Advances in Education 教育进展, 2026, 16(11), 1154-1161
Published Online November 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.16112486

Hans汉斯

大语言模型在医学生临床推理教学中的应用

蒋桢桢*，楼　夲

浙江大学医学院附属第四医院/国际医学院/国际健康医学研究院教学部，浙江　金华

收稿日期：2026年10月9日；录用日期：2026年11月9日；发布日期：2026年11月17日

摘　要

研究系统梳理大语言模型在医学生临床推理与临床思维教学中的应用现状，分析其核心应用场景、潜在挑战及未来发展方向。采用文献综述对大语言模型在虚拟病人交互、人机协同决策、形成性评价及教学方法融合等维度进行深入剖析。现有研究证实，大语言模型的应用显著提升了医学生的问诊技巧、鉴别诊断能力和临床决策水平，但其在准确性验证、伦理合规、长期教学效果追踪等方面仍存在局限。大语言模型为临床思维教育提供了可及性强、个性化的创新工具，未来需通过多中心大样本实证研究验证其长期教育效益，并构建规范化的应用框架。

关键词

大语言模型，人工智能，医学生，临床推理，临床思维

Application of Large Language Models in Clinical Reasoning Teaching for Medical Students

Zhenzhen Jiang*, Tao Lou

Department of Education, The Fourth Affiliated Hospital/International School of Medicine/International Institutes of Medicine, Zhejiang University School of Medicine, Jinhua Zhejiang

Received: October 9, 2026; accepted: November 9, 2026; published: November 17, 2026

Abstract

This study systematically reviews the current applications of large language models (LLMs) in clinical reasoning and clinical thinking education for medical students, analyzing their core implementation scenarios, potential challenges, and future development directions. Employing a literature review

*通讯作者。

文章引用: 蒋桢桢, 楼夲. 大语言模型在医学生临床推理教学中的应用[J]. 教育进展, 2026, 16(11): 1154-1161.
DOI: 10.12677/ae.2026.16112486

methodology, this paper performs a comprehensive analysis of LLMs across multiple dimensions, including virtual patient interactions, human-machine collaborative decision-making processes, formative assessment frameworks, and pedagogical methodology integration. Existing studies confirm that LLMs implementation significantly improves medical students' clinical history-taking proficiency, differential diagnosis capabilities, and clinical decision-making competencies. However, limitations persist regarding model accuracy validation, ethical compliance standards, and longitudinal evaluation of educational effectiveness. As highly accessible, personalized, and innovative educational tools, LLMs offer support for clinical thinking instruction. Future research necessitates multicenter empirical investigations with large cohort samples to rigorously assess their sustained educational impacts and to establish a standardized application framework.

Keywords

Large Language Models, Artificial Intelligence, Medical Student, Clinical Reasoning, Clinical Thinking

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

医学生在成为执业医师前，需要在培训期间培养安全且有效的临床推理能力。临床推理被定义为一种技能、过程和结果——即临床医生通过观察、收集及解读数据来对患者进行诊断与治疗的过程[1]。在传统教学中，培养临床推理能力的主要方法是基于案例的学习，如教学查房或病例讨论等，但临床教师大多采用模式识别讨论，缺少对学生逻辑思维的训练以及对其思维成长的观察和评价[2]。现实中由于临床教学资源短缺、教师教学与反馈的标准化不足，临床推理训练的有效性受到极大挑战。近年来，以大语言模型(Large Language Models, LLMs)为代表的生成式人工智能技术(GAI)取得了突破性进展，其强大的自然语言理解、情境模拟与即时反馈能力，为重塑医学教育范式提供了新的路径[3]。LLMs 不仅能够模拟高度逼真的医患交互场景，还能根据学生的认知水平动态调整病例难度，从而在临床推理早期培养阶段展现出独特的优势[4]-[6]。本文系统梳理了 LLMs 在医学生临床思维教学中的应用现状，探讨其在虚拟问诊、人机协同、评估反馈等环节的深度融合，并针对当前面临的挑战提出未来展望，以期为临床推理教学的数字化改革提供理论参考与实践指引。

