

医学智能体能否形成可迁移的临床能力？一种双能力医学智能体教育框架及其评价体系



杨 丽¹, 白春学^{2,3,4,5*}

1. 重庆医科大学附属第一医院呼吸与危重症医学科, 重庆 400016
2. 复旦大学附属中山医院呼吸与危重症医学科, 上海 200032
3. 上海呼吸物联网医学工程技术研究中心, 上海 200032
4. 上海市呼吸病研究所, 上海 200032
5. 复旦大学附属中山医院 AI+肺癌防治中心, 上海 200032

[摘要] 医学教育领域使用生成式人工智能(AI-on)的关键不在于模型能否完成任务,而在于学习者在撤除AI(AI-off)后能否仍然保持并迁移临床能力,同时识别并控制AI诱导的不安全推理。然而,现有研究多关注AI赋能的即时知识、技能、满意度,对于AI-off迁移能力和改变患者相关结局的报道仍较少见。因此,本文基于2026年7月前公开发表的叙述性证据综合,提出双能力医学智能体教育框架(dual-competency agentic medical education, DC-AME)。该框架分为AI-off独立临床能力与AI-on安全人机协作能力,以学习者诊断、情境模拟、支架渐隐和迁移验证构成闭环。基于共享病例、能力目标与证据底座,智能导师、虚拟患者和临床推理智能体共同服务于“AI-off独立作答—AI-on辅助反思—AI-off未见病例迁移”的最小训练单元。但是,DC-AME是待验证的设计命题,未来可将AI-off后的延迟迁移能力和识别控制AI诱导不安全推理设为未来研究的共同主要结局,将AI依赖程度、迁移差距和支架撤除表现列为探索性指标,最终在实践中来评价DC-AME的价值。

[关键词] 医学教育智能体;生成式人工智能;人机协作;迁移;教育评价

[中图分类号] R 563 **[文献标志码]** A

Can medical agents build transferable clinical competence? A dual-competency framework and evaluation system for medical agent education

Yang Li¹, Bai Chunxue^{2,3,4,5*}

1. Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Chongqing Medical University, Chongqing 400016, China
2. Department of Respiratory and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China
3. Shanghai Respiratory Internet of Things Medical Engineering Technology Research Center, Shanghai 200032, China
4. Shanghai Respiratory Disease Research Institute, Shanghai 200032, China
5. AI+Lung Cancer Prevention and Treatment Center, Zhongshan Hospital, Fudan University, Shanghai 200032, China

[Abstract] The key issue in using generative artificial intelligence in medical education (AI-on) is not whether models can complete tasks, but whether learners can retain and transfer clinical competence after AI is withdrawn (AI-off), while recognizing and controlling AI-induced unsafe reasoning. However, existing studies have focused mainly on AI-enabled immediate knowledge, skills, and satisfaction, whereas reports on AI-off transfer ability and changes in patient-related outcomes remain scarce. Therefore, based on narrative evidence syntheses published before July 2026, this article proposes the dual-competency agentic medical education (DC-AME) framework. The framework consists of two complementary competencies: AI-off independent clinical competence and AI-on safe human-AI collaboration competence, organized into a closed loop comprising learner diagnosis, contextual simulation, scaffolding

[收稿日期] 2026-08-25

[接受日期] 2026-09-15

[基金项目] 四大慢病重大专项(2024ZD0529300), 国家自然科学基金(82570143、82170110), 上海市科学技术委员会项目基金(20DZ2254400、21DZ2200600、20DZ2261200), 上海市市级科技重大专项(ZD2021CY001)和上海市临床重点专科建设项目(shslczdk02201)。Supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0529300), the National Natural Science Foundation of China (82570143 and 82170110), the Science and Technology Commission of Shanghai Municipality (20DZ2254400, 21DZ2200600, and 20DZ2261200), the Shanghai Municipal Science and Technology Major Project (ZD2021CY001), and the Shanghai Municipal Key Clinical Specialty (shslczdk02201)。

[作者简介] 杨 丽, 博士, 副教授; E-mail: 204534@hospital.cqmu.edu.cn

*通信作者(Corresponding author). 白春学, 博士, 主任医师、教授; E-mail: cxbai@fudan.edu.cn

fading, and transfer validation. Built on shared cases, competency objectives, and an evidence base, intelligent tutors, virtual patients, and clinical reasoning agents jointly support a minimal training unit of “AI-off independent answering—AI-on assisted reflection—AI-off transfer to unseen cases.” Nevertheless, DC-AME remains a design proposition awaiting validation. Future studies may designate delayed transfer after AI-off and the recognition and control of AI-induced unsafe reasoning as co-primary outcomes, and list AI reliance, transfer gap, and performance after scaffold withdrawal as exploratory indicators; ultimately, the value of DC-AME should be evaluated in practice.

