Original Paper

Medical Interview Performance Using an AI-Based Standardized Patient vs a Human Standardized Patient: Randomized Crossover Trial

Yasutaka Yanagita[1], MD, PhD; Daiki Yokokawa[1], MD, PhD; Fumitoshi Fukuzawa[2], MD; Shiichi Ihara[1], MD; Ryo Yoshida[1], MD; Rurika Sato[1], MD, PhD; Takanori Uehara[1], MD, PhD

[1]Department of General Medicine, Chiba University Hospital, Chiba, Japan

[2]Department of Medicine, Mount Sinai Morningside, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Yasutaka Yanagita, MD, PhD
Department of General Medicine
Chiba University Hospital
1-8-1 Inohana
Chiba, 260-8677
Japan
Phone: 81 43 222 7171
Email: y.yanagita@gmail.com

Abstract

Background: In Japan, standardized patients (SPs) have traditionally been recruited to teach medical interview competencies. Recent advances in large language models have enabled increasingly natural conversational interactions, raising the possibility of using AI-based SPs (AI-SPs) for medical interview training. However, medical interview performance during interactions with AI-SPs compared with human SPs remains insufficiently understood.

Objective: This randomized single-center crossover trial aimed to compare medical interview performance and learner-reported experiences between encounters with a large language model–based AI-SP and a human SP.

Methods: Participants were fifth-year medical students undertaking clinical clerkships at Chiba University Hospital. The AI-SP was a text-based simulated patient implemented using GPT-4o. Participants were randomized to AI-SP followed by human SP, or the reverse. A washout period of at least 3 days was set between conditions. Each participant completed 2 cases, depression and appendicitis, in both conditions. For the AI-SP interviews, chat logs were retained, and a physician provided feedback after completion of the 2 cases. Human SP interviews were video recorded, and the physician portraying the SP provided feedback immediately after each interview. Medical interview performance was assessed using selected items from the Mini-Clinical Evaluation Exercise (Mini-CEX), with ratings performed independently by 2 general medicine specialists. Participants completed a self-rated questionnaire (7-point Likert-type scale) regarding learning satisfaction, perceived rapport formation, and active engagement in history taking.

Results: Forty-eight participants were analyzed. The Mini-CEX–based mean total score was significantly higher for the human SP than for the AI-SP (23.9, SD 1.7 vs 22.7, SD 1.7, respectively; mean difference 1.24; t_{47}=3.7; P=.001). Scores for history taking and overall clinical competence were higher in the human SP condition. Interrater reliability (intraclass correlation coefficient of 2,1) was 0.76 (95% CI 0.66-0.83). No significant differences were found between conditions for learning satisfaction (human SP: mean 5.9, SD 1.1; AI-SP: mean 6.0, SD 0.8) or perceived rapport formation (human SP: mean 5.4, SD 1.3; AI-SP: mean 5.3, SD 1.3). Active engagement in history taking was significantly higher in the AI-SP condition (human SP: mean 5.8, SD 1.1; AI-SP: mean 6.2, SD 0.7; t_{47}=2.7; P=.009).

Conclusions: Medical interview performance differed between the 2 SP conditions, with higher history taking and overall clinical competence scores in the human SP condition. The text-based nature of the AI-SP and the absence of nonverbal cues may have contributed to these differences. Although no significant differences between conditions were observed in most other assessment outcomes, learning satisfaction, or perceived rapport, self-rated active engagement in history taking was higher with the AI-SP condition. AI-SPs may offer a complementary modality for medical interview practice; however, this study assessed performance during simulated encounters rather than subsequent learning gains or performance with real patients.

XSL•FO
RenderX

Trial Registration: UMIN Clinical Trials Registry UMIN000056228; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000064153

(JMIR Med Educ 2026;12:e111030) doi: 10.2196/111030

KEYWORDS

AI standardized patient; large language model; medical interview education; artificial intelligence; AI

Introduction

Medical interviewing is an essential clinical skill that enables health care professionals to accurately elicit the necessary information from patients to establish diagnosis, develop a good patient-clinician relationship, and guide subsequent processes such as treatment decision-making. It is important for professionals, especially novice learners, to know that medical interviewing goes beyond making predetermined questions, involving also the eliciting of patients' concerns, understanding the course of symptoms and their impacts on daily life, responding empathetically to patients' anxieties, and exploring patients' explanatory models. Some effective measures for improving medical interviewing skills among novice learners include appropriate feedback and opportunities for repeated practice in settings that closely resemble actual clinical encounters [1].

The Mini-Clinical Evaluation Exercise (Mini-CEX) has been widely used to observe and assess learners' clinical competence in contexts similar to real clinical practice and provide them with feedback [2,3]. In Japan, practical training using standardized patients (SPs) has traditionally played a central role in medical interview education [4,5]. Indeed, SPs allow learners to experience patient-clinician interactions in a safe environment and acquire clinical competencies (eg, history taking, communication, and professionalism) in a more integrated manner. However, many SP programs rely heavily on volunteers, and their implementation requires substantial effort both for SP recruitment and to train SPs to understand case scenarios and standardize their performances [6]. Additional burdens associated with these programs include securing faculty members, managing the program, and covering operational costs. Moreover, with the need to support practical examinations such as the objective structured clinical examination, which is conducted across Japanese medical schools, it has become increasingly difficult for many medical education institutions to continuously secure a sufficient number of SPs [7,8].