2. 大语言模型概述

大语言模型(LLMs)是一类人工智能(AI)系统，它们通过 Transformer 架构获得了类人类的理解与推理能力[7]。借助深度学习和先进的人工神经网络，这类模型能够解析字符与词语之间的关系，并生成连贯的文本。这些模型需经过包含数十亿参数的训练，来自动识别海量数据的复杂模式与关联，因此具备强大的文本生成、语义理解和逻辑推理能力[8]。近年来，已有大量研究探讨了 LLMs 在不同语言和国家医学执业资格考试题目上的表现，其在考试中的准确率已超越大多数人类考生，展现出令人瞩目的医学知识储备和临床决策潜力[9]。此外，在医学教育领域，LLMs 不仅是知识检索的增强工具，更是能够模拟患者叙事、生成鉴别诊断假设、提供结构化教学反馈的“智能导师”[10]。

3. 大语言模型在医学生临床思维教学中应用的意义

在临床实践中，医学生要成长为一名成熟医生，需要经过大量的案例实践，将需要启动分析型决策

的案例逐步内化为直觉型决策范围，并达到两种决策灵活切换与整合的效果。LLMs 的引入深刻契合了这一认知发展规律。而且，LLMs 打破了传统教学的时空壁垒，提供 24 小时可及的“伴随式”训练环境，学生可在无风险的前提下反复进行问诊和鉴别诊断练习，符合刻意练习理论的核心要求[11]。此外，基于检索增强生成和提示词工程优化的 LLMs，能够生成涵盖罕见病或复杂并发症的动态病例，弥补了临床实践中病例同质化高、病种单一的短板[12]。同时，LLMs 提供的即时且结构化的反馈一定程度上促成了医学生“疾病脚本”的构建。另一方面，LLMs 在教学中的常态应用也能帮助临床教师从“知识传授者”向“思维引导者”转型，推动医学教育真正向以能力为导向的教育转变[13]。

4. 文献检索和分析

4.1. 文献获取与初步整理

本研究采用系统性文献检索方法，检索数据库包括中文数据库(知网、万方)和英文数据库 Web of Science 核心合集。中文检索词包括“大语言模型”、“生成式人工智能”、“AI”、“ChatGPT”、“医学生”、“临床推理”、“临床思维”、“临床决策”。英文检索词包括“Large Language Models”、“Artificial Intelligence”、“Generative AI”、“ChatGPT”、“Medical Students”、“Undergraduate Medical Education”、“Clinical Decision-Making”、“Clinical Reasoning”、“Diagnostic Reasoning”、“Virtual Patients”。采用布尔逻辑运算符组合关键词，检索时限为 2017 年 1 月~2026 年 7 月。初步检索共获得文献 122 篇。文献筛选流程参考 PRIAMA 流程图，两名研究者独立对检索结果进行筛选，首先通过标题与摘要进行初筛，随后对符合条件的文献进行全文评估。若存在分歧，则邀请第三位研究者参与讨论共同决定。最终排除重复、不符合主题、无法获取全文、会议类、非教育类共 20 篇文献，纳入 102 篇文献进行整理。数据提取涵盖研究基本信息，包括作者、年份、地区、关键词、摘要。

4.2. 文献阅读分析

鉴于医学教育类文献同属于社科类文献，因此本研究引用“菲利普阅读密码表”作为批判性文献精读工具。结合本研究核心问题“大语言模型在医学生临床思维教学中的应用与挑战”，设置 6 项编码类别对每篇文献进行解码，分别是“他们要做什么(WTD)”、“研究结果(ROF)”、“现有文献综述(SPL)”、“现有文献批评/空白(CPL/GAP)”、“未来研究建议(RFW)”、“批评点/待探讨的相关问题(POC/RPP)”。由两名研究者通过对 WTD、ROF、SPL 前三部分和 CPL/GAP、RFW、POC/RPP 后三部分分别进行语义聚类和主题提炼，共同归纳 LLMs 在医学生临床思维教学中应用的四大场景以及应用挑战的五大方面。

5. 大语言模型在医学生临床思维教学中应用现状

5.1. LLMs 驱动的虚拟病人与问诊推理训练

传统的虚拟病人多依赖预设的疾病脚本，交互不够灵活，难以应对开放式的临床问诊。LLMs 凭借其优秀的上下文理解与自然语言生成能力，重塑虚拟病人的交互逻辑。研究表明，基于 GPT-4 的大语言模型所驱动的虚拟病人，在病史采集训练中的角色扮演和回答具有超过 99%的医学合理性，与人类在问诊评分上具有高度一致性(Cohen $\kappa = 0.832$) [14]。学生通过与虚拟病人进行多轮对话，不仅能够锻炼信息获取的完整性和逻辑性，还能在精神、妇产科等敏感领域的问诊中建立信心，提高病史采集质量[15] [16]。为进一步提升沉浸感，研究者将 LLMs 与多智能体架构相结合，赋予虚拟患者更细腻的情感表达和非语言反馈。例如，开发的多智能体框架通过功能划分的子智能体协同工作，有效模拟了虚拟病人的情绪变化、症状演进及角色表现力。结果表明，在标准对话场景下，该框架的信息准确性与角色代入感均显著

