 DOI: https://doi.org/10.12345/yzlcyxzz.v9i1.35830

A Preliminary Study on the Comparison Between SI-SP and Human Scoring

Liying Wang[1] Dongying Ni[2] Haifeng Gao[2]

1. Medical Simulation Center, Shanghai Medical College, Fudan University, Shanghai, 200032, China
2. Academic Affairs Office, Shanghai Medical College, Fudan University, Shanghai, 200032, China

Abstract

Objective: To optimize the scoring items and accuracy of an AI-based virtual standardized patient (AI-SP) by comparing AI and human examiner scores, and to explore potential pathways for improving human-machine integrated scoring models in the future. **Methods:** An Objective Structured Clinical Examination (OSCE) was conducted among undergraduate students majoring in clinical medicine in the 2025 academic year. During history taking, AI-SP was employed, and objective scoring items in the cases were double-scored by both AI and human examiners. **Results:** Across various cases, the proportion of score discrepancies (≥12 points) between AI and human examiners generally exceeded 40%, indicating relatively large scoring deviations. Among these, cases in internal medicine and pediatrics showed smaller scoring deviations compared to those in surgery and obstetrics and gynecology. AI-SP scores were generally lower than human scores. **Conclusion**s: The current AI scoring system still has several limitations and can't fully replace human scoring. Further exploration is needed to develop an intelligent medical education model that integrates AI and human scoring.

Keywords

Artificial Intelligence; Medical History Taking; Virtual Standardized Patients (AI-SP); Scoring System; Medical Education

基于 AI 的虚拟标准化病人与人工评分比较初探

王莉英[1] 倪东颖[2] 高海峰[2]

1.复旦大学上海医学院医学模拟教育中心，中国 · 上海 200032
2. 复旦大学上海医学院教务处，中国 · 上海 200032

摘 要

目的：通过AI与人工考官评分对比，对基于AI的虚拟标准化病人（AI-SP）评分条目设置、评分精准性方面进行优化，探讨未来人机融合评分模式的优化路径。**方法：**对2025年度临床医学专业本科毕业生进行客观结构化临床考试（OSCE），在病史采集使用AI-SP，将案例中客观性评分条目由AI和人工考官进行双重评分。**结果：**各案例中AI与人工评分差值≥12分的普遍占比40%以上，评分偏差相对较大。其中内科、儿科案例相较外科、妇产科案例的评分偏差小。AI-SP评分普遍低于人工评分。**结论：**目前AI评分系统仍存在诸多局限，尚不能完全替代人工评分，需进一步探索AI与人工评分融合的智能化医学教育新模式。

关键词

人工智能；病史采集；虚拟标准化病人（AI-SP）；评分系统；医学教育

1 引言

病史采集不仅是诊断的重要基础，更是医患沟通能力培养的关键训练内容。目前在医学教学与考核中，广泛采用标准化病人配合人工教师或考官的方式，对学生进行病史采集教学与评估。标准化病人（Standardized Patients，简称SP），指经过标准化、系统化培训后，能准确表现病人的实际临床问题的正常人或病人。SP培训周期长，使用成本高，维护难度大，流失率和更新率高，基层医学院校使用困难较大，且在使用过程中，存在同质化程度低、评分主观性强、重复性差等问题。随着 AI 与自然语言处理技术的发展，基于 AI 的虚拟标准化病人（AI-SP）被逐步引入医学教育场景，其成本相对低廉、应用简单，内置的自动评分系统更为标准化、规模化教学与考核提供了新思路[1-5]。

本研究的目的是进一步聚焦 AI-SP 的应用，并将 AI 与人工考官评分进行对比分析，以期在 AI-SP 应用过程中对评分条目设置、评分精准性方面进行优化，探讨未来人机融合评分模式的优化路径，使病史采集的教学与评估成本更低廉、应用更简便、评估更标准。

【作者简介】王莉英（1978–），女，中国上海人，硕士，副教授，从事医学教育与医学模拟教育研究。

2 研究方法

研究对象：复旦大学上海医学院临床医学本科毕业生（19级临床医学八年制和20级临床医学五年制）共384人。

研究时间：2025 年 5 月 13-16 日

研究地点：复旦大学上海医学院医学模拟教育中心

研究方法：对 2025 年度临床医学专业本科毕业生进行客观结构化临床考试（OSCE），首次在病史采集使用 AI-SP，将案例中客观性评分条目由 AI 和人工考官进行双重评分。

病史采集站共考核内科案例：2 型糖尿病、冠心病、甲状腺功能亢进症、哮喘；外科案例：急性阑尾炎、急性胆囊炎、消化道穿孔、急性胰腺炎；妇产科案例：前置胎盘、滴虫性阴道炎、异位妊娠、子宫肌瘤；儿科案例：小儿肺炎、小儿惊厥、小儿先天性心脏病、小儿腹泻病。每名考生随机选取 1 个内科或儿科案例 +1 个外科或妇产科案例进行病史采集考核。