More recently, advances in AI, particularly in large language models, have generated a growing interest in the educational use of AI systems capable of engaging in natural dialogue with humans. Medical education has recently begun to see the emergence of AI-based medical interview simulations and educational support tools. Prior research has demonstrated that AI-based SPs (hereinafter referred to as AI-SPs) can understand learners' questioning intentions with high accuracy and provide clinically appropriate responses, to the point that preliminary evidence has accumulated regarding both the feasibility of implementing such systems as dialogue-based educational tools and the accuracy of their basic conversational capabilities [9,10].

The problem that remains, then, is that the bulk of the literature has focused primarily on technical validation, including speech recognition accuracy, question understanding, and response accuracy. Meanwhile, a void persists of studies directly investigating the impact of AI-SPs on learners' medical interviewing skills and learning experiences in authentic medical education settings [11]. Related investigations are important because AI-SPs offer several potential advantages, including enabling learners to practice without fear of failure in a psychologically safe environment; engage repeatedly with the same case; simultaneously access training across large numbers of learners; and experience configuration flexibility for patient characteristics such as disease, age, sex, social background, and chief concern [12,13].

In this study, we conducted a randomized crossover trial to compare medical interview performance and learner-reported experiences between encounters with a large language model–based AI-SP and a human SP. Medical interview performance was assessed using selected items from the Mini-CEX, and participants' experiences were assessed using self-rated questionnaires. Rather than evaluating subsequent learning gains or transfer to real patient encounters, this study focused on differences in performance and learner experience under the 2 simulated patient conditions. If effective, AI-SPs could provide a flexible and efficient approach to medical interview education, particularly in settings with limited human resources. This approach may also be applicable in the near future to interprofessional education involving medical students, nurses, pharmacists, and other health care professionals.

Methods

Study Design

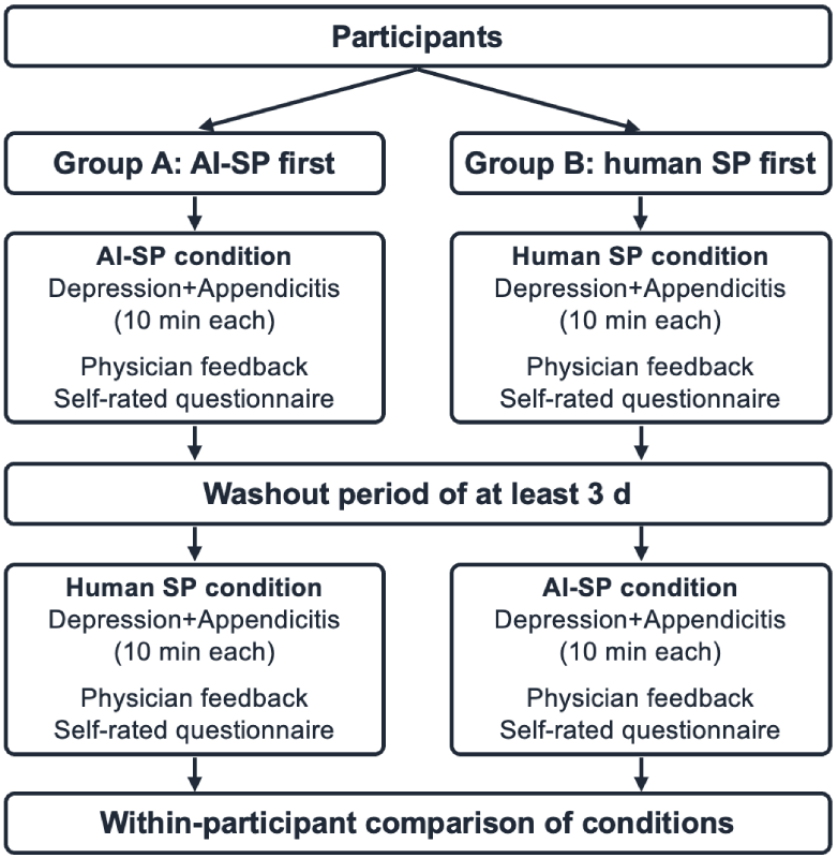
This was a single-center crossover trial. To examine whether AI-SPs could complement human SPs in medical interview education, we compared chat-based medical interviews using an AI-SP with face-to-face medical interviews using a human SP. Data were collected from December 2024 to September 2025. The participants were medical students from Chiba University School of Medicine who participated in a clinical clerkship at the Department of General Medicine, Chiba University Hospital.

Participants were randomly assigned to 1 of 2 groups. In the AI-SP–first group, participants first conducted chat-based medical interviews with the AI-SP; after a washout period of at least 3 days, they conducted face-to-face medical interviews with the human SP. In the human SP–first group, participants first conducted face-to-face medical interviews with the human SP; after the same washout period as in the other group, they conducted chat-based medical interviews with the AI-SP. Each

XSL•FO
RenderX

participant completed medical interviews for 2 cases, depression and appendicitis, under both the AI-SP and human SP conditions. This design enabled within-participant comparisons of medical interview performance between the AI-SP and human SP conditions. After each interview, a physician provided feedback to the participant. Medical interview performance was independently assessed by 2 general medicine specialists using selected items from the Mini-CEX. Participants then rated 3 aspects of their learning experience on a 7-point Likert scale. Figure 1 shows the flowchart of this study.

Figure 1. Study flow. AI-SP: AI-based standardized patient; SP: standardized patient.



