·教学与科研·

生成式人工智能在住院医师理论及病例分析考核中的应用与挑战

周元凯[1]，孙　俊[2]，刘圣均[1]，杨莹莹[1]，袁思依[1]，何怀武[1]，隆　云[1]

中国医学科学院北京协和医院 [1] 重症医学科 [2] 院办公室，北京 100730

通信作者：隆　云，E-mail：ly_icu@aliyun.com

【摘要】生成式人工智能（generative artificial intelligence，GAI）在医学领域的应用是目前研究的热点，医学教育是其重要方向之一。在个性化教学、考核题目创建、题库更新、智能评分等方面，GAI 有助于提升住院医师培训效果。然而，目前 GAI 在生成内容的准确性、一致性等方面还存在一定局限性。为克服上述局限性，GAI 本身需不断升级，同时还需制订 GAI 使用指南、加强数据质量控制、建立生成内容的人工审核制度等。住院医师应主动学习 GAI 的使用方法，强化理论知识的实践应用能力。未来，GAI 有望成为提高住院医师培养效率和质量的有益工具。

【关键词】生成式人工智能；住院医师规范化培训；理论及病例分析考核；医学教育

【中图分类号】R4319；G511　　【文献标志码】A　　【文章编号】1674-9081(2025)05-1352-05

DOI：10.12290/xhyxzz.2024-0795

The Applications and Challenges of Generative Artificial Intelligence in Theoretical and Case Analysis Assessment for Resident Physician Education

ZHOU Yuankai[1], SUN Jun[2], LIU Shengjun[1], YANG Yingying[1], YUAN Siyi[1], HE Huaiwu[1], LONG Yun[1]

[1]Department of Critical Care Medicine, [2]General Office, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China

Corresponding author: LONG Yun, E-mail: ly_icu@aliyun.com

【Abstract】 Generative artificial intelligence (GAI) represents a prominent research focus in medicine, with medical education being a key application area. GAI demonstrates potential to enhance residency training efficacy through personalized instruction, automated assessment item generation, question bank updating, and intelligent scoring systems. However, current limitations exist regarding output accuracy and content consistency. To address these constraints, strategic measures are required: continuous GAI model refinement, development of standardized usage guidelines, enhanced data quality control, and implementation of human verification protocols for generated content. Concurrently, residents should proactively acquire GAI utilization skills to strengthen the practical application of theoretical knowledge. With these advancements, GAI is anticipated to evolve into a valuable asset for improving the efficiency and quality of residency training programs.

基金项目：北京协和医学院 2023 年中央高校教育教学改革专项资金支持项目（2023zlgl022）；北京协和医学院 2023 年青年医学教育学者培训项目（2023zlgc0710）；中央高水平医院临床科研专项（2022-PUMCH-B-115，2022-PUMCH-D-005）；中国医学科学院医学创新基金（2021-I2M-1-062）

引用本文：周元凯，孙俊，刘圣均，等. 生成式人工智能在住院医师理论及病例分析考核中的应用与挑战［J］. 协和医学杂志，2025，16(5)：1352-1356. doi：10.12290/xhyxzz.2024-0795.

【Key words】 generative artificial intelligence; standardized residency training; theoretical and case analysis assessment; medical education

Funding: Central University Education and Teaching Reform Special Funding of Peking Union Medical College in 2023 (2023zlgl022); 2023 Medical Education Scholar Program of Peking Union Medical College (2023zlgc0710); National High Level Hospital Clinical Research Funding (2022-PUMCH-B-115, 2022-PUMCH-D-005); CAMS Innovation Fund for Medical Sciences (2021-I2M-1-062)

Med J PUMCH, 2025,16(5):1352-1356

生成式人工智能（generative artificial intelligence，GAI）对医学产生了重大影响，并显示出较大的应用潜力[1-3]。GAI已在医学领域的多个方面取得了初步成果，如辅助诊断、制订治疗计划、实现个性化教育等[4-7]。在国外，通用型GAI以OpenAI的对话生成预训练模型（chat generative pre-trained transformer，ChatGPT）为代表；国内的GAI也发展迅速，如阿里云研发的大语言模型“通义千问”以及百度研发的预训练语言模型“文心一言”等。

