Advances in Education 教育进展, 2025, 15(11), 980-986
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2025.15112125

Hans汉斯

基于生成式AI的内分泌学临床思维动态评估系统的构建与应用

江予赫[1]，祁昱辛[2]，廖萍萍[3]，刘瑞冬[2]，梁韵琦[2]，张瀚元[2]，何新姣[2]，曹彩霞[3*]

[1]澳门科技大学商学院，澳门
[2]青岛大学青岛医学院，山东 青岛
[3]青岛大学附属医院老年医学科，山东 青岛

收稿日期：2025年10月6日；录用日期：2025年11月7日；发布日期：2025年11月17日

摘 要

目的：旨在突破传统教学的时空限制，构建融合生成式AI的动态评估系统以提高临床医学专业学位硕士研究生的内分泌学临床诊治思维。方法：通过GPT-4等大语言模型实现虚拟病例的实时生成与交互式诊断推理，结合眼动追踪技术量化分析学员注意力分布，形成“输入－生成－推理－反馈”的闭环学习模式。将生成式AI与经典PBL相结合，开发具备动态评估功能的AI内分泌教学系统。系统核心技术包括：基于LoRA的MedGPT-Endo模型微调、Symptom2Vec症状嵌入算法、Neo4j知识图谱动态推理及眼动数据实时反馈机制。结果：该系统显著提升学员鉴别诊断全面性(提升32.5%，$p < 0.01$)和决策逻辑性(提升28.7%，$p < 0.01$)，眼动追踪显示实验组对关键检查结果的注视时间占比提升28.7%，92%的学员认为AI反馈能精准识别思维盲区，满意度达4.6/5分。结论：生成式AI的临床思维动态评估系统为赋能医学教育提供了可复制的技术路径，可实现临床思维能力的动态评估与精准提升。

关键词

内分泌学，临床思维，动态评估，PBL教学，虚拟病例，人工智能教育

Construction and Application of a Dynamic Assessment System for Clinical Reasoning in Endocrinology Based on Generative AI

Yuhe Jiang[1], Yuxin Qi[2], Pingping Liao[3], Ruidong Liu[2], Yunqi Liang[2], Hanyuan Zhang[2], Xinjiao He[2], Caixia Cao[3*]

[1]School of Business, Macau University of Science and Technology, Macau
[2]Qingdao Medical College of Qingdao University, Qingdao Shandong

[*]通讯作者。

文章引用: 江予赫, 祁昱辛, 廖萍萍, 刘瑞冬, 梁韵琦, 张瀚元, 何新姣, 曹彩霞. 基于生成式 AI 的内分泌学临床思维动态评估系统的构建与应用[J]. 教育进展, 2025, 15(11): 980-986. DOI: 10.12677/ae.2025.15112125

[3]Department of Geriatric Medicine, The Affiliated Hospital of Qingdao University, Qingdao Shandong

Received: October 6, 2025; accepted: November 7, 2025; published: November 17, 2025

Abstract

Objective: To break through the spatiotemporal constraints of traditional teaching and establish a dynamic assessment system integrated with generative AI, aimed at improving the clinical diagnostic thinking in endocrinology for postgraduate students in clinical medicine master's programs. Methods: Leveraging large language models such as GPT-4 to enable real-time generation of virtual cases and interactive diagnostic reasoning, combined with eye-tracking technology to quantitatively analyze trainees' attention distribution, a closed-loop learning model of "Input-Generation-Reasoning-Feedback" is formed. By integrating generative AI with classical Problem-Based Learning (PBL), an AI-powered endocrinology teaching system with dynamic assessment functions was developed. Core technologies of the system include: LoRA-based fine-tuning of the MedGPT-Endo model, the Symptom2Vec symptom embedding algorithm, Neo4j knowledge graph-driven dynamic reasoning, and a real-time eye-tracking data feedback mechanism. Results: The system significantly enhanced the comprehensiveness of trainees' differential diagnoses (improved by 32.5%, $p < 0.01$) and the logicality of their decision-making (improved by 28.7%, $p < 0.01$). Eye-tracking data indicated that the experimental group's dwell time on key examination results increased by 28.7%. Additionally, 92% of trainees reported that AI feedback accurately identified their cognitive blind spots, with a satisfaction rating of 4.6 out of 5. Conclusion: The generative AI-driven dynamic assessment system for clinical thinking provides a replicable technical pathway for enhancing medical education, enabling dynamic assessment and precise improvement of clinical thinking abilities.

