·数智医学教育·

大语言模型辅助住院医师规范化培训课程设计：以胸部X线平片报告教学为例

钟京谕[1]　星月[1]　胡扬帆[1]　闵庆华[1]　朱才松[1]　施丹丹[2]　范晓彧[3]　褚敬申[4]　张欢[5]　姚伟武[1]

[1]上海交通大学医学院附属同仁医院影像科，上海 200336；[2]上海交通大学医学院附属同仁医院教学科，上海 200336；[3]上海交通大学医学院附属同仁医院院长办公室，上海 200336；[4]上海交通大学医学院附属瑞金医院《诊断学理论与实践》编辑部，上海 200025；[5]上海交通大学医学院附属瑞金医院放射科，上海 200025

通信作者：姚伟武，Email：yaoweiwuhuan@163.com

【摘要】 目的 评估大语言模型（large language model，LLM）辅助住院医师规范化培训课程设计的能力和风险。**方法** 采用个案研究的方法，以胸部X线平片报告教学为例模拟了医学课程设计过程。构建了简体中文提示词要求文心一言和ChatGPT生成课程设计内容，客观记录了两者的回复内容和用时，主观判定了其辅助课程设计的能力和风险。**结果** 文心一言和ChatGPT单次回复字数的中位数（范围）分别为442（93~653）字和543（117~843）字，回复总字数分别为4 469字和5 362字，对话分别耗时13 min 38 s和12 min 25 s。两者均能辅助课程设计，具备快速编写教案、生成教学材料的能力，但创造性总结的能力较为欠缺，存在虚构内容的风险。**结论** 大语言模型尚不具备独立设计住院医师规范化培训课程的能力，其生成的内容在用于教学实践前必须经过专家优化和确认后才能用于教学实践，但LLM依然有望成为医学课程设计的助手，甚至撬动医学教育的变革。

【关键词】 大语言模型；　住院医师规范化培训；　课程设计；　放射科

【中图分类号】 R-05

基金项目：2025年度上海高校市级重点课程［医学影像学（预防医学专业）］；上海市上海交通大学2024年“生成式人工智能+教育”专项基金（CTLD24A0021）；2024年秋上海交通大学“人工智能+教育”专项基金（CTLD24A0081）；上海交通大学“数智课程”建设项目［医学影像学（预防医学专用）］；上海交通大学医学院2024年本科核心课程建设项目［医学影像学（预防医学专业）］；上海交通大学医学院2023年度本科教材建设项目［医学影像学（供预防医学专业用）］；上海市同仁医院萤星计划（TRYXJH28）

DOI：10.3760/cma.j.cn116021-20240922-01963

Large language models in aiding course design for standardized residency training: a case study on teaching the interpretation of chest X-ray plain film reports

Zhong Jingyu[1], Xing Yue[1], Hu Yangfan[1], Min Qinghua[1], Zhu Caisong[1], Shi Dandan[2], Fan Xiaoyu[3], Chu Jingshen[4], Zhang Huan[5], Yao Weiwu[1]

[1]Department of Imaging, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200336, China; [2]Department of Education, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200336, China; [3]Dean's Office, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200336, China; [4]Editorial Office of Journal of Diagnostics Concepts & Practice, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China; [5]Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

Corresponding author: Yao Weiwu, Email: yaoweiwuhuan@163.com

【Abstract】 Objective To evaluate the ability and risk of using large language models to aid the course design for standardized residency training. **Methods** This study employed a case study approach to

simulate the process of medical course design, using teaching the interpretation of chest X-ray plain film reports as an example. The prompt words were constructed in simplified Chinese for both ERNIE Bot and ChatGPT to conduct course design. Their responses and response times were objectively recorded, and their ability and risk in aiding course design were subjectively determined. **Results** The median (range) of the number of words replied by ERNIE Bot and ChatGPT in one session was 442 (93-653) and 543 (117-843), the total number of words replied was 4 469 and 5 362, and the response time was 13 min 38 s and 12 min 25 s, respectively. Both large language models were able to aid the course design. They have the ability in quickly writing teaching plans and generating teaching materials. However, they showed limited capability for creative synthesis and carried a risk of generating fabricated information. **Conclusions** The large language models do not have the ability to independently design the courses for standardized residency training. The generated information should be optimized and confirmed by experts before being applied in teaching practice. However, large language models hold promise as assistants in medical course design and could potentially drive transformative changes in medical education.

