·人工智能专题·

大语言模型辅助教学模式在消化内科临床见习教学中的应用

李若璞[1],吕 勇[2],秦妮平[2],郑 荣[2],田祖宏[2],王 钧[3],潘阳林[2],樊代明[2]

空军军医大学:1.基础医学院学员一大队;2.西京消化病医院,国家消化系疾病临床医学研究中心,消化系肿瘤整合防治全国重点实验室,整合医学发展战略研究院,西安 710032;3.西京医院 986 医院消化内科,西安 710054

【摘要】目的 评估大语言模型(large language model,LLM)辅助教学模式对消化内科临床见习教学的作用。**方法** 采用随机对照试验设计,纳入首次参与消化内科见习的临床医学四年级学生 156 人,随机分为试验组(传统教学+LLM 辅助,n=78)和对照组(传统教学,n=78)。比较两组的临床胜任力总分,包括理论考核、客观结构化临床考试(objective structured clinical examination,OSCE)及病例分析。次要结局指标包括知识深度、临床思维质量,以及系统易用性与认知负荷。**结果** 试验组临床胜任力总分高于对照组[(89.5±5.9)分 vs. (76.2±8.0)分,均值差 Δ=13.3 分,95%CI[11.0,15.5],F=138.678,P<0.001]。试验组理论考核[(36.6±2.3)分 vs. (32.8±3.7)分,F=60.880,P<0.001]、OSCE 操作[(36.9±2.8)分 vs. (31.2±5.7)分,F=78.912,P<0.001]及病例分析[(15.9±4.3)分 vs. (12.1±5.6)分,F=19.799,P<0.001]得分均高于对照组。在知识深度方面,试验组概念图评分≥4 分的比例高于对照组(85.9% vs. 70.2%,χ^2=4.457,P=0.035),指南更新识别率也高于对照组(78.2% vs. 52.6%,χ^2=11.331,P<0.001)。在临床思维质量方面,试验组鉴别诊断关键病种覆盖率高于对照组[(89.1±4.7)% vs. (67.8±7.3)%,F=181.576,P<0.001],误诊罕见病比例低于对照组[(2.1±1.5)% vs. (6.9±2.8)%,F=199.090,P<0.001]。试验组在病例分析时注视关键信息的平均时间短于对照组[(5.2±1.1)s vs. (6.7±1.5)s,F=50.758,P<0.001],信息回看次数少于对照组[(1.8±0.6)次 vs. (2.9±0.9)次,F=80.679,P<0.001]。在系统易用性与认知负荷方面,试验组的系统可用性量表(system usability scale,SUS)评分高于对照组[(85.6±7.2)分 vs. (72.3±9.1)分,F=102.505,P<0.001],认知负荷评分低于对照组[(48.3±6.5)分 vs. (56.7±7.8)分,F=53.332,P<0.001]。亚组分析显示,在各亚组内试验组临床胜任力总分均高于对照组。其中,理论成绩前 30% 学生中,试验组与对照组临床胜任力总分均值差(Δ=15.14 分,95%CI[13.36,18.93]),高于后 30% 学生相应的均值差(Δ=9.17 分,95%CI[4.52,13.83]),差异有统计学意义(F=212.008,P<0.001)。但计算机操作熟练亚组的临床胜任力总分均值差(Δ=13.68 分,95%CI[11.29,16.06])与不熟练亚组的均值差(Δ=12.09 分,95%CI[6.54,17.64])、普通典型病例亚组的临床胜任力总分均值差(Δ=12.81 分,95%CI[10.38,15.24])与复杂疑难病例亚组的均值差(Δ=14.62 分,95%CI[9.31,19.93])的差异均无统计学意义(P>0.05)。**结论** LLM 辅助教学可提升消化内科见习生的知识整合与临床决策能力,值得推广。

【关键词】大语言模型;临床见习教学;临床胜任力;消化内科

【中图分类号】G642.4;R-4 **【文献标志码】**A **【文章编号】**1004-5287(2026)05-0579-09

【DOI】10.13566/j.cnki.cmet.cn61-1317/g4.202605001

基金项目:国家自然科学基金面上项目(82370619);西京医院医务人员培养助推项目(XJZT24LY33)

收稿日期:2025-09-29

作者简介:李若璞,助教,本科,研究方向为消化系统疾病。

通信作者:吕勇,讲师,主治医师,博士,研究方向为消化系统疾病。E-mail:lvyongxj@yeah.net

Application of large language model-assisted teaching mode in clinical internship teaching of gastroenterology

LI Ruopu[1], *LÜ Yong*[2], *QIN Niping*[2], *ZHENG Rong*[2], *TIAN Zuhong*[2], *WANG Jun*[3], *PAN Yanglin*[2], *FAN Daiming*[2]

Air Force Medical University: 1. *First Cadet Brigade, School of Basic Medicine*; 2. *Xijing Digestive Diseases Hospital, National Clinical Research Center for Digestive Diseases, State Key Laboratory of Integrated Prevention and Treatment of Digestive System Tumors, Institute of Holistic Integrative Medicine Development Strategy, Xi'an* 710032, *China*; 3. *Department of Gastroenterology*, 986 *Hospital of Xijing Hospital, Xi'an* 710054, *China*

