大语言模型在医学多项选择题命题中的应用进展

李珂[1]　闫辉[2]　金哲[3]　佘康[4]　齐建光[1,2]
[1]北京大学第一医院教育处，北京 100034；[2]北京大学第一医院儿童医学中心，北京 100034；[3]北京大学第一医院呼吸与危重症医学科，北京 100034；[4]北京大学第一医院介入血管外科，北京 100034
通信作者：齐建光，Email：qjg2006@126.com

【摘要】 使用大语言模型（large language model，LLM）生成医学多项选择题（multiple-choice question，MCQ）已成为国外医学教育评价领域的研究热点之一。本文对LLM用于MCQ命题的应用现状、影响因素及评价方法进行综述。现有研究显示，LLM在缩短命题周期、生成结构规范的题目初稿方面优势明显，部分题目在内容准确性与拟真度上已接近专家命题水平；但在高阶临床推理考查、事实准确性、干扰项质量及心理测量学属性稳定性等方面仍存在局限。提示词设计、模型选择以及外部权威知识源的引入，均为影响生成题目质量的重要因素。在此基础上，本文提出了AI命题提示词框架和初步的AI命题质量评估框架，并提出专家与AI协同命题的模式，以期为LLM辅助医学MCQ命题提供参考。

【关键词】 大语言模型；　多项选择题；　命题；　医学教育；　提示词
【中图分类号】 R-0；R-4；R78
基金项目：2025–2026年度国家医学考试中心创新性研究型项目（基于生成式AI的医师资格考试高质量命题模式与评价体系研究）
DOI：10.3760/cma.j.cn116021-20260221-02386

Progress in the application of large language models to medical multiple-choice question generation

Li Ke[1], Yan Hui[2], Jin Zhe[3], She Kang[4], Qi Jianguang[1,2]
[1]Education Department, Peking University First Hospital, Beijing 100034, China; [2]Children's Medical Center, Peking University First Hospital, Beijing 100034, China; [3]Department of Respiratory and Critical Care Medicine, Peking University First Hospital, Beijing 100034, China; [4]Department of Interventional and Vascular Surgery, Peking University First Hospital, Beijing 100034, China
Corresponding author: Qi Jianguang, Email: qjg2006@126.com

【Abstract】 The application of large language models (LLMs) to generate medical multiple-choice questions (MCQs) has become one of the research hotspots in medical education assessment abroad. This article reviews the current application status, influencing factors, and evaluation methods of using LLMs for MCQ generation. Existing studies show that LLMs have clear advantages in shortening the item-writing cycle and generating structurally standardized first drafts of questions, and some generated questions have approached the level of expert-written ones in terms of content accuracy and authenticity. However, limitations still remain in the assessment of higher-order clinical reasoning, factual accuracy, distractor quality, and the stability of psychometric properties. Prompt design, model selection, and the introduction of external authoritative knowledge sources are all important factors affecting the quality of generated questions. On this basis, this paper proposes an artificial intelligence prompt framework for question generation and a preliminary quality evaluation framework for artificial intelligence-generated questions, along with a model of expert-artificial intelligence collaborative item writing, in order to provide a reference for LLM-assisted medical MCQ generation.

【Key words】 Large language model; Multiple-choice question; Item writing; Medical education; Prompt

Fund program: 2025-2026 National Medical Examination Center Innovative Research Project: Research on High-Quality Item Generation and Evaluation Systems for Medical Licensing Examinations Based on Generative AI

DOI: 10.3760/cma.j.cn116021-20260221-02386

多项选择题（multiple-choice question，MCQ）具有评分客观、便于标准化实施、能够在大规模考试中覆盖广泛知识内容等特点[1-2]。多项选择题常用于本科教学期末考试、执业医师资格考试和住院医师规范化培训结业理论考试等，可以考查学习者的知识掌握程度，评价临床推理、诊断决策和处理原则等综合能力。在我国，MCQ是执业医师资格考试以及研究生入学考试中的西医综合能力考试等高风险考试的主要题型[3]。目前的人工命题方式多依赖专家经验，面临开发周期长、耗费人力资源、题目更新慢等问题[4-5]。近年来出现的大语言模型（large language model，LLM）擅长自然语言处理与结构化文本输出，已在医学教育评价领域得到应用，并用于MCQ命题工作[6-7]。医学考试往往涉及高风险决策，在题目的内容准确性、临床情境的真实性、认知层次的适配程度以及伦理合规性等方面具有较高要求[3]。因此，将LLM用于MCQ命题，需要全面评估生成题目的质量、适用范围以及人类专家与模型的分工协作方式。国外的相关研究进展较快，已涉及本科教育、住院医师与专科医师培训等多种场景，国内则刚刚起步。本文综述LLM在MCQ命题中的研究进展，分析其应用场景、初步效果、提示词策略对题目质量的影响、引入外部权威知识源的技术路径、AI生成题目的质量评价方法以及人机协同命题路径，以期为利用LLM辅助命题提供参考。