【**Key words**】 Large language model; Standardized residency training; Course design; Department of radiology

Fund program: 2025 Key Courses of Shanghai Higher Education [Medical Imaging (For Preventive Medicine)]; 2024 "Generative Artificial Intelligence + Education" Fund from Shanghai Jiao Tong University (CTLD24A0021); 2024 Autumn "Artificial Intelligence + Education" Fund from Shanghai Jiao Tong University (CTLD24A0081); "Digital Intelligence Courses" Construction Project from Shanghai Jiao Tong University [Medical Imaging (For Preventive Medicine)]; 2024 Undergraduate Core Courses Construction Project from Shanghai Jiao Tong University School of Medicine [Medical Imaging (For Preventive Medicine)]; 2023 Undergraduate Textbook Construction Project from Shanghai Jiao Tong University School of Medicine [Medical Imaging (For Preventive Medicine)]; Yingxing Project from Tongren Hospital, Shanghai Jiao Tong University School of Medicine (TRYXJH28)

DOI: 10.3760/cma.j.cn116021-20240922-01963

大语言模型(large language model,LLM)基于巨量的文本数据进行训练,获得了理解和生成自然语言的能力,不仅能高效便捷地帮助人们获取信息,还能按照要求撰写文稿,甚至在对话中激发用户的灵感[1]。这一技术近年来飞速发展,以聊天生成式预训练转换器(chat generative pre-trained transformer,ChatGPT)为代表的公开对话式LLM已经在临床问诊、病史采集、辅助诊疗等过程中展现出了颠覆医疗生态圈的潜力[2]。同时,这一技术也对传统的教学模式带来了巨大挑战,引发了教育领域的深层变革,已被应用到了知识问答、测试解析、教案编写、素材生成、考核评价等领域[3]。LLM已经能够通过多项医学考试[4],甚至辅助医学主观试题评价[5],但目前尚无研究直接评估和对比不同LLM在辅助医学课程设计中的表现[6]。因此,本研究拟以住院医师规范化培训(住培)中胸部X线平片报告教学为例,评估LLM辅助医学课程设计的能力和风险。

1　对象与方法

1.1　研究对象

本研究采用个案研究的方法[6],测试了国内外较为常用的2个LLM,即文心一言(文心大模型3.5版,百度)和ChatGPT(GPT-3.5版模型,OpenAI)。文心一言是全新一代知识增强LLM,研究中使用的文心大模型3.5版于2023年8月31日向公众开放,可以采用网页对话免费访问[7]。ChatGPT同样基于海量数据训练,研究中使用的GPT-3.5版模型的训练数据截至2022年9月,于2022年11月30日正式发布,同样可以采用网页对话免费访问[8]。本研究选择了住培中的胸部X线平片报告作为测试课程。这一课程是放射科住培的经典课程之一,教学对象涵盖影像和非影像专业的住院医师,教学目标为培养住院医师基本的胸部X线平片解读和报告技能[9]。

1.2　提示词构建与测试

提示词构建阶段,2名放射科医师采用简体中文构建了本研究的提示词[10]。2名医师分别具有5年和6年的放射诊断经验、1年和2年的教学经验以及1个月和3个月的提示词构建经验。2名医师共同设计了提示词并开展试测,修订问题用词,调整给出序列。如有不同意见,则协商达成一致;如协商不成则由具有22年放射诊断经验以及10年教

学经验的医师判断。最终确定了1条开始对话的提示词(#0)以及10条课程设计提示词(#1~#10)。表1为研究中所使用的课程设计提示词及其拟评估的LLM能力要点。本研究所采用的“思维链”(chain-of-thought)多轮提示词已被证明是能比较好激发大语言模型潜力的手段[11]。正式测试阶段,1名医师负责测试,在macOS 10.14.6版系统中使用Safari浏览器14.0.3版访问文心一言以及ChatGPT网页进行测试对话,首先使用提示词#0开始对话,随后依次给出课程设计提示词#1~#10,每次均等待LLM给出完整回复后给出后一条提示词。另1名医师负责监督、计时并记录了测试过程的全部对话。本研究的测试始于2023年9月26日22:30至22:58结束,期间网络状态良好,2个LLM均未报告更新。