【Abstract】Objective To evaluate the effect of a large language model (LLM)-assisted teaching mode in clinical internship teaching in gastroenterology. **Methods** A randomized controlled trial was conducted, enrolling 156 fourth-year clinical medical students participating in their first gastroenterology internship. They were randomly assigned to the experimental group (traditional teaching + LLM assistance, $n=78$) and the control group (traditional teaching, $n=78$). The total scores of clinical competence, including theoretical assessment, objective structured clinical examination (OSCE), and case analysis, were compared between the two groups. Secondary outcome measures included depth of knowledge, quality of clinical reasoning, system usability, and cognitive load. **Results** The total clinical competency score of the experimental group was significantly higher than that of the control group [(89.5±5.9) vs. (76.2±8.0), mean difference $\Delta=13.3$, 95%CI [11.0, 15.5], $F=138.678$, $P<0.001$]. The experimental group also scored higher in the theoretical assessment [(36.6±2.3) vs. (32.8±3.7), $F=60.880$, $P<0.001$], OSCE performance [(36.9±2.8) vs. (31.2±5.7), $F=78.912$, $P<0.001$], and case analysis [(15.9±4.3) vs. (12.1±5.6), $F=19.799$, $P<0.001$]. Regarding knowledge depth, a higher proportion of students in the experimental group achieved a concept map score ≥ 4 compared with the control group (85.9% vs. 70.2%, $\chi^2=4.457$, $P=0.035$), and the guideline update recognition rate was also higher (78.2% vs. 52.6%, $\chi^2=11.331$, $P<0.001$). In terms of clinical reasoning quality, the experimental group showed higher coverage of key diseases in differential diagnoses [(89.1±4.7)% vs. (67.8±7.3)%, $F=181.576$, $P<0.001$] and a lower proportion of misdiagnoses involving rare diseases [(2.1±1.5)% vs. (6.9±2.8)%, $F=199.090$, $P<0.001$]. During case analysis, the average fixation time on key information was shorter in the experimental group [(5.2±1.1)s vs. (6.7±1.5)s, $F=50.758$, $P<0.001$], and its number of information reviews was lower [(1.8±0.6) times vs. (2.9±0.9) times, $F=80.679$, $P<0.001$]. In terms of system usability and cognitive load, the experimental group reported higher system usability scale (SUS) scores [(85.6±7.2) vs. (72.3±9.1), $F=102.505$, $P<0.001$] and lower cognitive load scores [(48.3±6.5) vs. (56.7±7.8), $F=53.332$, $P<0.001$]. Subgroup analysis showed that in each subgroup, the total clinical competence score of the experimental group was higher than that of the control group. Among students in the top 30% of theoretical scores, the mean difference in clinical competency total scores between the experimental group and the control group was 15.14 (95%CI [13.36, 8.93]), which was higher than the corresponding difference of 9.17 (95%CI [4.52, 13.83]) among students in the bottom 30%, with a statistically significant difference ($F=212.008$, $P<0.001$). However, no statistically significant differences were observed between the mean differences in the computer technology proficiency subgroup ($\Delta=13.68$, 95%CI [11.29, 16.06]) and the non-proficiency subgroup ($\Delta=12.09$, 95%CI [6.54, 17.64]), or between the mean differences in the common typical cases subgroup ($\Delta=12.81$, 95%CI [10.38, 15.24]) and the complex difficult cases subgroup ($\Delta=14.62$, 95%CI [9.31, 19.93]) (all $P>0.05$). **Conclusion** LLM-assisted teaching can enhance knowledge integration and clinical decision-making abilities among medical students in gastroenterology internship and is worthy of further promotion.

【Keywords】 large language models; clinical internship teaching; clinical competency; gastroenterology

临床见习是医学生向合格医师转型的关键阶段，尤其在消化内科等专科领域，学生需在有限时间内掌握复杂疾病谱的诊疗思维与操作规范。消化内科疾病以症状隐匿性强、检查复杂性高、鉴别诊断多维为显著特征，其临床教学不仅要求医学生掌握食管、胃肠、肝胆胰等多系统病变的诊疗逻辑，更需培养整合病史采集、辅助检查与循证决策的综合能力[1]。传统教学模式依赖床旁带教与标准化病例讨论，但其效能常受制于多重瓶颈：①病例资源时空限制，罕见病与急重症接触机会不均；②师资同质化不足，不同教师

对指南更新与临床推理的传授存在差异;③个性化反馈滞后,学生难以及时识别知识盲区并调整学习路径[2]。尽管虚拟仿真与在线课程等数字化手段已部分缓解上述问题,但其交互性与动态适应性仍显局限,难以满足临床思维培养的高阶需求[3]。

近年来,大语言模型(large language model,LLM)的突破性进展为医学教育革新提供了新契机。研究表明,LLM可通过自然语言处理生成高度拟真的临床病例,并模拟病情动态演变,同时提供循证医学依据的实时决策支持[4]。这一特性使其在临床思维训练中展现出独特优势:一方面,动态病例库可突破真实场景的病例稀缺性[5];另一方面,即时反馈机制能加速"诊断—反思—修正"的学习闭环[4]。然而,现有研究多聚焦于理论知识的AI辅助传授,对临床胜任力(含操作技能与决策质量)的干预效果仍缺乏高质量证据,且较少关注专科教学的特定需求。例如,在消化内科教学中,如何通过LLM强化疾病机制与诊疗指南的整合应用,尚未有严格设计的随机对照试验予以验证。

