Original Paper

Evaluating an AI-Powered Patient Simulator for Communication Skills Training: One-Year Randomized Controlled Trial

Xining Wang[1,2], PhD; Sandhya Duggal[2], PhD; Alun Hughes[2], PhD; Sayed Murad[2], MSc; Andrew O'Malley[2], PhD

[1]School of Education, City University of Macau, Macao SAR, China

[2]School of Medicine, University of St Andrews, St Andrews, Fife, Scotland, United Kingdom

Corresponding Author:
Andrew O'Malley, PhD
School of Medicine
University of St Andrews
College Gate, North Street
St Andrews, Fife, Scotland, KY16 9AJ
United Kingdom
Phone: 44 (0)1334 47 6161
Email: aso2@st-andrews.ac.uk

Abstract

Background: Clinical communication skills are a core yet complex competency in medical education and practice. However, early-stage medical students often struggle to apply communication frameworks in practice due to limited opportunities for structured simulation training. Generative AI (GenAI)–supported training has the potential to offer a scalable and innovative approach for clinical communication skills training.

Objective: This study examined the longitudinal development of medical students' clinical communication skills over 1 academic year and investigated whether AI-powered simulation contributed to differential improvement relative to conventional training. In addition, the study explored motivational, behavioral, and affective mechanisms associated with communication skills development.

Methods: A nonblinded randomized controlled design was used at a medical school in the United Kingdom, with students assigned either to the control group (n=28, 56%) or to the AI group (n=22, 44%), who practiced with SimPatient, a patient-simulator chatbot accessed through a web browser. Communication skills performance was assessed using standardized objective structured clinical examinations (OSCE) scores relative to the pass marks to enable comparison across time points. Paired *t* tests and difference-in-differences (DID) analyses were conducted between groups and over time to examine changes in communication skills. Latent change score modeling was used to identify predictors of communication skills change. Sentiment analysis was conducted on 6120 comments from AI-student conversations. Finally, a system quality survey was used to assess participants' feedback on the AI-powered simulator.

Results: Latent change score modeling showed that gains in intrinsic motivation and behavioral learning were significant predictors of students' communication skills change. The AI group showed a greater descriptive increase in communication skills than the control group, yet the between-group difference was not statistically significant. Within the AI group, consultations whose lexical tone trended more strongly toward resolution (consultation tone) were descriptively associated with communication skills improvement; given the small subsample (n=17), this finding is reported as exploratory. The AI group's confidence level in AI use was negatively associated with communication skills improvement, while the group nevertheless reported a generally high level of satisfaction with the system.

Conclusions: In this study, medical students' clinical communication skills increased over the academic year, with this development most strongly associated with their motivation and behavioral engagement; the AI-supported group did not differ significantly from conventional training. The simulator responded appropriately and provided a usable environment for practicing foundational communication behaviors but was limited in capturing the affective complexity, vulnerability, and narrative unpredictability of real clinical encounters. In conclusion, this study suggests that GenAI holds promise for supporting foundational clinical communication training and providing positive practice environments, thereby fostering motivation and behavioral engagement.

(J Med Internet Res 2026;28:e91886) doi: 10.2196/91886

XSL•FO
RenderX

KEYWORDS

generative AI; communication skills; medical education; objective structured clinical examinations; OSCE; sentiment analysis

Introduction

Background

Clinical communication is vital to medical practice. Effective clinical communication functions as a strategic, evidence-informed intervention that improves health outcomes at the individual, community, and societal levels [1,2]. In this context, effective communication refers to the observable consultation behaviors through which these outcomes are achieved—structured information gathering and history-taking; rapport and trust building; empathic and emotionally responsive engagement; and clear, shared information-giving—and it is these behaviors that constitute the core skills targeted by the training evaluated here. The importance of communication skills is reflected in medical curricula worldwide through the integration of structured frameworks such as the Calgary-Cambridge framework. Medical schools invest heavily in simulation-based communication training, standardized patients, and objective structured clinical examinations (OSCEs). However, communication is a complex, nontechnical competency that develops over time, alongside clinical reasoning. Early-stage medical students usually report difficulty in translating taught communication frameworks into real or simulated encounters, as opportunities for deliberate, structured simulation practice remain limited in many medical curricula [3,4]. Additionally, educational capacity for learning communication skills in clinical environments is finite, while student numbers continue to increase [5]. These issues highlight the need for scalable, adaptive training methods that are emotionally supportive for students and sustain their engagement while enabling repeated practice and feedback.

Generative AI (GenAI)–powered simulation presents a promising response to this challenge [6]. Recent developments in GenAI have enabled the creation of virtual standardized patients (VSPs) that simulate authentic patient dialogues, generate real-time feedback, and adapt to learners' responses [7]. AI-mediated environments can offer repeated, on-demand practice while maintaining consistency and psychological safety [8].

A growing body of empirical work has evaluated AI chatbots and VSPs for clinical communication training, with modest but mixed results. In a randomized controlled trial (n=56), Wang et al [9] found that students trained with GPT-simulated patients achieved higher structured clinical examination scores than peers trained through conventional role play (86.79 vs 73.64; $P<.001$), whereas Tyrrell et al [10], who randomized 378 students to AI- or actor-based training, found that although both modalities improved self-reported communication skills, the AI condition was inferior to human actors (mean difference −0.36; $P=.04$). Findings from nonrandomized designs point in a similar direction: Suárez-García et al [11] reported a 36.7-point increase (95% CI 31.4-42.1) in blind-rated diagnostic communication in a pre-post study (n=30), and Yamamoto et al [12] observed modestly higher OSCE interview scores among AI-trained students in a nonrandomized controlled trial (n=145; 28.1 vs 27.1; $P=.01$).