1 LLM在MCQ生成领域的应用场景及初步效果

1.1 应用场景广泛，覆盖多层次与多专科

LLM在多个培训层次和学科领域被用于MCQ命题，包括本科阶段的基础解剖学[8]、牙周病学[9]，住院医师培训阶段的病理学[10]、眼科学[11]、放射学[7, 12]与妇产科学[13]，以及专科医师培训阶段的重症医学[14]、急诊医学[15]等。

1.2 在命题效率与基础质量方面显示出初步优势

在命题效率方面，LLM能够有效节约时间[16-18]。Cheung等[17]发现，AI生成50道非病例型MCQ仅需0.34 h，明显低于人类专家所需的3.5 h。Safadi等[14]发现，AI生成15道MCQ仅需1 h，而人工命题时约需10 h。在基础质量方面，部分研究显示，在题目结构及题目临床情境设置等维度上，AI题目已接近人类命题水平[2, 9, 19-21]。Ahmed等[20]发现，69.00%的AI题目无需修改或仅需少量修改即可使用。Karahan和Emekli[19]的研究显示，72.50%的AI题目可直接用于教学。在心理测量属性方面，部分研究显示，AI题目的难度和区分度已经接近人类命题水平[13, 18, 20]。

1.3 命题效果存在一定的局限性

在题目考查的认知层次方面，LLM命题存在偏重考查知识记忆、缺少对临床思维等高阶能力考查的问题。例如，Karahan和Emekli[19]提出，高达79.31%的MCQ在考查知识记忆。同时，有研究者提出，LLM生成复杂MCQ题目的能力不足[22]，这可能与AI现阶段无法进行临床推理有关[15]。

在内容准确性方面，AI题目会出现内容错误、内容偏离等问题。Law等[15]发现，AI组的内容错误比例为6.00%，高于人类专家组的4.00%。Ahmed等[20]发现，31.00%的AI题目因内容错误或偏离学习目标在初筛阶段被剔除。在难度与区分度方面，AI题目可能整体偏易（难度指数为答对该题的考生比例，取值0~1，数值越大表示题目越容易）。Ma等[9]发现，AI试卷整体难度指数为0.791，高于人工试卷的0.725；简单题占比由30%升至40%，AI组学生平均分为79.34分，明显高于人工试卷的73.17分。Law等[15]同样发现，AI题目平均难度指数高于人工题目（0.78比0.69，$P<0.01$）。鉴于目前命题效果的局限性，命题人员应避免依赖人工智能技术命制MCQ，而应通过人机协作的形式开展命题工作。

2 影响MCQ生成质量的因素

2.1 LLM的选择

不同LLM之间存在差异，主要涉及答案准确性、难度调控与内容质量。在答案准确性方面，

Mistry 等[12]发现，Llama 2 的答案准确率仅为69.00%，而 GPT-4 为 100.00%。在难度调控方面，Karahan 和 Emekli[19]发现，Gemini 生成的题目较易（难度指数为0.70），Claude 3 生成的题目偏难（难度指数为0.67），同时全部 AI 生成题目的平均难度指数为 0.62。在内容质量方面，Abouzeid 等[22]指出，Bing 的表现最差，其基于通用一致法的量表水平内容效度指数（scale-level content validity index, universal agreement method，S-CVI/UA）仅为0.83，低于 ChatGPT-4 和 Gemini 的0.91 及 ChatGPT-3.5 的0.90。

目前，MCQ 命题研究多使用 ChatGPT、Gemini、Claude、Bing 等通用模型[1-24]，尚未见医学专用 LLM 用于 MCQ 命题并评价题目质量的研究报道。现有医学专用模型（如 Med-PaLM[25-26]、PMC-LLaMA[27]、Me-LLaMA[28]等）多集中于医学问答与文本处理。模型答题表现好，并不能直接推断其具备命题能力；命题除确保知识准确外，还涉及学习目标对齐、临床情境设计、干扰项构造、认知层次控制及心理测量学验证等环节。未来的研究中，应比较通用 LLM 与医学专用 LLM 的命题效果。