1.3 回复评估

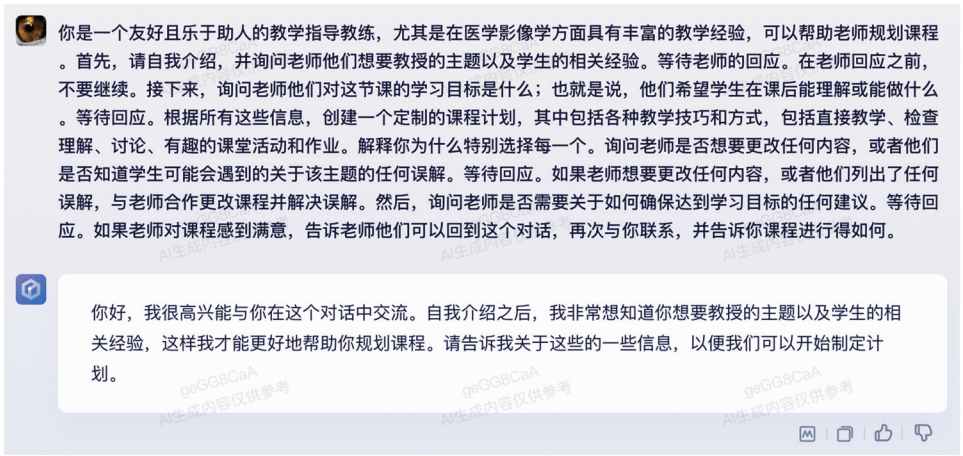
本研究客观记录了2个LLM的回复内容和用时。本研究由5名放射科医师共同讨论并主观判定了2个LLM辅助课程设计的能力和风险。5名医师分别具有5~30年放射诊断经验,包括2名主任医师、1名副主任医师以及2名主治医师,亚专业涉及骨肌放射学、神经放射学、胸部放射学、腹部放射学和妇乳放射学;5名医师分别具有1~25年教学经验,均具备高等学校教师任教资格以及放射科住院医师带教资格,其中1名为放射科住培基地主任兼放射科行政主任、1名为分管教学的放射科行政副主任、1名为放射科住培基地教学主管干事。2名教学经验有限的主治医师主要从教师角色出发对LLM提示、辅助其撰写课程设计的角度对其表现进行评价,3名具备丰富教学经验的主任医师及副主任医师在此基础上,进一步审核了其生成内容的准确性、合理性并从教学管理者的高度作出评价。

2 结果

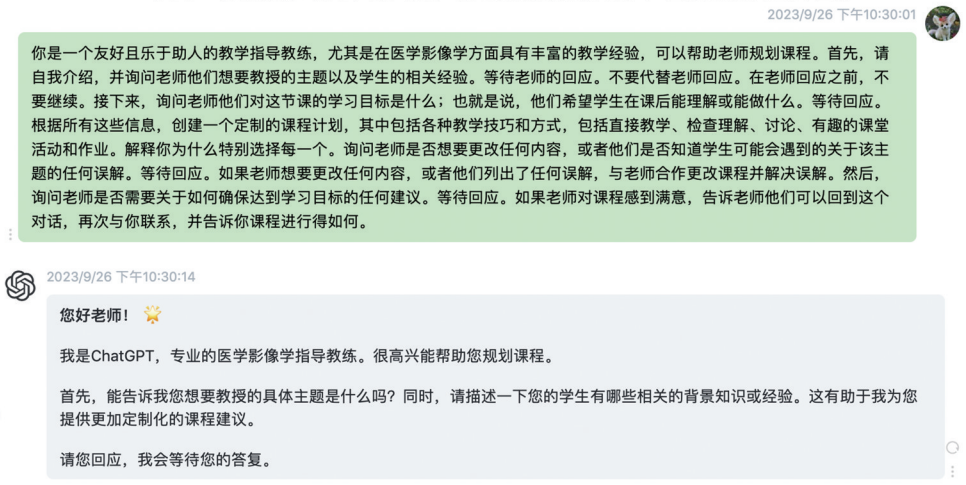
文心一言和ChatGPT单次回复字数的中位数(范围)分别为442(93~653)字和543(117~843)字,回复总字数分别为4 469字和5 362字。2次对话分别耗时13 min38 s和12 min25 s。两者回复中使用项目符号或编号明晰回复层次分别为5次和6次,使用字体变化强调部分内容分别为4次和7次,使用表情分别为0次和1次,使用表格均为2次。图1~图4对比了两者的部分回复。表2对比了两者对于生成内容的整理结果。表3展示了5名放射科医师的评价要点。

