MEDICAL TEACHER
2025, VOL. 47, NO. 1, 40–42
https://doi.org/10.1080/0142159X.2024.2376879

MEDICAL TEACHER
Taylor & Francis
Taylor & Francis Group

NEW WAVE

Creating virtual patients using large language models: scalable, global, and low cost

David A. Cook

Mayo Multidisciplinary Simulation Center, Mayo Clinic College of Medicine and Science, Rochester, MN, USA

ABSTRACT

Virtual patients (VPs) have long been used to teach and assess clinical reasoning. VPs can be programmed to simulate authentic patient-clinician interactions and to reflect a variety of contextual permutations. However, their use has historically been limited by the high cost and logistical challenges of large-scale implementation. We describe a novel globally-accessible approach to develop low-cost VPs at scale using artificial intelligence (AI) large language models (LLMs). We leveraged OpenAI Generative Pretrained Transformer (GPT) to create and implement two interactive VPs, and created permutations that differed in contextual features. We used systematic prompt engineering to refine a prompt instructing ChatGPT to emulate the patient for a given case scenario, and then provide feedback on clinician performance. We implemented the prompts using GPT-3.5-turbo and GPT-4.0, and created a simple text-only interface using the OpenAI API. GPT-4.0 was far superior. We also conducted limited testing using another LLM (Anthropic Claude), with promising results. We provide the final prompt, case scenarios, and Python code. LLM-VPs represent a 'disruptive innovation' – an innovation that is unmistakably *inferior* to existing products but substantially more *accessible* (due to low cost, global reach, or ease of implementation) and thereby able to reach a previously underserved market. LLM-VPs will lay the foundation for global democratization *via* low-cost-low-risk scalable development of educational and clinical simulations. These powerful tools could revolutionize the teaching, assessment, and research of management reasoning, shared decision-making, and AI evaluation (e.g. 'software as a medical device' evaluations).

ARTICLE HISTORY
Received 27 March 2024
Accepted 2 July 2024

KEYWORDS
Simulation Training; artificial intelligence; computer-assisted instruction; clinical decision-making; clinical reasoning

Educational challenge

Virtual patients (VPs) – 'a specific type of computer program that simulates real-life clinical scenarios; learners emulate the roles of health care providers … and make diagnostic and therapeutic decisions' [1] – have long been recognized as a useful approach to teach and assess clinical reasoning. In theory, VPs can be programmed to simulate authentic patient-clinician interactions and to reflect a variety of contextual permutations. In practice, their use has historically been limited by the high cost and logistical challenges of large-scale implementation. This article describes a novel globally-accessible approach to develop low-cost VPs at scale.

Clinical reasoning develops through exposure to example cases – lots of them, representing a wide range of contextual variation [2]. When supplementing real patient cases, paper cases and simple VPs can often suffice for diagnostic reasoning, since accurate diagnosis can be achieved using static information (written or multimedia information about the history, exam, and test findings). However, a static case is insufficient when the goal is management reasoning – 'the cognitive processes by which clinicians integrate clinical information, preferences, medical knowledge, and contextual (situational) factors to make decisions about the management of an individual patient' [3]. Why? Because management reasoning takes place largely in the 'space between' the clinician and the patient (i.e. shared decision-making [4]), and requires contextualization of care [5] (i.e. unique case features influence optimal management). For example, a patient with elevated blood glucose is *diagnosed* with diabetes, whether they like it or not. However, *managing* diabetes depends on patient preferences and numerous other factors including illness severity, comorbid conditions, social determinants of health, and healthcare system constraints; and moreover, management decisions must be *negotiated* with the patient.

Thus, although some features of management reasoning can be examined in isolation, it cannot be taught, assessed, or studied in its entirety outside the context of a patient-clinician interaction. In short, management reasoning requires interactive *conversations* – not static cases. This limits the application of instructional approaches known to be effective in promoting clinical reasoning skills, retention and transfer of knowledge and skills, and motivation to learn, such as case-based learning, testing for learning, spaced practice, and deliberate practice [1,2].

Sophisticated VPs can simulate such conversations, but are prohibitively expensive and require anticipation and manual programming of each node in the patient-clinician interaction; this prohibits scaling up, and precludes the case-to-case contextual variability needed to promote management reasoning.

Solution

Emerging artificial intelligence (AI) technologies offer potential solutions. Sophisticated generative AI-based

CONTACT David A. Cook cook.david33@mayo.edu Mayo Multidisciplinary Simulation Center, Mayo Clinic College of Medicine and Science, 200 First Street SW, Rochester, MN 55905, USA.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/0142159X.2024.2376879

'chatbots' now replace rule-based systems, using natural language processing and large language models (LLMs) to provide near-human responses. This approach has never been applied to VPs, but the potential to replace rule-based (manually-programmed) VPs is promising. We describe a novel approach to developing VPs using LLMs.

Implementation

We leveraged the OpenAI Generative Pretrained Transformer (GPT [openai.com]) to create and implement two interactive VPs, and then created permutations that differed in contextual features. We created two clinical scenarios – one for diagnosis (chronic cough), one for management (routine follow-up of type 2 diabetes).