医学是一门涉及多学科且对知识的准确性和应用能力要求极高的学科，因此医学教育过程中面临诸多复杂的挑战。不仅要培养住院医师掌握众多疾病的发病机制和诊疗特点等理论知识，还需锻炼其在实践中对理论知识的灵活应用能力，培养其在高压环境下快速决策的能力。在实现这些教学目标的过程中，模拟考核是目前较为重要的评估手段之一。模拟考核可使住院医师更清楚地认识到目前知识体系中的短板，进而有针对性地加强学习；还可通过模拟病例训练，培养住院医师的临床思维能力。然而，传统教学流程中的模拟考核方法在GAI冲击下面临较大挑战[8-10]。传统作业模式易受GAI生成内容的影响，如学生将GAI生成内容进行二次加工后作为考试答案，可能难以体现其真实能力。同时，传统医学模拟考核也存在诸多问题，如题库数量有限且内容陈旧、既往题目与当前临床脱节。传统考试形式还依赖医学教育专家手动组卷、评卷，占用其大量时间和精力，且缺乏个性化反馈机制。上述问题均限制了教学质量和效率，无法有效激发学生的学习潜力。

目前，GAI应用于多个行业，有助于提高工作效率，其在医学教育领域也引起了一定变革，如在考核方式、考核内容、答案评价等多个环节，均进行了初步探索[3,11-13]。GAI有望提升医学教学考核的效率及公平性，还可用于优化住院医师的业务学习过程、对考核结果进行分析解答等[8,14-15]，这些变革有助于提升医学教育质量。本文梳理目前GAI在医学教育考核领域的应用探索，并分析其在重症医学住院医师考核中应用的优势和局限性，以期为临床教学提供指导。

1 GAI在住院医师理论及病例分析考核中的应用

1.1 考核题目创建

通过学习需考核的理论知识，GAI在恰当提示词的引导下可自动生成选择题的考核题目，供医学专家筛选并审核，该应用已被诸多研究证实[16-19]。目前，病例分析类选择题也可由GAI生成，其可模拟真实临床场景考核住院医师的临床诊疗思维和应急处理能力。然而，若仅使用简单指令生成病例，GAI生成的模拟病例往往缺乏真实临床情景所需的丰富细节，难以全面反映临床实践的复杂性，其改进方向在于设计提示词时强调生成案例型试题，确保试题包含丰富的情境描述[16,20]。需注意的是，目前GAI在处理高级认知技能任务（如临床思维推理过程和细节性治疗计划制订）时还存在明显不足[17,21-22]，限制了其生成试题的考核难度，以及对于医学知识在临床中应用的考核能力。

1.2 题库更新

在合适提示词的帮助下，GAI能够快速生成高质量的医学试题，从而减轻命题人员的工作负担，并显著加快试题库建设。随着医学的快速发展，每年大量研究成果不断涌现，加之各类疾病诊疗指南的定期更新，使得知识库和题库的维护成为一项繁重任务。应用GAI构建动态题库更新系统，教学人员仅需定期筛选重要医学文献，GAI可分析、识别知识库和题库中的相关内容，并生成更新建议。更新建议仅需人工审核即可，从而大幅减少题库建设、维护中所需的时间和精力投入，进一步提升了教学工作效率。

1.3 智能评分

选择题是目前住院医师考核最主要的题型，而简答题和病例分析题等主观题的占比较低，主要原因之一是主观题的评分过程消耗较多教学资源。但主观题可更全面地考核住院医师的基础知识掌握情况，同时

也提高了对住院医师临床思维能力的考核。

学习了正确答案的 GAI，可顺利完成主观题的评分任务[11,22-23]。GAI 能够参照正确答案对住院医师的回答内容进行精准评分，得益于以下机制：（1）对语言的理解。GAI 能够理解主观题回答内容中的语法、词汇及上下文复杂多变性，理解回答内容的关键点，并与标准答案进行比对。（2）评分标准的自定义。医学教育人员可自定义一系列评分准则，并纳入 GAI 评价中，其根据自定义要求评估答案，从而实现对答案的准确性、完整度、逻辑性等的多维度评价。（3）评分的标准化。通过预先设定的正确答案和评分标准，GAI 可建立一个标准化的评分流程，减少评分过程的主观性，提高评分的一致性和公平性。

GAI 不仅能够阅卷，减少医学教育人员的工作负担，还可保证评分的一致性和公平性，这对于大规模考试或智能化教育平台而言非常重要。但需注意的是，尽管 GAI 的智能化水平提高迅速，若正确答案比较复杂或存在多样化，仍需进行人工审核。