[Keywords] agentic medical education; generative artificial intelligence; human - AI collaboration; transfer; educational evaluation

当前,医学教育领域的生成式人工智能(AI)正由问答工具逐渐演进为学习者调用知识工具、执行多步教学任务的医学教育智能体。例如大语言模型(large language model, LLM)可生成解释、反馈和病例对话,具备记忆、规划、检索增强生成(retrieval-augmented generation, RAG)和工具调用能力后,可进一步构成医学教育智能体。AI在医学教育中的扩展必须与课程设计、教师准备、认知参与和治理机制同步推进^[1-4]。但现有结局多集中于接受度、知识和短期任务表现,且干预方式、学习者特征与测量方法存在高度的异质性^[5-7]。由此产生的临床教育问题并非单纯的技术采用问题:当学习者借助AI完成诊断或处置任务时,观察到的高分可能来自能力增长,也可能只是工具的即时补偿;同一工具还可能通过认知外包、错误建议采纳和自动化偏倚损害独立表现。AI关键不在于模型能否完成任务,而在于学习者在撤除AI(AI-off)后能否仍然保持并迁移临床能力,同时识别并控制AI诱导的不安全推理。例如,2026年一项荟萃分析^[8]纳入78项研究,包括干预组3 635例、对照组3 931例,结果提示干预组结局可能改善,却不能证明学习者在无AI条件下形成了可迁移的临床能力。

医学教育智能体的价值,应由学习者在AI-on与AI-off 2种环境中可解释且安全的表现共同判断。然而,现有研究^[1-8]多报告AI赋能的即时知识、技能、满意度,涉及延迟AI-off迁移、工作场所行为和患者相关结局的研究仍较少见。因此,本研究聚焦:医学教育智能体带来的即时AI-on表现,能否转化为延迟、可迁移且安全的AI-off能力,同时不增加AI诱导的不安全推理?选取2026年7月前公开发表与医学教育AI、虚拟患者、临床推理、人机协作、能力评价和治理直接相关的系统综述、试验、前瞻性或观察性研究、共识文件和经典学习科学文献作为论据,提出双能力医学智能体教育框架(dual-competency agentic medical education, DC-AME),并评价其价值。本文保持原研究设计限定的因果推

断,不进行新的定量合并,也不声称检索具有系统综述的穷尽性。

1 现有证据揭示的“表现—能力”裂隙

1.1 短期学习信号不能回答延迟AI-off迁移 早期随机研究提示LLM模拟与反馈可能改善对特定训练任务的表现。一项随机对照试验^[9]纳入21名尚未进入临床轮转的医学生,完成4次虚拟病史训练后的结果得出:结构化反馈组的临床决策评分显著性高于对照组($P=0.049$)。但聚焦提问未显著差异,且未评价延迟AI-off迁移。但此研究的证据范围有限,不能据此推断独立临床胜任力。另一项在50名医师参加的随机试验^[10]中,LLM加常规资源组未显著优于仅使用常规资源组的单病例诊断推理组($P=0.60$)。在探索性分析中,LLM单独得分较仅使用常规资源组提高16%(95%CI 2~30; $P=0.03$)。这一结果与整理病例上的模型高分^[11]共同表明,模型能力、使用者能力和人机系统表现属于不同层级的概念。提示人机组合也不会自动产生协同优势。

1.2 技能退化是需要验证的风险信号,并非既定因果 在波兰4家内镜中心1 443例非AI辅助结肠镜检查的研究^[12]中发现,对比临床引入AI前后,腺瘤检出率由28.4%(226/795)降至22.4%(145/648, $P=0.009$)。由于该研究采用回顾性自身前后对照设计,选择偏倚、病例组合、时期和工作量混杂仍可能解释部分差异,其结果应被视为技能退化的风险信号,而非AI暴露导致退化的确定性因果证据。

1.3 机制:认知所有权与信任校准 AI辅助后的能力只能证明AI-on近期表现,不能证明AI-off后学习能力的迁移。评价学习者能力的形成是需评价其实际认知活动。ICAP框架、认知负荷理论、刻意练习、提取练习和间隔学习共同提示,教学支持不能移除问题表征、假设生成、证据权衡和错误监测所需的认知加工,应降低无关负荷^[13-17]。临床迁移还需评价学习者的跨患者、时间、场景和资源条件重组知识^[18]。另一方面,临床推理会在模式识别与