优于单一 LLMs 基线方案[17]。此外，基于动态提示词模型的“动态病例”系统能够根据学生的临床决策实时推进疾病病程，实现临床决策链条的连续推演，显著提升了学生的临床思维能力和批判性思维能力[18]。这些进展标志着 LLMs 驱动的虚拟病人从单一模型向多智能体协同、从静态病例向动态病例演进的逐步改进。

5.2. 人机协同促进临床推理与决策训练

人机协同是 LLMs 在复杂临床决策训练中的另一大应用前沿。临床推理往往面临信息不确定性和认知偏差的双重挑战，LLMs 可作为“第二助手”或“认知脚手架”辅助学生拓展鉴别诊断思维。实验研究证实，在具有挑战性、复杂性的真实临床场景中，人机协作显著提高了参与者的诊断准确率(从独立工作的 45%提升至协同工作的 60%)，并有效降低了认知负荷[19]。它的核心机制在于 LLMs 能够迅速检索并整合大量医学文献，提供基于循证医学的鉴别诊断列表，从而弥补低年资学生的医学知识盲点。例如，在眼科学、风湿病学等复杂专科领域，使用 ChatGPT-4o 联合传统资源的学生在准确识别首要诊断和累积诊断分数上显著高于仅使用传统资源的对照组(mean [SD]: 12.3 [2.3] vs. 6.7 [3.2]; $P < 0.001$)，且表现甚至超过 LLMs 独立运行的水平[20]。另外，对于 ChatGPT-4o 等模型进行的针对性教学及反馈环节，反而能使模型通过迭代反馈来优化诊断输出，逐步模拟出医学生典型的学习曲线，这也证明了 LLMs 作为思维训练工具的潜在价值[21]。然而，人机协同的效果也受任务难度调节，在任务难度较低时协同效应显著，而在极高难度任务中，由于人类判断能力的局限性，LLMs 的辅助效果相对有限，比如学生接受了错误答案或拒绝了正确的反馈[19]。

5.3. LLMs 辅助临床思维评估与个性化反馈

形成性评价是临床思维培养不可或缺的环节。传统评价依赖教师主观打分，反馈滞后且难以规模化。LLMs 的引入使得自动化、即时性的临床能力评估成为可能。在脚本一致性测试(SCT)的生成与评估中，ChatGPT-4o 和 Claude 3.5 Sonnet 等模型展现出与人类专家高度接近的生成能力，总体质量标准符合率超过 90% [22]。LLMs 不仅能高效生成高质量的 SCT 试题，还能对学生的回答进行多维度解析，指出其思维过程中的逻辑漏洞或知识盲区。在客观结构化考试(OSCE)准备中，LLMs 可根据学生的问诊查体表现帮助学生反思自身技能，如遗漏的既往史询问、不恰当的体检顺序等[23]。另外，LLMs 在自动批改临床思维作业中也展现出极高的效率，构建了“学–练–评–改”的完整闭环，大幅提升了教学反思的时效性[24]。有研究专门开发了 AI 辅助的形成性评估框架，强调 AI 在提供规模化、个性化反馈方面的潜力，同时指出必须培养学习者对 AI 输出的批判性解读能力，并将其转化为具体的学习行动[25]。

5.4. LLMs 与教学方法的深度融合

LLMs 并非孤立的技术工具，而是与众多成熟的教学方法如基于问题的学习(PBL)、基于团队的学习(TBL)、基于案例的学习(CBL)等深度融合，逐步产生 AI 联合教学方法赋能医学教育的各种应用场景。例如，在传统 PBL 教学中，启发性“触发点”的设计和导师辅导指南的编写耗时费力。有研究表明，利用 Gemini Advanced 等模型能够有效生成与学习目标高度一致的临床情景，并优化了导师指南的启发性问题结构，尽管在视觉元素，如图片、图表和教学适宜度上仍需教师评审，但已显著提升了 PBL 教学素材的生成效率和质量[26]。与 PBL 相似，CBL、TBL 均采用真实的案例来促进以学生为中心的小组学习，因此，在编写学习材料时运用 LLMs 也可使这些方法受益[27]。