Two limitations of this evidence base stand out. First, in a systematic review of 11 studies, Gilbert et al [13] found that most research measures self-reported satisfaction or self-assessed skill rather than objective, examiner-rated performance, with only a minority using rigorous outcome measures such as OSCEs or blinded rater assessment, raising questions about the transferability of reported gains to clinical practice. Second, in our own systematic review of 16 studies of GenAI-integrated virtual agents and simulations in health professions education, we found that although such systems show potential to enhance clinical communication and learner confidence, evidence of educational effectiveness remains insufficient and long-term efficacy is largely untested [4].

Most studies of AI for communication skills training mainly focus on the performance of the AI system, the accuracy of the diagnostic conversation engine, and the feedback that AI provides for medical students' learning outcomes [8,14]. However, these studies fail to adopt a human-centered perspective, which aligns with the core concept that physicians should be able to communicate with warmth, empathy, and relational presence to build trust with patients. In medical education, the development of students' communication skills depends on reflection on their learning, practice, and experience, as well as emotional engagement with the simulated patient [15].

Beyond system performance, learner-side factors shape how communication skills develop in simulation-based training: motivation, confidence, self-efficacy, and engagement have all been linked to skill development. Across 21 studies, Nojima et al [16] identified realism, opportunities for repetition, structured debriefing, and emotionally engaging environments as key drivers of intrinsic motivation and self-efficacy, and a synthesis of 43 papers by Donovan et al [17] likewise identified realism, authenticity, and psychological safety as the mechanisms through which simulation builds self-efficacy and confidence. Confidence gains have been demonstrated among trainees practicing with virtual patients (n=258; $P<.001$), although Yu et al [18] found that human standardized patients increased communication confidence more than high-fidelity simulators or virtual reality (n=43). Theoretically, these gains have been explained through social cognitive and constructivist accounts, in which self-efficacy develops through performance accomplishment and reflective engagement [19]. Critically, however, this evidence is largely cross-sectional: longitudinal, repeated-practice designs are rare, and direct measurement of learners' affective engagement with the simulated patient is seldom reported.

This study, therefore, aims to track the trajectory of medical students' communication skills development and examine whether use of an AI-powered patient simulator can (1) improve clinical communication outcomes and (2) reveal the mechanism shaping students' communication performance in simulated

XSL•FO
RenderX

encounters. In doing this, the study addresses 4 research questions (RQs):

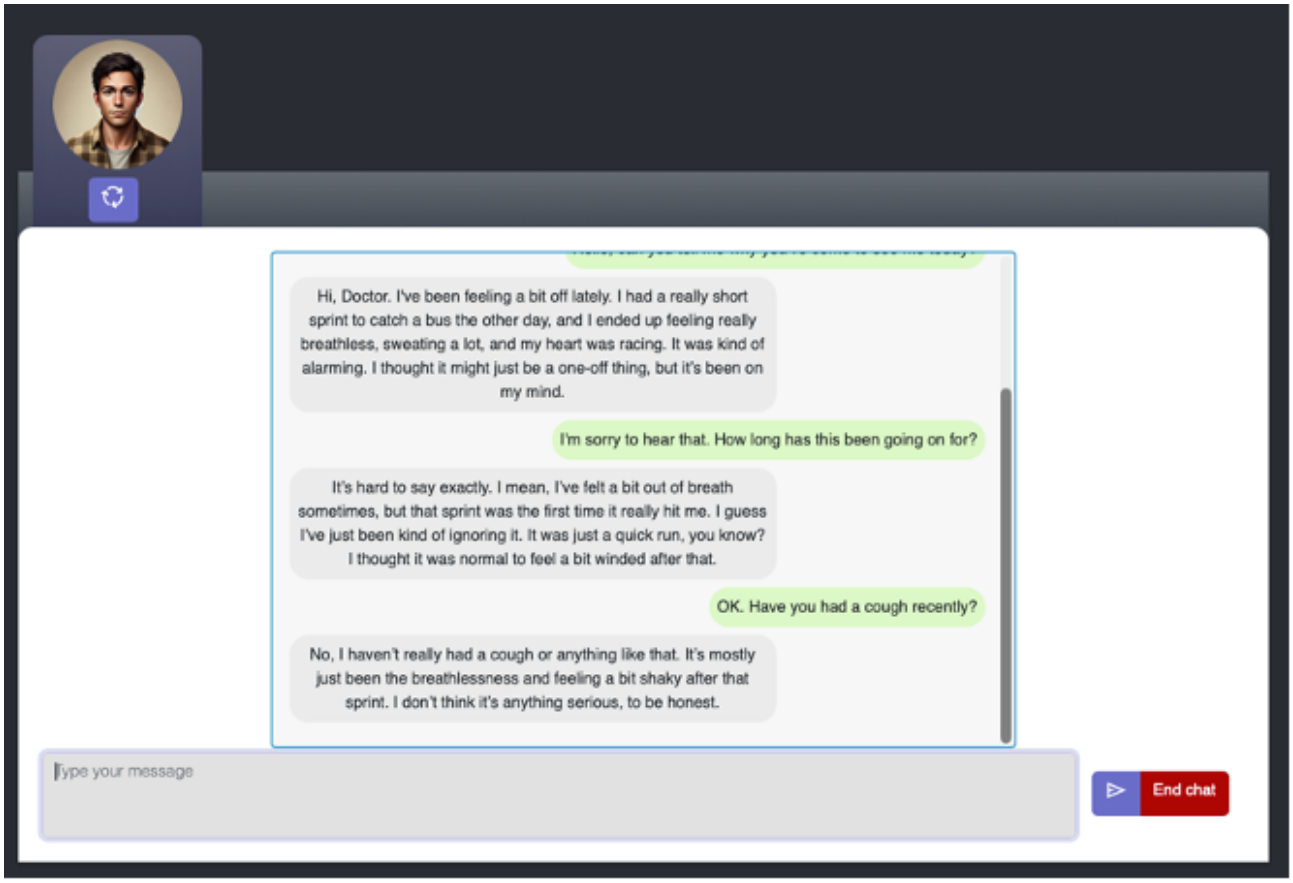
1. What is the overall trajectory of participants' clinical communication skills from year 1 to year 2 (both AI and control groups)?
2. Which factors significantly influence the trajectory of participants' clinical communication skills (both AI and control groups)?
3. How do changes in participants' communication skills differ between the AI group and the control group?
4. Within the AI group, what is the mechanism of interaction between AI, students, and communication skills development?

