·理论探索与实践·

AI幻觉识别与验证的批判性思维教学对神经外科轮转住培医师AI素养的影响

张风江，李中军，丁大领，陈 煜，李 恩
郑州大学第一附属医院，郑州 450052

【摘要】目的 探讨“AI(artificial intelligence，AI)幻觉识别—验证—正确重构”的批判性思维教学策略对神经外科轮转住培医师AI素养、批判性思维能力和专业知识掌握的影响。**方法** 采用单中心平行随机对照教育干预研究设计，将2025年9月至2026年3月在郑州大学第一附属医院神经外科接受住院医师规范化培训的48名轮转住培医师随机分为实验组(n=24)和对照组(n=24)。实验组接受“正向知识框架导入、AI输出错误识别、权威证据验证、正确答案重构和教师反馈总结”教学干预；对照组接受相同时长的常规AI工具使用培训。在干预前、干预结束后即时及干预后4周分别采用中文版批判性思维倾向评估量表(critical thinking disposition assessment，CTDA)、AI幻觉识别能力测试和神经外科专业知识测验进行评估。**结果** 干预前两组基线资料比较差异无统计学意义(均P>0.05)。干预后即时及干预后4周，实验组CTDA总分、AI幻觉识别准确率、错误分类准确率、专家论证评分及专业知识测验得分均高于对照组(均P<0.05)。重复测量方差分析显示，CTDA总分、AI幻觉识别准确率、错误分类准确率和专业知识测验得分的时间与分组交互效应有统计学意义；专家论证评分的时间与分组交互效应未达统计学显著(P=0.06)。**结论** 基于AI幻觉识别与验证的批判性思维教学可作为神经外科住培中常规AI工具培训和正向专业知识教学的补充模块，有助于提升住培医师对AI输出的批判性评估和验证能力。

【关键词】人工智能；大语言模型；AI幻觉；批判性思维；医学教育
【中图分类号】R-4；C975 **【文献标志码】**A **【文章编号】**1004-5287(2026)05-0611-07
【DOI】10.13566/j.cnki.cmet.cn61-1317/g4.202605005

The impact of critical thinking teaching based on AI hallucination recognition and evidence-based verification on AI literacy of neurosurgical residents

ZHANG Fengjiang, LI Zhongjun, DING Daling, CHEN Yu, LI En
The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, *China*

【Abstract】Objective To explore the impact of a critical thinking teaching strategy based on “AI (artificial intelligence) hallucination recognition, evidence-based verification, and correct reconstruction” on AI literacy, critical thinking ability, and professional knowledge among neurosurgical residents. **Methods** A single-center, parallel-group randomized controlled educational intervention was conducted. From September 2025 to March 2026, 48 neurosurgical residents undergoing standardized training at the First Affiliated Hospital of Zhengzhou University were randomly assigned to the experimental group (n=24) and the control group (n=24). The experimental group received a teaching intervention consisting of “positive knowledge framework introduction, AI-output error recognition, authoritative evidence verification, correct answer reconstruction, and teacher feedback”; the control group received conventional AI tool training of the same duration. Assessments were per-

基金项目：国家自然科学基金委员会青年科学基金项目(82303344)；河南省卫生健康委员会市厅级项目(LHGJ20220345)
收稿日期：2026-04-27
作者简介：张风江，副主任医师，硕士，研究方向为脊柱神经外科常见疾病的临床研究。
通信作者：李恩，副主任医师，硕士，研究方向为脊柱神经外科常见疾病的临床研究。E-mail：lien9_79@163.com

formed before the intervention, immediately after the intervention, and 4 weeks after the intervention using the Chinese version of the Critical Thinking Disposition Assessment (CTDA), an AI hallucination recognition test, and a neurosurgical knowledge examination. **Results** There was no statistically significant difference in baseline data between the two groups before the intervention (all $P>0.05$). Immediately after the intervention and at 4-week follow-up, the experimental group showed higher CTDA scores, AI hallucination recognition accuracy, error classification accuracy, expert argumentation scores, and neurosurgical knowledge scores than the control group (all $P<0.05$). Repeated-measures ANOVA showed significant time-by-group interactions for CTDA score, recognition accuracy, classification accuracy, and professional knowledge score; the interaction for expert argumentation score did not reach statistical significance ($P=0.06$). **Conclusion** Critical thinking teaching based on AI hallucination recognition and evidence-based verification may serve as a supplementary module to conventional AI tool training and positive professional knowledge teaching in neurosurgical residency training, which helps to enhance residents' ability to critically evaluate and validate AI-generated outputs.