6. 大语言模型在医学生临床思维教学中应用的挑战与不足

尽管 LLMs 在临床推理教学中展现出广阔的应用前景，但其在教育场景中的大规模落地仍面临众多

挑战。

6.1. “幻觉”风险与准确性问题

LLMs 的核心机制是基于概率的文本生成，而非严格的知识检索，这导致它在缺乏训练数据的边缘领域容易产生“幻觉”(Hallucination)，即生成看似合理但实际错误的医学信息[28]。研究表明，尽管一些先进的模型在医学考试中的准确率令人瞩目，但在处理复杂影像报告、罕见病细节或最新临床指南时，输出质量显著下降，且存在信息过时或可重复性差等问题[29]。在临床推理训练过程中，若 LLMs 提供的鉴别诊断建议或用药指导存在错误，而受训学生又缺乏足够的批判性思维能力对其进行有效甄别，将可能导致错误的医学认知建构，甚至在未来实际的临床实践中引发医疗安全隐患。因此，构建能保障 AI 输出医学准确性的校验机制，是当前该领域亟待突破的技术瓶颈。

6.2. 过度依赖与思维能力退化

以 LLMs 为代表的 AI 模型辅助学习的便捷性可能引发“认知卸载”(Cognitive Offloading)风险。实证调查显示，医学生在面对 AI 提供的答案时，有超过 20%的概率会修改自己的初始判断以迎合 AI 的建议，且这种顺从行为在病理生理等基础学科或自身不确定时更为显著[30]。当学生习惯于向 AI 直接索取结论而非经历完整的临床推理过程，他的独立思考能力、鉴别诊断能力和循证医学素养可能被削弱。长此以往，这种“算法依赖”可能会阻碍医学生从“分析性推理”向“模式识别”的思维培养，导致临床思维能力退化、认知结构碎片化。因此，教学中必须强调“以我为主、AI 为辅”的原则，设计强制性独立推理环节，鼓励医学生积极参与与 AI 输出的持续互动，防止技术反客为主。

6.3. 缺乏长期教学效果追踪与有效的教学实践转化

当前，关于 LLMs 辅助临床推理训练的研究多为短期干预和小样本探索性试验，缺乏对长期教学效果的纵向追踪。学生在使用 LLMs 后表现出的短时成绩提升是否能够转化为长期内化的临床思维能力，以及在处理真实病例时的迁移应用能力，目前尚无确凿证据[31]。此外，模拟环境中的临床推理与真实临床场景存在显著差异，AI 虚拟病人无法完全复现真实临床的复杂情境，如情感压力、时间紧迫性、多学科协作等[32]。未来需要开展多中心、大样本的随机对照试验，并引入真实世界数据，以验证 LLMs 对医学生长期临床胜任力的实质性影响。

6.4. 实证基础有限与研究方法异质性

现有文献在研究设计上存在高度异质性，缺乏统一的评价标准。不同研究采用的 LLMs 版本、提示词策略、干预时长和结局指标各不相同，导致研究结果难以进行横向比较[33]。部分研究仅依赖学生的主观满意度或自我评估的自信心提升作为主要结局，缺乏客观的行为学指标或临床推理过程数据。此外，针对 AI 辅助形成性评价的信效度检验等研究尚不充分，AI 评分与教师评分的一致性在不同专科和病例复杂度下波动较大。这些基础尚薄弱的实证研究一定程度限制了 LLMs 在临床教学中应用的有效性推广。

6.5. 伦理合规、数据隐私与学术诚信挑战

LLMs 在医学教育中的应用也触及伦理红线。如果将患者的真实脱敏数据或临床案例输入公共 LLMs，可能存在数据泄露，违反医学伦理。GAI 的算法偏见可能导致它在处理特定种族、性别或社会经济背景的患者案例时，输出不公平或有偏见的诊疗建议，进而将偏见内化为学生的临床态度[34]。此外，学生可能滥用 LLMs 直接生成病例分析报告或考试答案，而非将其作为思维辅助工具，存在学术诚信隐患[35]。因此，教育相关部门需制定明确的 AI 使用政策，进一步明确区分“辅助学习”与“学术不端”的边界，