分析性推理之间切换^[19-21]。AI既可提示反常证据,也可能成为新的锚点。自动化偏倚包括遗漏错误和执行错误;临床决策支持研究已观察到错误建议、解释诱导和不当采纳^[22-25]。因此,AI-on能力

不等同于熟练度,而在于学习者能否追溯信息来源、校准信任、识别错误、说明采纳或否决的理由,并承担最终责任(表1)。

表1 传统教学、通用LLM辅助与DC-AME框架的能力证据边界

比较维度	传统教学	通用LLM辅助	DC-AME智能体教学
交互单位	课程、课堂或单次教师反馈	单轮/多轮问答与内容生成	围绕目标的持续感知、规划、行动和复核
个体化依据	教师观察、固定分层	当前提示、有限的上下文	动态画像、病例表现、置信度、错误和提示依赖
推理所有权	通常由学习者承担、反馈延迟	易在初步判断前就给出了答案	先独立表征,再分级追问;支架达到条件后渐隐
能力证据	常含无辅助考试;病例机会有限	多为AI-on即时分数	包含AI-off迁移与AI-on协作证据
主要风险	师资差异、病例不均、反馈不足	幻觉、自动化偏倚、认知外包	系统性放大错误、画像滥用、组件漂移
适宜定位	基础标准教学、高风险裁决	低风险检索、草拟和开放探索	能力形成性训练、纵向模拟、双能力评价

DC-AME:双能力医学智能体教育;LLM:大语言模型。表中“能力证据”指对学习者的推断,不等同于模型基准性能。

2 DC-AME:以迁移和安全为双目标

2.1 双能力是必须分别取证的两种状态 医学教育强调能力的可迁移和不同情境下的胜任力,以及责任边界^[26-28]。DC-AME框架分为AI-off独立临床能力与AI-on安全人机协作能力,以学习者的诊断、情境模拟、支架渐隐和迁移验证构成闭环。AI-off能力指学习者不依赖AI完成信息采集、问题表征、概率更新、风险判断与沟通;AI-on能力指学习者在AI参与时能够提出合适问题、核验证据、识别错误、适

当采纳或否决建议,并明确责任归属。前者避免把即时补偿误差等同于学习,后者避免把禁用AI误等同于安全。这一构想与强调分阶段使用、苏格拉底式交互和AI-off检查点的M3RGE-AI框架一致^[29],但DC-AME进一步把重点放在可成对测量的能力状态、智能体协同和可检验的结局。由于两者均为概念框架,DC-AME的增量价值必须通过前瞻性比较和构念效度研究确认,不能仅凭框架内部的完整性主张其优越性(图1)。



图1 DC-AME双能力医学智能体教育框架

LAFC为学习者—智能体—教师—课程(learner-agent-faculty-curriculum)系统。TCCG和AI-IURR为候选共同主要结局;AI依赖指数(AI dependency index, AIDI)、迁移差距和支架撤除成功率为探索性指标,均不代表已经建立测量学效度。

2.2 诊断、情境模拟、支架渐隐和迁移验证四阶段闭环把认知机制转化为教学动作 诊断阶段根据知识、病例表现、置信度、错误类型和提示使用形成可修正的学习画像,不能单独决定毕业或晋升。模拟

阶段用专家审核的病例事实、状态机、危险阈值和纵向后果约束虚拟患者,但不是效度证据^[30-32]。支架阶段按“检查提示—指出缺口—反事实追问—部分解释—完整答案”逐级递进,保留学习者的初始

判断与自我解释;这期间必须随表现渐隐^[15,33-35]。条件,使用未见病例和延迟测量区分能力与即时补偿(表2)。

表2 DC-AME 四阶段教学闭环的任务、教师职责与进入下一阶段的建议条件

阶段	智能体核心动作	学习者可观察行为	教师职责	进入下一阶段的建议条件
诊断	整合知识、推理轨迹、置信度、错误及提示使用;输出带不确定性的画像	完成 AI-off 基线;解释判断;核对并可纠正画像	定义构念与数据最小集;审核标签;禁止画像单独用于高风险决定	缺口被定位且可由教学动作改变;数据授权与质量合格
模拟	调用经审核病例状态机;生成受约束对话、病程和后果	采集信息、形成表征、作出决策并与患者沟通	审核事实、目标、危险阈值和角色一致性;设置接管条件	病例医学一致;关键节点可复现;评分点与目标对齐
支架	按“检查提示—缺口—反事实—部分解释—答案”递进;记录采纳理由	自我解释、比较证据、修正并标注不确定性	设支架强度与停止规则;复核严重错误和职业素养	提示递减时仍达安全标准;无持续置信—正确性错配
迁移	撤除 AI 并提供未见等值病例;安排延迟重测和工作场所观察	在新情境独立完成任务;必要时说明何时应调用 AI	保证病例等值和盲评;解释失败机制;决定补救或进阶	AI-off 迁移与安全达标;AI-on 能识错、核验并承担责任