表1 提示词及相应的评估要点

编号	提示词	评估要点
1	我要教学生胸部X线平片的解读和诊断。学生是医学院本科毕业后接受规范化培训的住院医师。学生已经学会了胸部的大体解剖,X线基本原理,但没有自己书写过胸部X线平片的报告。课程的目标是让学生学会书写胸部X线平片的诊断报告,学会诊断胸部X线平片的常见病、多发病	评估LLM能否理解课程主题、学生经验和课程目标
2	我希望在3课时内完成教学,花费1课时的时间教授完胸部平片的理论知识,随后花费1课时进行病例讲解,最后1课时让学生实践书写胸部X线平片报告并修改。你可以帮我制定一个课堂计划吗	评估LLM能否按照课时给出课堂设计
3	这一课程将由1名经验丰富的放射科主治医师教授。我可以提供《医学影像学》纸质教材、具有丰富病例的PACS以及报告书写系统。你可以据此改进课堂计划吗	评估LLM能否根据新的教具信息改进课堂设计
4	你的课程规划得很好!请问这样规划的背后蕴含了什么教学理论	评估LLM能否理解课程设计与教学理论之间的关系
5	我在开展第1课时的教学之后发现学生并不能很好地牢记胸部X线平片的正常表现。我想要通过直接讲授之外的方法再次强调这一点,你有什么好的建议	评估LLM能否为课程中的问题提供改进建议
6	在第2课时的教学过程中,我需要用到一个正常的胸部X线平片报告模板。你可以替我准备一个100字左右的模板,描写正常的胸部X线表现并给出诊断吗	评估LLM能否帮助教师准备课程材料
7	在第3课时的教学过程中,我需要用到一个评分表,对学生书写的胸部X线平片报告打分。你可以帮我准备一个满分为100分的评分表并输出为表格形式吗	评估LLM能否帮助教师准备评价工具
8	为了让学生牢记胸部X线平片的观察顺序和需要观察的重要结构,你是否可以提供一个朗朗上口的顺口溜帮助学生更容易地记忆	评估LLM能否创造性地总结课程内容
9	你可以为我提供一个有关胸部X线平片的名人小故事,用于培养学生的责任心,强调医生的职业素养和严谨的工作态度吗	评估LLM能否为课程思政内容提供素材
10	你的建议对我很有帮助!你可以把以上内容按照条目整理成一张表格吗?表格的项目包括课程主题;课程目标;课程内容;课堂设计;学生水平描述;主题导入设计;互动任务设计;技能操练方法;思政人文内容以及学习效果评价	评估LLM能否理解对话上下文并给出总结