3 研究结果

16 个案例，每个案例的考生均为 48 人，共 768 人次。每个案例客观性评分条目均为 60 分。

表 1 各案例 AI–SP 与人工评分差值

		AI 与人工评分差值 ≤6 分（所占百分比）	AI 与人工评分差值 6~12 分（所占百分比）	AI 与人工评分差值 ≥12 分（所占百分比）
内科案例	2 型糖尿病	14（29.2%）	10（20.8%）	24（50%）
	冠心病	10（20.8%）	18（37.5%）	20（41.7%）
	甲状腺功能亢进症	5（10.4%）	12（25.0%）	31（64.6%）
	哮喘	14（29.2%）	20（41.7%）	14（29.2%）
外科案例	急性阑尾炎	7（14.6%）	21（43.8%）	20（41.7%）
	急性胆囊炎	0（0%）	4（8.3%）	44（91.7%）
	消化道穿孔	8（16.7%）	7（14.6%）	33（68.8%）
	急性胰腺炎	2（4.2%）	4（8.3%）	42（87.5%）
妇产科案例	前置胎盘	1（2.1%）	6（12.5%）	41（85.4%）
	滴虫性阴道炎	0（0%）	0（0%）	48（100%）
	异位妊娠	2（4.2%）	12（25.0%）	34（70.8%）
	子宫肌瘤	0（0%）	1（2.1%）	47（97.9%）
儿科案例	小儿肺炎	19（39.6%）	9（18.7%）	20（41.7%）
	小儿惊厥	13（27.1%）	8（16.7%）	27（56.2%）
	小儿先天性心脏病	10（20.8%）	7（14.6%）	31（64.6%）
	小儿腹泻病	20（41.7%）	19（39.6%）	9（18.7%）

表 2 各案例 AI–SP 与人工评分比较

	AI-SP 评分 ＞人工评分	AI-SP 评分 = 人工评分	AI-SP 评分＜人工评分
2 型糖尿病	6	0	42
冠心病	2	0	46
甲状腺功能亢进症	0	0	48
哮喘	1	1	46
急性阑尾炎	0	0	48
急性胆囊炎	0	0	48
消化道穿孔	0	0	48
急性胰腺炎	0	0	48
前置胎盘	0	0	48
滴虫性阴道炎	0	0	48
异位妊娠	0	0	48
子宫肌瘤	0	0	48
小儿肺炎	5	0	43
小儿惊厥	2	0	46
小儿先天性心脏病	1	0	47
小儿腹泻病	10	0	38
合计	27	1	740

4 结果分析

本次研究中，各案例中 AI 与人工评分差值 ≥12 分的普遍占比 40% 以上，评分偏差相对较大。其中内科、儿科案例相较外科、妇产科案例的评分偏差小。AI-SP 评分普遍低于人工评分。

评分偏差较大的原因，可能是 AI 在学生问诊信息匹配、匹配程度评分过程中存在问题如下 [6-8]：

4.1 可能存在问诊信息匹配错误的情况

如：学生使用不同提问方式（如“有没有什么导致你不舒服的？”代替“你发烧前有没有接触过感冒的人？”），可能语义正确但未被识别为“诱因”问询；匹配错误发生率主要集中在学生使用模糊提问或提问结构不清晰的情况。

表 3 典型问题归纳

表达方式	AI 评分结果	人工评分判断	问题分析
“你发烧之后有什么反应？”	未归类	归类为“发展演变”	多意图句子，AI 无法精确划分
“你家人有类似症状吗？”	未识别	归类为“家族史”	缺少背景上下文建模
“你怀孕过吗？”	错归为“现病史”	正确归为“生育史”	归类边界模糊

4.2 语言理解与上下文推理能力有限

AI 难以理解模糊、跳跃、隐含语义的提问，尤其在中文非结构化对话中更易误判；例如：“你一直发烧吗？”、“发烧之前有没有什么不舒服？”这类问题，AI 需结合上下文判断是获取“发热持续时间”还是“诱因”。

4.3 对医学专业语言的理解深度不够

模型对专业术语与临床提问结构的理解常常依赖有限语料；同一个意思的不同表达（如“咳嗽几天了？”和“咳嗽是从什么时候开始的？”）有时不能被准确归类为同一问诊点。

4.4 匹配程度评分：对于某一项分值，评分尺度不一致

如：诊疗经过（5 分）。患者上周在社区医院检查了血常规，结果白细胞偏高，医生说是有炎症，开了 *** 药，吃了 3 天，没有效果。学生只问诊了“是否就诊”，AI 和考官的给分会不一致。AI 根据系统匹配度给分，考官则根据自己的临床经验加以判断后给分。

4.5 评分标准缺乏灵活性和临床容忍度

规则引擎或关键词匹配方法“死板”，不能容忍合理但非标准提问表达；例如学生提出了变通但临床合理的问题，但系统仍可能不给分。

4.6 语音识别误差传导

系统在嘈杂环境或口音影响下易出错，影响后续评分；如学生有连续性发问或者回答，AI 可能只提取到其中一个点进行评分；医学生表达不清晰或语言逻辑跳跃也会导致评分偏差。