Participants

A total of 76 fifth-year medical students from a single academic year at Chiba University School of Medicine participated in this study. Students were eligible if they were fifth-year medical students participating in the clinical clerkship at the Department of General Medicine during the study period and had completed the common preclinical examinations, including computer-based testing and the objective structured clinical examination. Eligible students were consecutively approached for participation during their clinical clerkship. The purpose and procedures of the study were explained to them, and students who agreed to participate were enrolled after providing verbal informed consent. Participation was voluntary and did not affect their clerkship evaluation. Because this study used a crossover design, participants were required to complete both medical interview sessions, with a washout period of at least 3 days between the sessions. The 76 participants were randomly assigned in equal numbers to the AI-SP–first sequence (n=38) or the human SP–first sequence (n=38). Participants who did not complete both medical interview sessions were excluded from the analysis. Of the 76 participants, 28 did not complete both medical interview sessions and were therefore excluded from the analysis, leaving 48 participants for the final analysis (AI-SP first: n=25; human SP first: n=23).

Case Scenarios

In this study, we selected 2 cases, depression and appendicitis, as conditions commonly encountered in primary care outpatient settings and that require multidimensional history taking. For each case, we developed a scenario that included the chief concern, age, sex, medical history, medication history, social history, circumstances at presentation, and symptom course. The complete case scenarios used in this study are provided in Multimedia Appendix 1.

Generation and Validation of the AI-SP

GPT-4o (OpenAI) was used to create the AI-SP. The AI-SP interviews were conducted using GPT-4o through the ChatGPT web interface with a research-specific account from December 2024 to September 2025. Temporary Chat mode was used, and a new chat session was initiated for each participant and each case. Thus, conversation content from previous participants or cases was not carried over to subsequent sessions. No additional ChatGPT tools were used during the interviews. The model available during the study was identified as GPT-4o; a specific model snapshot or version identifier was not recorded. Prompts were developed to enable the AI to respond as an SP based on the case settings. The prompts included the patient's basic information, chief concern, symptom details, social history, attitude during responses, and information to be disclosed

XSL•FO
RenderX

depending on the participant's questions. Each case-specific prompt instructed the model to remain in the role of the simulated patient and disclose information according to the questions asked by the participant. The complete prompts used for the 2 cases are provided in Multimedia Appendix 2. The AI-SP was designed to avoid inappropriate expressions during the conversation and terminate the interview if the participant made inappropriate remarks. The AI-SP scenarios and responses were reviewed by 2 board-certified general internists, who evaluated their medical validity and educational appropriateness.

Procedure

For the AI-SP interviews, case-specific prompts were entered into GPT-4o in advance, and the model was instructed to behave like an SP. Participants conducted 10-minute chat-based medical interviews with the AI-SP for each of the 2 cases, depression and appendicitis. After both AI-SP cases were completed, a physician reviewed the chat logs with the participant and provided feedback on the interviews.

For the human SP interviews, participants conducted 10-minute face-to-face medical interviews for each of the same 2 cases, and the interviews were recorded. The human SPs were portrayed by 7 general physicians. Before participating in the study, all physicians reviewed the standardized case scenarios and received instructions regarding patient characteristics, clinical information, and the information to be disclosed in response to participants' questions to promote consistency across encounters. The physicians portraying the human SPs provided feedback immediately after each case. Medical interview performance was subsequently assessed by 2 general medicine specialists.

Evaluation

Medical interview performance was evaluated using the Mini-CEX, a tool used to assess clinical competence. Because the AI-SP condition did not involve any type of physical examination, the relevant item for this topic was excluded, and 6 items were adopted for the Mini-CEX–based assessment (each rated on a 9-point scale; maximum total score of 54 points). The assessment items included history taking, communication, diagnosis, management, professionalism, and overall clinical competence.

Medical interview performance was evaluated after participants completed both the depression and appendicitis cases in each condition (AI-SP and human SP). For the AI-SP condition, chat logs were used for assessment; for the human SP condition, recorded videos were used. Two specialist physicians engaged in general medicine who were not involved in the medical interviews or feedback sessions independently evaluated the participants' performance; they assessed each participant independently without consulting each other. For each condition, the depression and appendicitis cases were not scored separately. Instead, each evaluator reviewed the participant's performance across both cases and assigned a single score for each Mini-CEX item for that condition. The scores assigned by the 2 independent evaluators were then averaged to obtain the final score for each participant in each condition.

Participants used a 7-point Likert-type scale to provide self-evaluations on 3 items in the participant questionnaire, namely, satisfaction with the simulated interview, degree of rapport formation with the SP, and active engagement in history taking.

Statistical Analysis

The required sample size was calculated using G*Power (version 3.1) and was determined to be 34 participants while considering a 2-sided test, a significance level of .05, power of 0.8, and an effect size of 0.5. For each participant, medical interview performance was evaluated once in each condition after completion of the two cases, depression and appendicitis. Thus, each participant had 1 Mini-CEX–based assessment score for the AI-SP condition and 1 for the human SP condition. The scores were assigned by the 2 independent evaluators, and the mean score was used as the final score for each participant in each condition. The Mini-CEX–based assessment scores and questionnaire responses on the 7-point Likert scale were compared between the AI-SP and human SP conditions within the same participants using 2-tailed paired *t* tests. All statistical analyses were performed using SPSS Statistics for Windows (version 26.0; IBM Corp).

Ethical Considerations

This study was approved by the ethics committee of Chiba University Hospital (approval 545). The study protocol, including the use of verbal informed consent rather than written informed consent, was approved by the ethics committee. Verbal informed consent was obtained from all participants before enrollment in the study. Participants were instructed not to enter personally identifiable information, including their names or student identification numbers, during interactions with the AI-SP. A research-specific ChatGPT account was used for the AI-SP sessions. Chat logs used for study evaluation were stored on a university-managed computer, and access to the study data was restricted to the investigators involved in the study and assessment. No financial compensation was provided for participation. This study was registered with the University Hospital Medical Information Network Clinical Trials Registry on November 21, 2024 (unique trial number UMIN000056228).