SimPatient

SimPatient is an AI-powered patient simulator designed to support deliberate practice of clinical communication skills [20]. Learners interact with VSPs via free-text or speech in time-limited consultations that mirror authentic primary and secondary care encounters. The system uses a fine-tuned large language model (LLM) trained on anonymized clinical dialogue to generate contextually coherent patient responses. Encounters are delivered on demand through a web interface, allowing scalable, standardized simulation with immediate formative feedback integrated into the curriculum.

The system was implemented in Python and built on the Django web framework, which provided a stable server architecture and facilitated user authentication, data handling, and interaction logging. To enable dynamic conversational flow, the platform integrated LangChain, which managed dialogue memory and contextual state and allowed model-specific training instructions to persist across user turns. During early development, 2 LLMs were evaluated for text-based interaction: OpenAI's gpt-4o-mini and Hugging Face's google/gemma-1.1-7b-it. Both models were tested for their ability to generate coherent patient responses, maintain vignette consistency, and appropriately withhold or disclose information based on conversational cues. The model used in this study was gpt-4o-mini (temperature 0.4). Prompt refinement was conducted iteratively through internal pilot testing. The final prompt structure positioned the AI as a patient role player, stating: "You, the AI, are going to play the patient...You are medically naïve and reluctant to volunteer information...Do not provide all information at once." This configuration, presented as a chatbot (Figure 1), encouraged learners to practice sequential questioning, rapport building, and information gathering. The same model, system configuration, and prompt structure were applied identically to every AI-group participant and held constant throughout the trial period, standardizing the simulated patient condition across participants and over time. Subsequent versions of the system have become multimodal (using both voice and video avatars), incorporate a feedback engine to evaluate learner performance against established consultation frameworks such as the Calgary-Cambridge framework, and include a proprietary diversity engine that systematically varies patient demographics and social context to support representative case exposure (Figure 1).

Figure 1. SimPatient chatbot screenshot.



XSL•FO
RenderX

Methods

Overview

This study used a postpositivist and quantitative analytic framework with a nonblinded randomized controlled trial design to evaluate the intervention's efficacy through objective performance metrics and use data [21]. A total of 50 early-stage medical students were recruited for participation. Of this initial cohort, 13 participants were excluded from the analyses requiring complete intervention and outcome data because of incomplete adherence to the required intervention protocols, resulting in a complete-case sample of 37. Analyses drawing on this sample therefore follow a per-protocol (complete-case) rather than an intention-to-treat approach; the resulting potential for attrition bias is addressed in the Limitations section. This study used the following four categories of data (Table 1): (1) communication skills OSCE scores, (2) questionnaires, (3) user feedback on SimPatient, and (4) SimPatient use data.

Table 1. Data types and derived study variables.

Data sources	Category	Participants, n	Timeline	Derived variables	Variable type
Communication skills OSCE[a] results	1	50 (pretest and posttest; both groups)	May 2024 to May 2025	Learning outcome	Numeric data from scales
Pretest questionnaire (22 items)	2	37 (both groups)	February 2025	AI experience; intrinsic motivation; confidence; behavioral learning; apply, evaluate, and create; and AI attitude	Numeric data from scales
Posttest questionnaire (21 items)	2	37 (pretest and posttest; both groups)	May 2025	Intrinsic motivation; confidence; behavioral learning; apply, evaluate, and create; and AI attitude	Numeric data from scales
System Feedback Survey (18 items)	3	17 (posttest, AI group only)	May 2025	User feedback on SimPatient	Numeric data from survey
AI-student conversation logs (6120 comments)	4	17 (posttest, AI group only)	May 2, 2025	Mean sentiment score, sentiment SD, and number of sessions	Characters from system transcripts
System logs	4	17 (posttest, AI group only)	May 2, 2025	Log-in time and duration from each participant	Time-series data converted to numeric format

[a]OSCE: objective structured clinical examination.

Trial Registration

This randomized educational study was not prospectively registered because it met the JMIR exemption for education-related randomized trials involving learners rather than patients, in which the intervention and outcome measures are solely education-focused and not health-related [22]. This is consistent with the JMIR Publications trial registration requirements and the International Committee of Medical Journal Editors (ICMJE) definition of clinical trials [23].

Communication Skills OSCE Scores

The OSCE is a standardized, station-based method of assessing clinical competence, in which each student rotates through a series of timed stations that simulate defined clinical tasks and is scored by a trained examiner against a predetermined marking scheme. All participants undertook OSCEs as part of their standard medical training. For this study, we used their communication skills station scores from their OSCE in year 1 (pretest) and year 2 (posttest). Because the pretest and posttest assessments were conducted on different OSCE stations with distinct grading scales, raw communication scores were standardized relative to each station's pass mark. Both the year 1 (pretest, May 2024) and year 2 (posttest, May 2025) communication stations were standard set using the borderline regression method. Station-level reliability (α) was 0.88 for the year 1 station and 0.84 for the year 2 station, and the examiner pool and assessment blueprint were comparable across the 2 years. We note, however, that standardization relative to a pass mark yields a pass mark deviation score rather than formal psychometric linking or equating; it does not adjust for residual differences in station difficulty, examiner stringency, or blueprint coverage. Cross-time OSCE comparisons are therefore interpreted as indicative rather than as directly equated measures (for more information, refer to the Limitations section).

Questionnaires

The questionnaires included 22 pretest items and 39 posttest items. First, the AI Literacy Questionnaire [24] was used to measure participants' AI-related behavioral reflection and learning motivation. For the purposes of this study, we extracted 4 relevant subscales (16 items) that aligned most closely with the scope and learning outcomes of the intervention:

- Intrinsic motivation (4 items) captures participants' internal drive and interest in engaging with AI-related learning activities (Cronbach α was 0.87 at pretest and 0.85 at posttest)
- Confidence (4 items) evaluates participants' perceived ability in mastering AI-related tasks (Cronbach α was 0.83 at pretest and 0.87 at posttest)
- Behavioral learning (4 items) assesses participants' AI-related learning behaviors. (Cronbach α was 0.85 at pretest and 0.86 at posttest)

XSL•FO
RenderX

- Apply, evaluate, and create (4 items) evaluates participants' higher-order cognitive skills in critically assessing AI solutions (Cronbach α was 0.88 at pretest and 0.82 at posttest)

Second, students' general attitudes toward AI were assessed using the attitudes toward AI scale (items=5) [25]. Cronbach α was 0.83 at pretest and 0.89 at posttest.

User Feedback and Use Data

The AI System Evaluation Survey was administered exclusively to collect feedback on the AI group's user experience with SimPatient (posttest only, Cronbach α=0.96; section 1 in Multimedia Appendix 1) [26]. SimPatient AI-student interaction logs were used to perform sentiment analysis. These data were used to quantify students' interaction patterns and affective responses during AI-supported training. Consultation resolution tone (hereafter, "consultation tone") was operationalized as the mean Valence Aware Dictionary and Sentiment Reasoner (VADER) sentiment score across each participant's AI-student exchanges. Because the AI patient presents medical complaints using negatively valenced lexical items (eg, "pain," "anxious," "worried") by task design, the VADER score is interpreted here as a marker of how strongly the exchange trends toward resolution-oriented language (eg, reassurance and acknowledgment) rather than as a direct measure of student affective engagement. VADER itself is a lexicon- and rule-based model designed to assess affective valence in short, informal text [27]. In addition, tone variability (the standard deviation of consultation tone scores) and interaction time were measured as the cumulative chatting hours between the participant and the AI system.

Ethical Considerations

This study was reviewed and approved by the institution's ethics committee (MD18121) on October 10, 2024. All participants were informed about the purpose and procedures of the study, and written informed consent was obtained prior to participation. Participation could be withdrawn at any time, and all data were anonymized prior to analysis to ensure confidentiality. After the experiment, participants received a £50 (£1=US $1.32, as of October 8, 2026) shopping voucher as compensation for their participation in the study. The experiment was conducted at a university in the United Kingdom within an undergraduate medical education program. The intervention and assessments took place in an educational simulation setting, including AI-powered communication practice and OSCE.

Procedure and Timeline

Participants completed their year 1 OSCE in May 2024. Participants were recruited to the study in February 2025, when their pretest OSCE results were collected to establish the baseline measure. In March 2025, participants received an AI tutorial introducing GenAI, which covered the basic functioning of LLMs to provide an understanding of the use of AI in medical training (section 2 in Multimedia Appendix 1). Simple randomization (coin toss) was used to assign participants to groups. Consequently, the resulting group sizes were unequal (AI group, n=22; control group, n=28) [28]. Both groups were then asked to complete the pretest questionnaires.

Over the next 3 months, the AI group engaged with SimPatient as part of their clinical communication skills education, which was supported by the lecturer's guidance note and operational manual (section 3 in Multimedia Appendix 1). The AI group was asked to complete at least 10 practice sessions; each session should be no shorter than 20 minutes. Meanwhile, the control group continued with the standard communication skills curriculum. Finally, in May 2025, the intervention concluded, and both groups completed the posttest questionnaires prior to the exams, from which the OSCE communication skills station scores were collated.

Both groups received the standard communication curriculum, including instructor-led instruction, communication practice, instructor feedback, and debriefing. The AI group additionally received access to SimPatient for AI-supported communication practice during the final 3 months of the academic year. SimPatient supplemented rather than replaced any component of the standard communication curriculum. The control group received the same standard curriculum, including instructor feedback and debriefing but did not receive additional communication practice through SimPatient.

Results

This study addressed 4 RQs using nested analytic samples drawn from the 50 randomized participants, and we designate the communication skills OSCE comparison (RQ 1 and RQ 3) as the primary analysis. RQ 1 (change in communication skills over time) and RQ 3 (differential change between arms) used the full randomized sample of 50 participants (AI group: n=22; control group: n=28), for whom OSCE scores were available as part of standard assessment. RQ 2 (latent change score modeling [LCSM]) used the complete-case sample of 37 participants with full questionnaire data, and RQ 4 (consultation tone regression) used the 17 AI-group participants for whom AI-student interaction logs were available. The 13 participants excluded for incomplete adherence were removed only from analyses requiring complete intervention or outcome data; the primary RQ 1–RQ 3 comparison retained all randomized participants. RQ 2 and RQ 4 are complete-case analyses and are interpreted as secondary and exploratory. The flow of participants through recruitment, allocation, exclusion, and analysis is shown in Figure 2. Baseline information for the AI and control groups is summarized in Table 2.

XSL•FO
RenderX

Figure 2. Experimental flow diagram. OSCE: objective structured clinical examination.

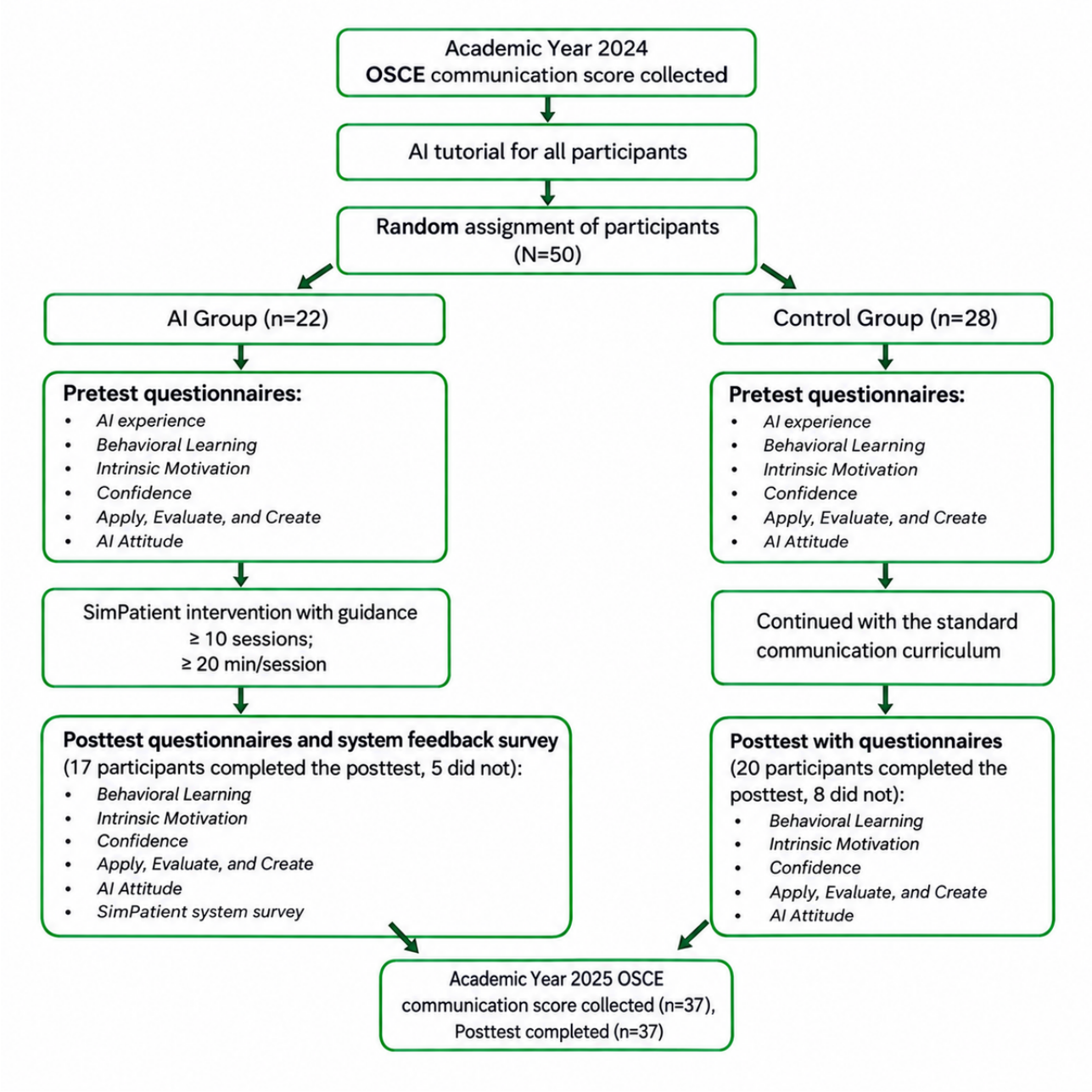


Table 2. Baseline characteristics of participants (N=50).

Baseline characteristics	AI group (n=22)	Control group (n=28)
Communication skills (pretest), mean (SD)	117.95 (24.86)	115.07 (21.42)
Previous AI experience, n (%)		
Never used AI	3 (13.6)	4 (14.3)
Familiar with AI but no use	5 (22.7)	8 (28.6)
Occasionally used AI	10 (45.5)	9 (32.1)
Frequently used AI	4 (18.2)	7 (25)

To address RQ 1, a paired-samples 2-tailed *t* test was conducted to examine changes in clinical communication skills performance between the pretest and posttest among all participants (N=50). Results showed that students' posttest scores (mean 124.03, SD 22.20) were higher than their pretest scores (mean 116.10, SD 22.77), with an average increase of 7.93 points. This difference was not statistically significant (t_{49}=1.77; *P*=.081; Cohen *d*=0.25). Although the mean gain did not reach significance, the pattern suggested general positive development in students' communication skills performance

XSL•FO
RenderX

over the academic year, although the statistical results were nonsignificant.

To answer RQ 2, which aimed to identify which factors contributed to changes in students' communication skills (n=37), an LCSM was estimated in R (version 4.4.2; The R Project for Statistical Computing) using the *lavaan* package [29]. The model estimated 13 free parameters. The predictive factors included AI experience and changes in *apply, evaluate, and create*, intrinsic motivation, confidence, behavioral learning, and AI attitude. The model showed an adequate fit to the data (χ^2_5=6.7; *P*=.25; comparative fit index=0.97), indicating that the proposed latent change structure was well supported. As shown in Table 3, both changes in *intrinsic motivation* (β=2.847; *P*=.03) and *behavioral learning* (β=2.849; *P*=.001) significantly predicted the latent slope, suggesting that students who reported greater increases in intrinsic motivation and behavioral engagement achieved greater improvement in communication skills. To address RQ 3, a difference-in-differences (DID) analysis was conducted to examine whether students' standardized communication skills scores increased differentially between the AI group (n=22) and the control group (n=28) from pretest to posttest. Descriptive results indicated that the control group increased by 6.11 points from pretest (mean 115.07, SD 21.42) to posttest (mean 121.18, SD 20.78), whereas the AI group increased by 10.96 points from pretest (mean 117.95, SD 24.86) to posttest (mean 128.91, SD 23.67; section 4 in Multimedia Appendix 1). The DID model revealed that the interaction effect between group and time (time×group) was positive but not statistically significant (β=4.42, 95% CI −13.61 to 22.44; *P*=.63). This suggested that although both groups improved over time, the additional gain in the AI group compared with the control group was not statistically significant.

Table 3. Latent change score model results and model fit.