2.2　提示词的重要作用及组成要素

提示词（prompt）质量在很大程度上决定了 LLM 的输出质量[5]。宽泛的指令可能导致结果不稳定，而结构清晰、内容具体的提示词更易于产出高质量 MCQ。提示词通常包括明确学习目标、规定题目结构、设定角色、提供参考资料以及答案解析要求等[2,5]。为梳理 MCQ 命题提示词的组成，本文从已纳入的文献中选取 16 篇明确报告了提示词内容与结构的研究进行对比分析，详见表1。

由表1可见，题目类型（16/16）、答案要求（13/16）与问题结构/格式（13/16）出现频率最高，构成提示词的骨架；核心主题/知识点（11/16）、解释/理由要求（8/16）与知识来源/范围（8/16）等内容性要素也被频繁纳入；角色/身份设定（2/16）、难度/认知水平（5/16）与高级指令/链式提示（4/16）等优化策略尚处于探索阶段。除组成要素外，提示词的组织逻辑同样影响质量。理想的提示词并非简单罗列要求，而应按照先界定目标，再约束知识、规范题型、控制认知层次、限定输出格式，最后自检迭代的逻辑展开。江哲涵等[3]比较了不同提示词设计方式，结果表明结构化、分步骤的提示词更利于提升题目的规范性与可用性。

Karahan 和 Emekli[19]主张细化提示词，如规定题干含入院原因、病史、放射学检查与问题主干4个部分，并按专科特点定制；Elzayyat 等[8]在提示词中标明考试层级与属性，使内容与美国医学执照考试（United States Medical Licensing Examination, USMLE）及国家医学考试委员会（National Board of

表1　MCQ命题提示词组成要素分布

作者	角色/身份设定	目标受众	知识来源/范围	核心主题/知识点	学习目标	问题结构/格式	题目类型	题目数量	难度/认知水平	答案要求	解释/理由要求	示例参考	负向指令/排除项	输出格式/规范	高级指令/链式提示
Safadi 等[14]		√			√	√	√		√	√	√				
Cheung 等[17]		√	√	√		√	√		√	√					
Laohawetwanit 等[10]				√			√						√		√
Karahan 和 Emekli [19]	√	√	√	√		√	√		√	√	√		√	√	
Nielsen 等[6]				√			√	√	√	√			√	√	
Emekli 和 Karahan[7]						√	√			√					
Wu 等[21]		√		√	√	√	√			√	√	√	√		√
Jackson 等[13]		√	√	√		√	√	√		√		√	√	√	√
Law 等[15]			√	√	√	√	√					√			
Ahmed 等[20]					√	√	√	√		√	√				
Ma 等[9]			√	√			√		√	√		√			
Gholami 等[11]			√	√		√	√	√				√		√	
Abouzeid 等[22]	√	√		√	√	√	√			√	√			√	
Mistry 等[12]			√			√	√	√		√	√	√			
Elzayyat 等[8]					√	√	√	√		√	√				
Sawamura 等[24]		√	√	√		√	√			√	√		√	√	√

注：“√”表示该研究明确使用或报告了相应提示词要素，空白表示文献中未明确报告此要素。

Medical Examiners, NBME)要求对齐;有研究[16]通过生成5个美国皮肤科应用考试(American Board of Dermatology Applied Examination, ABD-AE)风格的提示词框定考试范式;Sawamura等[24]引入选项随机化以增加变体、避免重复。上述研究提示,AI命题提示词正朝着结构化、精细化与场景化方向发展。基于上述研究,本文归纳提出了AI命题提示词框架(表2)。

2.3 结合外部权威知识

为解决LLM自身知识的局限性并降低幻觉风险,部分研究采用检索增强生成(retrieval-augmented generation, RAG)技术,即从指定的外部权威来源检索相关信息,再据此生成题目,以提升内容的准确性与教学适用性。教科书、临床指南、历年题库或官方命题指南均可作为外部知识来源。Sawamura等[24]从历年物理治疗师国家执照考试真题中提取340道MCQ构建知识库,其所生成题目的答案正确率为100.00%。Jackson等[13]使用美国妇产科医师学会(American College of Obstetricians and Gynecologists, ACOG)官方参考书 *PROLOG: Obstetrics* 的部分章节生成MCQ,所生成题目的结构与内容相关性接近专家水平。Karahan和Emekli [19]以讲义与笔记为依据生成题目,确保题目与课程内容及认知层次相匹配。Gholami等[11]使用美国眼科学会的教材 *Basic and Clinical Science Course* 及官方命题指南,生成了11个亚专科题目,结果显示其在适当性、清晰性、临床相关性、区分度及学员适配度5个维度上均与专家命题相当。引入外部权威知识源(如RAG技术)的方法在生成时调用教材、指南、题库或大纲等权威资料,强调内容与特定来源的锚定及时效性;医学专用LLM则通过医学语料预训练、微调或领域适配来增强模型自身的医学理解能力。当需严格对齐大纲或最新指南时,RAG更具优势;面对高度专科化任务时,则可借助医学专用LLM。