【Keywords】 artificial intelligence; large language model; AI hallucination; critical thinking; medical education

近年来,生成式人工智能和大语言模型在医学教育、文献检索、病例讨论和临床决策支持中的应用迅速增加[1-2]。对临床医师而言,AI(artificial intelligence,AI)素养不应仅限于工具操作能力,还应包括理解模型局限、识别输出风险、评价信息可信度、核查证据来源以及负责任使用AI的能力[3]。神经外科专业知识复杂、诊疗风险高,若住培医师未经训练即直接采信AI生成内容,可能导致诊疗思路偏差和医学证据误用。因此,在住院医师规范化培训中建立面向生成式AI的批判性评估训练具有现实必要性。

大语言模型(large language model,LLM)幻觉通常指模型生成表面流畅、语义上看似合理但与事实、证据或逻辑不一致的内容。在医学场景中,其表现形式不仅包括事实性错误,还可包括临床推理跳跃、文献或指南引用错误以及虚构研究、术式、概念或专业术语等[4]。这些错误往往具有较强迷惑性,尤其在专业门槛较高的神经外科场景中,学习者需要依靠教材、指南、PubMed原始文献和专家判断进行多源验证,而不能将AI输出直接视为知识来源。

从学习理论看,认知冲突理论认为,当学习者发现外部信息与既有知识框架不一致时,可能启动主动检索、比较、验证和概念重构过程[5]。错误管理训练和生产性失败理论进一步提示,在教师控制、及时反馈和明确纠错的前提下,错误案例可作为促进深度学习、迁移应用和反思性判断的教学资源[6]。因此,AI错误案例的教学价值并不在于呈现错误本身,而在于通过受控情境促使学习者回到权威证据和正向知识框架中完成验证与重构。

基于上述理论,本研究以教师控制下的AI错误案例作为认知冲突材料,构建"正向知识框架导入—AI输出错误识别—权威证据验证—正确答案重构—教师反馈总结"的教学流程。本研究采用随机对照教育干预设计,评估该策略对神经外科住院医师AI素养、批判性思维能力和专业知识掌握的影响,为高风险临床学科开展规范化AI教育提供初步证据。

1 对象与方法

1.1 研究设计

本研究采用单中心平行随机对照教育干预研究设计,评估基于AI幻觉识别与验证的批判性思维教学对神经外科轮转住培医师批判性思维能力、AI输出批判性评估能力及专业知识掌握水平的影响。研究时间为2025年9月至2026年3月。

1.2 研究对象

研究对象为2025年9月至2026年3月在郑州大学第一附属医院神经外科接受规范化培训的住院医师。研究期间,郑州大学第一附属医院神经外科住培基地同期在培及轮转住培医师共57名,其中第1年20名、第2年19名、第3年18名。纳入标准:①正在接受神经外科住院医师规范化培训或神经外科轮转培训,且培训年限为第1~3年;②至少使用过1种大语言模型工具(如ChatGPT、文心一言、通义千问、DeepSeek等),且使用频率不低于每月1次。排除标准:①既往接受过系统AI批判性使用培训或同类教学干预;②同期参与其他AI相关教学研究,可能产生干预混杂;③基线或任一随访节点数据缺失≥20%;④未完成全部5周干预及3次评估;⑤干预期间出勤率<80%;⑥自行接受系统AI错误识别训练或使用非指定AI教学工具且无法排除混杂影响。根据纳入和排除标准筛选后,最终纳入48名轮转住培医师。所有纳入对象均为外科专业神经外科方向住培医师和神经外科轮转住培医师。本研究为单中心先导性随机对

照教育干预研究，样本量主要受研究期间住培基地可纳入对象数量限制，采用同期符合标准者连续纳入。

1.3　干预设计方案

对照组接受相同时长（10 学时）的常规 AI 工具使用培训，内容限于AI工具基本操作、提示词基本规范、文献检索辅助、病例书写辅助和医学AI工具发展现状；不讲授AI幻觉分类框架，不提供预植入错误案例，不进行结构化错误识别、证据核查、纠错论证和小组纠错讨论训练。