Results

Participant Characteristics

Of the 76 students who participated in this study, 38 were randomly assigned to the AI-SP–first sequence, and 38 were assigned to the human SP–first sequence. Twenty-eight participants did not complete both medical interview sessions and were excluded from the analysis. Of these, 13 were in the AI-SP–first sequence, and 15 were in the human SP–first sequence. Reasons for noncompletion included clinical clerkship scheduling and personal circumstances, such as leave or illness. Consequently, 48 participants were included in the final analysis (AI-SP first: n=25; human SP first: n=23). Among these 48 fifth-year medical students, the median age was 22.0 years (IQR 22.0-23.0); 33 (68.8%) were male, and 15 (31.3%) were female. Participant characteristics are shown in Table 1 according to intervention sequence. No adverse events, privacy breaches,

XSL•FO
RenderX

major technical problems, or other unintended incidents related to either study condition were observed. The participant flow is shown in Figure 2.

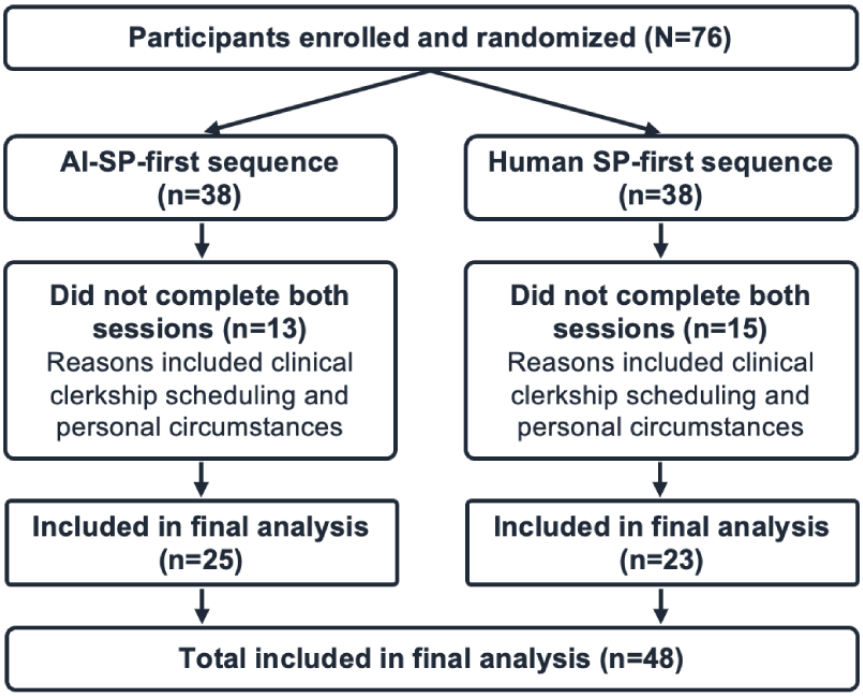
Table 1. Participant characteristics.

Characteristics	AI-SP[a] first (n=25)	Human SP[b] first (n=23)
Age (y), median (IQR)	22.0 (22.0-23.0)	22.0 (22.0-23.0)
Male, n (%)	16 (64)	17 (73.9)
Female, n (%)	9 (36)	6 (26.1)

[a]AI-SP: AI-based standardized patient.

[b]SP: standardized patient.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of study participants. AI-SP: AI-based standardized patient; SP: standardized patient.



Medical Interview Performance

Total scores were significantly higher in the human SP condition than in the AI-SP condition (mean 23.9, SD 1.7 vs mean 22.7, SD 1.7; t_{47}=3.7; *P*=.001), and the interrater reliability of the Mini-CEX–based assessment was good, with an intraclass correlation coefficient (2,1) of 0.76 (95% CI 0.66-0.83). Among the individual assessment items, history taking scores were significantly higher in the human SP condition than in the AI-SP condition (mean 4.3, SD 0.5 vs mean 3.5, SD 0.5; t_{47}=6.9; *P*<.001), and so were overall clinical competence scores (mean 3.9, SD 0.2 vs mean 3.7, SD 0.5; t_{47}=3.2; *P*=.002). No significant differences were observed between conditions for communication, diagnosis, professionalism, or management (Table 2).

XSL•FO
RenderX

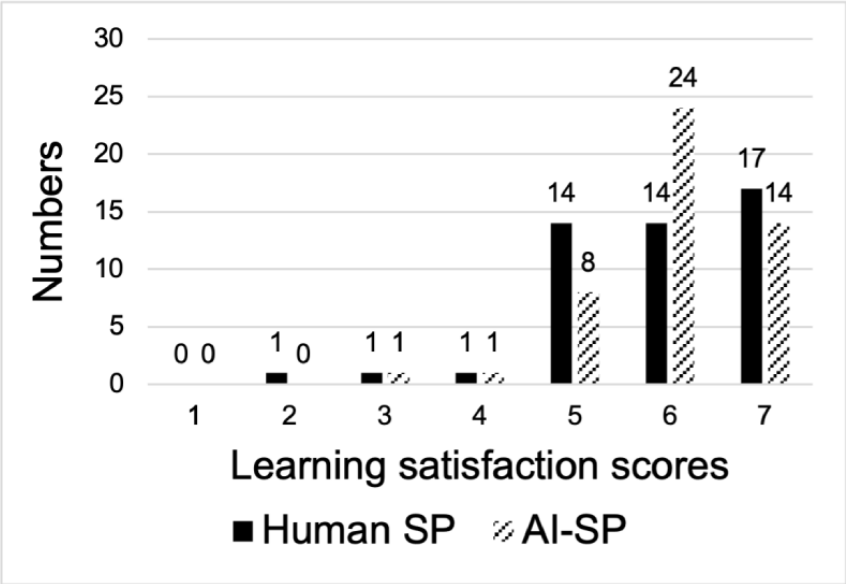
Table 2. Comparison of evaluation items between human standardized patients (SPs) and AI-based SPs (AI-SPs)[a].