1.4 考试结果解析及交互式发散学习

目前，考试题库中客观选择题的解析往往较为简略，未能详尽阐述临床思维的逻辑过程，尤其是涉及病例分析的选择题，导致住院医师对复杂问题的理解难度较大。经不断迭代，GAI 的逻辑分析能力得到了显著提升。如 OpenAI 最新 O1 模型可在数学问题的解答中展示详细的解题思路和步骤，国内 k0-math 模型也实现了相应功能。GAI 生成的答案，其完整解析过程仅需医学教育人员审核即可，显著提升了工作效率。

通过对医学基础知识、最新文献、药物信息和治疗指南等知识库的学习，以及对考核问题和答案的训练，GAI 不仅可输出详细的答案解析步骤，还可通过交互式反馈，辅助住院医师高效获取相关知识库内容，通过交互式知识检索助力住院医师发散思维，实现举一反三，提高学习效率。

1.5 个性化教学

在商业领域，GAI 可个体化记录客户的偏好、行为等，用于自动化客户画像。此功能也可应用于教学领域，通过特定的训练和教学设置，GAI 可对住院医师的学习进度进行跟踪，分析其学习轨迹，动态生成契合不同水平住院医师的考核试题，并对其考核过程中的错误进行总结，生成个体化考试测评报告；同时，提供个性化建议，帮助住院医师弥补知识盲点，优化学习路径，提高学习效率[24]。通过个体化考试测评报告，医学教育人员能够掌握不同住院医师的学习水平，实施个体化辅导，提高教学质量。个体化考试、测评和教学辅导共同组成了 GAI 的个性化教学内涵。

2 GAI 在理论及病例分析考核中应用的局限性

2.1 信息的准确性和前沿性待验证

尽管目前 GAI 通过学习具备一定的医学水平，并在部分国家的执业医师考核中获得不俗成绩[3,21]；同时，其生成的考核试题具有一定价值[19-20]，但所学信息的准确性和前沿性仍存在不足。主要原因：（1）大模型训练数据的时效性不足。GAI 通常基于历史数据进行训练，而医学知识更新迅速，因此其无法涵盖最新研究成果或临床指南。（2）数据来源的局限性。GAI 依赖于公开可用的数据进行训练，这些数据并非全面，且存在相互矛盾的可能。大部分医学文献和会议数据库处于非公开状态，如 PubMed、Medline、中国知网、万方数据知识服务平台等国内外医学数据库，其资源很难被纳入到训练数据集。（3）模型的非精确性。在回答复杂医学问题时，目前较为领先的模型 ChatGPT 也无法做到百分之百可靠，甚至在联系上下文的过程中，捏造毫无依据的信息或数据。因 GAI 的生成内容基于对已有知识的推测，而非基于确切的事实。因此，通用型 GAI 生成内容的准确性务必由医学专家进行审核。（4）医学专业知识的复杂性。医学领域知识较为复杂，涵盖大量专业理论、临床经验和不断更新的诊疗原则。针对相同的医学问题，其相关研究资料可能存在自相矛盾的结论。在临床实践中，这种复杂性需医学专家对研究的纳入人群、研究方法和统计学方法等进行仔细甄别、比对，方可得出谨慎结论，甚至需反复讨论，而 GAI 目前无法做到这一点。为提高 GAI 在医学领域应用的准确性和前沿性，医学教育人员需及时更新训练数据内容，同时结合医学专家的监督，辅助 GAI 在不断发展的医学实践中实现同步发展。

2.2 GAI 生成内容的一致性不足

针对相同的问题反复提问，GAI 生成内容可能出现答案不一致的情况。研究发现，当用户质疑 GAI 生成内容时，其回答可能随之发生改变，导致答案缺乏稳定性[3]。而医学领域的信息具有科学性、严谨性、精确性特点，GAI 可能使学习阶段的住院医师产生困惑，甚至学习到错误信息，这一局限性是由通用型 GAI 本身的特点所决定。首先，以 ChatGPT 为代表的 GAI 在设计中引入一定的随机性和多样性，使其在与

人对话时显得更加自然。然而，这种特性在应对医学等科学问题时，可能导致回答缺乏一致性。其次，GAI 训练数据的局限性。GAI 是基于互联网上的大量文本数据进行训练，这些数据来源不同，且分布于不同时间段，可能包含不同观点，当大模型从不同的数据中提取信息时可能导致生成内容的不一致性。