实验组接受为期5周、总时长10学时的基于AI幻觉识别与验证的批判性思维教学干预。总体流程为“正向知识框架导入—AI输出错误识别—权威证据验证—正确答案重构—教师反馈总结”。每次训练开始前，教师首先围绕相关神经外科主题进行正向知识框架讲授，明确核心诊疗原则、关键鉴别诊断、常见误区和指南推荐。随后，学员在教师控制的模拟材料中识别AI输出错误，并通过教材、指南和PubMed原始文献进行验证。训练结束时，教师对所有错误进行统一纠正，形成规范化答案。错误案例仅作为触发认知冲突和训练证据核查能力的材料，不作为临床诊疗依据。

1.3.1　AI幻觉分类依据和识别要点

本研究参考既往LLM幻觉研究中关于事实性幻觉、推理错误、引用/证据幻觉和虚构性输出的分类，并结合神经外科临床教学特点，按照AI输出错误发生的主要层级，将神经外科场景下AI幻觉归纳为4类：知识陈述层错误、临床推理层错误、证据支持层错误和信息生成层错误[3-4,7-10]。对应到教学实践中，分别表现为事实性错误、逻辑推理错误、数据引用错误和幻觉性编造。4类AI幻觉的分类依据、定义、识别要点和验证来源如表1所示。

表1　神经外科场景下AI幻觉类型、分类依据及识别要点

幻觉类型	分类依据	定义	普适性识别要点	神经外科示例	推荐验证来源
事实性错误	知识陈述层	AI输出与公认医学事实、解剖结构、病理生理或诊疗原则不符	核对解剖、病理、生理、诊断标准、适应证、禁忌证、剂量、并发症等是否与教材或指南一致	错误描述硬膜外血肿常见出血来源	神经外科学教材、临床指南、权威综述
逻辑推理错误	临床推理层	AI从症状、体征、影像或实验室结果推导诊断/治疗时出现跳跃或断裂	检查是否存在诊断跳跃、因果倒置、忽略鉴别诊断、忽略关键阴性证据、治疗路径不符合临床流程	仅凭头痛和乳头水肿直接诊断脑膜瘤	诊疗流程、指南算法、病例讨论规范
数据引用错误	证据支持层	AI引用的文献、指南、年份、样本量、推荐等级或结论与原文不符	核对DOI/PMID、指南年份、研究对象、样本量、主要结局和结论是否真实一致	误引GBM中位生存期或指南推荐等级	PubMed、指南官网、原始文献
幻觉性编造	信息生成层	AI生成不存在的研究、术式、概念、指南或专业术语	多个权威来源均无法证实；术语看似专业但无规范出处；文献题名、作者、期刊、年份无法对应	编造不存在的WHO分类文献或术式名称	PubMed、专业教材、指南数据库、专家复核

1.3.2　AI幻觉识别训练和验证

第2～4周训练材料采用命题蓝图设计。每次训练围绕1个神经外科核心主题设置1段AI生成病例分析或诊疗建议，并预先植入4～5个错误，确保事实性错误、逻辑推理错误、数据引用错误和幻觉性编造均被覆盖。每名学员需完成错误定位、错误类型判断、权威来源验证、正确答案重构和纠错依据说明。教师在讨论环节按照4类错误分别总结识别方法和验证路径。训练和验证对应设计如表2所示。

表2　AI幻觉识别训练与验证设计

阶段	训练主题	事实性错误（个）	逻辑推理错误（个）	数据引用错误（个）	幻觉性编造（个）	训练任务
第2周	颅脑损伤与脑疝	1～2	1	1	1	教材/指南核对、诊疗流程验证
第3周	脑血管病	1	1～2	1	1	指南算法核对、文献原文验证
第4周	颅内肿瘤	1	1	1～2	1	WHO分类、指南和PubMed核查
第5周	综合考核（每类）	2～3	2～3	2～3	2～3	独立识别、分类、验证、纠正

1.3.3　综合考核

第5周进行综合考核。综合考核病例材料依据同一命题蓝图编制，包含10个预植入错误，其中事实性错误、逻辑推理错误、数据引用错误和幻觉性编造各2～3个。考核要求学员分别完成错误识别、错误类型分类、验证来源记录、正确答案重构和纠错理由说明。评分时分别计算总识别准确率、总分类准确率、4类幻觉分类型识别准确率和专家论证评分。另设置标准化临床沟通场景，要求学员判断AI建议中哪些信息可参考、哪些需独立验证，并向标准化病人