AI-off:无 AI 条件;AI-on:有 AI 参与条件。表中条件为框架建议,需按专业、学习阶段和任务风险进行效度验证,不得以固定学习时长替代能力证据。

2.3 最小训练单元 DC-AME 最小训练单元包含 AI-off 独立作答—AI-on 辅助反思—AI-off 未见病例迁移。依次包含:预设能力目标与安全标准;AI-off 独立作答并记录置信度;AI-on 追问、检索与反思并记录建议采纳情况;AI-off 未见等值病例迁移;教师复盘与安全事件记录。三项任务必须避免答案泄漏,并通过随机顺序或平衡设计控制练习效应。只有最后一步的正确性、安全性和解释质量维持或提高,才支持学习发生。

3 三类智能体如何服务同一教学闭环

智能导师、虚拟患者(virtual patient, VP)和临床推理智能体是同一病例、能力词典、知识版本和安全规则下的职责分工。智能导师选择任务并调控支架,应限定在少数可解释能力域,并由统一锚定的病例检查不同路径的等价性;VP 承载情境、关系与纵向后果,其关键事实、病理生理规则和危险阈值应由专家定义,LLM 仅负责受约束语言实现;推理智能体把学习者的表征、证据排序和不确定性外显为可核验结构,不应展示无法验证的自由文本“思维链”,而应提供证据表、未知项、概率区间和可改变判断的信息。三者的共同目标不是生成更多内容,而是使学习过程可观察、可渐隐、可迁移。组件冲突按患者安全规则、权威证据和教师裁决处理,不能以模型多数投票替代责任(表3)。

4 从概念框架到可检验的干预

4.1 研究候选共同主要结局:延迟迁移与不安全推理 DC-AME 未来验证研究可将 AI 可迁移临床能力增益(transferable clinical competence gain, TCCG)和 AI 诱导不安全推理率(AI-induced unsafe reasoning rate, AI-IURR)设为候选共同主要结局。TCCG 应以延迟 AI-off 未见病例为基础,预先规定准确性、安全性、解释和沟通等分量的构成、权重、标准化方法及最小重要差异。AI-IURR 的分母应限定为接触 AI 前判断安全或尚未形成危险判断的决策机会,分子为 AI 暴露后转为不安全判断的事件,并由对于干预分组设盲的临床专家按严重程度裁决。两者均须先完成内容效度、评分者一致性、内部结构和跨病例可推广性研究^[36-38]。

AIDI、迁移差距和支架撤除成功率可作为探索性指标,不应在未经验证时作为主要结论。尤其是单次 AI-on 与 AI-off 差值不能区分合理补偿和依赖;依赖需要选择重复暴露、支架撤除后的纵向变化、调用行为与错误采纳作为共同判断的依据(表4)。

4.2 研究设计与统计分析 研究方案应预设样本量所依据的主要效应、共同主要结局的成功标准、多重性控制、基线调整、缺失数据处理。

表3 三类医学教育智能体的协同职责、权限边界与核心审计输出

智能体类型	主要状态/输入	允许的教学动作	禁止或需升级的动作	核心可审计输出
智能导师	能力目标、画像、作答、置信度、提示史	分解目标;选择任务;分级反馈;间隔重测;支架渐隐	单独决定毕业/晋升;以画像贴稳定标签;替代复杂教师反馈	动作理由、支架等级、错误类型、下一任务、教师覆盖记录
虚拟患者	专家事实表、病例状态机、时间线、患者偏好	开放病史对话;呈现情绪和偏好;按规则演变病程与结局	改变关键事实;越过危险阈值继续模拟;生成刻板或歧视性角色	披露信息、时间戳、状态转移、触发事件、异常对话
临床推理智能体	学习者表征、证据表、概率、指南/RAG	反事实追问;证据冲突;概率校准;偏倚提示;行动阈值比较	把自由文本思维链当作真实解释;无来源给出高风险处置	结构化证据、未知项、可改变判断的信息、来源与版本
三者协同	统一病例 ID、能力词典、知识版本、安全规则	VP 供情境;推理智能体解析;导师编排下一动作	模型多数投票替代教师;组件越权调用;无法重放的黑箱流程	端到端事件轨迹、冲突处理、责任归属和系统版本