同时建议将 AI 素养作为早期基础课程纳入医学课程体系，以便更好地帮助学生做好应对 AI 辅助实践的准备。

7. 总结与展望

大语言模型的兴起，为医学教育，尤其是临床推理训练提供了新的可能性。从现有研究来看，LLMs 通过构建高度仿真的虚拟问诊环境、提供即时个性化的形成性反馈，以及人机协同下的鉴别诊断练习，可以提升医学生的病史采集和临床决策能力。LLMs 在降低教学成本、突破时空限制以及实现规模化、个性化教学方面展现出不可忽视的优势，已成为传统标准化病人教学和床旁带教的有力补充。然而，该领域的研究仍处于早期，存在的挑战需要重视，如模型幻觉带来的认知误导风险、算法依赖导致的独立思考能力削弱、长期纵向大样本研究缺乏以及研究方法的异质性和伦理合规的不确定性，这些问题整体制约了模型在教学中的广泛应用。

基于上述挑战和问题，后续研究的重点建议从以下几个方面开展进一步探索：在技术层面，开发面向医学教育的垂直领域大模型，深度融合检索增强生成与知识图谱技术，并建立标准、规范的思维训练数据库，从源头上减少幻觉，提升输出的准确性；在教学评价上，构建标准化、多维度的 AI 辅助临床思维评价体系，使用过程性数据，如交互日志、思维链轨迹、动作捕捉等，并结合结局性指标，开展大规模、多中心的纵向实证研究；在教学实践中，倡导探索 AI 增强而非替代的教学范式，设计强制独立推理与人机辩证交互的课程模块，建立提问－结果－验证－反思的闭环教学过程，教师需要扮演思维引导者和结果审判员的角色，引导学生会提问题、提好问题，并提高学生对 AI 产出内容的证据核验能力，培养学生的自主学习能力和 AI 批判性素养；在治理层面，建议建立涵盖数据隐私保护、算法透明度审查、学术诚信界定的治理框架，推动跨机构的资源共享与国际合作。相信，随着技术的不断迭代和制度体系的逐步完善，大语言模型将成为赋能卓越临床医师培养的重要手段，推动医学教育向更高阶的精准化、智能化迈进。

致　谢

感谢浙江省教育厅、浙江大学医学院第四临床医学院提供的基金资助，感谢所有引用文献的作者及对本文作者提供研究指导的所有人员。

基金项目

浙江省教育厅一般科研项目(Y202558277)；浙江大学医学院第四临床医学院教育教学改革项目(JG20250206)。

参考文献

[1] Daniel, M., Rencic, J., Durning, S.J., Holmboe, E., Santen, S.A., Lang, V., *et al*. (2019) Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. *Academic Medicine*, **94**, 902-912. https://doi.org/10.1097/acm.0000000000002618

[2] 兰学立, 向阳. 临床推理——从入门到实践[M]. 北京大学医学出版社, 2022: 4.

[3] 王茹, 李毅恒, 孙世仁, 等. 人工智能在医学教育中的应用前景与挑战[J]. 中国医学教育技术, 2025, 39(3): 306-310.

[4] Borg, A., Georg, C., Jobs, B., Huss, V., Waldenlind, K., Ruiz, M., *et al*. (2025) Virtual Patient Simulations Using Social Robotics Combined with Large Language Models for Clinical Reasoning Training in Medical Education: Mixed Methods Study. *Journal of Medical Internet Research*, **27**, e63312. https://doi.org/10.2196/63312

[5] Hryszko, J., Michałek, A. and Roman, A. (2026) Application of Large Language Models in Medical Interview Training: A Study with Medical Students. *BMC Medical Education*, **26**, Article No. 586.

https://doi.org/10.1186/s12909-026-08819-2

[6] Fąferek, J., Kononowicz, A.A., Bogutska, N., Da Silva Domingues, V., Davydova, N., Frankowska, A., *et al*. (2025) Applying ChatGPT to Plan and Create a Realistic Collection of Virtual Patients for Clinical Reasoning Training. *BMC Medical Education*, **25**, Article No. 1277. https://doi.org/10.1186/s12909-025-08006-9

[7] Hagendorff, T., Fabi, S. and Kosinski, M. (2023) Human-Like Intuitive Behavior and Reasoning Biases Emerged in Large Language Models but Disappeared in ChatGPT. *Nature Computational Science*, **3**, 833-838. https://doi.org/10.1038/s43588-023-00527-x

[8] 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.