	Human SP, mean (SD)	AI-SP, mean (SD)	Mean difference (95% CI)	Cohen *dz*	*t* test (*df*)	*P* value
History taking	4.3 (0.5)	3.5 (0.5)	0.81 (0.58 to 1.05)	1.00	6.9 (47)	<.001
Communication	3.9 (0.5)	3.9 (0.5)	0.00 (−0.19 to 0.19)	0.00	0.0 (47)	>.99
Diagnosis	4.0 (0.5)	3.9 (0.5)	0.11 (−0.02 to 0.25)	0.26	1.8 (47)	.09
Professionalism	3.9 (0.4)	3.9 (0.4)	−0.01 (−0.16 to 0.14)	0.01	0.1 (47)	.89
Management	3.9 (0.3)	3.9 (0.3)	0.08 (−0.03 to 0.20)	0.22	1.5 (47)	.15
Comprehensive evaluation	3.9 (0.2)	3.7 (0.5)	0.24 (0.09 to 0.39)	0.46	3.2 (47)	.002
Total	23.9 (1.7)	22.7 (1.7)	1.24 (0.56 to 1.92)	0.53	3.7 (47)	.001

[a]Scores for each Mini-Clinical Evaluation Exercise item were rated on a 9-point scale, with a maximum total score of 54. Mean differences were calculated as human SP minus AI-SP.

Questionnaire results showed no significant differences between the human SP and AI-SP conditions in learning satisfaction (mean 5.9, SD 1.1 vs mean 6.0, SD 0.8) or perceived rapport formation (mean 5.4, SD 1.3 vs mean 5.3, SD 1.3). However, participants reported significantly greater active engagement in history taking in the AI-SP condition than in the human SP condition (mean 6.2, SD 0.7 vs mean 5.8, SD 1.1; t_{47}=2.7; *P*=.009; Figures 3-5).

Figure 3. Learning satisfaction scores in the Human SP and AI-SP conditions. AI-SP: AI-based standardized patient; SP: standardized patient.



XSL•FO
RenderX

Figure 4. Degree of rapport formation scores in the Human SP and AI-SP conditions. AI-SP: AI-based standardized patient; SP: standardized patient.

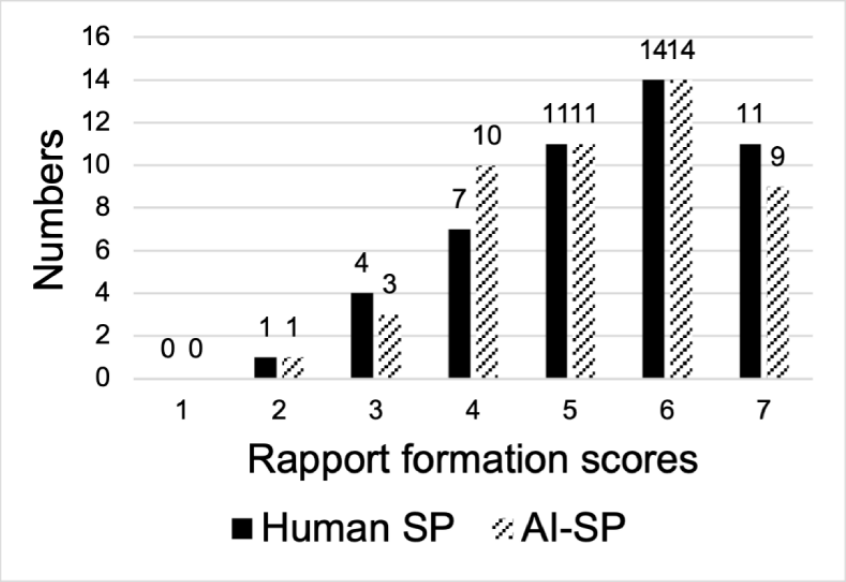
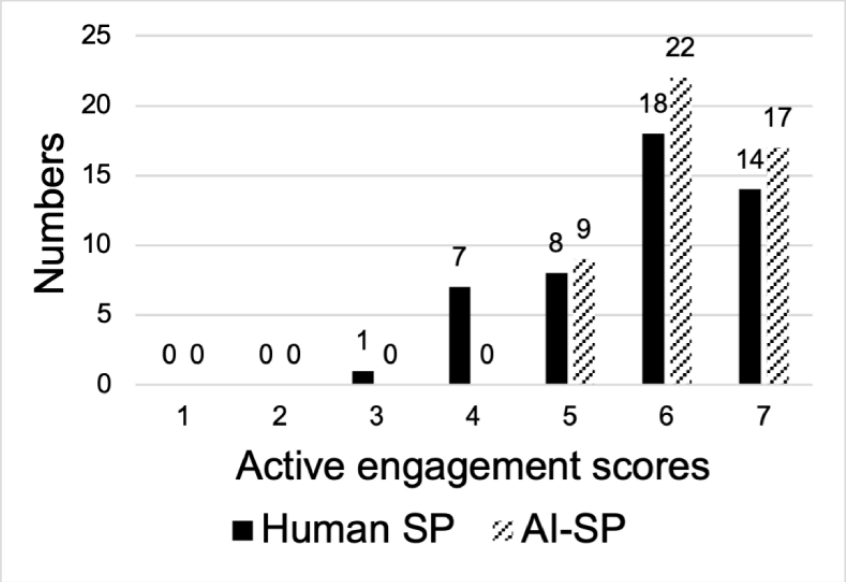


Figure 5. Active engagement in history taking scores in the Human SP and AI-SP conditions. AI-SP: AI-based standardized patient; SP: standardized patient.