基于此,本研究采用随机对照设计,旨在探讨LLM辅助教学模式对消化内科见习生临床胜任力的影响。

1　研究对象与方法

1.1　研究设计

本研究为一项优效性随机对照试验,采用平行组设计,旨在评估大语言模型(LLM)辅助教学模式在消化内科见习教学中的应用效果。研究严格遵循《赫尔辛基宣言》原则,并已通过空军军医大学西京医院伦理审查委员会审批。所有参与学生在试验前均签署书面知情同意书,并被告知可随时退出研究,且其退出行为不会对学业评价产生任何影响。本研究中进行的各项考核均属教学研究性质的附加评估,不纳入学生课程最终成绩的评定。学生的结业成绩仍依据原有教学大纲,由教研室统一评定,不受本研究干预。为保障教育公平,在研究数据收集完成后,我们为对照组学生提供了与试验组相同的LLM辅助系统,安排为期一个月的补充学习。

1.2　研究对象

以空军军医大学临床医学专业四年级本科生为研究对象。研究对象的纳入标准为:①首次参加消化内科见习;②已完成"内科学"基础课程且成绩≥75分;③自愿签署知情同意书。排除标准包括:①既往参与过人工智能辅助教学项目;②具有消化专科实习经历;③拒绝使用电子设备完成学习任务。

1.3　样本量计算

基于预实验数据,设定主要结局指标为临床胜任力总分(优效性界值$\Delta=10\%$),效应量Cohen's $d=0.8$,检验效能$1-\beta=0.8$,双侧检验水准$\alpha=0.05$。通过PASS 15.0软件计算,每组需62人,考虑20%脱落率,最终确定每组纳入78人,总样本量156人。

1.4　随机化与盲法

随机分组时,依据理论成绩(前30%、中间40%、后30%)和计算机操作熟练度(国际电脑使用执照,International Computer Driving Licence,ICDL;评分≥70分为熟练组,<70分为非熟练组)进行分层,确保各组在这些关键基线特征上均衡。分组信息密封于不透光编号信封中,由独立研究协调员在见习前24 h揭盲并通知受试者。由于干预措施差异,受试者与带教教师无法设盲,但禁止组间交流教学内容。结局评估者及数据分析人员对分组信息设盲。

1.5　干预方案

1.5.1　对照组

对照组采用传统教学方法,即采用床旁教学+理论授课。

(1)床旁教学

每周3次(每次2 h),按《消化科见习大纲》实施标准化教学,师生比1∶5,覆盖消化道出血、肝硬化、胰腺炎等典型病例。

(2)理论授课

每周2次(每次1.5 h),使用统一的PPT模板,内容涵盖解剖生理、病理机制至诊疗规范。

1.5.2　试验组

试验组采用传统教学+LLM辅助的方法,即在对照组基础上增加LLM辅助教学模块。LLM的技术底座为DeepSeek-R1。LLM辅助教学系统整合本院消化内科2018—2023年电子病历(脱敏后)1 200例、权威教材(《消化病学》第9版)及18部指南(如《中国急性胰腺炎诊治指南》),构建专科知识库。系统采用检索增强生成架构,基于DeepSeek-R1架构微调,采用医学文本强化训练策略,通过3名主任医师对500条生成内容进行盲审修正。在试验开始前,已邀请3名消化内科副主任医师及以上职称的临床专家,对系统生成的10个典型病例的"解读逻辑链"和"虚拟主治医师"提问序列进行了盲审评估(采用5分制Likert量表)。评估结果显示,逻辑合理性的平均分为4.6分,提问序列临床相关性的平均分为4.7分。LLM功能模块如图1所示。

(1)模块 1——虚拟问诊实验室

基于真实消化系统疾病病例(如慢性胃炎、肝硬化、炎症性肠病等),LLM 生成动态虚拟患者。学生通过语音/文字交互进行病史采集,系统实时标注遗漏项(如未询问饮酒史,需警惕酒精性肝病),并生成标准化问诊模板。以上腹痛 2 个月患者、学生未问及夜间痛为模拟案例,系统提示:十二指肠溃疡典型表现为饥饿痛、夜间痛,建议补充询问。

(2)模块 2——辅助检查智能解析

上传胃镜、肠镜、腹部 CT 等影像及检验报告(如 HBsAg 阳性、AFP 升高)后,LLM 自动生成解读逻辑链。例如:AFP>400 ng/mL 合并肝内占位,首先考虑肝癌,需结合增强 CT 的“快进快出”强化特点。

(3)模块 3——疑难病例讨论工坊

LLM 扮演“虚拟主治医师”,引导学生逐层分析。例如,针对不明原因腹水,系统依次抛出问题:漏出液还是渗出液?结核性还是肿瘤性?如何设计诊断性穿刺方案?并实时点评逻辑漏洞。

