Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialog with feedback

David Cook, Joshua Overgaard, V. Shane Pankratz, Guilherme Del Fiol, Chris A. Aakre

Submitted to: Journal of Medical Internet Research
on: November 06, 2024

Table of Contents

Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialog with feedback

David Cook[1]; Joshua Overgaard[1] MD; V. Shane Pankratz[2] PhD; Guilherme Del Fiol[3] MD, PhD; Chris A. Aakre[1] MD

[1]Division of General Internal Medicine Mayo Clinic College of Medicine and Science Rochester US
[2]University of New Mexico Health Sciences Center Albuquerque US
[3]Department of Biomedical Informatics University of Utah School of Medicine Salt Lake City US

Corresponding Author:
David Cook
Division of General Internal Medicine
Mayo Clinic College of Medicine and Science
200 First St SW
Rochester
US

Abstract

Background: Virtual patients (VPs) are computer screen-based simulations of patient-clinician encounters. VP use is limited by cost and low scalability.

Objective: Show proof-of-concept that VPs powered by large language models (LLMs) generate authentic dialogs, accurate representations of patient preferences, and personalized feedback on clinical performance; and explore LLMs for rating dialog and feedback quality.

Methods: We conducted an intrinsic evaluation study rating 60 VP-clinician conversations. We used carefully engineered prompts to direct OpenAI Generative Pre-trained Transformer (GPT) to emulate a patient and provide feedback. Using 2 outpatient medicine topics (chronic cough [diagnosis] and diabetes [management]), each with permutations representing different patient preferences, we created 60 conversations (dialogs plus feedback): 48 with a human clinician, and 12 "self-chat" dialogs with GPT role-playing both the VP and clinician. Primary outcomes were dialog authenticity and feedback quality, rated using novel instruments meticulously grounded in empirical and conceptual work. Each conversation was rated by 3 physicians and also by GPT. Secondary outcomes included patient preferences represented in the dialogs, cost, and user experience.

Results: The average cost per conversation was $0.51 for GPT-4.0-turbo and $0.02 for GPT-3.5-turbo. Conversation ratings (maximum 6) were mean (SD) overall authenticity 4.7 (0.7); overall user experience 4.9 (0.7); and average feedback 4.7 (0.6). For dialogs created using GPT-4.0-turbo, physician ratings of patient preferences aligned with intended preferences in 20-47 of 48 dialogs (42-98%). Subgroup comparisons revealed higher ratings for dialogs using GPT-4.0-turbo vs GPT-3.5-turbo, and for human-generated vs self-chat dialogs. Feedback ratings were similar for human-generated vs GPT-generated ratings, whereas authenticity ratings were significantly lower.

Conclusions: LLM-powered VPs can simulate patient-clinician dialogs, demonstrably represent patient preferences, and provide personalized performance feedback. This approach is scalable, globally-accessible, and inexpensive. LLM-generated ratings of feedback quality are similar to human ratings. Clinical Trial: None

(JMIR Preprints 06/11/2024:68486)
DOI: https://doi.org/10.2196/preprints.68486

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialog with feedback

David A. Cook, MD, MHPE[1,2] cook.david33@mayo.edu; ORCID: 0000-0003-2383-4633
Joshua Overgaard, MD[1] overgaard.joshua@mayo.edu
V. Shane Pankratz, PhD[3] vpankratz@salud.unm.edu; ORCID: 0000-0002-3742-040X
Guilherme Del Fiol, MD, PhD[4] guilherme.delfiol@utah.edu; ORCID: 0000-0001-9954-6799
Chris A. Aakre, MD[1] aakre.christopher@mayo.edu; ORCID: 0000-0001-9817-8533

[1] Mayo Clinic College of Medicine and Science and Division of General Internal Medicine, Mayo Clinic, Rochester, MN, USA.
[2] Mayo Clinic Multidisciplinary Simulation Center, Mayo Clinic College of Medicine and Science, Rochester, MN, USA.
[3] University of New Mexico Health Sciences Center, Albuquerque, NM, USA.
[4] Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, USA.

Running title: Virtual patients with LLM

Corresponding Author
David A. Cook, MD, MHPE
Division of General Internal Medicine
Mayo Clinic College of Medicine
Mayo 17-W, 200 First Street SW
Rochester, MN 55905

Phone: 507-284-2269
FAX: 507-284-5370
E-mail: cook.david33@mayo.edu * Twitter: @CookMedEd

Word count: Text – 3440; Abstract – 289
Tables: 6
Supplemental Digital Content: Appendix, eBox, 4 eTables, Example Case

Key words: Simulation Training; Natural Language Processing; Computer-Assisted Instruction; Clinical Decision-Making; clinical reasoning; Machine Learning; virtual patient; natural language generation

Abstract

Background: Virtual patients (VPs) are computer screen-based simulations of patient-clinician encounters. VP use is limited by cost and low scalability.

Objective: Show proof-of-concept that VPs powered by large language models (LLMs) generate authentic dialogs, accurate representations of patient preferences, and personalized feedback on clinical performance; and explore LLMs for rating dialog and feedback quality.

Methods: We conducted an intrinsic evaluation study rating 60 VP-clinician conversations. We used carefully engineered prompts to direct OpenAI Generative Pre-trained Transformer (GPT) to emulate a patient and provide feedback. Using 2 outpatient medicine topics (chronic cough [diagnosis] and diabetes [management]), each with permutations representing different patient preferences, we created 60 conversations (dialogs plus feedback): 48 with a human clinician, and 12 "self-chat" dialogs with GPT role-playing both the VP and clinician. Primary outcomes were dialog authenticity and feedback quality, rated using novel instruments meticulously grounded in empirical and conceptual work. Each conversation was rated by 3 physicians and also by GPT. Secondary outcomes included patient preferences represented in the dialogs, cost, and user experience.

Results: The average cost per conversation was $0.51 for GPT-4.0-turbo and $0.02 for GPT-3.5-turbo. Conversation ratings (maximum 6) were mean (SD) overall authenticity 4.7 (0.7); overall user experience 4.9 (0.7); and average feedback 4.7 (0.6). For dialogs created using GPT-4.0-turbo, physician ratings of patient preferences aligned with intended preferences in 20-47 of 48 dialogs (42-98%). Subgroup comparisons revealed higher ratings for dialogs using GPT-4.0-turbo vs GPT-3.5-turbo, and for human-generated vs self-chat dialogs. Feedback ratings were similar for human-generated vs GPT-generated ratings, whereas authenticity ratings were significantly lower.

Conclusions: LLM-powered VPs can simulate patient-clinician dialogs, demonstrably represent patient preferences, and provide personalized performance feedback. This approach is scalable,

globally-accessible, and inexpensive. LLM-generated ratings of feedback quality are similar to human ratings.

Introduction

Translating advances in biomedical knowledge and knowledge synthesis into data-driven, patient-centered, contextualized management decisions remains a wicked challenge. As we seek to prevent errors in clinical practice(1, 2) and promote high-value care,(3, 4) we need to better understand clinical reasoning and how to support its development and application.(2, 5) Because clinical reasoning is case-specific,(6) and since educationally-opportune encounters with real patients are finite, education and research in this field require a scalable approach to emulating authentic patient-clinician interactions. Virtual patients (VPs) powered by large language models (LLMs) offer a potential solution.