Discussion

Principal Findings

This study compared the educational effects of the AI-SP and human SP conditions using the Mini-CEX and a self-rated questionnaire. Physician assessments based on the Mini-CEX showed significantly higher total scores in the human SP condition than in the AI-SP condition, with significantly higher scores observed for history taking and overall clinical competence; no significant differences between the 2 conditions were observed in communication, diagnosis, professionalism, or management. In the self-rated questionnaire, no significant differences were observed between conditions in learning satisfaction or perceived rapport formation, whereas active engagement in history taking was significantly higher in the AI-SP condition.

The higher history taking and overall clinical competence scores observed in the human SP condition should be interpreted as differences in performance under the two interaction conditions rather than as evidence of greater educational effectiveness of human SP training. Human SPs provide rich nonverbal cues, including facial expressions, vocal tone, pauses, and hesitation, which can prompt learners to explore and elicit clinically relevant information [14]. Moreover, nonverbal behaviors such as eye contact, laughter, and gestures are known to be critical for shaping the quality of physician-patient interactions during simulated medical interviews [15]. Advanced interpersonal skills, including waiting for patients to elaborate on their concerns, probing ambiguous concerns, and gathering information while responding sensitively to emotional distress [16], may be supported by interactions that incorporate such nonverbal information. In contrast, the AI-SP used in this study was limited to text-based communication. The absence of vocal

XSL•FO
RenderX

tone, facial expressions, pauses, and other nonverbal cues may have affected not only participants' information gathering but also their interpretation of the simulated patient's responses. In addition, the AI-generated responses may have been more coherent and logically structured, which may have reduced the extent to which emotional nuances and patient affect were conveyed. These differences in interactional characteristics may therefore have contributed to the observed differences in performance between the 2 conditions.

No significant differences were observed between the AI-SP and human SP conditions regarding communication, diagnosis, management, and professionalism. As this study was not designed as an equivalence or noninferiority trial, these findings should not be interpreted as evidence that AI-SPs are equivalent or noninferior to human SPs in these domains. These findings suggest that AI-SPs can provide specific learning opportunities for medical interview education but not to all learning domains. Further research is required to determine the extent of their educational effectiveness and roles relative to human SPs.

Another factor that should be considered is the potential for evaluator bias. Because blinding was not feasible, the evaluators were necessarily aware of whether they were assessing a face-to-face interaction with a human SP or a text-based chat with an AI-SP. This awareness may have implicitly influenced their judgment. In particular, the domain of overall clinical competence likely encompasses tacit and less formalized evaluation criteria such as interpersonal skills, situational awareness, and the ability to interpret the social and emotional context of the encounter. Evaluators may have perceived interactions with human SPs as being more "clinical" or "authentic," and this may have led them to assign higher ratings accordingly. Previous research has shown that when experts evaluate the outputs generated by humans and AI, they may exhibit a perception bias favoring content believed to be human generated even when objective output quality is equivalent between the 2 conditions [17]. Therefore, evaluator bias may have contributed, at least partially, to the observed differences between the AI-SP and human SP conditions in this study.

Participant self-assessments revealed no significant differences between the AI-SP and human SP conditions regarding learning satisfaction or degree of rapport formation. Notably, there was a similar degree of rapport formation with the AI-SP despite the interaction being conducted entirely through text. What this indicates is that a well-designed AI-SP can provide meaningful interactive experiences for learners even through a text-based format [18]. Furthermore, active engagement in history taking was significantly higher in the AI-SP condition. Although conventional interactions with SPs may be associated with psychological stress for some learners [19], psychological safety, anxiety, interpersonal tension, and fear of failure were not directly measured in the present study. Therefore, the higher self-rated active engagement observed in the AI-SP condition should not be interpreted as evidence of greater psychological safety. One possible hypothesis is that interaction with an AI-SP may reduce perceived interpersonal pressure and thereby facilitate active participation; however, this mechanism should be directly examined in future studies using validated measures of psychological safety and anxiety.

Additionally, AI-SPs may provide meaningful learning opportunities in domains evaluated within a structured framework of medical interviewing, such as communication and professionalism. In the context of communication training using AI-based chatbots, most of the sample in a previous study found the feedback useful and perceived improvements in physician-patient communication and recognition of mental health conditions [20]. As AI systems can deliver consistent responses according to predefined patterns, they may be particularly effective in helping learners acquire the fundamental structure and flow of medical interviews. As AI-SP technology continues to evolve, advances in multimodal systems that integrate voice, text, and visual information may enable more natural and realistic interactions that incorporate emotional expressions and nonverbal cues [21]. Such developments could help address some of the current limitations of AI-SPs in history taking and rapport building highlighted by the present study, further expanding their potential role in medical interview education.