上述各种原因均可造成 AI 评分相对较低，而人工评分可以弥补很多误差而较 AI 评分高。内科、儿科案例相较外科、妇产科案例的评分偏差小，可能一方面是偏内科的案例，问诊客观性评分条目设置更清晰、更明确，AI 匹配程度相对较高。另一方面，偏内科的案例学生问诊也会更详细、更全面。

5 讨论

人工评分的优势在于人工考官在理解模糊表达、合理容忍表达多样性方面具备明显优势，能够根据语言语境与学生整体逻辑进行主观判断。人工评分可以弥补很多误差，因而较 AI 评分高。此外，在沟通技巧、表达风格等非结构化维度的评估中，人工评分更具灵活性和解释力。

AI 评分的优势在于其标准化、可重复、反馈即时，尤其适用于初学者的规范训练与自我评估。然而，目前 AI 系统在以下方面仍面临诸多挑战，例如对多样表达方式的归类能力不足、无法识别模糊语言与非语言行为、缺乏主观判断机制与语境理解力、评分结果解释性弱、反馈难以个性化等。因此目前阶段，AI 评分与人工评分还存在较大的差异性。

AI-SP 较传统真人 SP 培训简单、训练周期短、使用和维护成本低、人为因素影响小，能够极大地减少医学院校培训和使用 SP 的负担。现代医学生也更乐于接受人机交流模式，因此，在医学生病史采集的教学与评估中，AI-SP 的应用也能更好地发挥其作用。因此可以与 AI-SP 研发公司合作，对其进行进一步优化，使界面更友好，患者外形、年龄、疾病表现、动作表情等动画设计更真实，使医患交流更贴近实际、体验感更好 [9-11]。

在评分表条目的内容和设置上进一步优化，在客观性条目上避免有歧义、多重含义的内容，使 AI 充分发挥其优势。

在病史采集评估中，建议发展人机融合评分模式，结合 AI 的覆盖性与人工评分的灵活性，构建双维度评价体系。AI-SP 在表述明确、答案唯一的结构化评分项目中打分较为精准，这类评分可以由 AI 打分直接得出；主观性评分交由人工评分完成。人机融合评分模式，不仅可以节约人工 SP 的使用，还可以减轻考官打分工作量，实现人力和成本的双重节约 [12-15]。

6 研究结论

AI-SP 成本低廉，使用简便。AI 评分系统在病史采集教学中展现出良好的标准化与效率优势，能够有效辅助早期训练与过程性评估。然而，当前技术在语言理解深度、多意图归类及人际沟通评分等方面仍存在局限，尚不能完全替代人工评分。未来应着力于提升 AI 评分系统的表达容忍度、语义理解能力及反馈可解释性，探索 AI 与人工评分融合的

智能化医学教育新模式。

参考文献

[1] 李小波,曾丹.SP结合虚拟问诊系统应用于诊断学教学的体会[J].继续医学教育,2022,9(36):17-20.

[2] 陈娜平,唐陆禛,黄贤生,等.基于ChatGPT-4的虚拟标准化病人应用研究[J].中华医学教育杂志，2025,1(45):44-49.

[3] 胡雪晴，韩琪，付磊等.AI虚拟医生在医疗行业的应用研究综述[J].电脑知识与技术，2024，20（29）：15-17.

[4] 邓玉华.ChatGPT的现况及人工智能在医学教育中的应用展望[J].中国继续医学教育，2024，16（5）：195-198.

[5] 孙康妮，柳桂良.ChatGPT在医学生问诊教学中的应用可行性研究[J].卫生职业教育,2024，42（2）：72-75.

[6] 陈喜，阮国竹，胡海岩.DxR Clinician结合SP在医学生临床思维教学中的应用[J].科教导刊，2024，17（6）：147-149.

[7] 田甜，张飞，张璇等.基于ChatGPT的人工智能技术在全科医疗服务中的应用研究[J].中华全科医学，2025，23（5）：721-725.

[8] 李翔，陈晓云，张慧群等.基于VSP的病史采集思维训练系统在问诊技能教学中的应用研究[J].中国医学教育技术，2023，37（2）：186-190，201.

[9] 贾栗，黄祖辉.数智化驱动下医学人文教育的创新改革探索[J].广东医科大学学报，2025，43（1）：34-39.

[10] 汪国栋，李月，秦建星等.虚拟标准化病人在精神医学教学中的应用与展望[J].精神疾病与精神卫生，2025，25（1）：68-72.

[11] 李翔，梁杰，陈晓云.虚拟仿真实验教学平台在病史采集教学中的应用[J].医学研究与教育，2025，42（3）：74-80.

[12] 马腾，张梅静，韩雅蕾.虚拟现实技术在循环系统疾病教学中的应用[J].中华医学教育杂志，2021，4（5）：403-406.

[13] 孙晓彤，张晓敏，周荣佼.医学教育中标准化病人队伍建设困境及对策[J].中国继续医学教育，2024，12（6）：86-88.

[14] 王珍珍，向巴卓玛，赵岩松.以ChatGPT为代表的大型语言模型在医学教学中的应用[J].医学教育管理,2024,10（6）：692-697.

[15] 吴玉奇.基于大模型的中医问诊平台研究综述.安徽科技报，2025，5（16）：012.