(4)模块 4——医患沟通模拟器

模拟“告知胃癌诊断”“解释肠镜风险”等场景,LLM 根据学生回应生成情感反馈(如患者表现出焦虑,需加强共情),并提供沟通话术建议。

所采用的整合方式为床旁教学后 30 min 内完成 LLM 强化训练,系统运行于统一配置的平板电脑(5G 网络优先)。

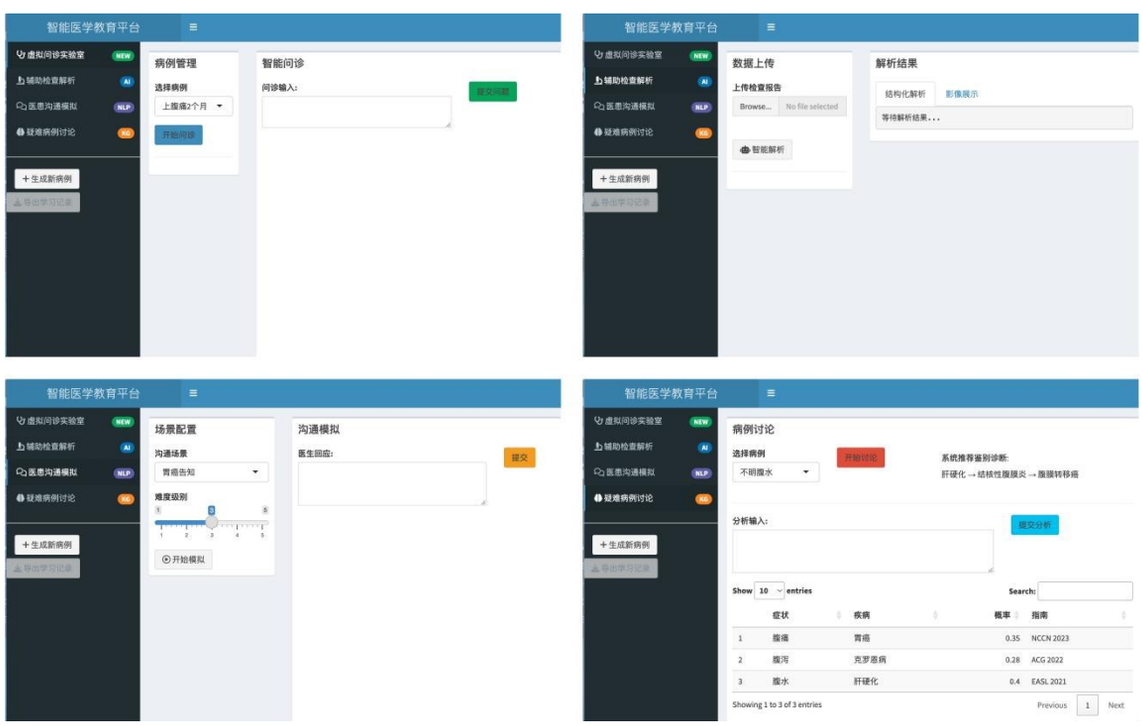


图 1 LLM 辅助教学模块

1.6 评价指标

1.6.1 主要结局指标

临床胜任力总分(0~100 分)由理论考核(40 分)、客观结构化临床考试(40 分)、病例分析(20 分)三部分构成。①理论考核:采用国家执业医师考试题库,经 IRT(item response theory,项目反应理论)模型校准难度(难度系数 0.65~0.75)。②客观结构化临床考试(objective structured clinical examination,OSCE):包含标准化病人问诊、腹部查体及医患沟通场景,由 3 名外院考官独立评分[6]。③病例分析:应用改良 RIME 框架(reporter-interpreter-manager-educator)评估病例处理能力[7-8]。

1.6.2 次要结局指标

(1)知识掌握深度

知识掌握深度由疾病机制理解度和指南更新识别率两部分构成。①疾病机制理解度。通过概念图评分法量化知识网络复杂度(概念图评分 0~5 分的综合评级标准:0 分代表“完全错误/缺失”,即未表现出任何相关理解;1 分代表“严重不足”,即仅有极少量、孤立的正确知识;2 分代表“有待提高”,即知识零散、不完整且存在错误;3 分代表“及格”,指具备基础核心知识但缺乏整合;4 分代表“良好”,即知识结构良好并有一定整合度;5 分代表“优秀”,即知识结构完整、深刻且高度整合),比较两组评分≥4 分的比例。

②指南更新识别率。设置5%的陷阱题(基于2023年肝硬化腹水诊疗指南修订内容),其余95%为非陷阱题(指南未修订的常规知识点)。识别率以正确回答题目占比(0～100%)计算。

(2)临床思维质量

临床思维质量由鉴别诊断质量和决策响应时间两部分构成。①鉴别诊断质量。由2名消化科专家盲评,评估关键病种覆盖率,即学生列出的鉴别诊断中包含由专家预先确定的该病例所对应的3～5个核心或易漏诊病种的比例(%)。同时,评估误诊罕见病的比例,即鉴别诊断中错误纳入罕见病的病例数占全体病例的比例(%)。②决策响应时间。采用Tobii Pro Nano眼动仪记录病例分析时的注视轨迹,响应时间以s为单位。

(3)系统易用性与认知负荷

系统易用性通过系统可用性量表(system usability scale,SUS)评估,评分范围0～100分。

认知负荷是反映学习过程中心理努力和投入度的重要指标,采用NASA任务负荷指数(NASA task load index,NASA-TLX)测量[9]。该量表从脑力需求、体力需求、时间需求、努力程度、业绩水平和受挫感六个维度进行评估,总分范围0～100分,分数越高代表认知负荷越大,说明心理努力和投入度越大。