VPs – computer screen-based simulations of patient-clinician encounters(7) – have demonstrated efficacy in teaching, assessing, and studying clinical reasoning,(8) and could also support validation of decision-support tools prior to clinical implementation.(9, 10) VPs may be particularly important for *management reasoning*, which is a subset of clinical reasoning. In contrast with diagnostic reasoning, management reasoning is arguably more difficult, more complex to study, and more important.(11, 12) Yet it has received scant investigation due to challenges in replicating management tasks (most notably patient-clinician conversations), which necessarily involve shared decision-making(13-16) and contextualization of care (consideration of social determinants of health, patient preferences, comorbid conditions, etc).(17-20)

VP use to-date has been limited by the high cost and logistical challenges of large-scale implementation. One survey found that 85% of bespoke VPs cost >$10,000 per case and required >16 months to produce.(21) Commercial VP libraries exist, but subscriptions are expensive (approximately $100/student/yr). Hence, VP implementations typically comprise few cases and lack

case-to-case variability in salient features (e.g., diagnosis, illness severity, preferences, ethnic diversity).(8, 21, 22)

Providing performance feedback to clinicians is also essential in clinical skill development,(23) yet is commonly of low quality or simply absent.(24-27) Specific, actionable feedback(28-30) on VP-clinician interactions could promote clinical reasoning and communication skills.

LLMs represent a disruptive technology(31) offering an unprecedented opportunity to transform VP production and use, enabling scalable, accessible (inexpensive and low-expertise), interoperable, and reusable(32) simulations of patient-clinician encounters. Our aim was to show proof-of-concept that VPs powered by OpenAI Generative Pre-trained Transformer (GPT) generate authentic preference-sensitive dialogs and high-quality feedback. We hypothesized that human ratings of *observed* patient preferences would agree with corresponding *planned* preferences (i.e., that GPT would perceptibly represent the intended preference). We compared GPT-4.0-turbo against the earlier, cheaper GPT-3.5-turbo, hypothesizing that GPT-4.0-turbo would be superior. We also piloted GPT to role-play the clinician, hypothesizing that conversations involving human clinicians would be superior.

As a sub-study, we aimed to pilot LLMs for rating the quality of VP-clinician dialogs and feedback. Artificial intelligence (AI) has long been used to rate narrative text,(33-37) but this typically requires supervised machine learning – using human-graded texts to train the AI system. We explored the use of LLMs without any training exemplars ("zero-shot" learning).

Methods

We conducted an intrinsic evaluation study (a study that evaluates the quality of computer-generated

outputs on specific predefined [not real-world] tasks), rating the quality of 60 conversations (dialog and LLM-generated performance feedback) between an LLM-powered VP and a clinician. We created 3 novel instruments to rate dialog authenticity and feedback quality. Three physicians and GPT rated all conversations.

LLM-powered VP interface

We used Python to create a text VP interface (previously described(38)) that accesses GPT through the OpenAI application programming interface (API). We iteratively, rigorously engineered detailed "prompts" guiding GPT to emulate a diagnosis-focused or management-focused VP and provide feedback. To instantiate a specific VP the interface accesses a 1-page case description (see Example in supplemental materials).

Conversation creation

We used the interface to create 48 simulated conversations between the VP and a human clinician. We selected as topics 2 common problems in ambulatory medicine (chronic cough [diagnosis] and diabetes [management]) and for each topic created a written description of a prototypical scenario (for this pilot study we did not base scenarios on specific real patients). We generated 4 permutations per topic by varying the patient preferences or GPT model (details in eTable 1):

- Case 1: Patient has good insurance, wants to avoid tests or new medications (GPT-4.0-turbo);
- Case 2: Patient has financial concerns (limited income, poor insurance) (GPT-4.0-turbo);
- Case 3: Patient is anxious, pushes for more tests and more aggressive treatments (GPT-4.0-turbo);
- Case 4: Same as Case 1 (GPT-3.5-turbo).

The dialogs were further permuted for 3 clinician personas (role-played by a board-certified physician): an average third-year medical student, a poor third-year medical student, and an average second-year internal medicine resident. We ran each permutation twice (2 topics, 4 case variations, 3 clinician personas, 2 replications = 48 conversations). Permutation identifiers were not disclosed during the dialog. After the clinician ended a dialog, the VP interface offered detailed performance feedback.

For each topic-persona combination we created an interview guide that was adjusted according to VP response. One investigator role-played all conversations for cough, another investigator role-played those for diabetes.

Additionally, we employed GPT-4.0-turbo to play the role of an "excellent physician," and "self-chat" as both the VP and clinician (using independent GPT threads) for Cases 1-3, with 2 replications each (total 12 GPT-GPT self-chats).

Each conversation was saved verbatim, along with time spent, word count, and GPT "tokens" used. We calculated costs using GPT pricing.

Instrument creation

We created 3 novel instruments for rating the quality of VP dialogs and feedback (Table 1), and 1 item to flag potential bias. We conducted a validation study collecting validity evidence from 4 of 5 potential sources:(39, 40) content (grounding of the instruments in theory and prior empirical work), internal structure (rating reproducibility), relations with other variables (sensitivity of ratings to case differences, including expectation of higher ratings for more advanced LLM models and human

clinician personas), and response process (clarification on why raters responded as they did). The Appendix further describes instrument development and validation planning.

Table 1. Constructs, items, and reproducibility for rating scales*

Construct	Item	Operational clarifications	ICC: Human N=3	ICC: GPT† N=3
Dialog authenticity: Human-like	The virtual patient's responses were human-like.	Sensible, natural, conversational; uses appropriate word choice, phrasing, and tone	0.34	0.29
Coherent	The virtual patient's responses were coherent.	Contextually appropriate and internally consistent (logical) over the course of the dialog	0.40	0.45
Personal	The virtual patient's responses were personal.	Reflects preferences, opinions, values, and priorities; is not overly agreeable or pleasing	0.22	0.35
Relevant	The virtual patient's responses were relevant and meaningful.	Meaningful, useful, helpful as a clinically-relevant simulation; requires/supports clinical reasoning; stimulates appropriate emotions and empathy	0.30	0.20
Overall	The dialog as a whole mirrored a real-life patient-clinician conversation.		0.34	0.49
User	This was an authentic	Similar to a real-world situation	0.37	0.29

experience: Realness	representation of a real-world experience.			
Cognitive authenticity	I had to continuously revise my mental image of the problem using new information.	Requires or stimulates the same mental activities, same decisions as in real situation; real professional demand	0.24	-[†]
Variability	The interaction seemed unscripted and appropriately complex.	Natural variation in responses; spontaneous, unstructured, unplanned, flexible; complex, multidimensional (not superficial); not robot-like or prefabricated	0.19	-[†]
Involvement[‡]	I was fully engaged in this conversation.	Immersed, focused (not distracted), captivated; stimulated empathy and authentic emotions	-[‡]	-[†]
Overall	I felt as if I were the doctor.		0.17	-[†]
Feedback: Evidence-based	The feedback correctly identifies important weaknesses and strengths in the clinician's performance.	Specific observations of behavior; accurately interpreted; well prioritized	0.15	0.09
Actionable	The feedback contains suggestions that are specific and actionable.	Specific, actionable suggestions for behavior change	0.17	0.26
Connected	The feedback correctly connects each suggestion with specific strengths and weaknesses.	Explicit and logical connection between the observed behaviors and suggested changes	0.22	0.25
Balanced	The feedback balances corrective and reinforcing statements appropriate to	Includes both praise and critique; balance of positive and negative statements matches actual	0.08	0.16

	the clinician's performance.	performance.		
Bias (overall)	Did you detect any indication of bias or stereotyping in the dialog or feedback?	Includes stereotyping, disparagement, dehumanization, erasure, and inequitable performance.	1[§]	-[†]

* Items were presented in the sequence shown above; for final ratings, the operational clarifications were also included in a smaller font. A "conversation" refers to the VP-clinician dialog plus feedback. During conversation creation, each dialog was rated before feedback was provided. Response options for all rating scale items ranged from 1 = strongly disagree to 6 = strongly agree. For Authenticity and Experience, a rating of 6 was operationally defined as "This is exactly what I would expect in a real conversation; this could have come from a human patient." For Feedback, a rating of 6 was operationally defined as "This is surprisingly good, better than I would expect from a trained human clinician-supervisor." See eBox for additional details. Response options for Bias were: Yes, No. ICC is an intraclass correlation coefficient representing the overall reproducibility coefficient for a single rating. Human ratings were provided by 3 board-certified internal medicine physicians.