3 LLM生成MCQ的质量评价

现有研究显示,AI命题的质量评价指标如表3所示。

表3显示,文献报道中的AI命题的质量评价指标分布呈现一定规律。高频出现的指标包括:题干清晰完整(14/16)、难度适宜性(13/16)与干扰项质量(10/16),提示此部分已初步达成共识。次高频出现的指标为:临床相关性(9/16)、难度(8/16)、区分度(8/16)、内容准确性(7/16)、答案解释充分性(7/16)、正确答案准确性(6/16)及教学目标一致性(5/16)。低频出现的指标包括:教育启发性(3/16)、原创性(2/16)与来源鉴别测试(2/16)。现有研究显示,AI命题的质量评价应是一种客观化、多维度组合、来源于实证数据的综合评价方式。基于此,本文初步提出了AI命题的质量评价指标框架(表4)。

表2 AI命题提示词框架

分类	关键要素	功能描述	提示词示例
目标界定	学习目标	对齐教学大纲及课程目标	题目应对应学习目标:学生能够根据胸痛患者的病史、心电图和肌钙蛋白结果进行初步诊断
内容约束	知识来源/范围	锚定于特定、权威知识来源	依据《内科学》中“慢性阻塞性肺疾病”章节内容命题,不得使用超出教材范围的信息
	核心主题/知识点	聚焦预定考查内容	围绕“急性心肌梗死的诊断与初始处理”生成题目;
结构格式	问题结构/格式	明确题干组成	采用临床案例题形式,题干包括患者年龄、性别、主诉、病史、体格检查和辅助检查结果;
	题目类型	明确题型	生成最佳单项选择题;
	答案要求	明确选项数量、干扰项特征	每题设置5个选项,其中1个正确答案和4个具有临床迷惑性的干扰项;
	输出格式/规范	明确输出细节	按“题干、选项、正确答案、答案解析、干扰项解析”的格式输出
情境设定	角色/身份设定	限定特定专业与视角	以具有10年命题经验的心血管内科临床教师身份进行命题;
	目标受众	校准题目难度与语言风格	题目对象为临床医学专业五年级本科生,难度相当于毕业综合考试水平;
	难度/认知水平	要求达到特定难度水平	生成考查应用分析题目,避免单纯考查记忆性问题
高级优化	解释/理由要求	校对答案,补充学习资源	说明正确答案为什么最佳,并逐一解释每个干扰项;
	示例参考	提供范例	参考以下示例题目的结构和语言风格;
	负向指令/排除项	规避常见命题缺陷	不要出现“哪项是错误的”等否定式表述;
	高级指令/链式提示	通过多步指令或复杂逻辑实现精细控制	先生成初稿,再检查是否存在事实错误、线索词、选项长度不均和干扰项不合理,并给出修订版;
	题目数量	明确批量生成要求	一次生成10道题,每道题考查不同知识点

表3　AI命题的质量评价指标

作者	内容质量		结构形式				教学价值			测量属性		其他	
	内容准确性	临床相关性	题干清晰完整	正确答案准确性	干扰项质量	答案解释充分性	教育启发性	教学目标一致性	难度适宜性	难度	区分度	来源鉴别测试	原创性
Safadi 等[14]			√	√	√			√	√			√	
Laohawetwanit 等[10]	√	√	√		√	√	√			√	√		
Karahan 和 Emekli[19]	√	√							√	√	√		√
Nielsen 等[6]		√	√		√	√			√				
Emekli 和 Karahan[7]			√		√			√	√	√	√		
Wu 等[21]	√	√	√	√	√	√			√				
Jackson 等[13]										√	√	√	
Magzoub 等[5]			√		√			√	√	√			
Law 等[15]	√	√	√						√	√	√		
Ahmed 等[20]	√		√	√				√		√	√		
Ma 等[9]			√				√	√	√	√	√		
Gholami 等[11]		√	√						√		√		√
Abouzeid 等[22]	√		√	√	√	√			√				
Mistry 等[12]		√	√	√	√	√			√				
Elzayyat 等[8]	√	√	√		√	√	√		√				
Sawamura 等[24]		√	√	√	√	√			√				