Factors	β	SE	*z* score	*P* value[a]	Standardized β
AI experience	2.664	2.20	1.21	.23	0.36
Behavioral learning	2.849	0.863	3.32	*.001* [b]	0.78
Intrinsic motivation	2.847	1.323	2.17	*.03*	0.71
Confidence	–3.272	1.962	–1.67	.095	–0.48
Apply, evaluate, and create	–1.273	1.021	–1.24	.21	–0.35
AI attitude	–0.461	0.663	–0.70	.48	–0.21

[a]*P* values are calculated based on 2-tailed Wald z tests of the regression coefficients.

[b]*P* values <.05 (statistically significant).

To answer RQ 4, VADER sentiment analysis was conducted via Python (version 3.12.0; Python Software Foundation) on 6120 AI-student interaction comments generated by the AI group (n=17). Results showed that 54.2% of the comments expressed positive sentiment, 8.2% were neutral, and 37.6% were negative, suggesting generally positive engagement with the AI system (section 5 in Multimedia Appendix 1).

Next, mean sentiment and sentiment variability (ie, SD) were then computed for each participant and used with other related variables in an exploratory hierarchical regression analysis to examine how affective tone predicted communication skills change; this produced 3 models (Table 4).

In model 1 (the AI use layer), participants' prior AI experience and hours spent conversing with SimPatient did not significantly predict communication change ($F_{2,14}$=0.23; *P*=.79). Note that these hierarchical models examine variation within the AI group only (n=17); they do not compare the AI and control groups, which are compared in the DID analysis above.

In model 2, consultation tone and tone variability were added. This change significantly improved model fit ($F_{2,12}$=5.90; *P*=.03). Consultation tone emerged as a positive predictor (β=272.18; *P*=.01), indicating that consultations whose overall lexical tone trended more strongly toward resolution were associated with greater communication skills change. As discussed in the Limitations section, this association is interpreted as descriptive of consultation outcome rather than as evidence that affective engagement caused skill gains.

In model 3, 5 psychological variables were added to the AI use and consultation tone variables in an exploratory hierarchical regression (Table 4). Given the small sample size (n=17) relative to the 9 predictors (residual *df*=7), the model was considered underpowered for confirmatory inference. The overall model was not statistically significant ($F_{9,7}$=2.31; *P*=.14), although it explained 74.8% of the variance (adjusted R^2=0.42). To characterize the uncertainty of the coefficient estimates, bias-corrected and accelerated 95% CIs based on 10,000 bootstrap resamples were examined alongside the regression coefficients, including those for consultation tone (β=447.56, 95% CI −788.10 to 2890.40), tone variability (β=−368.40, 95% CI −3660.30 to 1323.30), confidence change (β=−22.99, 95% CI −84.96 to 232.26), and behavioral learning change (β=−8.93, 95% CI −47.54 to 63.71). Model 3 coefficients are interpreted as exploratory and hypothesis-generating rather than confirmatory evidence.

System feedback results showed that the AI group valued SimPatient as a usable, informative, and supportive AI tool for clinical communication training.

XSL•FO
RenderX

Table 4. Hierarchical regression results.

Predictors	Model 1[a], β (*P* value)	Model 2 [a], β (*P* value)	Model 3 [a], β (*P* value)	Bias-corrected and accelerated 95% CI[b]
AI use layer				
Intercept	22.35 (.53)	97.84 (.17)	194.15 (.04)	—[c]
AI experience	0.49 (.95)	3.71 (.56)	–3.73 (.61)	−59.42 to 237.06
Chat hours	–3.21 (.55)	–3.21 (.48)	1.49 (.77)	−79.50 to 43.91
Consultation tone layer				
Consultation tone (mean)	—	272.18 *(.01)*[d]	447.56 *(.02)*	−788.10 to 2890.40
Tone variability (SD)	—	–216.79 (.13)	–368.40 (.06)	−3660.30 to 1323.30
Psychological layer				
Confidence change	—	—	–22.99 *(.04)*	−84.96 to 232.26
Intrinsic motivation change	—	—	4.56 (.17)	−34.60 to 42.83
Behavioral learning change	—	—	–8.93 (.08)	−47.54 to 63.71
Apply, evaluate, and create change	—	—	–0.27 (.92)	−17.00 to 57.15
AI attitude change	—	—	0.64 (.69)	−9.01 to 19.54

[a]The models explained 3.0%, 46.0%, and 74.8% of the variance in communication skills change, respectively (model 1: R^2=0.03, adjusted R^2=−0.11, *P*=.79; model 2: R^2=0.46, adjusted R^2=0.28, *P*=.03; model 3: R^2=0.75, adjusted R^2=0.42, *P*=.14).

[b]BCa 95% CIs are for the regression coefficients in model 3.

[c]Not applicable.

[d]*P* values <.05 (statistically significant).

Discussion

This study examined how medical students' clinical communication skills developed over 1 academic year and assessed whether an AI-powered patient simulator influenced this development.

Student Trajectory of Communication Skills (RQ 1)

Results showed that students improved their communication skills OSCE scores relative to the standard-setting pass mark from year 1 to year 2, but this improvement did not achieve statistical significance. Because the year 1 and year 2 OSCE stations were not formally equated, this cross-time comparison should be interpreted as indicative of the direction of change rather than as a directly comparable measure. Students' communication skills showed a small-to-moderate mean increase of 7.9 (SD 31.6) points but a highly heterogeneous trend. These findings demonstrate that the cohort's exam performance was relatively stable relative to a standard-setting pass mark, as the standard-setting pass mark is not an objective and consistent measure of communication skills.

Wider contextual factors should also be considered. Students' communication skills may have improved in absolute terms, as the year 2 OSCE was more difficult and complex than the year 1 OSCE. Specifically, in the year 2 OSCE, examiners evaluate sophisticated communication skills, including empathy, active listening, and patient-centered reasoning. These exams often involve demanding and nuanced clinical contexts, which often require the integration of both medical knowledge and interpersonal competence. These findings align with the literature, which suggests that medical students' communication skills show modest and inconsistent improvements during medical training, with significant variability across different studies and skill dimensions. Research has found that interventions had positive but small effects on communication skills, with evidence quality ranging from moderate to very low [30].

Critical Factors Associated With Changes in Communication Skills (RQ 2 and 3)

Results showed that the upward trajectory of students' communication skills was significantly shaped by intrinsic motivation and behavioral learning. Students who demonstrated stronger motivational growth and behavioral involvement were more likely to demonstrate greater increases in their communication skills. It may be that students who scored highly on these psychometric traits were genuinely interested in communicating with patients, kept practicing related skills, and had greater emotional presence during the conversation. This aligns with self-determination theory [31], which emphasizes that autonomous motivation enhances persistence, reflection, and skill acquisition. This interpretation is offered tentatively, however, because intrinsic motivation and behavioral learning were operationalized using the AI Literacy Questionnaire, which was designed to capture motivation and engagement toward AI technology rather than clinical communication practice; the applicability of a self-determination account to the communication learning domain is therefore assumed rather than demonstrated. Additionally, although the AI group's communication skills improved more than those of the control group in absolute terms, this between-group difference was not statistically significant and therefore cannot be attributed to the

XSL•FO
RenderX

intervention. We accordingly do not interpret it as evidence that SimPatient outperformed usual training. SimPatient may nonetheless offer a useful means of practicing fundamental communication behaviors, such as building the physician-patient relationship, information gathering, and information giving, but this possibility requires confirmation in an adequately powered trial. However, it cannot replace human simulated patients, particularly in real-time scenarios that offer nuanced emotional expression, interpersonal attunement, and unpredictable patient responses [32].

The Consultation Tone Is Descriptively Associated With Communication Skills Development (RQ 4)

The version of SimPatient used in this study was a first-generation design, situated at an early-intermediate level for clinical communication training. Although SimPatient can respond appropriately to user input, it lacks divergent thinking to generate emotionally salient cues and the scenario shifting capacity required for generating complex and evolving patient narratives. Participants' feedback via system quality surveys echoed this AI generative characteristic. Students reported moderately positive feedback on system quality, information quality, service quality, user satisfaction, and individual impact (mean 7.6/10), suggesting a generally positive, albeit not perfect, user experience. Without emotionally salient triggers or adaptive scenario shifts, participants had fewer opportunities to engage in deeper inquiry, manage clinical ambiguity, or practice emotionally responsive communication. This may also explain why the AI group did not present a significant differential increase in communication skills compared to the control group.

A descriptive pattern within the AI group was that consultations whose lexical tone trended more strongly toward resolution were associated with greater increases in communication skills. We interpret consultation tone here as a downstream marker of how the consultation unfolded rather than as a measure of student affective engagement: VADER is calibrated on social media text rather than clinical dialogue, the AI patient presents medical complaints using negatively valenced lexical items by task design, and a student who communicates competently will tend to elicit more resolution-oriented AI responses (eg, "that's helpful, thank you, doctor"), which may inflate the VADER score independent of the student's own affective state. Specifically, sentiment analysis showed that 54.2% of the 6120 comments from user-AI conversations were classified as positive in lexical valence. In addition, greater communication skills improvements were associated with decreased confidence in AI use (sections 5 and 6 in Multimedia Appendix 1).

These observations align with the behavioral characteristics of our AI system: SimPatient does not provide narratives that would prompt deeper clinical reasoning or uncertainty management; therefore, SimPatient may limit opportunities for more challenging clinical dialogue. This level of human-AI interaction involved low cognitive demand and few diagnostic challenges, which may have allowed students to experience easy practice and early success, co-occurring with a resolution-oriented consultation tone and declining self-rated confidence in AI use. Despite students' decreasing confidence in AI use, it should be noted that they did in fact demonstrate greater competence (ie, gains in communication skills). Students who demonstrate gains in communication skills and have positive interactions with patients while also recognizing the limits of their clinical abilities (ie, low confidence in AI use) may represent a desirable outcome in medical education and reflect current academic interest in competency-based medical education [33].

Limitations and Future Work

The limitations of sentiment analysis in this context mean that these results should be interpreted cautiously. Sentiment analysis is a measurement of conversation sentiment that reflects both AI (patient) output and user (student) input. Conversation sentiment therefore is fundamentally influenced by both the fact that a patient is presenting to a physician with a negative "medical complaint" (eg, "I feel anxious," "my stomach hurts," and "I am worried") and by positive, polite, or warm interactions as the consultation progresses toward resolution (eg, "That makes me feel better" and "Thank you, doctor"). Indeed, where positive sentiment was detected, it was usually attributable to AI responses rather than user questions (section 5 in Multimedia Appendix 1). In this case, the high proportion of positive comments therefore does not reflect a rich emotional exchange but rather the structure of the consultation. However, the structure of the consultation within the AI system could still influence the acquisition of communication skills. Indeed, our findings of students' better learning outcomes (ie, greater gains in communication skills) implied that students can elicit more positive responses from the system to continue the conversation. Future studies may further investigate how sentiment engagement and cognitive challenge jointly shape learning outcomes in AI-powered clinical communication training, particularly the balance between skill acquisition and learners' confidence when using AI systems.

In future studies, a larger sample size would increase the likelihood of detecting statistically significant results; similarly, more participants would also likely reduce the unequal group sizes resulting from simple randomization and mitigate against participants' lack of completion, which is common in educational studies such as this.

Relatedly, the 13 participants who did not adhere to the intervention protocol were excluded from the analyses requiring complete data, so the primary analyses are per-protocol (complete-case) rather than intention-to-treat. Because posttest outcome data were not available for these participants, a full intention-to-treat analysis with imputation was not feasible at this sample size. We therefore cannot exclude the possibility that participants who completed the intervention differed systematically from those who did not, and the between-group findings should be interpreted with this potential for attrition bias in mind. Future trials should prespecify an intention-to-treat primary analysis alongside a complete-case sensitivity analysis.

