JHPLT
Journal of Health Professions Learning and Technology

COMMENTARY

Vol. 1, No. 1 (Jul-Sep 2026) | e15
Published 1 September 2026
Published by Palmetto Innovative Education LLC, Mount Pleasant, SC | e-ISSN 3144-5454

Advances in Large Language Models and the Case for GPT-4o in Virtual Patient Simulation for Health Professions Education

Georgia Mappin[1*]

[1] The Citadel

[*] Correspondence: Georgia Mappin, mappin@musc.edu

ABSTRACT

Large language models (LLMs) have advanced rapidly, and the release of GPT-4o introduced a natively multimodal model capable of processing and generating text, audio, and images in near real time. These capabilities make LLMs newly practical for virtual patient simulation, a long-standing method in health professions education. This article argues that GPT-4o is a strong foundation for simulated patients by connecting three lines of evidence: the demonstrated medical knowledge of contemporary LLMs, the established educational value of simulation and standardized patients, and emerging studies of LLM-powered simulated patients. Conversational realism, the degree to which a simulated encounter reproduces the unscripted, responsive dialogue of a real clinical interview, is central to the educational value of these experiences, and it is precisely where earlier scripted virtual patients fell short and where generative models excel. Early studies show that GPT-based simulated patients produce largely plausible responses and are well received by students, and that pairing simulation with automated feedback improves clinical decision making. Realizing this potential requires attention to fabrication, bias, and the preservation of human interaction.

Keywords large language models; GPT-4o; virtual patients; simulated patients; conversational realism; health professions education; clinical communication

Introduction

Simulation is a foundational method in health professions education, allowing learners to develop clinical reasoning and communication skills without risk to real patients, and a large meta-analysis has shown that technology-enhanced simulation is consistently associated with improved knowledge, skills, and behaviors [6]. Standardized patients, actors trained to portray clinical cases, have long been used to teach and assess history taking and communication, and their value rests on the realism of the encounter [5]. Computer-based virtual patients extend this method with a scalability that human actors cannot match, and they too improve learning outcomes [7]. Until recently, however, virtual patients relied on branching scripts that constrained the dialogue and limited realism. The emergence of large language models, and in particular GPT-4o, changes this picture by enabling open-ended, responsive conversation [1].

Advances in Large Language Models and GPT-4o

Large language models have progressed quickly from text prediction toward systems that display substantial medical knowledge. ChatGPT performed at or near the passing threshold of the United States Medical Licensing Examination without domain-specific training [3], and analyses of GPT-4 documented strong performance across diagnostic and explanatory tasks alongside explicit cautions about its limits [2]. GPT-4o, released in 2024, is described by its developers as an omni model that accepts and generates text, audio, and images and responds to speech with latencies approaching human conversational timing [1]. Its medical knowledge has also improved: on the Chinese National Medical Licensing Examination, GPT-4o achieved accuracies of 84.2% and 88.2% across two examination years, substantially exceeding GPT-4 (74.7% and 73.2%) and GPT-3.5 (approximately 50%) [4]. This combination of stronger reasoning and native speech interaction is what makes GPT-4o notable for simulation.

Conversational Realism as the Core of Simulation Value

The educational power of a simulated encounter depends heavily on conversational realism, meaning the extent to which the simulated patient responds naturally, unpredictably, and in character to whatever the learner asks. In describing standardized patients, Barrows emphasized that the encounter must faithfully reproduce a real clinical interview if it is to teach and assess clinical skills validly [5]. Scripted virtual patients, by contrast, break down when a learner asks a question the author did not anticipate, interrupting both the illusion and the learning. Because generative models produce fluent, contextually appropriate language on the fly, they can sustain the unscripted give and take of a genuine history, which is the aspect of realism most relevant to communication training [7]. With GPT-4o, the addition of real-time voice further narrows the gap between a typed exercise and a spoken clinical conversation [1].

Evidence for AI-Simulated Patients in Health Professions Education

Direct evidence for LLM-powered simulated patients is emerging and broadly encouraging, though still preliminary. In a mixed-methods study, a GPT-powered chatbot acting as a simulated patient for history taking produced responses that were rated plausible in 97.9% of question and answer pairs, and students reported a positive user experience (77 out of 100); notably, when learners asked questions outside the case script, the system fabricated information in 56.4% of instances, usually offering socially desirable rather than clinically accurate replies [8]. In a randomized controlled trial, medical students who practiced with an LLM-simulated patient and received structured, AI-generated feedback showed significantly better clinical decision making than peers who practiced without feedback, although the sample was small and the authors positioned the tool as a supplement to, not a replacement for, human interaction [9]. Together these studies indicate that current models are already usable for communication practice and that structured feedback amplifies their value.