作出安全、准确且不过度依赖AI的解释。

1.3.4 干预一致性与污染控制

为降低组间干预污染，本研究采取以下措施：①两组使用不同教学材料，实验组案例库、标准答案和4类幻觉识别表不向对照组开放；②两组课程分时段、分教室进行；③授课教师依据预先制定的统一教案实施教学，对照组教师不得讲授AI错误分类、证据核查和纠错论证内容；④研究期间要求学员不跨组传播课程材料；⑤每次评估前记录学员课外AI使用频率及是否接触类似AI错误识别训练；⑥对明显违反研究要求或自行接受系统AI错误识别训练者，按照预设排除标准处理。

1.4 评估指标与工具

1.4.1 主要结局指标

批判性思维能力：采用中文版批判性思维倾向评估量表（CTDA）进行评估[11]。该量表包含系统性与分析性、探究性与熟悉度、成熟度与怀疑精神3个维度，共19个条目，采用7点Likert评分（1=非常强烈反对，7=非常强烈同意），总分范围19～133分，得分越高表明批判性思维倾向越强。

AI幻觉识别能力：AI幻觉识别能力测试采用3套平行测试材料，分别用于基线、干预后即时和干预后4周评估。每套材料均包含10个预植入错误，4类错误各2～3个，难度经3名高年资神经外科医师评估后保持相对一致。评分指标包括：①错误识别准确率；②错误分类准确率；③4类幻觉分类型识别准确率；④纠错论证评分。纠错论证由3名高年资神经外科专家在盲法状态下评分，评分包括证据准确性、纠错完整性和论证深度3个维度，每维度1～5分，总分3～15分。

1.4.2 次要结局指标

专业知识掌握水平：采用神经外科理论知识测验进行评估。测验涵盖脑疝、颅内肿瘤、脑血管病、颅脑损伤等核心内容，共20道单选题、10道多选题和5道简答题，满分100分，测验题目由3名高年资神经外科医师共同编制。

教学满意度：采用自编满意度问卷，包含“干预对批判性思维的提升作用”“对临床工作的帮助程度”“教学内容的实用性”“教学方法的有效性”等8个项目，采用1～5分Likert评分，总分8～40分。

1.5 数据收集流程

数据收集在3个时间点进行：①干预前（基线，T0）：所有研究对象完成CTDA量表、AI幻觉识别能力测试和神经外科专业知识测验；②干预结束后即时（T1）：所有研究对象完成上述所有量表和测试，并填写教学满意度问卷；③干预后4周（T2）：所有研究对象再次完成CTDA量表、AI幻觉识别能力测试和神经外科专业知识测验。

1.6 统计学处理

采用R 4.4.0软件进行统计分析，使用tidyverse包进行数据整理，rstatix包进行重复测量方差分析，effsize包计算Cohen's d效应量。计量资料以（均数±标准差）（$\bar{x}\pm s$）表示，组间比较采用独立样本t检验（方差齐性检验采用Levene检验）；组内比较采用配对样本t检验；重复测量数据采用重复测量方差分析，效应量采用Cohen's d表示。以$P<0.05$为差异有统计学意义。

2 结果

2.1 基线资料比较

研究期间共筛选住培医师57名，排除9名，最终纳入48名并完成随机分组。研究过程中无脱落，所有研究对象完成5周教学干预及3次评估。两组研究对象在年龄、性别、规培年级、AI使用频率、基线CTDA总分、AI幻觉识别准确率、错误分类准确率、专家论证评分及专业知识测验得分等方面比较，差异均无统计学意义（均$P>0.05$），具有可比性（表3）。