Importantly, AI-SPs should not be viewed as direct replacements for human SPs but as complementary educational tools that can support the repeated practice of medical interviewing with a lower psychological burden on participants. One of the greatest advantages of AI-SPs is their ability to provide opportunities for repeated practice without time or location constraints. However, AI-SPs remain limited in their ability to reproduce communication through nonverbal cues such as facial expressions, vocal intonation, and silence [14,15]. In this respect, human SPs continue to play important and irreplaceable roles. Therefore, AI-SPs and human SPs serve distinct educational functions, and a staged educational approach that combines both modalities may represent an effective implementation strategy. Such an approach could involve the use of AI-SPs during initial practice stages followed by progression to interviews with human SPs as participants gain proficiency. In educational settings where securing human SPs is challenging and there is a lack of human resources, AI-SPs may also be actively used within the scope of their capabilities. The appropriate integration of these tools into medical education has the potential to expand educational opportunities and contribute to the training of future physicians.

Limitations

This study had several limitations. First, although the crossover design enabled within-participant comparisons between the AI-SP and human SP conditions and reduced the influence of individual differences in medical interviewing and communication skills, only 2 case scenarios were included. Therefore, it remains unclear whether similar findings would be observed across various diseases or more complex clinical presentations. In addition, although a washout period of at least 3 days was implemented, potential order effects and carryover effects resulting from prior interview experiences and feedback cannot be completely excluded. Moreover, 28 of the 76 randomized participants did not complete both study sessions and were excluded from the final analysis. Although the number of noncompleters was similar between the 2 sequences, the possibility of attrition bias cannot be excluded.

XSL•FO
RenderX

Second, this study included fifth-year medical students from a single medical school, which may limit the generalizability of the findings. The AI-SP used was limited to a text-based chat interface and did not provide nonverbal cues such as vocal characteristics, facial expressions, or other affective signals. Consequently, participants may have had difficulty assessing patients' understanding, anxiety, or degree of acceptance, limiting their opportunities to learn how to adapt explanations to patient needs and establish rapport through emotionally responsive communication. As nonverbal communication constitutes an important component of real-world clinical encounters, human SPs currently retain a substantial educational value that may not be fully replicated by chat-based AI systems. Results may differ when multimodal AI-SPs capable of incorporating voice, facial expressions, and other nonverbal cues are used.

Third, AI technology is evolving rapidly, and both the language model and prompt design used here may become outdated over time. Future AI implementation in medical education will require systematic approaches to prompt engineering as well as ongoing evaluation of the educational effectiveness and validity of virtual SPs.

Fourth, the Mini-CEX–based assessment was originally developed for face-to-face clinical encounters such that its construct validity in the context of AI-SP–based interviews is inherently limited. Domains that depend primarily on the content of the interaction, such as history taking, diagnosis, and management, may be more directly assessable in a text-based format. In contrast, domains that incorporate nonverbal and interpersonal behaviors, including communication, professionalism, and overall clinical competence, may not be assessed equivalently between text-based AI-SP and face-to-face human SP encounters. The present study was not designed to distinguish whether the observed differences in assessment scores reflected differences in participants' clinical competence or in the interactional characteristics of the 2 conditions and the interactional characteristics provided by the AI-SP. Future studies should consider developing and validating assessment frameworks specifically tailored to AI-mediated clinical communication environments.

Conclusions

This study shows how human SPs outperformed AI-SPs in the assessment of history taking and overall clinical competence, although there were no significant differences between the 2 conditions in communication, diagnosis, management, or professionalism. Meanwhile, participants demonstrated greater active engagement in history taking when interacting with AI-SPs. These findings suggest that human SPs remain superior in teaching history taking skills that require nonverbal interaction and advanced interpersonal communication. Thus, AI-SPs should not be viewed as direct replacements for human SPs, but they could serve as complementary educational tools that promote active learning and support early-stage practice in the stepwise acquisition of medical interviewing skills. Future research should focus on establishing assessment frameworks tailored to AI-SP–based learning environments and validating their educational effectiveness across multiple institutions to strengthen the evidence base for the implementation of AI-SPs in medical education.

Data Availability

The case questions and datasets related to the results of this study are available from the corresponding author (YY) on reasonable request.

Funding

This work was supported by Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research JP25K20499. The article processing charge was funded by the Ministry of Health, Labour and Welfare, Japan, under its FY2026 physician training program.

Generative AI Use

During the revision of this manuscript, the authors used GPT-4o (OpenAI) to assist with language editing, restructuring, and refinement of the manuscript text. All AI-generated output was reviewed and revised by the authors, who take full responsibility for the final content.

Authors' Contributions

Conceptualization: YY, DY, TU
Formal analysis: YY, DY, FF
Investigation: YY, SI, RI, RS
Project administration: YY, DY, TU
Writing—original draft: YY
Writing—review and editing: DY, TU, FF
All authors have read and approved the final manuscript, agree to be accountable for all aspects of the work, and ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1

Original and English-translated cases.
[DOCX File , 23 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Original and English-translated prompts.
[DOCX File , 18 KB-Multimedia Appendix 2]