[9] Nouri, H., Mahdavi, A., Abedi, A., Mohammadnia, A., Hamedan, M. and Amanzadeh, M. (2025) Performance of Large Language Models in Medical Licensing Examinations: A Systematic Review and Meta-Analysis. *Journal of Educational Evaluation for Health Professions*, **22**, Article 36. https://doi.org/10.3352/jeehp.2025.22.36

[10] Abd-Alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P.M., Latifi, S., *et al*. (2023) Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Medical Education*, **9**, e48291. https://doi.org/10.2196/48291

[11] Benítez, T.M., Xu, Y., Boudreau, J.D., Kow, A.W.C., Bello, F., Van Phuoc, L., *et al*. (2024) Harnessing the Potential of Large Language Models in Medical Education: Promise and Pitfalls. *Journal of the American Medical Informatics Association*, **31**, 776-783. https://doi.org/10.1093/jamia/ocad252

[12] Amugongo, L.M., Mascheroni, P., Brooks, S., Doering, S. and Seidel, J. (2025) Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review. *PLOS Digital Health*, **4**, e0000877. https://doi.org/10.1371/journal.pdig.0000877

[13] 李飞岩, 招皓景, 黄绮翘, 等. 人工智能赋能本科医学教育形成性评价的路径与策略: 基于 CBME 框架的综述[J]. 中国医学教育技术, 2026, 40(3): 413-421.

[14] Holderried, F., Stegemann-Philipps, C., Herrmann-Werner, A., Festl-Wietek, T., Holderried, M., Eickhoff, C., *et al*. (2024) A Language Model-Powered Simulated Patient with Automated Feedback for History Taking: Prospective Study. *JMIR Medical Education*, **10**, e59213. https://doi.org/10.2196/59213

[15] Espírito Santo, D.S., Lott Bezerra, T., Viana Lopes Gomes, M.C., Augusto de Souza Fonseca, P. and Joviano-Santos, J.V. (2026) AI-Assisted Learning: ChatGPT for Anamnesis in Women's Health Education. *Medical Education Online*, **31**, Article ID: 2655562. https://doi.org/10.1080/10872981.2026.2655562

[16] Kilincel, S., Bulut, F., Goksel, P., Usta, M.B., Mutluer, T. and Kilincel, O. (2026) Effectiveness of AI-Enhanced Virtual Patients for Psychiatric Interview Training in Health Professions Education: A Meta-Analysis. *Frontiers in Medicine*, **13**, Article 1834636. https://doi.org/10.3389/fmed.2026.1834636

[17] Qu, Y., Xu, X., Long, Y., Wang, Y., Li, J. and Lu, X. (2026) A Large Language Model-Powered Multiagent Framework Emulating Standardized Patients in Clinical Communication Skills Training: Development and Evaluation Study. *Journal of Medical Internet Research*, **28**, e84747-e84747. https://doi.org/10.2196/84747

[18] 魏晓彤, 闻德亮. 基于大语言模型的动态病例在儿科学教学中的应用研究[J]. 中华医学教育杂志, 2026, 46(4): 269-274.

[19] Ong, K.T., Seo, J., Kim, H., Kim, J., Kim, J., Kim, S., *et al*. (2026) Success and Failure of Human-AI Collaboration in Clinical Reasoning: An Experimental Study on Challenging Real-World Cases. *International Journal of Medical Informatics*, **211**, Article ID: 106342. https://doi.org/10.1016/j.ijmedinf.2026.106342

[20] Roemer, A., Schlicker, N., Kernder, A., Albe, B., Hack, J., Hirsch, M., *et al*. (2026) Large Language Models Enhance Diagnostic Reasoning of Medical Students in Rheumatology: A Randomized Controlled Trial. *BMC Medical Education*, **26**, Article No. 579. https://doi.org/10.1186/s12909-026-09079-w

[21] Cabir, Y., Hornstein, R., Segal, G. and Shoham, R. (2026) Generative Pretrained Transformer (GPT) Algorithm Can Be Taught to Improve the Task of Differential Diagnosis According to Simple Principles of Clinical Reasoning. The Case of GPT-4o and a Lesson to Apply in Medical Education. *Digital Health*, **12**. https://doi.org/10.1177/20552076261416793