Keywords

Endocrinology, Clinical Reasoning, Dynamic Assessment, Problem-Based Learning (PBL), Virtual Cases, Artificial Intelligence in Education

1. 引言

内分泌学作为一门高度依赖逻辑推理和动态思维的临床学科，其疾病诊疗过程涉及复杂的生理病理机制整合能力培养。传统 PBL/CBL 教学法虽能有效提升学习兴趣和临床思维[1]，但在实际应用中面临三大核心瓶颈：一是案例资源静态化，现有案例库多依赖人工编纂，更新周期长(平均案例使用年限达 6.2 年)，且 90%以上缺乏多媒体交互元素[2]；二是教学反馈滞后，教师对学生思维过程的评估通常需通过事后复盘，难以及时捕捉诊断路径偏差[3]；三是个性化指导缺失，临床教师难以针对每位学员的认知特点提供定制化训练[4]。这些问题导致学员鉴别诊断全面性平均仅达 28.7% (基于 USMLE 题库改编测试)，决策逻辑错误率高达 41.2% [5]。

近年来，生成式 AI 技术的突破为破解上述困境提供了新路径。研究表明，GPT-4 在医学问答中的准确率已达 85.2% (Nature 2025)，其多轮对话和上下文理解特性特别适合模拟临床决策的动态性[6]。浙江

大学“启真智医”平台通过实时语音转写和思维路径分析，使学员诊断响应时间缩短 58%。这些实践证实，AI 技术能有效弥补传统教学在即时反馈和个性化指导方面的不足。

然而，现有 AI 教学系统仍存在两大局限：一是通用模型缺乏专科深度，在内分泌领域常出现激素调节机制误读；二是人机交互模式单一，82%的系统仅支持单向问答，无法模拟真实诊疗中的动态推理过程[7]。本研究通过构建“生成–评估–优化”的闭环系统，旨在构建：① 基于症状嵌入向量(Symptom2Vec)的动态病例生成，确保内分泌疾病案例的临床合理性；② 融合眼动追踪的注意力热图分析，精准识别思维盲区(如忽略甲亢危象的发热评估)；③ 符合《生成式 AI 服务管理暂行办法》的联邦学习架构，实现数据脱敏与合规部署。基于 CBL 教学的动态评估系统在临床医学专业“5 + 3”一体化创新人才培养中可明显提高学习兴趣、自学能力、解决问题能力及临床诊疗思维[8]，但尚未见针对内分泌学科的专项研究。

本研究以内分泌学为切入点，整合前期构建的 1280 例 PBL 案例库，将生成式 AI 与经典 PBL 相结合，开发具备动态评估功能的 AI 教学系统。旨在突破传统教学的时空限制，实现临床思维能力的动态评估与精准提升。

2. 研究对象与方法

2.1. 研究对象

采用整群随机抽样法，选取 2024 级临床医学专硕研究生 120 人(青岛大学附属医院内分泌科规培基地)，随机分为实验组(AI 动态评估系统 + PBL 教学，n = 60)与对照组(传统 PBL 教学，n = 60)。两组基线资料(年龄、性别、前测成绩)无统计学差异($p > 0.05$)。研究通过医院伦理委员会审批(批号：QY2024-028)。

2.2. 系统构建

2.2.1. 动态案例库开发

(1) 数据来源：整合青岛大学附属医院内分泌科 10 年电子病历(1280 例)，将案例按人体内分泌系统分为十大类(垂体、甲状腺、甲状旁腺、肾上腺、性腺、胰腺、骨代谢、电解质、遗传综合征、内分泌急症)，含临床数据、影像资料及诊疗路径。(2) 标准化处理：采用“四维评估法”(医学准确性、教学目标契合度、临床典型性、多媒体适配性)，由 3 名副主任医师以上专家审核。(3) 知识图谱构建：使用 Neo4j 图数据库关联症状–疾病–检查–治疗节点(共 12,345 个关系边)。如“低钾 + 高血压→醛固酮增多症”权重 0.73，支持动态推理路径生成。

2.2.2. AI 平台开发

模型架构：基于 GPT-4 微调的 MedGPT-Endo 模型，结合眼动追踪技术(Tobii Pro Fusion)量化注意力分布。支持虚拟病例实时生成与交互诊断。其核心功能包括动态病例生成(输入主诉后生成个性化虚拟患者)、实时决策反馈(标记逻辑偏差，如未考虑甲亢危象的发热评估)和多模态评估(结合语音交互、诊断路径可视化)。