References

1. Abraham RM, Singaram VS. Using deliberate practice framework to assess the quality of feedback in undergraduate clinical skills training. BMC Med Educ. Apr 11, 2019;19(1):105. [FREE Full text] [doi: 10.1186/s12909-019-1547-5] [Medline: 30975213]
2. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med. Nov 15, 1995;123(10):795-799. [doi: 10.7326/0003-4819-123-10-199511150-00008] [Medline: 7574198]
3. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Ann Intern Med. Mar 18, 2003;138(6):476-481. [doi: 10.7326/0003-4819-138-6-200303180-00012] [Medline: 12639081]
4. Akatsu H, Shiima Y, Gomi H, Hegab AE, Kobayashi G, Naka T, et al. Teaching "medical interview and physical examination" from the very beginning of medical school and using "escape rooms" during the final assessment: achievements and educational impact in Japan. BMC Med Educ. Jan 28, 2022;22(1):67. [FREE Full text] [doi: 10.1186/s12909-022-03130-2] [Medline: 35090459]
5. Harada Y, Kubota Y, Hirayama Y, Otaki J, Mitoma H. Medical school staff's standardized patient experience alters their understanding of student education. Tohoku J Exp Med. Jan 2022;256(1):63-71. [FREE Full text] [doi: 10.1620/tjem.256.63] [Medline: 35095029]
6. Hayashi M. The paradox caused by the standardized pre-clinical objective structured clinical examination in Japan. Korean J Med Educ. Mar 2025;37(1):85-87. [FREE Full text] [doi: 10.3946/kjme.2025.327] [Medline: 40049687]
7. Flanagan OL, Cummings KM. Standardized patients in medical education: a review of the literature. Cureus. Jul 17, 2023;15(7):e42027. [FREE Full text] [doi: 10.7759/cureus.42027] [Medline: 37593270]
8. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Med Educ. Oct 24, 2015;15:183. [FREE Full text] [doi: 10.1186/s12909-015-0468-1] [Medline: 26498479]
9. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. Nov 2024;58(11):1276-1285. [doi: 10.1111/medu.15402] [Medline: 38639098]
10. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med. Dec 2023;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
11. Maicher KR, Stiff A, Scholl M, White M, Fosler-Lussier E, Schuler W, et al. Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity. Med Teach. Mar 2023;45(3):279-285. [doi: 10.1080/0142159X.2022.2130216] [Medline: 36346810]
12. Gray M, Baird A, Sawyer T, James J, DeBroux T, Bartlett M, et al. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of ChatGPT: exploratory observational study. JMIR Med Educ. Feb 01, 2024;10:e50705. [FREE Full text] [doi: 10.2196/50705] [Medline: 38300696]
13. Aljedaani S, Sultan I, El-Tarhouny S. Knowledge-based chatbots in clinical teaching: AI-powered virtual patients to improve history-taking during internal medicine rotations. Int Med Educ. May 21, 2026;5(2):51. [FREE Full text] [doi: 10.3390/ime5020051]
14. Park KH, Park SG. The effect of communication training using standardized patients on nonverbal behaviors in medical students. Korean J Med Educ. Jun 2018;30(2):153-159. [FREE Full text] [doi: 10.3946/kjme.2018.90] [Medline: 29860781]
15. D'Agostino TA, Bylund CL. Nonverbal accommodation in health care communication. Health Commun. 2014;29(6):563-573. [FREE Full text] [doi: 10.1080/10410236.2013.783773] [Medline: 24138223]
16. Mast MS. On the importance of nonverbal communication in the physician-patient interaction. Patient Educ Couns. Aug 2007;67(3):315-318. [doi: 10.1016/j.pec.2007.03.005] [Medline: 17478072]
17. Franke Föyen L, Zapel E, Lekander M, Hedman-Lagerlöf E, Lindsäter E. Artificial intelligence vs. human expert: licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship. Internet Interv. Jun 3, 2025;41:100841. [FREE Full text] [doi: 10.1016/j.invent.2025.100841] [Medline: 40525210]
18. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ. Jan 16, 2024;10:e53961. [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]

XSL•FO
RenderX

19. Deinzer R, Kiupel S, Weik U. Endocrine and psychological stress response in simulated doctor-patient interactions in medical education. Psychoneuroendocrinology. Jul 2019;105:172-177. [doi: 10.1016/j.psyneuen.2018.09.028] [Medline: 30292650]
20. Herschbach L, Festl-Wietek T, Stegemann-Philipps C, Sonanini A, Herrmann B, Erschens R, et al. Evaluation of an AI-based chatbot providing real-time feedback in communication training for mental health care professionals: proof-of-concept observational study. J Med Internet Res. Nov 28, 2025;27:e82818. [FREE Full text] [doi: 10.2196/82818] [Medline: 41314643]
21. Jacobs C, Johnson H, Brownlie K, Joiner R, Thompson T. Scaling multimodal agentic AI in medical education: multisite cross-sectional study of simulation effectiveness in primary care. JMIR Form Res. Mar 23, 2026;10:e88905. [FREE Full text] [doi: 10.2196/88905] [Medline: 41871335]

Abbreviations

AI-SP: AI-based standardized patient
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
Mini-CEX: Mini-Clinical Evaluation Exercise
SP: standardized patient

Edited by T de Azevedo Cardoso; submitted 02.Sep.2026; peer-reviewed by H Akatsu, X Liang, N Nguyen; comments to author 05.Oct.2026; revised version received 08.Oct.2026; accepted 26.Oct.2026; published 16.Nov.2026

Please cite as:
Yanagita Y, Yokokawa D, Fukuzawa F, Ihara S, Yoshida R, Sato R, Uehara T
Medical Interview Performance Using an AI-Based Standardized Patient vs a Human Standardized Patient: Randomized Crossover Trial
JMIR Med Educ 2026;12:e111030
URL: https://mededu.jmir.org/2026/1/e111030
doi: 10.2196/111030
PMID:

©Yasutaka Yanagita, Daiki Yokokawa, Fumitoshi Fukuzawa, Shiichi Ihara, Ryo Yoshida, Rurika Sato, Takanori Uehara. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX