060603.
- [15] KANEDA Y. In the era of prominent AI, what role will physicians be expected to play?[J]. *QJM*, 2023, 116(10): 881. DOI: 10.1093/qjmed/hcad099.
- [16] AGHARIA S, SZATKOWSKI J, FRAVAL A, et al. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: an analysis of ChatGPT 3.5, ChatGPT 4, and Bard[J]. *J Orthop*, 2024, 50: 1-7. DOI: 10.1016/j.jor.2023.11.063.
- [17] Science Assistant. 智谱清言: AI对话的新选择, 与DeepSeek、ChatGPT的差异化优势[EB/OL]. (2025-04-15) [2025-08-18] https://mp.weixin.qq.com/s/Xl8mW_i4L-LGVCLz8WIGAQ.
- [18] VENKATASUBRAMANIAN V. Do large language models "understand" their knowledge?[J]. *AIChE J*, 2025, 71(3): e18661. DOI: 10.1002/aic.18661.
- [19] 李源, 罗碧如, FU MEI R, 等. 大语言模型在临床护理实践的潜在应用及障碍分析[J]. 护理学报, 2024, 31(21): 44-48. DOI: 10.16460/j.issn1008-9969.2024.21.044.
- [20] WANG T, MU J, CHEN J, et al. Comparing ChatGPT and clinical nurses' performances on tracheostomy care: a cross-sectional study[J]. *Int J Nurs Stud Adv*, 2024, 6: 100181. DOI: 10.1016/j.ijnsa.2024.100181.
- [21] LIU J, WANG C, LIU S. Applications of large language models in clinical practice: path, challenges, and future perspectives[EB/OL]. (2024-08-21) [2025-08-18]. <https://doi.org/10.31219/osf.io/82bjd>.

[本文编辑] 杨亚红