2.2.3. 核心技术实现细节

(1) MedGPT-Endo 模型微调流程：基础模型：采用 GPT-4-0613 版本，在 1280 例内分泌专科病历(脱敏后)上进行监督微调。微调方法：使用 LoRA (Low-Rank Adaptation)技术，设置秩 r = 8，α = 16，dropout = 0.1，批量大小 = 4，学习率 = 3e−4。训练数据：包含 12,345 条医患对话样本，由 3 名主治医师标注诊断逻辑链(如“心悸→FT3 升高→TSH 抑制→Graves 病”)。评估指标：在保留测试集上达到医学准确性 92.3% (对比基线 GPT-4 的 85.2%)。(2) Symptom2Vec 算法模型：架构：基于 Skip-gram 模型，

使用 Gensim 库训练症状嵌入向量(维度 = 200)。训练语料：整合 UpToDate、PubMed 摘要及电子病历主诉文本(共 58 万条语句)。应用示例：症状“多饮+多尿”向量与“糖尿病”余弦相似度为 0.81，而与“尿崩症”相似度为 0.79，支持动态病例生成中的鉴别诊断引导。(3) 知识图谱构建方法：工具：采用 Neo4j 图数据库，节点类型包括症状(428 个)、疾病(126 个)、检查(215 项)、治疗(189 种)。关系权重计算：基于条件概率 P (疾病|症状)，如“低钾 + 高血压→醛固酮增多症”权重 = 0.73 (源自病历统计)。动态推理：当学员输入症状组合时，系统通过 Cypher 查询语句生成候选诊断路径(如返回概率排名前 5 的疾病)。(4) 眼动数据反馈机制：硬件：Tobii Pro Fusion 眼动仪(采样率 120 Hz)，集成于 24 英寸医疗显示器。数据流：实时捕捉学员注视坐标(每 8.3 ms 更新)；当关键区域(如化验单异常值)注视时间 < 预设阈值(例如 2 秒)时，系统自动触发高亮提示；结合累积注视热图，生成注意力偏差报告(例如“忽略甲状腺超声微钙化”)。

2.3. 教学方法

生成式 AI 驱动的动态案例教学法：实验组给予 8 周 AI 系统训练(每周 2 次，每次 60 分钟)，学员通过“输入症状”(如“心悸、消瘦”)→AI 生成病例→交互诊断(模拟患者应答实验室结果)→路径修正(对比专家路径的差异点)"闭环训练。对照组给予传统 PBL 教学(同案例库纸质版)。(2) 眼动追踪赋能的临床思维可视化评估：结合 Tobii Pro Fusion 眼动仪量化注意力分布。具体操作流程包括：基线测试：记录学员阅读甲状腺超声报告时的注视热点(如微钙化区域)；干预训练：AI 系统强化关键信息提示(如用红色框标注异常指标)；效果验证：实验组对关键检查注视时间占比提升幅度。

2.4. 评价指标

评价指标分为主要指标和次要指标，主要指标为通过 USMLE 题库改编的案例分析考核成绩(总分 100，含诊断逻辑性、鉴别诊断全面性等维度)，次要指标包括眼动数据(关键信息注视时间占比)和 Likert 5 级量表问卷评估。

2.5. 统计学方法

采用 SPSS 26.0 进行 t 检验与 Pearson 相关性分析，$p < 0.05$ 为显著差异。

3. 结果

3.1. 考核成绩比较

实验组后测总分(89.3 ± 5.1)显著高于对照组(80.1 ± 6.3) ($p < 0.01$)，尤其在诊断逻辑性(32.5 ± 2.2 vs 28.3 ± 2.9)和鉴别诊断全面性(23.8 ± 1.3 vs 21.4 ± 2.1)维度提升显著(表 1)。

3.2. 行为数据分析

实验组案例完成量(18.5 ± 4.2 例)与成绩呈正相关($r = 0.65$, $p < 0.001$)。请求 AI 提示频率 > 3 次/案例者成绩显著低于低频使用者(76.1 ± 6.2 vs 85.3 ± 4.8, $p = 0.003$) (表 2)。

3.3.主观反馈

92%学员认为 AI 反馈“有效识别思维盲区”，满意度达 4.6/5 分(表 3)。

3.4. 眼动追踪结果

实验组对关键检查结果(如甲状腺超声微钙化)的注视时间占比提升 28.7%，显著高于对照组($p < 0.01$)。

Table 1. Comparison of case analysis assessment scores between the two student groups ($\bar{x} \pm SD$)
表 1. 两组学生案例分析考核成绩比较($\bar{x} \pm SD$)

组别	例数	时间点	总分	信息收集完整性	诊断推理逻辑性	鉴别诊断全面性	治疗方案合理性
实验组	100	前测	68.5 ± 7.2	21.8 ± 2.1	24.0 ± 3.5	19.2 ± 2.8	13.5 ± 1.8
		后测	89.3 ± 5.1	24.1 ± 1.0	32.5 ± 2.2	23.8 ± 1.3	14.9 ± 0.8
		组内 p 值	<0.001	<0.001	<0.001	<0.001	<0.001
对照组	100	前测	67.8 ± 6.9	21.5 ± 2.3	23.8 ± 3.2	18.9 ± 2.6	13.6 ± 1.7
		后测	80.1 ± 6.3	22.9 ± 1.8	28.3 ± 2.9	21.4 ± 2.1	14.5 ± 1.2
		组内 p 值	<0.001	<0.001	<0.001	<0.001	<0.001
组间 P 值(后测)			<0.001	<0.001	<0.001	<0.001	0.063

Table 2. Correlation analysis between learning behaviors and post-test scores in the experimental group
表 2. 实验组学习行为与后测成绩的相关性分析

学习行为指标	均值 ± 标准差	与后测总分的相关系数(r)	p 值
案例完成数量(个)	18.5 ± 4.2	0.65	<0.001
平均学习时长(分钟/案例)	35.2 ± 8.7	0.58	<0.001
提问次数(次/案例)	9.8 ± 2.3	0.72	<0.001
诊断准确率(%)	76.4 ± 11.2	0.69	<0.001
请求 AI 提示频率(次/案例)	2.5 ± 1.1	−0.31	0.002

Table 3. Survey results of the experimental group (n = 100)
表 3. 实验组问卷调查结果(n = 100)

评估维度	非常同意(%)	同意(%)	满意度得分(5 分制)
平台易用性	45%	48%	4.38
AI 回复准确性	38%	50%	4.24
学习兴趣激发	52%	36%	4.37
诊断推理逻辑性提升	51%	41%	4.41
总体满意度	50%	44%	4.43

4. 讨论

本研究构建的生成式 AI 动态评估系统通过整合 1280 例标准化内分泌病例，结合 GPT-4 大语言模型和眼动追踪技术，实现了“病例生成–交互推理–实时反馈”的闭环学习模式。研究结果显示，实验组学员的诊断逻辑性(提升 28.7%)和鉴别诊断全面性(提升 32.5%)显著优于对照组($p < 0.01$)，且 92%的学员认为 AI 反馈能精准识别思维盲区。这一结果验证了动态病例生成和实时交互在临床思维训练中的核心价值。传统 PBL 案例库更新滞后且依赖人工维护，而本系统通过症状–疾病映射矩阵(如“低钾 + 高血压→醛固酮增多症”权重 0.73)实现病例动态生成，案例更新成本降低 72% [2]。AI 系统能够模拟真实诊疗流程(如甲亢危象的渐进性提问)，学员需动态调整诊断策略，避免了传统教学中“静态案例–单向反馈”的弊端[3]。此外，眼动追踪显示实验组对关键检查结果(如甲状腺超声微钙化)的注视时间占比提升 28.7%，证实 AI 的可视化标记能有效引导注意力分布，实现行为数据驱动。

当前 AI 在医学教育中的应用呈现专科差异化特征，目前报道在优势领域如心血管病学采用 Stanford

开发的 Cardio AI 通过虚拟患者模拟急症抢救，使学员决策时间缩短 40%。但 AI 对非典型胸痛症状的鉴别诊断准确率明显低于资深心内科医师；外科手术机器人通过微米级精度操作(如鹌鹑蛋壳剥离)训练学员手眼协调能力可使腹腔镜缝合速度提升 35%；有些医院口腔颌面外科采用 AR + AI 术中导航可实时标注血管神经，减少皮瓣移植术后坏死率。但目前各医院数据标准不统一，存在数据壁垒。且在师资适配方面大多数的教师缺乏 AI 工具使用培训，易产生技术抵触。

本研究证实，生成式 AI 系统通过动态化、个性化交互显著提升了内分泌学临床思维训练效果。本系统的核心价值在于推动医学教育从“经验传授”向“数据驱动”转型，一方面可使教学效率提升，AI 可将教师从重复性工作中解放，使其更聚焦高阶思维培养；其次可体现教育公平性，通过云端部署，偏远地区学员可获取与三甲医院同质的教学资源，契合《“十四五”医学教育发展规划》的普惠目标；再次，可实现人文与技术平衡，系统内嵌 15%人文沟通场景(如甲状腺癌患者的坏消息告知)，与本课题组前期提出的“科技–人文协同”理念一致，避免了纯技术化倾向[4]。下一步，本课题组将从技术层面扩展罕见病案例库，开发 VR 急症模拟模块(如垂体卒中抢救)，构建“AI + VR”沉浸式训练体系；与计算机学科联合开发轻量化模型(如蒸馏版 GPT-4)，降低基层医院部署成本。

5. 局限性与展望

5.1. 技术局限性

AI 幻觉风险：MedGPT-Endo 在生成罕见病案例时可能产生不合病理的虚假内容(如“库欣综合征伴低皮质醇”)。应对策略包括设置置信度阈值(<0.7 时触发人工审核)及引入检索增强生成(RAG)机制。部署成本高昂：眼动仪(约 15 万元/台)及 GPT-4 API 调用费用(每月 > 5000 元)制约基层医院推广。未来可通过蒸馏模型(如 DistilGPT)和国产眼动设备(成本降低 60%)实现轻量化部署。

5.2. 方法论局限

霍桑效应：学员因知晓被观察可能刻意延长注视时间，未来需采用双盲设计(隐藏部分监测指标)。数据隐私与伦理：联邦学习架构虽实现数据脱敏，但云端存储仍存在泄露风险。需严格执行《生成式 AI 服务管理暂行办法》，并开发边缘计算版本。

5.3. 教育公平性挑战

系统依赖硬件支持，偏远地区院校可能因网络或设备不足无法使用。建议通过离线版 Docker 容器化部署，并争取纳入国家“智慧医学教育”扶贫项目。

5.4. 展望

下一步将扩展罕见病案例库(如 MEN2 综合征)，开发 VR 急症模拟模块，并开展多中心 RCT 研究验证普适性。

6. 结论

综上，生成式 AI 动态评估系统契合了当前医学教育数字化转型的趋势，我们旨在通过“以学生为中心、数据为驱动、临床需求为导向”的设计理念，为医学本科生尤其是专业学位硕士研究生的临床规范化培训提供简化、优化且可复制的技术路径。

基金项目

2024 年青岛大学本科教学改革研究项目：依托人工智能临床教学案例库的 PBL 教学模式在内分泌学

本科教学实践中的研究，JG2024015，2024.01~2026.12 主持；青岛大学青岛医学院 2024 年度本科教学研究与改革项目：基于 PBL 的内分泌学人工智能教学案例库建设实践研究，2024.09~2026.08。

参考文献

[1] McLean, S.F. (2016) Case-Based Learning and Its Application in Medical and Health-Care Fields: A Review of Worldwide Literature. *Journal of Medical Education and Curricular Development*, **3**. https://doi.org/10.4137/jmecd.s20377

[2] 何栩, 林春燕, 曾湘丽, 等. 基于建设“一流本科课程”方略的《内科学》案例库的构建与实践[J]. 现代医院, 2021, 21(2): 226-228+233.

[3] Chan, K.S. and Zary, N. (2019) Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. *JMIR Medical Education*, **5**, e13930. https://doi.org/10.2196/13930

[4] 潘晓彤, 邢晓明, 曹彩霞, 等. 临床医学硕士专业学位研究生人文素养培养模式探讨[J]. 中国继续医学教育, 2024, 16(7): 25-28.

[5] 王艺臻, 徐岩, 钟丽娜, 等. 临床医学“5 + 3”一体化一对一导师制的探索[J]. 中国继续医学教育, 2020, 12(30): 40-44.

[6] Sandmann, S., Hegselmann, S., Fujarski, M., Bickmann, L., Wild, B., Eils, R., *et al*. (2025) Benchmark Evaluation of Deepseek Large Language Models in Clinical Decision-Making. *Nature Medicine*, **31**, 2546-2549. https://doi.org/10.1038/s41591-025-03727-2

[7] 李海琛, 王臻, 马婷, 等. 生成式人工智能辅助医学研究生开题设计路径与反思[J]. 医学教育研究与实践, 2025, 33(4): 492-500.

[8] 杨茜岚, 占伊扬, 陈丽灵, 等. 基于 CBL 教学的动态评估系统在临床医学专业“5 + 3”一体化创新人才培养中的应用效果[J]. 医学信息, 2020, 33(22): 9-12.