1.6.3　评估时点

基线评估于见习前1周完成,终期评估于见习结束后48 h内实施。

1.7　统计分析方法

所有分析使用R 4.3.1软件完成,显著性水平设定为双侧检验水准$\alpha=0.05$。计量资料以($\bar{x}\pm s$)表示,组间比较采用独立样本t检验;计数资料以n(%)表示,组间比较采用χ^2检验或Fisher确切概率法;等级资料采用n(%)描述,组间比较采用Mann-Whitney U检验。主要分析遵循意向性分析原则,即将所有随机分组的受试者纳入分析。对主要结局指标临床胜任力总分进行优效性检验:以临床胜任力总分为因变量,组别为固定效应,基线理论成绩及计算机操作熟练度为协变量,建立协方差分析模型,计算试验组与对照组校正后的均值差及其95%CI,若其下限大于预先设定的优效性界值(Δ=总分的10%),则判定试验组优效。按理论成绩(前30%、中间40%、后30%)、计算机操作熟练度(ICDL评分≥70分为熟练组,<70分为非熟练组)及病例复杂程度(专家评议分为基础典型病例与复杂疑难病例)对主要结局指标临床胜任力总分进行亚组分析,比较不同亚组试验组和对照组研究对象的临床胜任力总分的差异。亚组分析采用与主分析一致的协方差分析。敏感性分析采用按方案集分析,即排除缺席>2次或完全退出的受试者,仅对完成全部教学与评估(缺席≤2次见习课程)的受试者进行分析,以评估干预在依从性完好人群中的效果。

2　结果

2.1　研究对象的基本特征

本研究共纳入符合标准的临床医学专业四年级学生156人,其中试验组(传统教学+LLM辅助组)78人,对照组(传统教学组)78人。基线数据显示,两组学生在性别、年龄、理论成绩分层及计算机操作熟练度分层分布均无统计学差异(均$P>0.05$)(表1)。研究期间试验组脱落3人(因设备操作困难退出2人、病假1人),对照组脱落5人(因学习压力大退出3人、个人原因2人),最终完成率分别为96.2%和93.6%($\chi^2=0.515$,$P=0.472$)。后台交互日志数据显示,试验组学生在2周见习期内,人均使用系统进行案例训练(12.5±3.2)次,每次平均交互轮次(问答)为(8.7±4.1)轮。研究过程中,试验组2人报告短暂性眼疲劳,经调整后缓解;对照组3人因学习压力出现轻度焦虑,经心理疏导后完成研究。未发生数据泄露等重大不良事件。

表1　研究对象的基本特征

特征	试验组(n=78)	对照组(n=78)	$\chi^2/t/z$	P
男性[n(%)]	32(41.0)	34(43.6)	0.123	0.735
年龄[($\bar{x}\pm s$),岁]	215±1.2	21.3±1.1	1.082*	0.282
理论成绩分层[n(%)]			−0.347**	0.856
前30%	23(29.4)	24(30.7)		
中间40%	31(39.7)	32(41.0)		
后30%	24(30.7)	22(28.2)		
计算机操作熟练度分层[n(%)]			0.130	0.718
熟练	58(74.4)	56(71.8)		
不熟练	20(25.6)	22(28.2)		

注:*为独立样本t检验的统计量t;**为Mann-Whitney U检验的统计量z。

2.2　主要结局指标

调整基线理论成绩和计算机操作熟练度后,试验组临床胜任力总分高于对照组[(89.5±5.9)分 vs.(76.2±8.0)分,均值差Δ=13.3分,95%CI[11.0,15.5],$F=138.678$,$P<0.001$],达到预设优效性界值(Δ>10%)。如表2所示,各维度分析显示如下。

理论考核:试验组得分高于对照组[(36.6±2.3)分

vs.(32.8±3.7)分],差异有统计学意义(F=60.880,P<0.001)。

OSCE考核:试验组OSCE考核得分高于对照组[(36.9±2.8)分 vs.(31.2±5.7)分],差异有统计学意义(F=78.912,P<0.001)。试验组在标准化病人问诊[(18.9±2.1)分 vs.(15.0±2.8)分,F=96.826,P<0.001]和腹部查体规范性[(8.9±1.1)分 vs.(7.3±1.6)分,F=52.998,P<0.001]得分高于对照组,差异均有统计学意义,但医患沟通得分无统计学差异[(9.1±1.3)分 vs.(8.9±1.5)分,F=0.735,P=0.374]。

病例分析:试验组应用改良RIME框架得分高于对照组[(15.9±4.3)分 vs.(12.1±5.6)分],差异有统计学意义(F=19.799,P<0.001)。

表2 两组临床胜任力总分及各维度得分比较

指标	试验组[n=75,($\bar{x}$±s),分]	对照组[n=73,($\bar{x}$±s),分]	调整后均值差*[Δ(95%CI),分]	F	P
临床胜任力总分	89.5±5.9	76.2±8.0	13.3[11.0,15.5]	138.678	<0.001
理论考核	36.6±2.3	32.8±3.7	3.8[2.9,4.8]	60.880	<0.001
OSCE考核	36.9±2.8	31.2±5.7	5.7[4.5,7.0]	78.912	<0.001
病例分析	15.9±4.3	12.1±5.6	3.8[1.9,5.7]	19.799	<0.001

注:*为采用协方差分析调整基线理论成绩和计算机操作熟练度后计算所得。

2.3 次要结局指标

2.3.1 知识掌握深度

疾病机制理解度:试验组概念图评分≥4分(良好和优秀)的比例高于对照组(85.9% vs. 70.2%),差异有统计学意义(χ^2=4.457,P=0.035)。

指南更新识别率:试验组对2023年肝硬化腹水诊疗指南修订内容的识别准确率高于对照组(78.2% vs. 52.6%),差异有统计学意义(χ^2=11.331,P<0.001),但对非陷阱题的识别率差异无统计学意义(91.0% vs. 89.7%,χ^2=0.074,P=0.786)。