[22] Yapıcı Coşkun, Z., Kıyak, Y.S., Coşkun, Ö., Budakoğlu, I.İ. and Özdemir, Ö. (2025) Large Language Models for Generating Script Concordance Test in Obstetrics and Gynecology: ChatGPT and Claude. *Medical Teacher*, **47**, 1767-1771. https://doi.org/10.1080/0142159x.2025.2497888

[23] Brügge, E., Ricchizzi, S., Arenbeck, M., Keller, M.N., Schur, L., Stummer, W., *et al*. (2024) Large Language Models Improve Clinical Decision Making of Medical Students through Patient Simulation and Structured Feedback: A Randomized Controlled Trial. *BMC Medical Education*, **24**, Article No. 1391. https://doi.org/10.1186/s12909-024-06399-7

[24] 朱晓燕, 杨健, 杨尚君, 等. AI 赋能寄生虫实验教学新模式提升学生形态识别与临床决策能力[J]. 医学教育研究与实践, 2026, 34(1): 106-113.

[25] Nguyen, Q.T., Ha, T.M. and Mai, T. (2026) AI-Assisted Formative Assessment in Clinical Education: From Algorithms to Agency. *JMIR Medical Education*, **12**, e93710. https://doi.org/10.2196/93710

[26] Arain, S.A., Akhund, S.A., Barakzai, M.A. and Meo, S.A. (2025) Transforming Medical Education: Leveraging Large Language Models to Enhance PBL—A Proof-of-Concept Study. *Advances in Physiology Education*, **49**, 398-404. https://doi.org/10.1152/advan.00209.2024

[27] Qian, C., Gao, C., Park, S., Gim, H., Hou, K., Cook, B., *et al*. (2025) Use of Large Language Models for Rapid Quantitative Feedback in Case-Based Learning: A Pilot Study. *Medical Science Educator*, **35**, 1169-1171. https://doi.org/10.1007/s40670-025-02343-6

[28] Gilbert, S., Kather, J.N. and Hogan, A. (2024) Augmented Non-Hallucinating Large Language Models as Medical Information Curators. *npj Digital Medicine*, **7**, Article No. 100. https://doi.org/10.1038/s41746-024-01081-0

[29] Tordjman, M., Liu, Z., Yuce, M., Fauveau, V., Mei, Y., Hadjadj, J., *et al*. (2025) Comparative Benchmarking of the DeepSeek Large Language Model on Medical Tasks and Clinical Reasoning. *Nature Medicine*, **31**, 2550-2555. https://doi.org/10.1038/s41591-025-03726-3

[30] Kirpalani, A. and Wang, P.Z.T. (2026) Medical Student Reliance on Artificial Intelligence in Nephrology Education. *Journal of Nephrology*, **39**, 1158-1164. https://doi.org/10.1093/joneph/aajag033

[31] Castillo, E.V., Durazo Enríquez, C.R., Valenzuela Dominguez, L.R., Ramírez, A.A.P. and López, S.T. (2026) Generative Artificial Intelligence for Clinical Reasoning Assessment in Medical Students: A Systematic Review of Early Experimental Evidence. *Educación Médica*, **27**, Article ID: 101197. https://doi.org/10.1016/j.edumed.2026.101197

[32] Berman, N.B., Durning, S.J., Fischer, M.R., Huwendiek, S. and Triola, M.M. (2016) The Role for Virtual Patients in the Future of Medical Education. *Academic Medicine*, **91**, 1217-1222. https://doi.org/10.1097/acm.0000000000001146

[33] Simoni, J., Urtubia-Fernandez, J., Mengual, E., Simoni, D.A., Royo, M., Egaña-Yin, D., *et al*. (2025) Artificial Intelligence in Undergraduate Medical Education: An Updated Scoping Review. *BMC Medical Education*, **25**, Article No. 1609. https://doi.org/10.1186/s12909-025-08188-2

[34] Omar, M., Soffer, S., Agbareia, R., Bragazzi, N.L., Apakama, D.U., Horowitz, C.R., *et al*. (2025) Sociodemographic Biases in Medical Decision Making by Large Language Models. *Nature Medicine*, **31**, 1873-1881. https://doi.org/10.1038/s41591-025-03626-6

[35] Zhao, L., Han, Q., Li, P., Dai, H. and Ju, X. (2026) Adoption of Generative AI Chatbots among Medical Postgraduates at Two Universities in China: Patterns, Attitudes, and Concerns. *BMC Medical Education*, **26**, Article No. 626. https://doi.org/10.1186/s12909-026-08947-9