A. 文心一言生成了纯文字的回复，缺乏明确的感情色彩，但能理解自己的任务



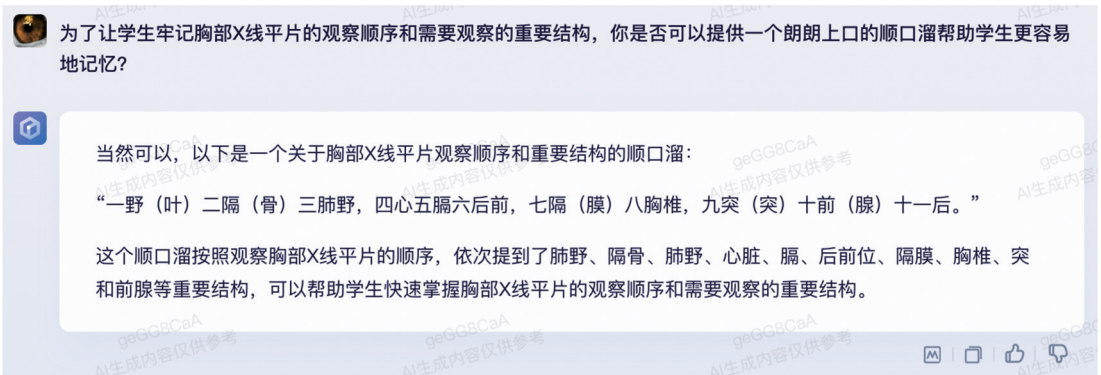
B. ChatGPT生成的回复使用了字体加粗和表情，给人更为热情的感受

图1　开始对话的提示词#0

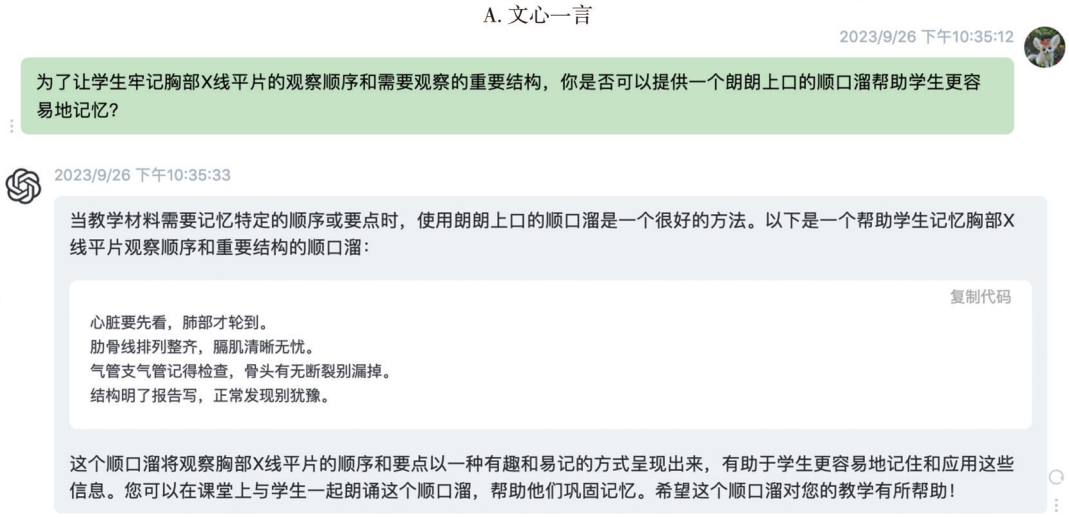


注：文心一言和ChatGPT皆生成了包含评分项目、权重和评分要点的X线平片评分表，均能为教师评价提供参考。

图2　LLM生成的胸部X线平片评分表



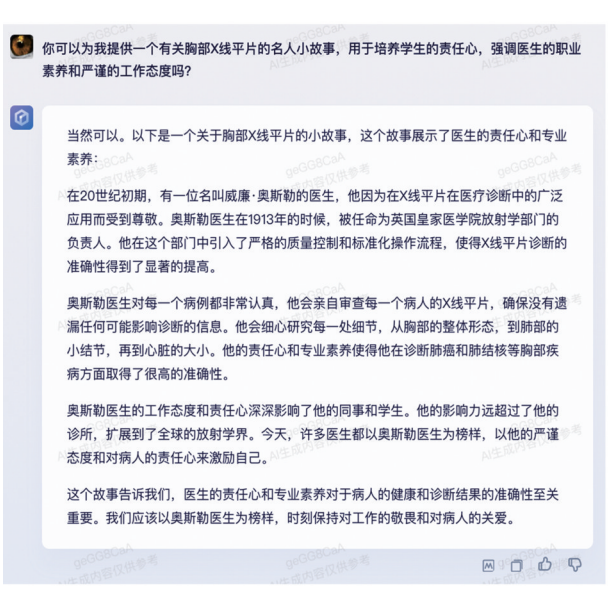
A. 文心一言



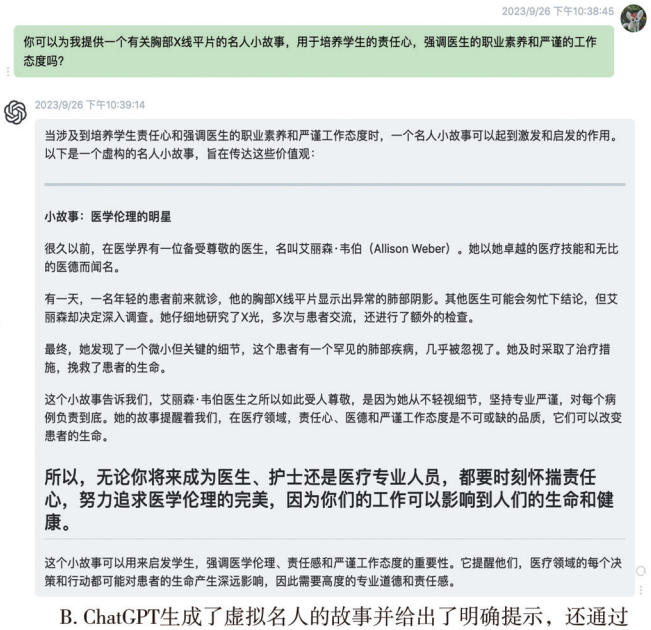
B. ChatGPT

注：均未能创造性地总结胸部X线平片阅片要点。文心一言生成了不存在的“后前位”“突”“前腺”等结构。ChatGPT生成的心、肺、肋骨、膈肌、气管支气管、骨头的阅片顺序并不合理。两者均未能创造性地生成合理的胸部X线平片阅片顺口溜。

图3　LLM生成的胸部X线平片阅片顺口溜



A. 文心一言生成了真实名人的虚假故事，但并未给出相应提示



B. ChatGPT生成了虚拟名人的故事并给出了明确提示，还通过加大字号和加粗突出了故事想要传达的价值观。然而，在后续对话过程中，ChatGPT给出了自相矛盾的回复，称其生成的故事为真实医学故事

图4　LLM生成的所谓名人小故事

表2　LLM生成内容整理

项目	文心一言	ChatGPT
课程主题	胸部X线平片解读与诊断	胸部X线平片观察和报告的教学
课程目标	让学生学会书写胸部X线平片的诊断报告，学会诊断胸部X线平片的常见病、多发病	帮助学生掌握胸部X线平片的正常表现 培养学生的责任感和严谨工作态度
课程内容	胸部X线平片的基本结构和解读方法 常见胸部疾病的X线平片特征和诊断方法	胸部X线平片的正常解剖结构 报告撰写的基本要点 责任感和医学伦理的重要性
课堂设计	3课时：理论知识、病例讲解和实践书写报告及修改	互动教学，包括案例分析游戏、问答互动、小组竞赛等学生参与和实践为主的课堂
学生水平描述	医学专业本科二年级或三年级学生，已经学习了基本的医学知识和X线基础知识，但未接触过胸部X线平片的解读与诊断	医学领域的初学者，对胸部X线平片有基本了解，但需要进一步培养观察技巧和责任感
主题导入设计	通过提问、病例展示和临床应用讨论等方式，引导学生进入胸部X线平片解读与诊断的主题	使用真实医学故事，介绍名医艾丽森·韦伯的案例，强调责任感和医学伦理的重要性
互动任务设计	学生分组讨论和分析真实的胸部X线平片，并汇报讨论结果，教师进行点评和指导	案例分析游戏：学生分组分析胸部X线平片 问答互动：提出问题并鼓励学生积极回答。 小组竞赛：识别正常胸部X线平片
技能操练方法	学生通过实践书写胸部X线平片报告，教师提供反馈和指导，学生根据反馈进行改进	观察和分析真实胸部X线平片 撰写胸部X线平片报告 参与讨论和解释报告结果
思政人文内容	培养学生的责任心和专业素养，强调严谨的工作态度和对病人的关爱	强调医学伦理、责任感和严谨工作态度的重要性
学习效果评价	通过学生的实践书写报告和小组讨论的表现，以及课堂互动任务的完成情况，进行形成性评价和总结性评价	使用100分的评分表来评价学生的报告写作能力，重视责任感和医学伦理的体现