表3 两组研究对象基线资料比较

指标	实验组（n=24）	对照组（n=24）	t/χ^2	P
性别（男/女，人）	19/5	20/4	0.00	1.00
规培年级（一年级/二年级/三年级，人）	9/6/9	9/9/6	1.20	0.55
AI使用频率（每周≥3次/1～2次/每月1～2次，人）	11/6/7	9/8/7	0.49	0.78
年龄[（$\bar{x}\pm s$），岁]	26.60±1.47	27.48±1.88	1.81	0.08
CTDA[（$\bar{x}\pm s$），分]	45.17±19.26	46.38±23.03	0.20	0.84
AI幻觉识别准确率[（$\bar{x}\pm s$），%]	38.44±10.06	38.24±12.00	0.06	0.95
错误分类准确率[（$\bar{x}\pm s$），%]	48.68±13.69	43.97±15.68	1.11	0.27
专家论证评分[（$\bar{x}\pm s$），分]	6.34±1.49	5.58±1.75	1.61	0.11
专业知识测验得分[（$\bar{x}\pm s$），分]	71.17±9.27	66.41±7.71	1.93	0.06

2.2　主要结局指标比较

2.2.1　批判性思维能力比较

干预前，两组CTDA总分比较差异无统计学意义（$P>0.05$）。干预后即时及干预后4周，实验组CTDA总分均高于对照组（均$P<0.01$）。重复测量方差分析显示，时间与分组的交互效应显著（$F=11.98$，$P<0.001$，偏$\eta^2=0.207$），提示实验组CTDA总分随时间的改善幅度大于对照组（表4）。

2.2.2　AI幻觉识别能力比较

干预前，两组AI幻觉识别准确率、错误分类准确率及专家论证评分比较差异均无统计学意义（均$P>0.05$）。干预后即时及干预后4周，实验组识别准确率、错误分类准确率及专家论证评分均高于对照组。重复测量方差分析显示，识别准确率（$F=9.93$，$P<0.001$，偏$\eta^2=0.18$）和分类准确率（$F=6.19$，$P=0.003$，偏$\eta^2=0.12$）的时间与分组交互效应有统计学意义；专家论证评分在干预后即时和4周随访时实验组均高于对照组，但重复测量交互效应未达到统计学显著（$F=2.86$，$P=0.06$，偏$\eta^2=0.06$），提示该指标仍需更大样本研究进一步验证（表5）。

表4　两组不同时间点CTDA总分比较

时间点	实验组[$n=24$，($\bar{x}\pm s$)，分]	对照组[$n=24$，($\bar{x}\pm s$)，分]	t	P	Cohen's d
干预前（T0）	45.17±19.26	46.38±23.03	0.20	0.84	0.06
干预后即时（T1）	92.00±29.92	64.25±27.72	3.33	0.001 7	0.96
干预后4周（T2）	91.88±18.52	65.12±17.67	5.12	<0.001	1.48

表5　两组不同时间点AI幻觉识别能力比较

指标	时间点	实验组($n=24$)	对照组($n=24$)	t	P	Cohen's d
识别准确率[($\bar{x}\pm s$)，%]	T0	38.44±10.06	38.24±12.00	0.06	0.95	0.02
	T1	77.66±25.07	43.52±17.07	5.51	<0.001	1.59
	T2	72.60±20.43	38.41±15.42	6.54	<0.001	1.89
分类准确率[($\bar{x}\pm s$)，%]	T0	48.68±13.69	43.97±15.68	1.11	0.27	0.32
	T1	70.45±21.96	42.45±13.51	5.32	<0.001	1.54
	T2	70.73±20.79	45.56±18.33	4.45	<0.001	1.28
专家论证评分[($\bar{x}\pm s$)，分]	T0	6.34±1.49	5.58±1.75	1.61	0.11	0.47
	T1	11.21±6.53	6.96±2.31	3.01	0.004	0.87
	T2	9.66±4.40	6.69±1.86	3.05	0.005	0.88

2.3　次要结局指标比较

2.3.1　专业知识掌握水平比较

干预前，两组专业知识测验得分比较差异无统计学意义（$P>0.05$）。干预后即时及干预后4周，实验组专业知识测验得分均高于对照组（均$P<0.001$）。重复测量方差分析显示，时间与分组的交互效应有统计学意义（$F=3.516$，$P=0.034$，偏$\eta^2=0.071$），提示实验组专业知识掌握水平随时间的改善幅度大于对照组（表6）。

表6　两组不同时间点专业知识测验得分比较

时间点	实验组[$n=24$，($\bar{x}\pm s$)，分]	对照组[$n=24$，($\bar{x}\pm s$)，分]	t	P	Cohen's d
干预前（T0）	71.17±9.27	66.41±7.71	1.93	0.06	0.56
干预后即时（T1）	86.72±5.75	73.78±7.80	6.54	<0.001	1.89
干预后4周（T2）	84.26±6.91	74.48±6.14	5.19	<0.001	1.50