2.3.2 临床思维质量

鉴别诊断质量:调整基线理论成绩和计算机操作熟练度后,试验组诊断条目中关键病种覆盖率高于对照组[(89.1±4.7)% vs.(67.8±7.3)%,F=181.576,P<0.001],校正后均值差Δ=21.3%,95%CI[19.1,23.5];且误诊罕见病的比例低于对照组[(2.1±1.5)% vs.(6.9±2.8)%,F=199.090,P<0.001],校正后均值差Δ=−4.8%,95%CI[−5.4,−4.2],差异有统计学意义。

决策响应时间:调整基线理论成绩和计算机操作熟练度后,试验组在病例分析时注视关键信息(如实验室异常值、影像学特征)的平均时间短于对照组[(5.2±1.1)s vs.(6.7±1.5)s,F=50.758,P<0.001],校正后均值差Δ=−1.5 s,95%CI[−1.9,−1.1];且信息回看次数少于对照组[(1.8±0.6)次 vs.(2.9±0.9)次,F=80.679,P<0.001],校正后均值差Δ=−1.1次,95%CI[−1.3,−0.9],差异有统计学意义。

2.3.3 系统易用性与认知负荷

系统易用性:调整基线理论成绩和计算机操作熟练度后,试验组SUS评分高于对照组[(85.6±7.2)分 vs.(72.3±9.1)分,F=102.505,P<0.001],校正后均值差Δ=13.5分,95%CI[11.0,15.9],差异有统计学意义。

认知负荷:调整基线理论成绩和计算机操作熟练度后,试验组NASA-TLX量表认知负荷总分低于对照组[(48.3±6.5)分 vs.(56.7±7.8)分,F=53.332,P<0.001],校正后均值差Δ=−8.4分,95%CI[−10.7,−6.1]。其中,在时间需求[(7.2±1.3)分 vs.(8.9±1.6)分,F=53.070,P<0.001]和努力程度[(6.8±1.1)分 vs.(8.1±1.4)分,F=41.653,P<0.001]两个维度,试验组得分均低于对照组,差异有统计学意义。

2.4 亚组分析

如表3所示,调整基线理论成绩和计算机操作熟练度后,在所有预设亚组(理论成绩、计算机操作熟练度、病例复杂程度)中,试验组的临床胜任力总分校正后均数均高于对照组,亚组内差异均有统计学意义(均P<0.001)。在理论成绩前30%的学生中,试验组与对照组临床胜任力总分的校正后均值差为15.14分(95%CI[11.36,18.93]),高于后30%学生中的相应差值9.17分(95%CI[4.52,13.83]),亚组间差异有统计学意义(F=212.008,P<0.001)。在计算机操作熟练亚组与不熟练亚组中,试验组与对照组临床胜任力总分的校正后均值差分别为13.68分(95%CI[11.29,16.06])和12.09分(95%CI[6.54,17.64]),亚

组间差异无统计学意义（$F=0.281$，$P=0.597$）。在普通典型病例亚组与复杂疑难病例亚组中，试验组与对照组临床胜任力总分的校正后均值差分别为12.81 分（95%CI[10.38，15.24]）和 14.62 分（95%CI[9.31，19.93]），亚组间差异无统计学意义（$F=0.106$，$P=0.745$）。

表 3　不同亚组内试验组与对照组临床胜任力总分的比较

亚组	试验组[$n=75$，($\bar{x}\pm s$)，分]	对照组[$n=73$，($\bar{x}\pm s$)，分]	调整后均值差*[Δ(95%CI)，分]	P
理论成绩				
前 30%	91.80±4.44	76.66±7.44	15.14[11.36，18.93]	<0.001
中间 40%	90.27±6.28	75.59±6.58	14.68[11.46，17.91]	<0.001
后 30%	85.63±4.80	76.46±9.78	9.17[4.52，13.83]	<0.001
计算机熟练程度				
熟练	89.65±5.87	75.97±7.05	13.68[11.29，16.06]	<0.001
不熟练	88.90±6.04	76.81±10.56	12.09[6.54，17.64]	<0.001
病例复杂程度				
普通典型病例	88.78±6.15	75.97±7.05	12.81[10.38，15.24]	<0.001
复杂疑难病例	91.41±4.64	76.81±10.56	14.62[9.31，19.93]	<0.001

注：*为采用协方差分析调整基线理论成绩和计算机操作熟练度后计算所得。

2.5　临床胜任力总分敏感性分析

按方案集分析，排除脱落病例后，试验组与对照组之间的临床胜任力总分差异仍有统计学意义（Δ=13.7 分，95%CI[11.2，16.1]，$P<0.001$），如表4所示。

表 4　基于不同数据集的临床胜任力总分敏感性分析

分析类型	试验组样本量	对照组样本量	调整后均值差*[Δ(95%CI)，分]	F	P
意向性分析（ITT）	78	78	13.3[11.0，15.5]	136.625	<0.001
按方案集（PP）	75	73	13.7[11.2，16.1]	99.182	<0.001

注：*为采用协方差分析调整基线理论成绩和计算机操作熟练度后计算所得。

3　讨论

3.1　主要研究发现与创新性

本研究通过分层随机对照试验显示，LLM 辅助教学对消化内科见习生临床胜任力具有积极作用。试验组采用传统教学联合基于 LLM 的动态病例系统，其临床胜任力总分较对照组高出 13.3 分，其中试验组理论考核、操作技能和病例分析得分均高于对照组。次要结局指标进一步显示，试验组在知识掌握深度、临床思维质量及系统易用性与认知负荷方面均优于对照组。亚组分析显示，基础理论成绩前 30% 的学生较基础理论成绩后 30% 临床胜任力提升幅度更大。然而，试验组与对照组在医患沟通能力方面未见显著差异。本研究支持 LLM 在复杂病例教学和个性化学习路径设计方面的应用潜力，为临床教学数字化转型提供了实证依据，未来可在分层教学与技术整合方面进一步优化实施策略。