Why GPT-4o Is Well Suited to Virtual Patient Simulation

Several features make GPT-4o particularly well suited to virtual patient simulation. Its native multimodality allows spoken history-taking practice, so learners can rehearse the verbal rhythm of a real interview rather than typing, which better approximates the target skill [1]. Its improved medical reasoning reduces the rate of clinically implausible responses relative to earlier models, addressing a key weakness identified in GPT-3.5-based simulated patients [4,8]. And its broad availability and comparatively low cost relative to specialized simulation infrastructure make it feasible to give every student unlimited, on-demand practice, extending the scalability advantage that made virtual patients attractive in the first place [7].

Limitations and Responsible Use

These strengths do not eliminate important risks. The tendency of LLMs to generate confident but incorrect content, observed as off-script fabrication in a simulated-patient setting [8] and cautioned against in medicine generally [2], means that unconstrained models can portray a patient inaccurately and must be guided by careful prompting, case grounding, and instructor oversight. Bias, data privacy, and the validity of automated assessment remain open concerns that require governance and faculty development as generative AI is adopted [10]. Finally, the value of simulation is partly relational: real encounters involve nonverbal cues and human rapport that a chatbot cannot fully reproduce, so AI-simulated patients are best used to complement, not replace, standardized patients and clinical placements [9].

Conclusion

Advances in large language models, culminating for now in GPT-4o, have made conversational realism achievable at scale in virtual patient simulation. The convergence of demonstrated medical knowledge, native voice interaction, and a maturing evidence base for LLM-simulated patients supports GPT-4o as a strong foundation for practice environments in health professions education. Used with appropriate safeguards and paired with structured feedback, GPT-4o-based simulated patients can give learners abundant, realistic opportunities to practice history taking and communication, while human interaction remains the standard these tools are designed to approach.

References

1. OpenAI. GPT-4o system card. arXiv:2410.21276. 2024. doi:10.48550/arXiv.2410.21276.

2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184.

3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198.

4. Luo D, Liu M, Yu R, Liu Y, Jiang W, Fan Q, Kuang N, Gao Q, Yin T, Zheng Z. Evaluating the performance of GPT-3.5, GPT-4, and GPT-4o in the Chinese National Medical Licensing Examination. Sci Rep. 2025;15:14119. doi:10.1038/s41598-025-98949-2.

5. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med. 1993;68(6):443-451. doi:10.1097/00001888-199306000-00002.

6. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hamstra SJ. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA. 2011;306(9):978-988. doi:10.1001/jama.2011.1234.

7. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, Tudor Car L, Carlstedt-Duke J, Car J, Zary N. Virtual patient simulations in health professions education: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res. 2019;21(7):e14676. doi:10.2196/14676.

8. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, Holderried M, Herrmann-Werner A, Festl-Wietek T, Mahling M. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ. 2024;10:e53961. doi:10.2196/53961.

9. Brugge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, Holling M, Lu MH, Darici D. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ. 2024;24:1391. doi:10.1186/s12909-024-06399-7.

10. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ. 2023;9:e48785. doi:10.2196/48785.

Ethics approval

n/a

Funding

na

Competing interests

n/a

Use of artificial intelligence

The authors used generative artificial intelligence in developing this manuscript, including for generating ideas, producing content, and writing portions of the text. The authors reviewed and edited all AI-generated material and take full responsibility for the accuracy and integrity of the published work.

Data availability

n/a

Acknowledgements

n/a

How to cite this article

Mappin, G. (2026). Advances in Large Language Models and the Case for GPT-4o in Virtual Patient Simulation for Health Professions Education. Journal of Health Professions Learning and Technology, 1(1), e15.

Open access

This is an open access article distributed under the terms of the CC BY 4.0 licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Permanent link	https://jhplt.org/pages/article.php?id=23
Publisher	Palmetto Innovative Education LLC, Mount Pleasant, SC
e-ISSN	3144-5454