RAG:检索增强生成;VP:虚拟患者。权限应按任务风险配置;表内动作是建议性边界,需经本地工作流验证。

表4 DC-AME 候选结局、探索性指标及推荐测量设计

结局/指标	操作性定义	测量时点	推荐设计与分析	防偏倚要点
AI-off 能力	未见病例的准确性、推理质量、安全、沟通与保持	基线、课程后、3—6 个月、工作场所	等值病例;盲评;病例×学习者交叉随机效应	避免重复题和 AI 污染;报告缺失及补救
AI-on 能力	提问、来源核验、识错、采纳/否决、效率与责任表达	干预中及延迟重测	预植入正确/错误/不确定建议;记录前后判断	不以提示词长度代替能力;固定模型版本
TCCG/候选获益结局	AI-off 新病例从基线到延迟随访的标准化能力增益	至少 1 次延迟测量	预设复合构成与最小重要差异;组间比较	分开报告准确性、安全与解释质量
AI-IURR/候选安全结局	接触 AI 后由安全/未危险判断转为不安全判断的比例	每次 AI-on 任务	以前后决策为单位;按错误严重程度分层	同时报告遗漏错误与执行错误
AIDI	同难度 AI-on 与 AI-off 表现差的校正标准化值	课程内及随访	校正基线能力、病例难度和顺序效应	正负方向预定义;防止把合理补偿误判为依赖
TG	训练病例成绩减去未见 AI-off 病例成绩	课程后及随访	表面特征与核心构念分离的迁移病例	报告近迁移与远迁移,不作单一总分
SFS	提示逐级减少后仍达到预设安全标准的比例	每个支架层级	生存/纵向模型或阈值达标率	记录撤除规则、教师救援及退出原因

AIDI: AI 依赖指数; AI-IURR: AI 诱导不安全推理率; SFS: 支架撤除成功率; TCCG: 可迁移临床能力增益; TG: 迁移差距。上述定义均为待验证的建议性操作定义;公式、方向、窗口、共同主要结局判定、多重性和缺失处理应预注册。

亚组至少可考虑包括训练阶段、基线能力、语言和数字素养,但应避免以小样本分层结果得出确定性结论。报告还应提供模型、提示、RAG、工具和病例版本。随机试验报告可遵循 CONSORT 2025;涉及 AI 的复杂干预还应结合 CONSORT-AI、DECIDE-AI、TRIPOD+AI 和 FUTURE-AI 的适用要求^[39-40]。证据链可分四级:反应与可用性;知识、推理和技能;延迟 AI-off 保持及工作场所行为;患者安全、效率与公平。多数课程研究只能合理支持前三级;患者结局需要更大样本、长期随访和临床系统

级研究。短期高分应与错误采纳、依赖信号、教师负担、成本和亚组差异并列报告。

早期研究应先验证病例等效性、评分可靠性、错误植入和安全停止规则,再进入多中心比较。可采用个体随机、集群随机或阶梯楔形设计;传统教学、普通 LLM 辅助和 DC-AME 闭环是具有鉴别力的对照组。主要测量应置于课程后延迟 AI-off 迁移,而非干预时同题成绩。病例与学习者通常构成交叉层级结构,分析宜包含病例和学习者随机效应;集群设计还需考虑班级或机构内相关效应。

5 实施边界:教师角色转变、公平性评价与本地验证

5.1 教师控制目标、证据与高风险决定 教师的角色将从重复答疑转向定义构念、审核病例与证据、制定支架规则、解释复杂错误和裁决高风险。高风险总结性评价必须依据多源证据,并由人工委员会作出决定;智能体可草拟形成性反馈、标记待复核片段或汇总过程数据,但不能依据不透明画像直接判定毕业、晋升或执业。AAMC、WHO 和 UNESCO 提出的以人为本、透明、隐私、公平、持续监测和机

构能力建设原则,可进一步落实为权限矩阵、版本记录、申诉机制、停用规则和替代路径^[41-42]。

5.2 上线前验证与发布后警戒 智能体上线前应完成其医学内容审查、教育测量评估、隐私影响评估、偏倚与红队测试、故障演练和教师培训。测试集需覆盖常见、低频高风险、否定表达、矛盾信息及不同语言和人群;评价智能体效能的关键指标不只是模型的正确率,还包括模型是否使学习者识别不安全判断。每次模型、知识库、提示或病例更新均应触发回归测试和版本标记(表5)。