针对上述缺陷，目前可能的解决方案为：（1）优化模型算法。改进算法或建立垂直模型以减少 GAI 通用模型的随机性，提高其回答科学问题的一致性。目前已有许多医学领域相关 GAI 出现，且随着不断发展，其逻辑性不足逐渐改善。如 O1 模型在处理数学和科学等复杂逻辑推理问题方面较之前的模型明显进步。相信未来随着 GAI 的优化，将逐渐接近人类的思考习惯和特性。（2）提高训练数据质量。医学专家对纳入训练的数据质量进行控制和筛选，可减少不准确和低质量信息的输入，降低生成答案的多样性。（3）人工审核机制。引入医学专家人工审核生成内容是提高 GAI 回答准确性的保障，可尝试探索审核效率更高的内容格式。

2.3 GAI 对主观题阅卷评分不足

虽然 GAI 在主观题阅卷评分方面的应用具有高效和公平等优势，但其局限性需警惕。首先，对于需深入分析和全面理解的复杂病例分析类主观题，现阶段 GAI 可能无法准确把握住院医师答案中的细微差别和深层逻辑关系，导致出现以评分高估为主的失准现象。其次，尽管可通过预设条件调整评分规则，但对于开放式答案的多样化逻辑表达，GAI 难以完全适应，易引发评分失准现象。最后，现阶段 GAI 在面对医师的创造性思维或批判性分析时，缺乏医学专家的经验和判断。随着逻辑思维能力更强的 GAI（如 O1 和 k0-math 模型）产生，此类问题将逐渐得以解决。

2.4 住院医师对人工智能的依赖及相关伦理问题

目前，医学专家普遍认为医学生在学习时依赖 GAI 可能导致自身知识水平下降[11-12,22,25]，尤其对于临床思维尚未建立、基础知识掌握尚不扎实的住院医师，更需通过自主记忆打牢基础，以便逐步形成正确的临床思维。因此，鼓励住院医师使用 GAI，同时又要避免对其过度依赖导致判断能力下降，这是目前迫切需要解决的问题。以下措施有助于解决上述矛盾：（1）制订住院医师使用 GAI 辅助学习的指南，教学过程中强调自主学习能力的培养，加强对知识的实际应用训练，如模拟病例和模拟场景训练等。（2）引入混合学习模式，将 GAI 与传统教学方法（如问题导向学习、团队基础学习）相结合，鼓励住院医师进行自主学习和实践训练，培养其独立思考和解决问题的能力[2,23,26]，建立 GAI 与传统教学互补的课程和教学平台可能是未来的探索方向。

此外，在使用 GAI 的过程中，医生和患者的隐私数据存在泄露风险，需提前制订严格的数据保护措施，确保所有输入 GAI 的数据和生成的内容均经过加密和匿名化处理，符合相关伦理规范。

3 小结

综上所述，GAI 技术可帮助医学教育人员在个性化教学、考核题目创建、题库更新、智能评分等多个方面提高效率。虽然目前 GAI 在生成内容的准确性等方面还存在局限性，但随着模型的不断优化，通过建立 GAI 使用指南、加强训练数据的质控、建立生成内容的人工审核制度等管理方法，GAI 有望成为提高医学教学质量的有效工具。

作者贡献： 周元凯负责论文框架设计、起草及修订；刘圣均、杨莹莹、袁思依协助文献收集和汇总；孙俊、何怀武协助论文起草和修订；隆云为论文框架设计、修订及终审提供指导。

利益冲突： 所有作者均声明不存在利益冲突

参 考 文 献

[1] Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers [J]. JMIR Med Educ, 2023, 9: e46885.

[2] Arif T B, Munaf U, Ul-Haque I. The future of medical education and research: Is ChatGPT a blessing or blight in disguise? [J]. Med Educ Online, 2023, 28 (1): 2181052.

[3] Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine [J]. N Engl J Med, 2023, 388 (13): 1233-1239.

[4] 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.
Yan J Z, He Y X, Luo Z Y, et al. Generative large language models in the medical domain: potential and typical applications and challenges [J]. J Med Inform, 2023, 44 (9): 23-31.

[5] 肖仰华，徐一丹. 大规模生成式语言模型在医疗领域的应用：机遇与挑战 [J]. 医学信息学杂志，2023，44（9）：1-11.

Xiao Y H, Xu Y D. The application of large generative language models in the medical field: opportunities and challenges [J]. J Med Inform, 2023, 44 (9): 1-11.

[6] Betzler B K, Chen H C, Cheng C Y, et al. Large language models and their impact in ophthalmology [J]. Lancet Digit Health, 2023, 5 (12): e917-e924.