[†] Agreement across 3 replications of ratings from GPT-4.0-turbo (user experience and bias not rated).

[‡] Not coded in this study, inasmuch as we investigators did not feel authentically "engaged" in the task when creating multiple conversations. This item could be used in future studies with real learners.

[§] There was 100% agreement across all raters on the Bias item.

Dialog ratings

Two instruments focused on the dialogs: dialog authenticity and user (clinician) experience. To rate dialog authenticity, we drew on literature on dialog systems and natural language generation(41-51) from which we distilled five repeatedly-emphasized constructs: responses are Human-like (sensible, natural, avoiding bias), Coherent (contextually appropriate, internally consistent), engaging or Personal (reflecting preferences, empathy, personality), helpful or Relevant (specific, useful, meaningful), and Correct (for knowledge-delivery systems). We dropped Correct since our purpose is dialog, not knowledge-delivery. We considered but omitted a domain for Fluency, because recent literature suggests that fluency can be presumed for contemporary AI models.(42, 44, 46) We created 1 item for each construct, plus an overall item, resulting in a 5-item instrument.

To rate user experience we merged 2 conceptual frameworks for measuring authenticity in VPs – one emphasizing decision-making and cognitive strategies,(52) the other highlighting realism, empathy, and variability.(22, 53) We added a third empirically-derived framework for evaluating "presence" in virtual reality (realness, involvement, and spatial ["physical"] presence).(54, 55) We synthesized these into four constructs: Realness (similarity to real-world situation); Cognitive Authenticity (real mental activities and decisions); Variability (case-to-case variation, spontaneous responses); and Involvement (user engaged, immersed). We created 1 item for each construct, plus an overall item, resulting in a 5-item instrument. We did not rate Involvement in this study, because we never felt "immersed" when creating and rating multiple conversations; we plan to rate this in future studies.

Feedback

To rate feedback, we integrated findings from focus group studies,(24, 28) published instruments(30, 56, 57) and other empirical and conceptual studies,(29, 58-61) and identified 4 recurrent constructs: Evidence-based (behavior-focused) observations; specific, Actionable suggestions; observations explicitly Connected with suggestions; and Balanced praise and critique. We created 1 item for each construct, resulting in a 4-item instrument. We did not rate feedback "overall," and instead calculated the average rating.

Shared features

Three experts in virtual patients or natural language generation reviewed the 3 instruments, and approved them with minor clarifications. Response options ranged 1=strongly disagree, 6=strongly agree. After case creation and before the final rating phase, we created brief operational criteria for each response option (see eBox).

Bias

Bias – "skew that produces a type of harm towards different social groups"(62) – is a well-know risk in AI generally and natural language generation specifically.(62-65) Bias can result in harms of stereotyping, disparagement, dehumanization, erasure, and inequitable performance.(62) We asked raters to identify any bias in the conversation.

Conversation appraisal procedure

Each conversation creator rated the dialog authenticity and user experience immediately following the dialog. They then received and rated the feedback.

Later, all conversations were rated again by all 3 investigators using all 3 instruments ("final ratings"). Raters also indicated patient preferences represented in the dialog regarding less vs more testing, the importance of cost, and prioritization of lifestyle or control of illness. Following the dialog ratings, and again after the feedback ratings, we asked, "What specific features of this [dialog, feedback] detracted from its authenticity?" and "What specific features enhanced its authenticity?" We inductively generated a list of features during conversation creation, and then marked these during final rating.

Conversations were randomized for appraisal (a unique sequence for each rater). Blinded human raters entered final ratings using an online data entry form (DistillerSR).

We also used GPT-4.0-turbo to rate each conversation 3 times. We created a Python interface to rate dialog authenticity (but not user experience), then feedback.

Data analysis

Reproducibility of final ratings

To appraise rating reproducibility we estimated variance components and calculated a single-rating intraclass correlation coefficient (ICC), interpreted using criteria from Landis and Koch(66) (0-0.2 slight, 0.21-0.4 fair, 0.41-0.6 moderate, 0.61-0.8 substantial).

Comparison across design features

We selected 5 outcomes (overall authenticity, human-like, overall experience, realness, average

feedback) as most aligned with study aims, and compared these across GPT models, topics, clinician personas, and human vs LLM raters. Using mixed models ANOVA, we conducted paired analyses that accounted for features of the factorial design and (for final ratings) repeated measures from multiple raters. We used SAS 9.4 for all analyses, and an alpha of 0.05.

Results

Conversation creation resources

We created 48 VP-clinician conversations (dialog plus feedback) with human physicians playing the clinician role, and 12 conversations with GPT as clinician. Each human-created conversation lasted mean 622 seconds, of which GPT's responses took 90 seconds and cost $0.50 USD (see Table 2 for estimates of measurement variability [i.e., standard deviation]). GPT-3.5-turbo was significantly faster than GPT-4.0-turbo (62 vs 100 seconds; difference 38 [95% confidence interval, 29, 47]) and much cheaper ($0.02 vs $0.51 per conversation), although quality was substantially lower (see below). Compared with diabetes, cough conversations required substantially more GPT time (122 vs 59 seconds) and tokens (72,745 vs 27,241) even though the dialog itself was only slightly longer (1165 vs 908 words). This was due to more back-and-forth turns in the dialog (mean 37 vs 14 turns), because each time GPT processes a clinician statement (even a short query like "Do you have heartburn?") the entire dialog is resubmitted to GPT as context.

The average time for the 12 GPT-GPT (self-chat) conversations was 113 (20) seconds: 62 seconds for the clinician, and 51 seconds for the VP. The average cost was $0.29, because these dialogs had fewer turns (mean 21).

Table 2. Conversation creation: Resource metrics and immediate ratings of conversation quality*

Metric	Human clinician					Self-chat
	All N=48	GPT-4.0 N=36	GPT-3.5 N=12	DM N=24	Cough N=24	All N=12
Resources & time						
Total time, sec†	622 (173); 611	653 (168); 669	551 (171); 508	617 (158); 611	627 (189); 619	113 (20); 107
Physician time, sec†	534 (166); 553	553 (162); 556	488 (173); 477	562 (151); 553	510 (178); 511	62 (14); 57
Virtual patient (GPT) time, sec	90 (38); 76	100 (36); 99	62 (28); 63	59 (16); 65	122 (24); 129	51 (8); 51
Words, dialog‡	1037 (302); 1003	1092 (304); 1059	871 (238); 810	908 (232); 942	1165 (313); 1165	1377 (351); 1291
Words, feedback‡	387 (118); 425	449 (50); 450	202 (38); 198	371 (94); 413	403 (138); 458	424 (43); 407
Tokens, total‡	49993 (25609); 47621	50788 (25788); 46826	47607 (26036); 47894	27241 (6139); 26205	72745 (14904); 66960	28628 (7997); 27209
Dialog turns‡	26 (13); 24	26 (13); 24	26 (13); 24	14 (3); 15	37 (7); 34	21 (6); 20
Cost per conversation, USD§	0.50	0.51	0.02	0.27	0.73	0.29
Dialog authenticity:‖ Overall	4.6 (0.6); 5	4.8 (0.6); 5	3.9 (0.3); 4	4.5 (0.6); 4.5	4.8 (0.7); 5	--
Human-like	4.8 (0.7);	5.1 (0.5);	3.9 (0.5);	4.7 (0.6);	4.9 (0.8);	--

	5	5	4	5	5	
Coherent	5.4 (0.6); 5	5.5 (0.6); 5.5	5.3 (0.7); 5	4.9 (0.3); 5	6.0 (0.2); 6	--
Personal	5.0 (0.7); 5	5.4 (0.5); 5	4.1 (0.3); 4	4.8 (0.4); 5	5.3 (0.8); 6	--
Relevant	5.3 (0.7); 5	5.4 (0.6); 5	4.7 (0.7); 5	4.9 (0.4); 5	5.6 (0.6); 6	--
User experience:‖ Overall	4.9 (0.6); 5	5.0 (0.5); 5	4.4 (0.5); 4	4.7 (0.5); 5	5.0 (0.6); 5	--
Realness	4.6 (0.7); 5	4.8 (0.7); 5	4.0 (0.4); 4	4.3 (0.8); 5	4.8 (0.6); 5	--
Cognitive authenticity	4.5 (0.8); 4	4.6 (0.8); 5	4.2 (0.7); 4	3.9 (0.4); 4	5.1 (0.5); 5	--
Variability	5.0 (0.7); 5	5.2 (0.6); 5	4.5 (0.7); 5	4.8 (0.4); 5	5.3 (0.8); 5	--
Feedback:‖ Average	4.6 (0.9); 5	4.9 (0.6); 5	3.7 (1.0); 4	4.4 (0.9); 5	4.9 (0.8); 4.6	
Evidence-based	4.3 (1.1); 4.5	4.6 (0.9); 5	3.5 (1.0); 3.5	4.4 (1.0); 5	4.3 (1.1); 4	--
Actionable	4.9 (0.8); 5	5.2 (0.5); 5	4.0 (1.0); 4	4.6 (0.8); 5	5.3 (0.7); 5	--
Connected	4.8 (0.9); 5	5.1 (0.6); 5	3.8 (0.9); 4	4.5 (0.8); 5	5.1 (0.9); 5	--
Balanced	4.5 (1.1); 5	4.8 (0.9); 5	3.6 (1.3); 4	4.2 (1.2); 5	4.8 (0.9); 5	--

* Results are mean (SD); median. The clinician was a human physician for the "human clinician" conversations, and GPT-4.0-turbo for the "self-chat" conversations. The virtual patient was GPT for all conversations.

† N=37 for total time and human physician time, after excluding 11 conversations in which the recorded time was inexact due to interruptions.

‡ Dialog was generated as an interaction between the virtual patient (GPT) and clinician (human or

GPT). Feedback was generated by GPT. A "conversation" refers to the dialog plus feedback. Tokens includes entire conversation (both dialog and feedback; and for self-chat, both patient and physician).

[§] Pricing (per OpenAI, 30 May 2024): $1.00/100,000 tokens for GPT-4.0-turbo; $0.05/100,000 tokens for GPT-3.5-turbo.

[‖] Ratings provided by the physician involved in creating the conversation, immediately following the dialog and feedback. Response options for all items ranged from 1 = strongly disagree to 6 = strongly agree.

Representation of patient preferences

Each case was written to represent patient preferences in testing/treatment, cost of care, and prioritization of illness control vs lifestyle. During final rating, we independently (blinded) indicated whether the VP represented such preferences in the dialog. Reproducibilities (ICCs) for these ratings were: testing/treatment, 0.59; cost of care, 0.75; and prioritization of control, 0.39.

VPs demonstrably represented planned preferences with high frequency (Table 3). For dialogs created using GPT-4.0-turbo, 5 of 6 non-neutral planned preferences were recognized as such in ≥54% of dialogs, and all 3 neutral planned preferences were rated as "no opinion" ≥90% of the time. We observed comparable results for GPT-3.5-turbo.

Table 3. Patient preferences reflected in dialogs: Planned vs perceived by raters

Preference category	Human rating	Case 1, no. (%) N=48*	Case 2, no. (%) N=48*	Case 3, no. (%) N=48*	Case 1 GPT-3.5*, no. (%) N=36	DM, no. (%) N=90	Cough, no. (%) N=90
Testing/ treatment	Less	**20 (42)**	**35 (73)**	0	**17 (47)**	41 (46)	31 (34)
	No opinion	27 (56)	13 (27)	11 (23)	17 (47)	25 (28)	43 (48)
	More	1 (2)	0	**37 (77)**	2 (6)	24 (27)	16 (18)
Cost	Lower	3 (6)	**47 (98)**	1 (2)	2 (6)	25 (28)	28 (31)
	No opinion	**43 (90)**	1 (2)	21 (44)	**29 (81)**	39 (43)	55 (61)
	Not an issue	2 (4)	0	**26 (54)**	5 (14)	26 (29)	7 (8)
Impact on life	Prioritize	3 (6)	3 (6)	0	1 (3)	6 (7)	1 (1)

	lifestyle						
	No opinion	**45 (94)**	**43 (90)**	19 (40)	**34 (94)**	65 (72)	76 (84)
	Prioritize illness control	0	2 (4)	**29 (60)**	1 (3)	19 (21)	13 (14)

* Bold text indicates dialogs in which the planned preference aligns with human ratings. Case 1 was written to reflect desire for less testing/treatment. Case 2 was written to reflect strong desire for lower cost, and hence less testing/treatment. Case 3 was written to reflect desire for more testing/treatment, cost not an issue, and prioritization of illness control over lifestyle. See eTable 1 for details.

Conversation quality: authenticity, experience, and feedback

Conversation quality was appraised by 1 rater at the time of creation, and later by all 3 investigators (final ratings).

Conversation creation

During creation, dialog mean ratings ranged 4.8 to 5.4 (maximum 6) for authenticity and 4.5 to 5.0 for user experience (Table 2). Feedback quality ranged 4.3 to 4.9. Ratings were significantly higher for GPT-4.0-turbo vs GPT-3.5-turbo (difference: dialog overall 0.92 [0.64, 1.19]; experience overall 0.58 [0.21, 0.96]; feedback average 1.33 [0.80, 1.87]).

Final ratings

The reproducibilities of authenticity and experience final ratings were typically "fair," with ICCs ranging 0.17 to 0.40 (Table 1). By contrast, reproducibilities for feedback ratings were "slight," with all but 1 domain ≤0.17. We examined the variance components (eTable 2), and found very small between-rater variances (≤5% of total variance for all except feedback evidence-based =18%). By contrast, we found large (≥60% of total) between-replication variances, which reflect a combination of true differences in GPT performances and within-rater variability.

Final ratings ranged 4.6 to 5.0 for authenticity, 4.6 to 4.9 for experience, and 4.5 to 4.9 for feedback (see Table 4 and eTable 4 for details including estimates of measurement variability).