注：“√”表示该文献报道使用或报告了相应评价指标，空白表示文献中未明确报告此要素。

表4　AI命题的质量评价指标框架

领域	维度
题目内容质量	题目内容准确性
	题目临床相关性
题目结构形式	题干清晰完整
	正确答案准确性
	干扰项质量
题目教学价值	答案解释充分性
	题目教学目标一致性
	题目认知层次适配性
题目测量属性	题目难度指数
	题目区分度指数
其他	题目来源鉴别
	题目原创性

4　小结

LLM用于MCQ命题具有可行性，能够提升命题效率。通过结构化的提示词并结合外部权威知识库，能够命制出一批质量良好的题目。但是目前的LLM仍有一些局限性，例如考查层次的认知深度不足、内容准确性存在风险等。因此，LLM不是命题专家的替代者，而是辅助工具之一。综合考虑LLM命题的质量和技术成熟度，AI生成的题目不应直接使用，而应进行与人工命题相同甚至更为严格的审核。未来研究者可以尝试在课堂练习、自主练习等低利害场景中进行试点，再在高利害考试中经过严格审核后引入使用。

随着国内人工智能技术日趋成熟，已有少量报道介绍了AI命制MCQ的经验，但仍缺少深度研究。因此，未来研究者应关注LLM辅助命题这一热点问题，开展相关研究与探索。国内医学教育研究人员有必要开发MCQ命题提示词模板，便于命题专家使用与参考。同时，应构建AI命题的质量评价指标，为审题专家提供评判标准。最终，形成一套人机协同的命题流程，服务于一线临床教师。此外，应积极开展相关师资培训，帮助临床教师全面认识LLM命题的优势与不足，介绍相关技术方案与路径，进而减轻命题工作负荷并提高命题质量。

利益冲突　所有作者声明不存在利益冲突

作者贡献声明　李珂：项目设计、文献检索、撰写论文；闫辉、金哲、佘康：资料整理、论文修订；齐建光：项目指导

参考文献

[1] 曾令铝，方天乐，李婧，等. 发达国家及地区国际医学毕业生参加临床执业医师资格考试的比较研究[J]. 循证医学，2025，25(3)：181-190. DOI: 10.12019/j.issn.1671-5144.202412058.

[2] Stadler M, Horrer A, Fischer MR. Crafting medical MCQs with generative AI: a how-to guide on leveraging ChatGPT [J]. GMS J Med Educ, 2024, 41(2): Doc20. DOI: 10.3205/zma001675.

[3] 江哲涵, 奉世聪, 王维民. 生成式大语言模型在医学考试题库建设中的实践探索[J]. 中华医学教育杂志, 2024, 44(8): 561-569. DOI: 10.3760/cma.j.cn115259-20240524-00520.

[4] 李丽, 朱朝阳, 李冬梅, 等. 智能题库模型quizLlama的构建与教学应用[J]. 基础医学教育, 2025, 27(7): 674-679. DOI: 10.13754/j.issn2095-1450.2025.07.16.

[5] Magzoub ME, Zafar I, Munshi F, et al. Ten tips to harnessing generative AI for high-quality MCQS in medical education assessment [J]. Med Educ Online, 2025, 30(1): 2532682. DOI: 10.1080/10872981.2025.2532682.

[6] Nielsen JPS, Mikkelsen AK, Kuenzel J, et al. Evaluation of multiple-choice tests in head and neck ultrasound created by physicians and large language models [J]. Diagnostics (Basel), 2025, 15(15): 1848. DOI: 10.3390/diagnostics15151848.

[7] Emekli E, Karahan BN. Artificial intelligence in radiology examinations: a psychometric comparison of question generation methods [J]. Diagn Interv Radiol, 2026, 32(5): 548-554. DOI: 10.4274/dir.2025.253407.

[8] Elzayyat M, Mohammad JN, Zaqout S. Assessing LLM-generated *vs*. expert-created clinical anatomy MCQs: a student perception-based comparative study in medical education [J]. Med Educ Online, 2025, 30: 2554678. DOI: 10.1080/10872981.2025.2554678.