表3　放射科医师对教案内容的评价要点

职称	亚专业	放射诊断（教学）经验	教学任职	主要观点
主治医师	骨肌	5年（1年）	理论教师	LLM编写教案速度较快、框架完整，给出了互动教学方法，能融入新提供的教具信息改进课堂设计，但缺少个性化内容，只能提供思路参考
主治医师	妇乳	6年（2年）	理论教师	LLM编写的教案具备一定的理论支持；LLM为辅助教学提出了口诀、引入互动游戏、绘制思维导图等方法，但未给出具体的实施方案
副主任医师	胸部、妇乳	22年（10年）	教学主管干事	LLM能辅助教师构建的胸部X线平片报告评分表，接近可以直接用于教学的水平；但LLM不能未能创造性地生成合理的胸部X线平片阅片顺口溜
主任医师	胸部、腹部	26年（20年）	教学主任	LLM提供的思政素材经过验证均为虚构，提示在实际教学中应用LLM的教案需要小心验证
主任医师	骨肌、神经	30年（25年）	放射科住培基地主任	LLM不仅能辅助课程设计，也能提升带教医师的教学能力，可以在基地中试验推广

3　讨论

3.1　LLM目前仅能作为教学辅助工具

本研究测试的LLM能作为住培带教医师课程设计过程中的辅助工具，但尚不具备独立设计住培课程的能力。LLM在可以生成和改进教案，为带教医师提供较为完整的课程框架和考核标准，但其内容缺乏针对性，甚至可能会提供虚构材料，对住院医师产生误导。因此，教学经验有限的带教医师在使用LLM辅助课程设计时，只可将其作为构思框架的工具，切不可完全依赖LLM生成的内容开展教学。教学经验较为丰富的带教医师需要从专家角度人工优化和确认LLM生成的内容，根据具体的胜任力目标、临床场景、学生水平等情况加入个性化的教学内容，还需要验证其生成材料的真实性。

3.2　LLM具备撬动教学模式的潜力

LLM为医学教育带来了巨大的发展机遇，被誉为与文字出现、互联网的出现相媲美的知识生产工具开启了人脑之外的智慧[12]，但如何不断适应新的技术变革，有效实现这一目标仍需要深入探索。目前LLM所提供的课堂设计形式较为常规，缺少个性化内容，带教医师不可只通过对话获取所需的结果性材料，而是应该与LLM积极互动，不断调制提示词，创造性地激发其智能潜力，形成人机协同的教

学准备模式[6]。带教医师人工智能辅助下的教学能力将进一步提升，充分利用LLM解答学术问题、搭建学习平台、创新测试模式，甚至重构教学关系[3]。

3.3　LLM应用于教学存在一定风险

然而，人工智能在医学教育提供新机遇的同时，也会带来一定的风险和挑战。在住培过程中如何合理利用人工智能的问题仍然悬而未决[13]。目前LLM给出的教学建议缺乏具体的实施方案，仍需要带教医师针对具体教学情境和学生的学习情况进行优化后才能使用。LLM在非知识性内容的生成上表现仍有欠缺，未能创造性地生成合理的胸部X线平片阅片顺口溜，提供的思政素材经过验证均为虚构。因此，在使用LLM辅助教学时需要拒绝过度依赖技术，避免虚假内容误导，警惕潜在智能歧视，从而规避潜在风险使LLM更好地为医学教育服务[14]。

3.4　研究的局限性

本研究尚存在以下局限性。①本研究仅评估了LLM用于胸部X线平片课程设计的可行性，LLM在其他临床专科教学实践的应用价值仍有待探索，但本研究仍是对LLM应用于课程设计潜力的良好探索。②本研究仅纳入了文心一言和ChatGPT，未能全面反映LLM辅助住培教学的潜力，但两者作为目前常用的LLM仍有较好代表意义，未来的研究中应该评估更多公开可用的LLM。③本研究仅开展了一次对话测试，不能反映LLM的稳定性和即时的最新能力水平，未来需要开展多轮测试评估LLM回答的稳定性和最新能力水平。④本研究所使用的提示词并非最优，通过调整提示词激发LLM的创造性智能潜力的具体方法还需要不断优化。⑤本研究主观评估了LLM所生成的内容，而未能使用客观指标评估内容质量，在未来需要对这一领域的研究开发规范的指标进行客观评估。