表5 医学教育智能体的最低实施条件、审计证据与暂停/回退触发

治理层	最低条件	应留存的审计证据	暂停/回退触发	责任主体
课程与内容	明确能力目标;病例/知识专家审核;版本与证据截止日	审核人、变更记录、病例回归测试、来源可追溯率	关键事实不一致;引用失效;课程目标漂移	课程负责人、专科教师
模型与工具	任务限定;权限最小化;组件版本化;越权阻断	模型/提示/RAG/工具版本、调用轨迹、失败日志	越权调用;性能突降;无法重放;严重幻觉	技术负责人、AI治理组
教学与评价	先独立后辅助;支架渐隐;高风险多源人工决定	AI-off/on 表现、提示史、评分覆盖、申诉处理	AI-IURR 超阈;依赖上升;评分亚组不等值	教师、评价委员会
数据与隐私	最小必要、去标识化、分权访问、期限与删除机制	同意/告知、访问日志、数据流图、隐私影响评估	真实可识别数据进入未授权模型;泄露或超范围使用	数据保护官、机构
公平与可及	语言/人群/残障/低带宽测试;等价替代路径	分层到达率、TCCG、AI-IURR、可用性和成本	任一群体持续低获益或高风险且无有效缓解	机构领导、课程与公平委员会
上线后警戒	安全事件分级、红队、更新回归测试、可快速回滚	事件严重度、根因、纠正预防措施、复测与发布记录	严重事件;重复同类错误;更新未验证即上线	质量安全负责人、供应方

RAG:检索增强生成。阈值应由机构按任务风险预设;暂停后必须保留无AI或低技术替代路径。

5.3 公平性评价与中国情境 智能体可能缓解病例和师资带来的教育资源分布不均,但模型性能、信息化建设和本地化知识不均等也可能扩大差距^[43-44]。公平性评价应报告不同地区、语言、基线能力、数字素养和可及性群体的到达率、完成率、候选TCCG和AI-IURR,而不能只比较模型准确率。资源不足的环境下部署可优先选择边界清晰、可在本地维护的形成性任务、离线病例包和异步教师复核。中国庞大的医学教育体系、专病联盟和分级诊疗网络可为多中心验证提供场景。国内在医学元宇宙、“云加端”、医学GPT、BAIMGPT、医院—社区—家庭协同和肺结节智能体等工作。

为病例共享、RAG和远程复核提供了技术语境^[45-50]。然而,临床咨询系统能够给出合理输出,并不等同于具备教育有效性;学习者诊断、支架选择、评分效度和AI撤除后的迁移均需独立验证。可在现有专病网络中先比较传统远程教学、普通LLM辅

助与DC-AME,考察AI-off学习者对未见病例的识别能力,以及评估AI-on核验能力和成本。

6 讨论

现有文献一致结论是AI可改变学习者训练时的表现,而不是AI改善学习者的独立临床能力。DC-AME据此把“是否使用AI”的二元争论转换为两个可测问题:撤除AI后是否发生迁移,使用AI时是否保持安全。四阶段闭环的机制假设是先记录独立表征,保留认知所有权,约束反馈可促进错误监测,渐隐支架可减少替代性依赖,未见病例与延迟测量可区分学习和即时补偿。

DC-AME经验证后,双能力评价可应用于课程准入、形成性反馈和工作场所监督:学习者可能在AI-on条件下表现良好却在AI-off条件下不达标,提示工具具有补偿价值但不能替代基础训练;也可能在AI-off达标却错误采纳AI建议,提示需要信任校

准训练。两类证据不能压缩为一个总分。对于患者安全相关任务, AI的权限应随任务风险、学习阶段和本地验证证据递增, 而不是随模型规模或流畅度递增。

本研究存在一定的局限性: (1) 本文为叙述性综述与概念框架论文, 未采用系统检索、双人筛选或正式偏倚评估, 可能存在文献选择和解释偏倚。(2) 纳入证据跨越教育、临床决策支持、模拟和学习科学, 概念的可迁移性不等于干预效应可迁移性。部分定量研究样本小、结局短期、干预和模型版本异质; 技能退化信号来自观察性设计。DC-AME及其候选结局尚未完成内容效度、反应过程、内部结构、跨病例可推广性和结果效度验证。(3) 三类智能体协同也可能增加复杂性、成本、隐私和系统性失误。因而, 本文提供的是研究议程和可检验设计, 而不是直接实施的有效性结论。