[7] 陈湘, 邓然, 吴川清. "ChatGPT+" 时代普通外科住培教育的探索与思考 [J]. 中国继续医学教育, 2024, 16 (12): 162-165.
Chen X, Deng R, Wu C Q. Exploration and reflection on general surgery standard training education of residents in the time of "ChatGPT+" [J]. China Contin Med Educ, 2024, 16 (12): 162-165.

[8] 袁欢欢, 全祉悦, 林祎, 等. 当医学课程考核遇上生成式人工智能: 价值、挑战与未来展望 [J]. 中国医学教育技术, 2023, 37 (3): 254-259.
Yuan H H, Quan Z Y, Lin Y, et al. Generative AI in medical education assessment: value, challenges, and future prospects [J]. China Med Educ Technol, 2023, 37 (3): 254-259.

[9] Pan G X, Ni J. A cross sectional investigation of ChatGPT-like large language models application among medical students in China [J]. BMC Med Educ, 2024, 24 (1): 908.

[10] 叶嘉欣, 罗林翼, 曾晖, 等. 住培医师执业医师资格考试考前培训模式的探索 [J]. 中国继续医学教育, 2023, 15 (6): 130-134.
Ye J X, Luo L Y, Zeng H, et al. Exploration of the pre-examination training mode for practicing physicians in the residency training [J]. China Contin Med Educ, 2023, 15 (6): 130-134.

[11] Boscardin C K, Gin B, Golde P B, et al. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity [J]. Acad Med, 2024, 99 (1): 22-27.

[12] Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review [J]. JMIR Med Educ, 2023, 9: e48785.

[13] Karabacak M, Ozkara B B, Margetis K, et al. The advent of generative language models in medical education [J]. JMIR Med Educ, 2023, 9: e48163.

[14] Jowsey T, Stokes-Parish J, Singleton R, et al. Medical education empowered by generative artificial intelligence large language models [J]. Trends Mol Med, 2023, 29 (12): 971-973.

[15] Lucas H C, Upperman J S, Robinson J R. A systematic review of large language models and their implications in medical education [J]. Med Educ, 2024, 58 (11): 1276-1285.

[16] Indran I R, Paranthaman P, Gupta N, et al. Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using Chat GPT [J]. Med Teach, 2024, 46 (8): 1021-1026.

[17] 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.
Jiang Z H, Feng S C, Wang W M. Exploratory practice of generative large language models in the construction of medical item banks [J]. Chin J Med Educ, 2024, 44 (8): 561-569.

[18] Stadler M, Horrer A, Fischer M R. Crafting medical MCQs with generative AI: a how-to guide on leveraging ChatGPT [J]. GMS J Med Educ, 2024, 41 (2): 1-5.

[19] Kiyak Y S, Kononowicz A A. Case-based MCQ generator: a custom ChatGPT based on published prompts in the literature for automatic item generation [J]. Med Teach, 2024, 46 (8): 1018-1020.

[20] Kiyak Y S, 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.

[21] Flores-Cohaila J A, García-Vicente A, Vizcarra-Jiménez S F, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: cross-sectional study [J]. JMIR Med Educ, 2023, 9: e48039.

[22] Mohammad B, Supti T, Alzubaidi M, et al. The pros and cons of using ChatGPT in medical education: a scoping review [J]. Stud Health Technol Inform, 2023, 305: 644-647.

[23] Jeyaraman M, K S P, Jeyaraman N, et al. ChatGPT in medical education and research: A boon or a bane? [J]. Cureus, 2023, 15 (8): e44316.

[24] 吴芳, 苏壮志, 孙峥, 等. 人工智能辅助教学在医学影像规范化培训中的应用及探索 [J]. 医学教育管理, 2023, 9 (6): 746-751.
Wu F, Su Z Z, Sun Z, et al. Application and exploration of the artificial intelligence assisted teaching in standardized training program for medical imaging [J]. Med Educ Manag, 2023, 9 (6): 746-751.

[25] Omiye J A, Gui H W, Rezaei S J, et al. Large language models in medicine: the potentials and pitfalls: a narrative review [J]. Ann Intern Med, 2024, 177 (2): 210-220.

[26] Armitage R C. ChatGPT: the threats to medical education [J]. Postgrad Med J, 2023, 99 (1176): 1130-1131.

(收稿: 2024-10-12 录用: 2025-01-06 在线: 2025-09-10)
(本文编辑: 李玉乐)