2.3.2　教学满意度比较

实验组教学满意度问卷平均得分为（36.5±2.7）分（满分40分），其中“干预对批判性思维的提升作用”“对临床工作的帮助程度”两项得分较高，分别为（4.7±0.5）分和（4.6±0.6）分。该结果提示学员对该补充训练模块接受度较高，但满意度为主观评价指标，仍需结合客观学习效果进行解释。

2.4　实验组四类AI幻觉识别准确率变化的探索性分析

实验组干预前后4类AI幻觉识别准确率均提高（均$P<0.001$），其中事实性错误识别准确率由43.11%提高至86.19%，幻觉性编造识别准确率由

31.10%提高至56.47%，提升后仍相对较低（表7）。

表7 实验组4类AI幻觉识别准确率变化的探索性分析

幻觉类型	干预前[($\bar{x}\pm s$),%]	干预后即时[($\bar{x}\pm s$),%]	t	P	Cohen's d
事实性错误	43.11±15.05	86.19±28.35	7.19	<0.001	1.86
逻辑推理错误	39.37±12.12	70.47±23.85	5.42	<0.001	1.66
数据引用错误	46.61±13.93	78.38±25.43	6.16	<0.001	1.50
幻觉性编造	31.10±9.87	56.47±26.90	4.51	<0.001	1.22

3 讨论

本研究尝试在神经外科住培场景中构建并评估基于AI幻觉识别与验证的批判性思维教学策略。结果显示，与相同时长的常规AI工具培训相比，实验组在干预后即时及干预后4周的CTDA总分、AI幻觉识别准确率、错误分类准确率和专业知识测验得分均较对照组更高。从理论机制看，该教学策略可能通过“认知冲突—主动验证—正确知识重构”的路径发挥作用。当学员发现AI输出与教材、指南或原始文献不一致时，需要主动检索证据、比较不同来源、判断证据质量并重构正确答案。该过程与概念转变、错误管理训练和生产性失败理论具有内在一致性。与单纯讲授AI工具操作相比，结构化的错误定位、类型判断和验证任务更强调学习者的主动加工和证据意识，有助于将AI使用能力从“会用工具”推进到“能评价工具输出”[5,11-13]。

本研究表5显示，实验组AI幻觉识别准确率由干预前38.44%提高至干预后即时77.66%，干预后4周为72.60%；错误分类准确率由48.68%提高至70.45%，干预后4周为70.73%。这些结果提示，经过系统训练后，学员不仅更容易发现AI输出中的可疑信息，也能进一步判断错误类型并寻找验证依据。专家论证评分在干预后即时和4周时实验组均高于对照组，但重复测量交互效应P=0.06，尚未达到统计学显著。考虑到纠错论证能力涉及证据检索、信息整合和临床推理等综合能力，其形成可能需要更长时间的训练与巩固，后续研究可在更大样本和更长随访条件下进一步观察其变化。4类AI幻觉的探索性分析显示，事实性错误识别准确率提升较明显，而幻觉性编造识别准确率提升后仍相对较低。可能原因在于事实性错误通常可通过既有专业知识或教材直接核对，而幻觉性编造常表现为看似专业的术语、文献或指南名称，需要通过PubMed、指南数据库和多个权威来源交叉验证[14]。该发现提示，后续训练应进一步强化文献真实性核查、指南版本确认和多源交叉验证训练[15]。需要强调的是，本研究所称AI幻觉识别训练并不是以错误替代正确知识教学，而是在正向规范教学基础上加入的批判性评估训练模块。该策略的核心目标不是让住培医师记忆错误，而是通过受控错误案例促使其回到教材、指南和原始文献中完成证据核查和正确知识重构。因此，该策略必须满足2个前提：第一，错误案例必须由教师预先设计并经专家审核；第二，课堂中必须及时纠错，避免错误信息残留。对于一年级住培医师，应选择基础性、低复杂度案例，并加强教师即时反馈和正确答案重构，以降低认知负荷和误导风险。