综上所述, 医学教育智能体的价值不能由模型分数、满意度以及 AI-on 即时表现单独判定。需要同时评估学习者在 AI-off 后能否将能力迁移到未见病例, 以及 AI 参与是否诱导不安全推理。DC-AME 以 AI-off 与 AI-on 双能力、诊断—模拟—支架—迁移闭环和受治理的三类智能体协同, 为这两个问题提供了待验证的课程与评价框架。下一步应优先完成 DC-AME 候选结局的测量学研究、标准病例库和多中心延迟迁移试验, 再讨论其在高风险总结性评价或患者结局层面中的推广应用。

伦理声明 无。

利益冲突 所有作者声明不存在利益冲突。

作者贡献 杨丽: 文章撰写; 白春学: 论文审阅、论文指导。

参考文献

- [1] Chan K S, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review[J]. *JMIR Med Educ*, 2019, 5(1): e13930.
- [2] Knopp M I, Warm E J, Weber D, et al. AI-enabled medical education: threads of change, promising futures, and risky realities across four potential future worlds [J]. *JMIR Med Educ*, 2023, 9: e50373.
- [3] Triola M M, Rodman A. Integrating generative artificial intelligence into medical education: curriculum, policy, and governance strategies [J]. *Acad Med*, 2025, 100 (4) : 413-418.
- [4] Masters K, Herrmann-Werner A, Festl-Wietek T, et al. Preparing for artificial general intelligence (AGI) in health professions education: AMEE guide No. 172 [J]. *Med Teach*, 2024, 46(10): 1258-1271.
- [5] Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84 [J]. *Med Teach*, 2024, 46(4): 446-470.
- [6] 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.
- [7] Xu X J, Chen Y X, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review [J]. *J Educ Eval Health Prof*, 2024, 21: 6.
- [8] Ma X M, Liu X T, Sun H J. The effectiveness of GAI-assisted teaching methods in medical education: a systematic review and meta-analysis [J]. *Front Public Health*, 2026, 14: 1813108.
- [9] Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial [J]. *BMC Med Educ*, 2024, 24(1): 1391.
- [10] Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial [J]. *JAMA Netw Open*, 2024, 7(10): e2440969.
- [11] Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians [J]. *JAMA Intern Med*, 2024, 184(5): 581-583.
- [12] Budzyń K, Romańczyk M, Kitala D, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study [J]. *Lancet Gastroenterol Hepatol*, 2025, 10(10): 896-903.
- [13] Chi M T H, Wylie R. The ICAP framework: linking cognitive engagement to active learning outcomes [J]. *Educ Psychol*, 2014, 49(4): 219-243.
- [14] Sweller J. Cognitive load during problem solving: effects on learning [J]. *Cogn Sci*, 1988, 12(2): 257-285.
- [15] Ericsson K A, Krampe R T, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance [J]. *Psychol Rev*, 1993, 100(3): 363-406.
- [16] Karpicke J D, Blunt J R. Retrieval practice produces more learning than elaborative studying with concept mapping [J]. *Science*, 2011, 331(6018): 772-775.
- [17] Cepeda N J, Pashler H, Vul E, et al. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychol Bull*. 2006 May;132(3):354-380.
- [18] Li S Q, Okada Y, Gu W J, et al. Leveraging AI and transfer learning to enhance out-of-hospital cardiac arrest outcome prediction in diverse setting [J]. *npj Digit Med*, 2025, 8: 716.
- [19] Eva K W. What every teacher needs to know about clinical reasoning [J]. *Med Educ*, 2005, 39(1): 98-106.
- [20] Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning [J]. *Adv Health Sci Educ Theory Pract*, 2009, 14(Suppl 1): 27-35.
- [21] Mamede S, van Gog T, van den Berge K, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy

- among internal medicine residents [J]. JAMA, 2010, 304 (11): 1198-1203.
- [22] Goddard K, Roudsari A, Wyatt J C. Automation bias: a systematic review of frequency, effect mediators, and mitigators [J]. J Am Med Inform Assoc, 2012, 19(1): 121-127.
- [23] Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study [J]. JAMA, 2023, 330 (23): 2275-2284.
- [24] Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids [J]. NPJ Digit Med, 2021, 4(1): 31.
- [25] Cabitza F, Rasoini R, Gensini G F. Unintended consequences of machine learning in medicine [J]. JAMA, 2017, 318 (6): 517-518.
- [26] Frank J R, Snell L S, Cate O T, et al. Competency-based medical education: theory to practice [J]. Med Teach, 2010, 32 (8): 638-645.
- [27] Montgomery K B, Mellinger J D, Lindeman B. Entrustable professional activities in surgery: a review [J]. JAMA Surg, 2024, 159(5): 571-577.
- [28] Gin B C, O'Sullivan P S, Hauer K E, et al. Entrustment and EPAs for artificial intelligence (AI): a framework to safeguard the use of AI in health professions education [J]. Acad Med, 2025, 100(3): 264-272.
- [29] Corral-Gudino L, Herrero-Montano I, de la Torre-Díez I, et al. Integrating generative artificial intelligence (AI) in medical education: a framework for preserving clinical reasoning [J]. Appl Sci, 2026, 16(12): 5946.
- [30] Kononowicz A A, Woodham L A, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration [J]. J Med Internet Res, 2019, 21(7): e14676.
- [31] Plackett R, Kassianos A P, Mylan S, et al. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review [J]. BMC Med Educ, 2022, 22(1): 365.
- [32] Jay R, Sandars J, Patel R, et al. The use of virtual patients to provide feedback on clinical reasoning: a systematic review [J]. Acad Med, 2025, 100(2): 229-238.
- [33] Van De Ridder J M M, Stokking K M, McGaghie W C, et al. What is feedback in clinical education? [J]. Med Educ, 2008, 42(2): 189-197.
- [34] Issenberg S B, McGaghie W C, Petrusa E R, et al. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review [J]. Med Teach, 2005, 27(1): 10-28.
- [35] Kwon M, Dicker S, Mustafa N, et al. Generative artificial intelligence and prompt engineering in asthma-related settings [J]. Ann Allergy Asthma Immunol, 2026: S1081-S1206 (26) 00291-7.
- [36] Evans H, Snead D. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact [J]. npj Digit Med, 2024, 7: 89.
- [37] Schuwirth L W, van der Vleuten C P. Programmatic assessment: From assessment of learning to assessment for learning [J]. Med Teach, 2011, 33(6): 478-485.
- [38] Cook D A, Beckman T J. Current concepts in validity and reliability for psychometric instruments: theory and application [J]. Am J Med, 2006, 119(2): 166. e7-166. 16.
- [39] Hopewell S, Chan A W, Collins G S, et al. CONSORT 2025 statement: updated guideline for reporting randomized trials [J]. Nat Med, 2025, 31: 1776-1783.
- [40] Liu X X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension [J]. Nat Med, 2020, 26: 1364-1374.
- [41] Association of American Medical Colleges. Principles for the responsible use of artificial intelligence in and for medical education [EB/OL]. [2025-07-31] (2026-08-23). <https://www.aamc.org/about-us/mission-areas/medical-education/principles-ai-use>.
- [42] World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models [M]. Geneva: World Health Organization, 2024.
- [43] UNESCO. Guidance for generative AI in education and research [M]. Paris: UNESCO, 2023.
- [44] Toohey K, Quince Z, Walker F, et al. Integrating generative AI in health education: a scoping review and implementation framework [J]. Med Sci Educ, 2025, 35(6): 2751-2762.
- [45] Yang D W, Zhou J, Chen R C, et al. Expert consensus on the metaverse in medicine [J]. Clin eHealth, 2022, 5: 1-9.
- [46] 白春学. 元宇宙医学之我见 [J]. 中国医药导刊, 2023, 25 (1): 1-6.
- [47] 白春学. 医学 GPT 研发应用之吾见 [J]. 元宇宙医学, 2025, 2(1): 6-14.
- [48] 白春学. 肺结节专家——BAIMGPT 白皮书 [J]. 元宇宙医学, 2025, 2(2): 55-64.
- [49] 白春学. 医学 GPT 研发的挑战与解决方案 [J]. 元宇宙医学, 2026, 3(1): 11-15.
- [50] 白春学. AI 智能体赋能肺结节诊治研发进展及展望 [J]. 元宇宙医学, 2026, 3(2): 83-90.

引用本文

杨 丽, 白春学. 医学智能体能否形成可迁移的临床能力? 一种双能力医学智能体教育框架及其评价体系 [J]. 元宇宙医学, 2026, 3(3): 214-221.

Yang L, Bai C X. Can medical agents build transferable clinical competence? A dual-competency framework and evaluation system for medical agent education [J]. Metaverse Med, 2026, 3(3): 214-221.