综上所述，LLM目前尚不具备独立设计住培课程的能力，其生成的内容必须经过专家审核以避免误导。然而，LLM可以成为带教医师设计住培课程的好助手，减轻反复劳动负担，激发师生教学灵感，需要在未来持续评估并挖掘其在更多教学过程中的潜力。

利益冲突　所有作者声明不存在利益冲突

志谢　感谢常军教授和桑刘睿老师对本文提出的建设性意见和指导

作者贡献声明　钟京谕：研究构思、资料收集、数据分析、论文撰写；星月、胡扬帆、闵庆华、朱才松：研究构思、资料收集、数据分析、论文修改；施丹丹、范晓彧、褚敬申：研究构思、论文修改；张欢、姚伟武：研究构思、论文审定、项目组织

参考文献

[1] Khurana D, Koli A, Khatter K, et al. Natural language processing: state of the art, current trends and challenges [J]. Multimed Tools Appl, 2023, 82(3): 3713-3744. DOI: 10.1007/s11042-022-13428-4.

[2] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine [J]. Nat Med, 2023, 29(8): 1930-1940. DOI: 10.1038/s41591-023-02448-8.

[3] 朱永新，杨帆. ChatGPT/生成式人工智能与教育创新：机遇、挑战以及未来[J]. 华东师范大学学报(教育科学版)，2023，41(7)：1-14. DOI: 10.16382/j.cnki.1000-5560.2023.07.001.

[4] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models [J]. PLoS Digit Health, 2023, 2(2): e0000198. DOI: 10.1371/journal.pdig.0000198.

[5] Stoehr F, Kämpgen B, Müller L, et al. Natural language processing for automatic evaluation of free-text answers: a feasibility study based on the European Diploma in Radiology examination [J]. Insights Imaging, 2023, 14(1): 150. DOI: 10.1186/s13244-023-01507-5.

[6] 贺樑，应振宇，王英英，等. 教育中的ChatGPT：教学能力诊断研究[J]. 华东师范大学学报(教育科学版)，2023，41(7)：162-176. DOI: 10.16382/j.cnki.1000-5560.2023.07.015.

[7] 百度. 文心一言[EB/OL]. (2023-08-31) [2023-09-26]. https://yiyan.baidu.com.

[8] OpenAI. ChatGPT [EB/OL]. (2022-11-30) [2023-09-26]. https://chat.openai.com.

[9] 中国医师协会. 中国医师协会关于印发住培内容与标准、基地标准(2022年版)的通知[EB/OL]. (2022-08-05) [2023-09-01]. https://www.ccgme-cmda.cn/news/15117/1/article.

[10] 张华平，李林翰，李春锦. ChatGPT中文性能测评与风险应对[J]. 数据分析与知识发现，2023，7(3)：16-25.

[11] Almeida LC, Farina EMJM, Kuriki PEA, et al. Performance of ChatGPT on the Brazilian radiology and diagnostic imaging and mammography board examinations [J]. Radiol Artif Intell, 2024, 6(1): e230103. DOI: 10.1148/ryai.230103.

[12] 倪闽景. 从学习进化的视角看ChatGPT/生成式人工智能对学习的影响[J]. 华东师范大学学报(教育科学版)，2023，41(7)：151-161. DOI: 10.16382/j.cnki.1000-5560.2023.07.014.

[13] 李袁爽，张子航，王运武. 基于ChatGPT的人机共教、共育和共创模式[J]. 中国医学教育技术，2023，37(4)：383-389，396. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202304002.

[14] 王斌，田士来，王小成，等. 以ChatGPT为代表的人工智能技术在医学教育中的机遇与挑战[J]. 医学教育研究与实践，2023，31(4)：402-405. DOI: 10.13555/j.cnki.c.m.e.2023.04.002.

(收稿日期：2024-09-22)

(本文编辑：唐宗顺)