本研究创新点在于：①开发融合时序症状演变的智能病例系统，模拟真实临床决策环境；②通过多模态评估（理论考核、OSCE、眼动数据等）量化 LLM 对知识掌握深度与临床思维的作用路径；③引入计算机操作熟练度分层，揭示技术适配性对教学效果的调节效应。

3.2　作用机制与理论解释

本研究设计为附加式试验，即试验组在接受与对照组完全相同的标准化床旁教学和理论授课基础上，额外使用 LLM 系统进行辅助练习。两组的主要授课教师、教学时长和内容完全一致，从而在最大程度上控制了教师因素的影响。本研究结果验证了认知负荷理论在 AI 教育应用中的有效性。试验组借助 LLM 实时交互系统和动态诊断树功能，使临床推理过程中的工作记忆需求显著降低（基于眼动追踪数据分析），从而释放更多认知资源用于高阶思维活动。这一现象符合认知负荷理论中的“脚手架效应”[10-11]，即 LLM 通过即时提示（如遗漏关键诊疗指征时触发警示），有效避免了因知识断层引发的认知过载。

试验组在指南更新识别率上的优势可能得益于系统的知识更新机制。LLM 每周自动整合 PubMed 最新文献，而传统教学组依赖人工更新课件（存在 4～6 周延迟）。这种实时更新特性在快速迭代的医学领域尤为重要，与数字教育中的动态知识建构理论相契合[12]。在个性化学习方面，LLM 基于适应性测试理

论构建的学习路径模块展现出独特优势。系统通过识别学生知识盲点生成靶向训练内容，避免了传统统一授课的重复学习问题。数据显示，试验组核心知识强化时间较对照组减少，但概念图复杂度等深度理解指标更高，这符合专业知识逆转效应理论[13]，即精准的个性化内容更适配高阶学习者的认知需求。

3.3 与现有研究的异同

本研究结果既印证了人工智能教育领域的研究共识，也揭示了专科医学教育的特殊性。与既往关于AI辅助医学教育的研究结论一致[14-16]，本研究发现智能系统在提升理论考核和操作规范性方面具有显著优势。但值得注意的是，两组在医患沟通能力培养上未显现组间差异，这与部分研究报道的虚拟病人训练效果存在差异[5,17]。可能原因包括：①本研究的LLM模块聚焦于临床决策支持，未专门设计沟通情境的深度模拟功能[18]；②传统教学组通过真实医患互动积累了情境化经验，这种面对面的情感交流具有不可替代性。这提示该方面仍存在局限性，未来需开发整合微表情识别和语音情感分析的多模态训练系统[19-20]。

研究还发现计算机操作熟练度对教学效果存在影响。这一发现深化了数字技术教育应用的理论认知[21-23]。在普遍具备较高数字素养的医学生群体中，技术适配性问题仍可能造成学习效果差异，这与技术接受模型的核心观点相契合[24]。因此，医学院在推进智能教学过程中，需要建立数字能力评估体系，并将基础技术培训纳入教学准备阶段，以保障教育公平性。

3.4 临床与教育实践意义

在实践层面，本研究为医学教育数字化转型提供了三大启示。

(1)教学场景重构

LLM可突破传统床旁教学的时间与空间限制。例如，学生在下班后仍可通过智能病例库反复训练罕见病（如Whipple病）的诊疗流程，而传统模式下此类病例接触机会不足。这种7×24 h沉浸式学习可能成为临床技能训练的新常态。

(2)评估体系升级

传统OSCE考核依赖标准化病人和人工评分，成本高且难以规模化[6]。本研究中，LLM通过自然语言处理技术对学生问诊录音进行自动评分（如关键词覆盖率、逻辑连贯性），其评估效率显著提升。未来可探索AI辅助考核的标准化方案。

(3)资源公平分配

在基层医院师资匮乏地区，LLM可作为虚拟教学助手，提供三甲医院水平的病例资源。例如，本研究病例库中的模拟病例涵盖十余种消化系统疾病亚型，而对照组仅能接触20～30个典型病例，这一差距在欠发达地区的教学中可能更为显著。

3.5 研究局限性

尽管研究设计严格，仍需正视以下局限。①样本代表性：受试者均来自同一医学院，且纳入标准要求理论成绩≥75分，可能高估LLM在更广泛群体中的效果。例如，对基础薄弱学生，LLM虽提升了理论分，但OSCE操作分无显著改善，提示该系统在技能转化上可能存在瓶颈。②随访时间：评估仅覆盖见习周期，缺乏长期效果数据。例如，试验组在指南更新识别率上的优势可能随时间衰减，需追踪3～6个月后的知识留存率。③技术依赖性：LLM系统依赖5G网络和高端设备，在欠发达地区推广面临硬件门槛。研究中试验组2人因设备操作困难退出，提示技术适配性需纳入实施方案考量。④盲法与霍桑效应：本研究虽为随机对照试验，但难以对参与者实施盲法，试验组可能因接触新技术而产生霍桑效应，即学习动机的暂时性提升，这可能是结果的一个潜在混杂因素。未来研究可设计更巧妙的盲法或设置主动控制组（如使用非AI辅助系统）以区分技术效应与动机效应。

3.6 未来研究方向

基于上述发现与局限，未来研究可从以下维度深入。①优化混合教学模式：探索LLM与传统教学（尤其是床旁人文沟通训练）的最佳整合方式与时间配比。②开发情感智能模块：增强LLM在医患沟通训练中对情感、共情等非语言信息的识别与反馈能力[19-20]。③验证跨专科普适性：在心内科、外科等不同专科中验证并优化该教学模式。④开展长期效果评估：进行纵向研究，评估LLM辅助教学对临床推理能力的长远影响。⑤构建伦理与数据规范：建立医学教育AI使用的数据安全、隐私保护及伦理审查框架。

4 结束语

本研究证实，基于大语言模型的辅助教学模式能显著提升消化内科见习生的临床胜任力，其核心优势体现在知识更新实时性、认知负荷优化和个性化学习路径三个方面。然而，其效果受学习者基础水平和技术素养的显著调节，提示需实施分层教学策略。尽管存在技术依赖和长期效果不确定等局限，LLM仍代表了医学教育范式变革的重要方向。未来研究应致力于构建“AI-教师-学生”三元协同机制，以实现教育资源最大化利用与人才培养质量提升的共赢。

参考文献

[1] 欧阳慧, 黄楚媚, 齐健. 消化内科见习带教效果的探讨[J]. 中国卫生产业, 2023, 20(14): 248-251.

[2] 徐艺可, 曹俊, 张凤兰, 等. 医学生见习面临的问题和对策: 构建和谐社会 促进患者参与临床实践教学[J]. 继续医学教育, 2011, 25(5): 4-6.

[3] 青青, 林春燕, 谢玥, 等. 内科学消化系统疾病在线见习实施策略与评价方法[J]. 中国教育技术装备, 2023(4): 43-47.

[4] 陈剑锋. 大语言模型在临床医学的可应用性探讨[J]. 医学与哲学, 2023, 44(21): 1-6.

[5] 周志浩, 宋佳琳, 刘嘉成, 等. 基于人工智能LLM技术的虚拟患者系统构建与临床教学应用[J]. 医学新知, 2024, 34(7): 833-842.

[6] 李翔, 梁杰, 张慧群, 等. 临床医学专业OSCE考核情况分析[J]. 高校医学教学研究(电子版), 2024, 14(4): 49-55.

[7] Pangaro L, Rodriguez R G, Hemmer P A. Reporter-interpreter-manager-educator: An observational framework, not a grading framework[J]. Academic Medicine, 2023, 98(1): 9-10.

[8] Desjardins V, Pageau P, Power B, et al. Comparison of clerkship directors' expectations of physical examination skills with point-of-care ultrasound skills using the RIME framework[J]. POCUS Journal, 2021, 6(2): 93-96.

[9] 梁丽玲, 赵丽, 邓娟, 等. NASA-TLX量表的汉化及信效度检验[J]. 护理研究, 2019,33(5): 734-737.

[10] 姜维政, 邵晓晖, 胡晓玲. 有关在线学习与认知负荷理论的综述及思考[J]. 知识经济,2023(33): 116-119.

[11] 杨希, 李峙, 张珣, 等. 基于认知负荷理论的中外学生麻醉学情境模拟教学设计[J]. 医学教育管理, 2024, 10(3): 344-349.

[12] 蒋纪平, 满其峰, 胡金艳, 等. 在线协作知识建构的知识进化: 内涵、本质与模型[J]. 开放教育研究, 2022, 28(6): 52-59.

[13] 肖玉琴, 丁道群. 专业知识逆转效应及其给多媒体学习的启示[J]. 心理研究, 2008,1(6): 32-35.

[14] 刘睿楠, 王余广, 胡海峰, 等. 人工智能在医学影像诊断学教学中的应用[J]. 中国卫生产业, 2024, 21(16): 144-146.

[15] 王妙苗, 王湛, 臧远胜. 人工智能辅助教学在肿瘤学继续教育培养中的作用探讨[J]. 当代教育实践与教学研究(电子刊), 2021(19): 61-63.

[16] 潘柔百, 陶蓉. 浅谈基于大型语言模型的聊天机器人对医学教育的影响[J]. 内科理论与实践, 2023, 18(6): 439-446.

[17] 杨秀华, 孟涛. 虚拟患者在产科临床见习中的应用[J]. 中国继续医学教育, 2019, 11(19): 46-49.

[18] 顾文君, 孟祥军. 情境模拟教学在消化内科教学的应用初探[J]. 科教导刊(电子版),2024(10): 106-108.

[19] 张蕊波. 人工智能时代大学生情感教育面临的挑战及其重构[J]. 高教学刊, 2023,9(20): 74-77.

[20] 侯应龙, 朱明琪, 宋启元, 等. ChatGPT之于临床医生是助手而非替手[J]. 医学与哲学, 2024, 45(14): 1-5.

[21] 姜晶波, 王启帆. 数字化赋能医学教育内在机理与现实路径[J]. 医学信息学杂志, 2024,45(2): 98-101.

[22] 郭颖. 教育数字化背景下医学数字教材建设思考[J]. 新闻研究导刊, 2025, 16(1):225-228.

[23] 徐晓巍, 唐琳, 归航, 等. 数字赋能的医学教育创新模式探析[J]. 中华医学教育杂志, 2025, 45(3): 161-166.

[24] 李继学, 巩见刚, 高照军. 技术接受模型理论研究的可视化分析[J]. 现代情报, 2015,35(12): 99-105.