本研究存在以下局限性。第一，本研究为单中心研究，样本量相对较小，仅纳入48名住院医师，并且研究对象均来自同一所三级甲等教学医院的神经外科，其教育背景、培训环境和临床实践模式可能与其他医疗机构存在差异，研究结果的外推性可能受到一定限制。未来研究应开展多中心、大样本的随机对照实验，纳入不同层级医院和不同地区的神经外科住院医师，以进一步验证“AI幻觉逆向利用”教学策略的有效性和可推广性。第二，本研究干预时长为5周，干预后随访时间仅为4周，仅能反映干预效果的短期保持情况，未能评估批判性思维能力和AI素养的长期变化趋势。后续研究应将随访时间延长至6个月至1年，设置多个随访时间点以评估干预效果的长期保持情况及衰减规律。第三，本研究采用纯定量设计，未结合定性研究方法，未能深入探索学员的主观体验和认知变化过程。未来研究采用混合方法设计，增加半结构化访谈或焦点小组讨论等定性研究方法。第四，本研究主要评估了教学效果层面的指标，尚未追踪干预对学员在真实临床场景中的AI使用行为、临床决策质量和患者安全的影响。后续研究可采用前瞻性队列设计，追踪接受过培训的住院医师在临床工作中的表现。第五，本研究虽然已经通过课程内容隔离、材料隔离和过程监督尽量降低组间污染，但教育干预研究无法完全排除学员之间交流或课外AI使用所造成的潜在影响。后续研究可采用集群随机、多中心分层随机或分时段交叉设计进一步降低干预污染。

4 结论

基于AI幻觉识别与验证的批判性思维教学可作为神经外科住培中常规AI工具培训和正向专业知识

教学的补充模块。该策略在教师指导、案例审核和即时纠错的前提下,通过“识别—验证—纠正—重构”流程,有助于提升住培医师对AI输出的批判性评估能力、验证能力和专业知识掌握水平。后续仍需多中心、大样本、长期随访研究进一步验证其推广价值和长期效果。

参考文献

[1] Iqbal U, Tanweer A, Rahmanti A R, et al. Impact of large language model (ChatGPT) in healthcare: An umbrella review and evidence synthesis[J]. Journal of Biomedical Science, 2025, 32(1): 45.

[2] Yu E, Chu X, Zhang W, et al. Large language models in medicine: Applications, challenges, and future directions[J]. International Journal of Medical Sciences, 2025, 22(11): 2792-2801.

[3] Tangsrivimol J A, Darzidehkalani E, Virk H U H, et al. Benefits, limits, and risks of ChatGPT in medicine[J]. Frontiers in Artificial Intelligence, 2025, 8: 1518049.

[4] Huang L, Yu W J, Ma W T, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[5] Cui L Y, Zhu Y X, Qu J L, et al. Psychometric properties of the critical thinking disposition assessment test amongst medical students in China: A cross-sectional study[J]. BMC Medical Education, 2021, 21(1): 10.

[6] Maynez J, Narayan S, Bohnet B, et al. On faithfulness and factuality in abstractive summarization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: ACL, 2020: 1906-1919.

[7] Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial[J]. JAMA Network Open, 2024, 7(10): e2440969.

[8] Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: Comparative analysis[J]. Journal of Medical Internet Research, 2024, 26: e53164.

[9] Ji Z W, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[10] Li H Z, Moon J T, Purkayastha S, et al. Ethics of large language models in medicine and medical research[J]. The Lancet Digital Health, 2023, 5(6): e333-e335.

[11] 臧伟进, 王渊. 以问题为导向的医学跨学科整合课程建设的初步探索与实践[J]. 中国医学教育技术, 2010, 24(6): 659-661.

[12] 朱锁珍, 朱粉琴, 苏丽丽,等. 以问题为导向的小组教学在“护理教育学”教学中的应用[J]. 中华医学教育探索杂志, 2023, 22(10): 1586-1589.

[13] 何蕾, 柴桦, 闫昱江. 中国医学教育领域人工智能应用研究与实证研究现状和热点分析(2022 — 2025年)[J]. 中国医学教育技术, 2026, 40(1): 1-6.

[14] Zhao Y, Yue Y, Sun Z, et al. Does generative artificial intelligence improve students' higher-order thinking? A meta-analysis based on 29 experiments and quasi-experiments[J]. Journal of Intelligence, 2025, 13(12): 160.

[15] 李健, 赵心怡. 人工智能与西方高等教育协同发展的态势、议题与理路:基于政策文本荟萃的分析[J]. 比较教育研究, 2026, 48(1): 56-67.