Table 4. Final ratings of conversation quality*

	Rater					Case			
Metric	**All human raters N=180**	**GPT rater N=180**	**Human rater 1 N=60**	**Human rater 2 N=60**	**Human rater 3 N=60**	**Case 1 N=48**	**Case 2 N=48**	**Case 3 N=48**	**Case 1, GPT-3.5† N=36**
Dialog authenticity: Overall	4.7 (0.7); 5	5.2 (0.6); 5	4.8 (0.8); 5	4.8 (0.6); 5	4.6 (0.7); 5	4.7 (0.7); 5	4.9 (0.8); 5	4.8 (0.6); 5	4.4 (0.8); 5
Human-like	4.6 (0.8); 5	5.6 (0.5); 6	4.8 (0.9); 5	4.6 (0.5); 5	4.5 (0.8); 5	4.5 (0.7); 5	5.0 (0.7); 5	4.8 (0.6); 5	4.1 (0.9); 4
Coherent	5.0 (0.6); 5	5.6 (0.5); 6	5.0 (0.8); 5	4.9 (0.5); 5	5.0 (0.6); 5	5.0 (0.5); 5	5.1 (0.5); 5	5.2 (0.4); 5	4.4 (0.9); 5
Personal	5.0 (0.6); 5	5.1 (0.6); 5	5.2 (0.9); 5	5.0 (0.2); 5	4.9 (0.7); 5	5.0 (0.5); 5	5.2 (0.6); 5	5.1 (0.7); 5	4.6 (0.6); 5
Relevant	4.9 (0.6); 5	5.8 (0.4); 6	4.8 (0.9); 5	4.9 (0.4); 5	4.9 (0.5); 5	4.9 (0.5); 5	5.0 (0.7); 5	5.0 (0.5); 5	4.6 (0.8); 5
User experience: Overall	4.9 (0.7); 5	--§	4.9 (1.0); 5	4.9 (0.3); 5	4.8 (0.6); 5	4.7 (0.7); 5	5.1 (0.7); 5	4.9 (0.7); 5	4.8 (0.6); 5
Realness	4.6 (0.8); 5	--§	4.7 (1.1);	4.6 (0.6); 5	4.5 (0.8); 5	4.6 (0.7);	4.9 (0.8);	4.8 (0.8);	4.1 (0.9); 4

	Rater					Case			
Metric	**All human raters N=180**	**GPT rater N=180**	**Human rater 1 N=60**	**Human rater 2 N=60**	**Human rater 3 N=60**	**Case 1 N=48**	**Case 2 N=48**	**Case 3 N=48**	**Case 1, GPT-3.5† N=36**
			5			5	5	5	
Cognitive authenticity	4.8 (0.7); 5	--§	5.0 (0.9); 5	4.9 (0.3); 5	4.6 (0.7); 5	4.8 (0.7); 5	5.0 (0.7); 5	4.9 (0.8); 5	4.7 (0.5); 5
Variability	4.8 (0.9); 5	--§	4.6 (1.2); 5	4.9 (0.3); 5	4.7 (0.8); 5	4.6 (0.9); 5	5.0 (0.8); 5	4.8 (0.9); 5	4.5 (0.9); 5
Feedback: Average	4.7 (0.6); 4.9	4.6 (0.2); 4.8	4.8 (0.8); 4.9	4.8 (0.4); 5	4.5 (0.6); 4.8	4.8 (0.5); 5	4.9 (0.6); 5	4.8 (0.5); 5	4.1 (0.7); 4.1
Evidence-based	4.5 (0.9); 5	4.9 (0.3); 5	4.7 (1.0); 5	4.8 (0.4); 5	4.1 (0.9); 4	4.7 (0.7); 5	4.7 (0.8); 5	4.6 (0.8); 5	3.9 (1.1); 4
Actionable	4.9 (0.6); 5	4.5 (0.5); 5	5.1 (0.8); 5	4.9 (0.3); 5	4.8 (0.6); 5	5.0 (0.5); 5	5.1 (0.5); 5	5.1 (0.5); 5	4.4 (0.7); 5
Connected	4.9 (0.7); 5	4.6 (0.5); 5	4.9 (0.9); 5	4.9 (0.4); 5	4.8 (0.6); 5	5.1 (0.5); 5	5.0 (0.7); 5	5.1 (0.4); 5	4.2 (0.8); 4
Balanced	4.5 (0.9); 5	4.5 (0.6); 5	4.5 (1.1); 5	4.6 (0.7); 5	4.4 (0.9); 5	4.6 (0.9); 5	4.7 (0.8); 5	4.6 (0.9); 5	4.0 (0.9); 4

* Results are unweighted mean (SD); median across all conversations (a "conversation" refers to the VP-clinician dialog plus feedback). Response options for all items ranged from 1 = strongly disagree to 6 = strongly agree. eTable 4 contains additional details with ratings by topic and clinician persona.

† "GPT-3.5" conversations used GPT-3.5-turbo as the virtual patient (N=36 because these did not include 12 self-chat conversations). All other conversations used GPT-4.0-turbo.

‡ MS3A was role-played (by a human board-certified internal medicine physician) to be an "average third-year medical student;" MS3B was a "poor third-year medical student;" PGY2 was an "average second-year internal medicine resident." For GPT clinician, GPT-4.0-turbo played the role of both the clinician (an "excellent physician") and the virtual patient using 2 independent GPT threads.

§ GPT did not rate user experience.

We report final ratings subgroup comparisons in Table 5. Differences between topics were small. All ratings were higher for GPT-4.0-turbo vs GPT-3.5-turbo (differences ranging 0.17 to 0.71) although differences did not always reach statistical significance. Conversations involving human clinicians had higher experience ratings than those with GPT as clinician (differences ≥0.57) but similar authenticity and feedback (differences ranging -0.05 to 0.31). Among human clinicians, the resident persona had higher ratings than the poor medical student; differences (≥0.48) were statistically significant for authenticity and experience.

No instances of potential bias were identified during creation or final rating.

Table 5. Subgroup comparisons*

Outcome	Topic: DM vs cough Difference (CI) N=180	GPT model: 4.0 vs 3.5 (case 1) Difference (CI) N=72	Clinician: Human vs GPT Difference (CI) N=180	Clinician: Resident vs MS3B Difference (CI)† N=144	Rater: Human vs GPT Difference (CI) N=360
Dialog authenticity: Overall	0.14 (-0.25, 0.54)	0.42 (-0.19, 1.02)	0.31 (-0.25, 0.88)	0.69 (0.26, 1.12)	-0.52 (-0.85, -0.19)
Dialog authenticity: Human-like§	0.09 (-0.32, 0.50)	0.50 (-0.16, 1.16)	0.12 (-0.46, 0.70)	0.71 (0.22, 1.20)	-0.98 (-1.24, -0.71)
User experience:	0.03	0.17	0.57	0.48	--†

Overall	(-0.37, 0.44)	(-0.39, 0.72)	(0.04, 1.11)	(0.07, 0.88)	
User experience: Realness[§]	0.18 (-0.28, 0.63)	0.58 (-0.08, 1.25)	0.69 (0.06, 1.33)	0.75 (0.24, 1.26)	--[†]
Feedback: Average	0.03 (-0.33, 0.38)	0.71 (0.13, 1.28)	-0.05 (-0.51, 0.41)	0.17 (-0.24, 0.59)	0.10 (-0.37, 0.58)

* Results reflect differences accounting for repeated measures on conversations and Tukey-adjusted 95% confidence intervals (CI). A "conversation" refers to the VP-clinician dialog plus feedback. Conversations included in each analysis were matched according to design features; non-matching conversations were excluded. Response options for all items ranged from 1 = strongly disagree to 6 = strongly agree.

[†] This contrast was selected post-hoc, after the omnibus test across all human clinician personas revealed statistically significant differences (P≤.03) for all outcomes except feedback. None of the other pairwise contrasts among human-played personas reached statistical significance.

[‡] GPT did not rate user experience.

[§] These outcomes were selected a priori for reporting because they closely aligned with the overarching study aim.

Features that detracted from or enhanced authenticity

We identified features that detracted from or enhanced conversation authenticity (Table 6). Across 180 dialogs, the most frequent detractors were that GPT was verbose or used atypical vocabulary (N=93 [52%]), was overly agreeable (56 [31%]), repeated the question as part of the response (47 [26%]), was too easily convinced by clinician suggestions (35 [19%]), or was not offended or confused by poor clinician performance (e.g., jargon, poorly worded questions; 32 [18%]). Frequent enhancers included expressing an explicit preference or choice (especially preferences contrary to the clinician's initial suggestion, 106 [59%]), expressing appropriate emotion (38 [21%]), and notably natural speech (38 [21%]).

For feedback, detractors included excessively positive feedback relative to actual performance (42 [23%]), failure to mention an important weakness or strength (41 [23%]), or inaccuracies due to claimed omissions that were actually done (39 [22%]) or suggested behaviors that were not really needed (32 [18%]). Enhancers included being notably specific or actionable (75 [42%]), suggesting a useful clinical action (63 [35%]), and recognizing a subtle aspect of clinician performance (46 [26%]).

Table 6. Features that detracted from or enhanced virtual patient conversation

	Feature	All, No. (%) N=180	DM, No. (%) N=90	Cough, No. (%) N=90

Dialog				
Detracted	Responses reflect atypical word choice, verbose	93 (51.7%)	50 (55.6%)	43 (47.8%)
	Overly agreeable	56 (31.1%)	35 (38.9%)	21 (23.3%)
	Repeated question as part of response	47 (26.1%)	16 (17.8%)	31 (34.4%)
	Easily convinced or manipulated by clinician	35 (19.4%)	23 (25.6%)	12 (13.3%)
	Not offended or confused by poor clinician performance (including jargon)	32 (17.8%)	20 (22.2%)	12 (13.3%)
	Clinician dialog was unrealistic	29 (16.1%)	14 (15.6%)	15 (16.7%)
	Volunteered too much information (without being asked)	28 (15.6%)	15 (16.7%)	13 (14.4%)
	Test ordering and reporting was unrealistic	23 (12.8%)	1 (1.1%)	22 (24.4%)
	Responses did not make sense	12 (6.7%)	2 (2.2%)	10 (11.1%)
	Offered excessive teaching support	10 (5.6%)	4 (4.4%)	6 (6.7%)
	Switched to playing role of doctor	6 (3.3%)	0 (0.0%)	6 (6.7%)
Enhanced	Expressed preference, challenged recommendations, made clear choice	106 (58.9%)	57 (63.3%)	49 (54.4%)
	Expressed appropriate emotion	40 (22.2%)	23 (25.6%)	17 (18.9%)
	Very natural flow; authentic word choice; fluent	38 (21.1%)	24 (26.7%)	14 (15.6%)
	Challenged clinician when vague or nonsensical	31 (17.2%)	6 (6.7%)	25 (27.8%)
Feedback				
Detracte	Too positive/insufficient critique (relative to	42	17	25

d	actual performance)	(23.3%)	(18.9%)	(27.8%)
	Omission: Behavioral weakness or strength not mentioned	41 (22.8%)	18 (20.0%)	23 (25.6%)
	Inaccurate: "Omitted" behaviors really *were* done	39 (21.7%)	19 (21.1%)	20 (22.2%)
	Inaccurate: "Needed" behaviors really not needed	32 (17.8%)	19 (21.1%)	13 (14.4%)
	Too long, unrealistically detailed	24 (13.3%)	9 (10.0%)	15 (16.7%)
	Too negative/ insufficient praise (relative to actual performance)	23 (12.8%)	13 (14.4%)	10 (11.1%)
	Inaccurate: "Observed" behaviors really *not* done	22 (12.2%)	15 (16.7%)	7 (7.8%)
	Too vague, brief	19 (10.6%)	11 (12.2%)	8 (8.9%)
	Omission: Inappropriate treatment plan not mentioned	17 (9.4%)	9 (10.0%)	8 (8.9%)
	Inaccurate: Suggested clinical test/treatment not really needed	15 (8.3%)	10 (11.1%)	5 (5.6%)
Enhanced	Notably specific, actionable, constructive, accurate	75 (41.7%)	41 (45.6%)	34 (37.8%)
	Suggested notably useful clinical action	63 (35.0%)	31 (34.4%)	32 (35.6%)
	Identified notably or subtly good/bad behavior	46 (25.6%)	22 (24.4%)	24 (26.7%)
	Notably well justified or prioritized	31 (17.2%)	14 (15.6%)	17 (18.9%)
	Notably balanced; limited praise for poor performance	12 (6.7%)	3 (3.3%)	9 (10.0%)

* We inductively iteratively developed a list of detracting and enhancing features throughout the process of conversation creation and final rating, and each rater then independently marked the

presence of each feature as it was noted.

Human vs LLM quality ratings

We used GPT-4.0-turbo to rate each conversation 3 times, requiring exactly 121,860 tokens ($1.22) per replication. GPT took 228 to 506 seconds to rate authenticity and 221 to 234 seconds to rate feedback. In contrast with human ratings, between-replication variance approached zero, such that all non-feature variance arose from run-to-run inconsistencies in GPT ratings (eTable 3). The resulting ICCs (Table 1) were on par with those of human raters.

In paired (feature-matched) analyses, authenticity ratings (Table 4) were significantly lower (Table 5) for human-generated vs GPT-generated ratings (-0.98 points for human-like, -0.52 points overall) whereas feedback ratings were similar (0.10 points higher).

Discussion

This study explored 4 applications of LLMs for clinical education: a low-cost, scalable LLM-powered interactive VP, LLM-generated feedback on clinician performance, LLM role-playing the clinician, and LLM-generated ratings of dialog and feedback. This is the first study to empirically evaluate LLM-powered VPs, and the results are overall favorable. According to blinded human raters, VPs approached a "very good approximation of a real conversation" with "easily overlook[ed] flaws," and LLM-generated personalized feedback was nearly "on par with [feedback] from a trained human clinician-supervisor" (quoting operational criteria for rating =5, see eBox). Moreover, the VP demonstrably represented distinct patient preferences, including often expressing opinions that opposed clinician suggestions. LLM-as-clinician dialogs had authenticity ratings similar to human-as-clinician dialogs. LLM-generated ratings of feedback

quality were similar to human ratings, whereas ratings of authenticity were much higher (which suggests inaccuracy). We also developed and validated instruments for rating dialog authenticity, VP user experience, and feedback quality.

Limitations

The most salient limitation is suboptimal rating reproducibility. Importantly, the high between-replication variances suggest that inconsistencies could come from real differences in GPT performance. Indeed, conversation creators noted significant differences in GPT responses on the second replication. High variances could also indicate within-rater idiosyncrasies and inconsistencies. Low reproducibility could further arise from restriction of range: we asked GPT to provide excellent feedback, and for the most part it delivered. Soliciting a wider range of performance (e.g., to include substandard feedback) might reveal higher agreement. We noted difficulty rating long conversations, especially when problems manifest in only part of an otherwise satisfactory conversation. User experience was difficult to rate from a written transcript; we surmise that rating experience as it dynamically unfolds in written text, or viewing a recorded performance, would be more meaningful. Importantly, our analyses adjusted for within-rater correlation, which helps mitigate rater inconsistencies for the purposes of this particular study. GPT-generated ratings also had low reproducibility, but variance arose from run-to-run inconsistencies rather than replications.

We changed the scoring rubric between conversation creation and final ratings, thus precluding a meaningful evaluation of intrarater test-retest reliability. These VPs used only written text; however, authenticity was high even with this limitation. Moreover, we note that much clinical

work now occurs using text communication. Recently-released LLMs now support live bidirectional audio and video. We implemented just 2 topics from outpatient internal medicine and a limited spectrum of patient preferences; however, our approach easily extends to other topics and contextualizing features. Finally, for this intrinsic evaluation study the clinician role was played by study investigators rather than real learners; real-world performance will be investigated in future extrinsic evaluations.

Implications

We demonstrated proof-of-concept for scalable, globally-accessible, low-cost LLM-powered VPs. The unscripted, responsive dialogs contrast sharply with most existing VPs, for which authentic, flexible dialog is notoriously difficult to replicate (and often not attempted). Such authenticity will facilitate training, assessment, and research on shared decision-making(13-16) and other management reasoning processes.(11, 12, 20) Although patient preferences were not always apparent, this parallels real life. A patient's preferences will not surface in every patient-clinician encounter, and often require elicitation by a skilled clinician.(67) Accordingly, the LLM's ability to perceptibly represent preferences is commendable. Using this LLM-powered approach, thousands of preference-sensitive VPs can be created with much higher efficiency, and potentially higher authenticity, than current labor-intensive methods. A VP is "created" as a 1-page document, and permutations are incorporated by changing a few sentences. Such permutations (preferences, comorbidities, social determinants of health, system constraints) will prove invaluable in training and assessing contextualized care.(17-19)

Our findings support the use of LLMs to deliver specific, actionable feedback to clinicians. This

fills an important, long-recognized gap in clinical training.(24-27) Although LLM-generated feedback was not perfect, it was very good. If future research can improve feedback quality (perhaps using defined rubrics), it could support education across the continuum of clinician training (beyond VPs), including audio-recorded encounters involving simulated or real human patients and encompassing practicing physicians (e.g., automated feedback on actual patient-clinician conversations for continuous professional development).

Subgroup comparisons clarify nuanced understanding. GPT-4.0-turbo outperformed GPT-3.5-turbo, at substantially greater cost. LLM-as-clinician generated a less realistic user experience even though dialog authenticity was similar. Dialogs for the poor medical student had low ratings; we attribute this to failure of the LLM to respond appropriately to poor performance (e.g., by volunteering information or not expressing confusion), and raters' perception that the student performance was unnatural.

We present evidence supporting the validity of scores from 3 instruments rating dialog authenticity, user experience, and feedback quality. Items (content) are well-grounded, and we confirmed expected relations with other variables. Reproducibility was suboptimal; however, our data suggest that inconsistencies arise, at least in part, from variation in LLM performance rather than rater idiosyncrasies. Additional research is warranted, including clarification of rater guidelines and operational criteria.

"Zero-shot" LLM-generated ratings were suboptimal. Feedback ratings were similar to pair-matched human-generated ratings but reproducibility was low. Dialog ratings were higher than

humans' (and presumably inaccurate), perhaps because GPT was rating itself. We speculate that a different LLM might be more objective. Providing examples ("few-shot" learning) may also be needed.

Our findings suggest additional avenues for research. All these innovations – the LLM-powered VPs, LLM-generated feedback, LLM-clinician, and LLM-generated ratings – would benefit from further-refined prompt engineering and iterative evaluation. As we found, LLMs respond differently every time; this is a strength (spontaneous, natural dialog) but also a liability (inconsistent conditions for assessment or training). What are the consequences of such variability, and how can variability be mitigated when needed (such as for standardized assessment)? VPs could help address or inadvertently propagate bias and stereotypes; this warrants ongoing attention.

Finally, we note diverse potential applications of LLM-powered VPs, including clinical reasoning in other contexts (e.g., inpatient and procedural settings), training non-clinicians (nurses, therapists, pharmacists, patients), education beyond clinical reasoning (basic knowledge [case-based learning], communication, teamwork, interprofessional education, tasks such as cognitive behavioral therapy or motivational interviewing, socialization into the clinical role), and generating transcripts for research (such as comparing different feedback approaches). LLM-powered VPs could also help test clinical interventions (novel workflows, informatics tools [software as a medical device], AI innovations) or rehearse specific high-stakes scenarios ("digital twin").

Acknowledgments

We thank Martin G. Tolsgaard, PhD, DMSc (Copenhagen University Hospital and Copenhagen Academy for Medical Education and Simulation [CAMES]), Grace C. Huang, MD (Harvard Medical School and Beth Israel Deaconess Medical Center), and David M. Howcroft, PhD (Edinburgh Napier University) for their review and suggestions on the rating instruments.

This study had no external funding. The funding organization (the primary investigator's institution) had no role in planning, executing, or reporting this study.

Computer tools played no role in the writing of the manuscript itself.

Conflicts of Interest

None declared.

References

1. Newman-Toker DE, Pronovost PJ. Diagnostic Errors - The Next Frontier for Patient Safety. JAMA. 2009;301:1060-2.
2. Norman GR, Monteiro SD, Sherbino J, et al. The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. Academic Medicine. 2017;92:23-30.
3. Owens DK, Qaseem A, Chou R, et al. High-value, cost-conscious health care: concepts for clinicians to evaluate the benefits, harms, and costs of medical interventions. Ann Intern Med. 2011;154:174-80.
4. Weinberger SE. Providing high-value, cost-conscious care: a critical seventh general competency for physicians. Ann Intern Med. 2011;155:386-8.
5. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005;39:98-106.
6. Norman GR, Eva KW. Diagnostic error and clinical reasoning. Medical Education. 2010;44:94-100.
7. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Medical Education. 2009;43:303–11.
8. Cook DA, Erwin PJ, Triola MM. Computerized Virtual Patients in Health Professions Education: A Systematic Review and Meta-Analysis. Academic Medicine. 2010;85:1589-602.
9. International Medical Device Regulatory Forum. *Software as a Medical Device (SAMD): Clinical Evaluation - Guidance for Industry and Food and Drug Administration Staff*: U.S. Department of Health and Human Services Food and Drug Administration; 2017.
10. Sutton RT, Pincock D, Baumgart DC, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
11. Cook DA, Sherbino J, Durning SJ. Management Reasoning: Beyond the Diagnosis. JAMA. 2018;319:2267-8.
12. Cook DA, Durning SJ, Sherbino J, et al. Management Reasoning: Implications for Health Professions Educators and a Research Agenda. Acad Med. 2019;94:1310-6.
13. 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;45:1025-37.
14. Elwyn G, Durand MA, Song J, et al. A three-talk model for shared decision making: multistage consultation process. Bmj. 2017;359:j4891.
15. Bomhof-Roordink H, Gärtner FR, Stiggelbout AM, et al. Key components of shared decision making models: a systematic review. BMJ Open. 2019;9(12):e031763.
16. Hargraves IG, Fournier AK, Montori VM, et al. Generalized shared decision making approaches and patient problems. Adapting AHRQ's SHARE Approach for Purposeful SDM. Patient Educ Couns. 2020;103:2192-9.
17. Weiner SJ, Schwartz A. Contextual Errors in Medical Decision Making: Overlooked and Understudied. Acad Med. 2016;91:657-62.
18. Weiner SJ, Schwartz A, Sharma G, et al. Patient-centered decision making and health care outcomes: an observational study. Ann Intern Med. 2013;158:573-9.