[9] Ma X, Pan W, Yu XN. Evaluating AI-generated examination papers in periodontology: a comparative study with human-designed counterparts [J]. BMC Med Educ, 2025, 25(1): 1099. DOI: 10.1186/s12909-025-07706-6.

[10] Laohawetwanit T, Apornvirat S, Kantasiripitak C. ChatGPT as a teaching tool: Preparing pathology residents for board examination with AI-generated digestive system pathology tests [J]. Am J Clin Pathol, 2024, 162(5): 471-479. DOI: 10.1093/ajcp/aqae062.

[11] Gholami S, Mummert DB, Wilson B, et al. Leveraging large language models to generate multiple-choice questions for ophthalmology education [J]. JAMA Ophthalmol, 2025, 143(11): 955-961. DOI: 10.1001/jamaophthalmol.2025.3622.

[12] Mistry NP, Saeed H, Rafique S, et al. Large language models as tools to generate radiology board-style multiple-choice questions [J]. Acad Radiol, 2024, 31(9): 3872-3878. DOI: 10.1016/j. acra. 2024.06.046.

[13] Jackson FI, Keller NA, Kouba I, et al. Large language model clinical vignettes and multiple-choice questions for postgraduate medical education [J]. Acad Med, 2025, 100(10): 1163-1166. DOI: 10.1097/acm.0000000000006137.

[14] Safadi S, Amirahmadi R, Tlimat A, et al. Quality of human expert vs large language model-generated multiple-choice questions in the field of mechanical ventilation [J]. Chest, 2025, 168(6): 1425-1432. DOI: 10.1016/j.chest.2025.07.005.

[15] Law AK, So J, Lui CT, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination [J]. BMC Med Educ, 2025, 25(1): 208. DOI: 10.1186/s12909-025-06796-6.

[16] Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review [J]. Postgrad Med J, 2024, 100(1189): 858-865. DOI: 10.1093/postmj/qgae065.

[17] Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-a multinational prospective study [J]. PLoS One, 2023, 18(8): e0290691. DOI: 10.1371/journal.pone.0290691.

[18] Zuckerman M, Flood R, Tan RJB, et al. ChatGPT for assessment writing [J]. Med Teach, 2023, 45(11): 1224-1227. DOI: 10.1080/0142159x.2023.2249239.

[19] Karahan BN, Emekli E. Comparison of applicability, difficulty, and discrimination indices of multiple-choice questions on medical imaging generated by different AI-based chatbots [J]. Radiography, 2025, 31(5): 103087. DOI: 10.1016/j.radi.2025.103087.

[20] Ahmed A, Kerr E, O′Malley A. Quality assurance and validity of AI-generated single best answer questions [J]. BMC Med Educ, 2025, 25(1): 300. DOI: 10.1186/s12909-025-06881-w.

[21] Wu H, Zerner T, Lee D, et al. GPT-4 versus human authors in clinically complex MCQ creation: a blinded analysis of item quality [J]. Med Teach, 2025, 47(12): 1961-1974. DOI: 10.1080/0142159x.2025.2505122.

[22] Abouzeid E, Wassef R, Jawwad A, et al. Chatbots' role in generating single best answer questions for undergraduate medical student assessment: comparative analysis [J]. JMIR Med Educ, 2025, 11: e69521. DOI: 10.2196/69521.

[23] Ch′en PY, Day W, Pekson RC, et al. GPT-4 generated answer rationales to multiple choice assessment questions in undergraduate medical education [J]. BMC Med Educ, 2025, 25(1): 333. DOI: 10.1186/s12909-025-06862-z.

[24] Sawamura S, Kohiyama K, Takenaka T, et al. Potential of large language models in generating multiple-choice questions for the Japanese national licensure examination for physical therapists [J]. Cureus, 2025, 17(2): e79183. DOI: 10.7759/cureus.79183.

[25] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620(7972): 172-180. DOI: 10.1038/s41586-023-06291-2.

[26] Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models [J]. Nat Med, 2025, 31(3): 943-950. DOI: 10.1038/s41591-024-03423-7.

[27] Wu C, Lin W, Zhang X, et al. PMC-LLaMA: toward building open-source language models for medicine [J]. J Am Med Inform Assoc, 2024, 31(9): 1833-1843. DOI: 10.1093/jamia/ocae045.

[28] Xie Q, Chen Q, Chen A, et al. Medical foundation large language models for comprehensive text analysis and beyond [J]. NPJ Digit Med, 2025, 8(1): 141. DOI: 10.1038/s41746-025-01533-1.

（收稿日期：2026-02-21）
（本文编辑：邬小梅）