We used iterative, systematic prompt engineering to refine a prompt instructing ChatGPT (a chat-based GPT interface) to emulate the role of the patient for a given case scenario, and then provide feedback on clinician performance after the case. Careful, creative wording of the prompt was required to maintain the patient role (GPT tended to give away the diagnosis and teach about the clinical problem), respond appropriately to inappropriate language (GPT was too forgiving of insensitive comments and medical jargon), report results of physical exam and clinical tests, and provide appropriate feedback at the end of the case but not sooner. We created several permutations of each case, differing in patient preferences and social determinants of health (e.g. desire for more vs fewer tests or medications, ability to pay) and further refined the prompt to authentically represent these case variations.

We implemented the prompts using GPT-3.5-turbo and GPT-4.0. We briefly tested these prompts using different clinical problems (kidney stone, management of hypertension) and using another LLM (Anthropic Claude 2.0 [anthropic.com]).

Using these cases in ChatGPT requires the user to paste the instructions and clinical case into the chat interface. This may work for simple self-directed learning activities, but does not work for teaching or assessment when the diagnosis or patient preferences should be hidden from learners. To circumvent this, we used Python and the OpenAI application programming interface (API) to create a simple text-only interface that runs the VP and saves a transcript of the conversation.

The final prompt, case scenarios, Python code, and sample transcript are available in Online Supplemental Materials.

Lessons learned

We found the patient-clinician conversations using GPT-4.0 to be authentic, engaging, and appropriately unpredictable. The VP made 'decisions' about management using preferences aligned with those outlined in the case scenario, and would (as instructed) invent details of personal life to add authenticity. Moreover, these decisions and details differed from one run to the next, further enhancing the authenticity. Feedback was typically rich and on-target. The only issue that precluded an authentic experience was the reporting of test results: the VP must tell the doctor what the test showed (e.g. 'In the methacholine challenge, the FEV1 dropped from 1.93 to 1.65 L.'), rather than vice versa.

By contrast, GPT-3.5-turbo was markedly inferior. The VP frequently broke character (despite explicit instructions otherwise), the narrative was less engaging, and the feedback was commonly off-target ('observing' behaviors not actually performed [usually information in the case scenario that was not requested], and suggesting 'improvements' for behaviors that were actually performed well). Anthropic Claude performed exceptionally well, rivaling GPT-4.0; however, testing was constrained by usage limits.

These VPs were not perfect. GPT-4.0 responses, while surprisingly good, often reflected language subtly atypical of a real patient (responses were longer, more polite, more cheerful/positive, and more detailed than normal). GPT-3.5-turbo was clearly inferior. Moreover, it was easy to 'break' the case in all LLMs: A user asking inappropriate questions will quickly steer the conversation off-course. However, the VPs were 'well behaved' if the learner was well behaved (took the interaction seriously).

LLM-VPs represent a classic 'disruptive innovation' – an innovation that is unmistakably *inferior* to existing products (existing VPs) but substantially more *accessible* (due to low cost, global reach, and/or ease of implementation) and thereby able to meet the needs of a new (previously underserved) market. We believe that LLM-VPs will lay the foundation for global democratization *via* low-cost-low-risk scalable development of educational and clinical simulations. These powerful tools could revolutionize the teaching, assessment, and research of management reasoning, shared decision-making, and AI evaluation.

Next steps

We are currently extending our model to represent other clinical problems in which patient-clinician interactions are essential, such as breaking bad news and motivational interviewing. We plan to create a case library comprised not only of different problems (e.g. diabetes, hypertension, fibromyalgia) but also permutations within each problem (different patient personalities, preferences, comorbidities, social determinants of health, etc). We anticipate that such permutations will help HPE learners appreciate the importance of customizing decisions to the individual patient (contextualized care) [5]. We also aim to explore their use as assessment tools.

LLM-VPs will immediately benefit learners and educators for teaching and (we hope) assessment. We also believe they will advance the study of management reasoning, since this field has historically been limited by the absence of research tools that replicate patient-clinician interactions at scale and with case-to-case variation. LLM-VPs could fill that gap. Finally, these tools could be used in clinical research, as a platform for studying shared decision-making and testing novel workflows including computer-based clinical solutions (e.g. 'software as a medical device' evaluations).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

Notes on contributor

David A. Cook is Professor of Medicine and Professor of Medical Education; Research Chair, Mayo Multidisciplinary Simulation Center, Mayo Clinic College of Medicine and Science, Rochester, MN, USA.

ORCID

David A. Cook http://orcid.org/0000-0003-2383-4633

References

1. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Med Educ. 2009;43(4):303–311. doi:10.1111/j.1365-2923.2008.03286.x
2. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005;39(1):98–106. doi:10.1111/j.1365-2929.2004.01972.x
3. Cook DA, Durning SJ, Sherbino J, et al. Management reasoning: implications for Health Professions Educators and a Research Agenda. Acad Med. 2019;94(9):1310–1316. doi:10.1097/ACM.0000000000002768
4. Cook DA, Hargraves IG, Stephenson CR, et al. Management reasoning and patient-clinician interactions: insights from shared decision-making and simulated outpatient encounters. Med Teach. 2023;45(9):1025–1037. doi:10.1080/0142159X.2023.2170776
5. Weiner SJ, Schwartz A. Contextual errors in medical decision making: overlooked and understudied. Acad Med. 2016;91(5):657–662. doi:10.1097/ACM.0000000000001017