19. Weiner SJ, Schwartz A, Weaver F, et al. Contextual errors and failures in individualizing patient care: a multicenter study. Ann Intern Med. 2010;153:69-75.
20. Cook DA, Stephenson CR, Gruppen LD, et al. Management reasoning: Empirical determination of defining features and a conceptual model. Academic Medicine. 2023;98::80-7.
21. Huang G, Reynolds R, Candler C. Virtual patient simulation at US and Canadian medical schools. Acad Med. 2007;82(5):446-51.
22. Peddle M, Bearman M, Nestel D. Virtual Patients and Nontechnical Skills in Undergraduate Health Professional Education: An Integrative Review. Clinical Simulation in Nursing. 2016;12(9):400-10.
23. Ende J. Feedback in Clinical Medical Education. JAMA. 1983;250:777-81.
24. Dudek NL, Marks MB, Wood TJ, et al. Assessing the quality of supervisors' completed clinical evaluation reports. Med Educ. 2008;42:816-22.
25. Norcini JJ, Blank LL, Duffy FD, et al. The mini-CEX: a method for assessing clinical skills. Annals of Internal Medicine. 2003;138:476-81.
26. Holmboe ES, Yepes M, Williams F, et al. Feedback and the mini clinical evaluation exercise. Journal of General Internal Medicine. 2004;19:558-61.
27. Fernando N, Cleland J, McKenzie H, et al. Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments. Medical Education. 2008;42:89-95.
28. Jackson JL, Kay C, Jackson WC, et al. The Quality of Written Feedback by Attendings of Internal Medicine Residents. J Gen Intern Med. 2015;30:973-8.
29. Marcotte L, Egan R, Soleas E, et al. Assessing the quality of feedback to general internal medicine residents in a competency-based environment. Can Med Educ J. 2019;10:e32-e47.
30. Chan TM, Sebok-Syer SS, Sampson C, et al. The Quality of Assessment of Learning (Qual) Score: Validity Evidence for a Scoring System Aimed at Rating Short, Workplace-Based Comments on Trainee Performance. Teach Learn Med. 2020;32:319-29.
31. Bower JL, Christensen CM. Disruptive Technologies: Catching the Wave. Harvard Business Review. 1995;January/February:43-53.
32. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
33. Dias RD, Gupta A, Yule SJ. Using Machine Learning to Assess Physician Competence: A Systematic Review. Acad Med. 2019;94:427-39.
34. Spickard A, 3rd, Ridinger H, Wrenn J, et al. Automatic scoring of medical students' clinical notes to monitor learning in the workplace. Medical Teacher. 2014;36:68-72.
35. Cianciolo AT, LaVoie N, Parker J. Machine Scoring of Medical Students' Written Clinical Reasoning: Initial Validity Evidence. Acad Med. 2021;96:1026-35.
36. Turner L, Hashimoto DA, Vasisht S, et al. Demystifying AI: Current State and Future Role in Medical Education Assessment. Acad Med. 2023.
37. Bond WF, Zhou J, Bhat S, et al. Automated Patient Note Grading: Examining Scoring Reliability and Feasibility. Acad Med. 2023;98(11S)(11s):S90-s7.
38. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach. 2024;Online early 2024/07/12:1-3.

39. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Validity. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014:11-31.
40. Cook DA, Beckman TJ. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. American Journal of Medicine. 2006;119:166.e7-16.
41. Howcroft DM, Belz A, Clinciu M-A, et al. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. Proceedings of the 13th International Conference on Natural Language Generation, December 2020. Dublin, Ireland. Association for Computational Linguistics: 169-82.
42. Roller S, Boureau Y-L, Weston J, et al. Open-Domain Conversational Agents: Current Progress, Open Problems, and Future Directions. arXiv e-prints. 2020:arXiv:2006.12442.
43. Adiwardana D, Luong M-T, So DR, et al. Towards a Human-like Open-Domain Chatbot. arXiv e-prints. 2020:arXiv:2001.09977.
44. Clark E, August T, Serrano S, et al. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, August 2021. Online. Association for Computational Linguistics: 7282-96.
45. Deriu J, Rodrigo A, Otegi A, et al. Survey on evaluation methods for dialogue systems. Artificial Intelligence Review. 2021;54(1):755-810.
46. Finch SE, Choi JD. Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols. Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, July 2020. 1st virtual meeting. Association for Computational Linguistics: 236-45.
47. Zellers R, Holtzman A, Clark E, et al. TuringAdvice: A Generative and Dynamic Evaluation of Language Use. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2021. Online. Association for Computational Linguistics: 4856-80.
48. Liu C-W, Lowe R, Serban I, et al. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, November 2016. Austin, Texas. Association for Computational Linguistics: 2122-32.
49. Smith E, Hsu O, Qian R, et al. Human Evaluation of Conversations is an Open Problem: comparing the sensitivity of various methods for evaluating dialogue agents. Proceedings of the 4th Workshop on NLP for Conversational AI, May 2022. Dublin, Ireland. Association for Computational Linguistics: 77-97.
50. Yeh Y-T, Eskenazi M, Mehri S. A Comprehensive Assessment of Dialog Evaluation Metrics. The First Workshop on Evaluations and Assessments of Neural Conversation Systems, November 2021. Online. Association for Computational Linguistics: 15-33.
51. van der Lee C, Gatt A, van Miltenburg E, et al. Best practices for the human evaluation of automatically generated text. Proceedings of the 12th International Conference on Natural Language Generation, oct–nov 2019. Tokyo, Japan. Association for

Computational Linguistics: 355-68.
52. Huwendiek S, De Leng BA, Kononowicz AA, et al. Exploring the validity and reliability of a questionnaire for evaluating virtual patient design with a special emphasis on fostering clinical reasoning. Med Teach. 2015;37(8):775-82.
53. Peddle M, Bearman M, McKenna L, et al. Exploring undergraduate nursing student interactions with virtual patients to develop 'non-technical skills' through case study methodology. Adv Simul (Lond). 2019;4:2.
54. Schubert T, Friedmann F, Regenbrecht H. The Experience of Presence: Factor Analytic Insights. Presence: Teleoper. Virtual Environ. 2001;10(3):266–81.
55. Fink MC, Reitmeier V, Stadler M, et al. Assessment of Diagnostic Competences With Standardized Patients Versus Virtual Patients: Experimental Study in the Context of History Taking. J Med Internet Res. 2021;23(3):e21196.
56. Clement EA, Oswald A, Ghosh S, et al. Exploring the Quality of Feedback in Entrustable Professional Activity Narratives Across 24 Residency Training Programs. J Grad Med Educ. 2024;16:23-9.
57. McGuire N, Acai A, Sonnadara RR. The McMaster Narrative Comment Rating Tool: Development and Initial Validity Evidence. Teach Learn Med. 2023:1-13.
58. Zelenski AB, Tischendorf JS, Kessler M, et al. Beyond "Read More": An Intervention to Improve Faculty Written Feedback to Learners. J Grad Med Educ. 2019;11:468-71.
59. Van Ostaeyen S, Embo M, Rotsaert T, et al. A Qualitative Textual Analysis of Feedback Comments in ePortfolios: Quality and Alignment with the CanMEDS Roles. Perspect Med Educ. 2023;12(1):584-93.
60. Gin BC, Ten Cate O, O'Sullivan PS, et al. Exploring how feedback reflects entrustment decisions using artificial intelligence. Med Educ. 2022;56:303-11.
61. Spadafore M, Yilmaz Y, Rally V, et al. Using Natural Language Processing to Evaluate the Quality of Supervisor Narrative Comments in Competency-Based Medical Education. Acad Med. 2024;Online early 2024/01/17.
62. Dev S, Sheng E, Zhao J, et al. On Measures of Biases and Harms in NLP. 2021:arXiv:2108.03362.
63. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. Language and Linguistics Compass. 2021;15(8):e12432.
64. Navigli R, Conia S, Ross B. Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality. 2023;15(2):Article 10.
65. Blodgett SL, Barocas S, Daumé III H, et al. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 2020. Online. Association for Computational Linguistics: 5454-76.
66. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.
67. Lee YK, Low WY, Ng CJ. Exploring patient values in medical decision making: a qualitative study. PLoS One. 2013;8:e80051.

Supplementary Files

Multimedia Appendixes

Supplemental Digital Content.
URL: http://asset.jmir.pub/assets/d743dd7615b834df946bc1171dc48a75.docx