We also did not conduct a formal, expert-rated assessment of the simulator's fidelity, such as blinded expert ratings of patient response authenticity and scenario adherence. Although the intervention was standardized at the level of the system specification and participants rated the simulation positively on perceived quality (mean 7.6/10), formal fidelity evaluation

XSL•FO
RenderX

would provide stronger assurance that SimPatient reproduced authentic clinical encounters, and we identify this as a priority for future work.

A further limitation concerns construct validity. The motivation and engagement constructs used in the hierarchical models were operationalized with the AI Literacy Questionnaire [24], an instrument developed to measure engagement with AI technology rather than clinical communication practice. Interpretations of these measures in terms of communication learning motivation, including those framed through self-determination theory, should therefore be read as tentative, as the transferability of AI literacy subscales to the clinical communication domain has not been established. Dedicated validation of these constructs in the communication learning context is warranted in future work.

Conclusions

This study presents an empirical examination of how an AI-powered patient simulator relates to medical students' clinical communication development across 1 year. The findings show that changes in AI-related intrinsic motivation and behavioral learning were significantly associated with changes in communication skills. Within the AI-group subsample, greater skill gains were descriptively associated with consultations whose lexical tone trended more strongly toward resolution; this exploratory pattern requires replication before any causal interpretation. This AI system presented an early form of conversational intelligence: capable of responding appropriately but limited in its ability to enact the full complexity of human affect, vulnerability, and narrative unpredictability that characterize real clinical encounters. The AI group's confidence change was negatively associated with communication development. Hence, AI-supported communication training offered both opportunities and limitations. Careful instructional and system design are required to ensure that the training develops not only fluency but also reflective and emotionally responsive communication skills. In conclusion, this study suggests that GenAI holds promise to support foundational clinical communication training and may provide opportunities for learners to engage with AI-related motivational and behavioral factors.

Acknowledgments

Generative AI (Claude Opus 5; Anthropic PBC) was used only to proofread the final manuscript.

Funding

This research was funded by a charitable donation made to the University of St Andrews School of Medicine by the Owen Silver Legacy. AH provided project administration,

Data Availability

Data are available from the University of St Andrews data repository.

Authors' Contributions

AO led the project and conceived the SimPatient system. SM led the technical development of SimPatient. XW led data collection, data analysis, and manuscript preparation. All other authors contributed equally to other aspects of the study.

Conflicts of Interest

None declared.

Editorial Notice

This randomized study was not registered, explained by authors as follows: educational intervention, meet JMIR exception. The editor granted an exception from ICMJE rules mandating prospective registration of randomized trials, because the risk of bias appears low. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Multimedia Appendix 1

Difference-in-differences analysis of communication scores, sentiment and lexical analyses, survey of satisfaction and usability, and educational materials.
[DOCX File , 7024 KB-Multimedia Appendix 1]

Multimedia Appendix 2

CONSORT 2025 checklist.
[PDF File (Adobe PDF File), 201 KB-Multimedia Appendix 2]

References

XSL•FO
RenderX

1. Guidi C, Traversa C. Empathy in patient care: from 'clinical empathy' to 'empathic concern'. Med Health Care Philos. Dec 2021;24(4):573-585. [FREE Full text] [doi: 10.1007/s11019-021-10033-4] [Medline: 34196934]
2. Dahm MR, Cattanach W, Williams M, Basseal JM, Gleason K, Crock C. Communication of diagnostic uncertainty in primary care and its impact on patient experience: an integrative systematic review. J Gen Intern Med. Feb 2023;38(3):738-754. [FREE Full text] [doi: 10.1007/s11606-022-07768-y] [Medline: 36127538]
3. McCarrick CA, Moynihan A, McEntee PD, Boland PA, Donnelly S, Heneghan H, et al. Impact of simulation training on communication skills and informed consent practices in medical students- a randomised controlled trial. BMC Med Educ. Jul 18, 2025;25(1):1078. [FREE Full text] [doi: 10.1186/s12909-025-07671-0] [Medline: 40682099]
4. Wang X, O'Malley A, Hughes A, Khalid MS. Generative AI-integrated virtual agents and simulations in health professions education: a systematic review. Educ Sci. Jun 18, 2026;16(6):973. [doi: 10.3390/educsci16060973]
5. O'Malley A. Reflections on confronting a capacity challenge with an AI-powered patient simulator (SimPatient). Simul Healthc. Jun 01, 2026;21(3):209-210. [doi: 10.1097/SIH.0000000000000890] [Medline: 41056031]
6. Li J, Yin K, Wang Y, Jiang X, Chen D. Effectiveness of generative artificial intelligence-based teaching versus traditional teaching methods in medical education: a meta-analysis of randomized controlled trials. BMC Med Educ. Aug 19, 2025;25(1):1175. [FREE Full text] [doi: 10.1186/s12909-025-07750-2] [Medline: 40830476]
7. 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. Nov 08, 2022:1-7. [doi: 10.1080/0142159X.2022.2130216] [Medline: 36346810]
8. Chance EA. The combined impact of AI and VR on interdisciplinary learning and patient safety in healthcare education: a narrative review. BMC Med Educ. Jul 11, 2025;25(1):1039. [FREE Full text] [doi: 10.1186/s12909-025-07589-7] [Medline: 40646457]
9. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC. Feasibility study of using GPT for history-taking training in medical education: a randomized clinical trial. BMC Med Educ. Jul 10, 2025;25(1):1030. [FREE Full text] [doi: 10.1186/s12909-025-07614-9] [Medline: 40640776]
10. Tyrrell EG, Sandhu SK, Berry K, Ghannam SF, Lewis SA, Crowfoot D, et al. Web-based AI-driven virtual patient simulator versus actor-based simulation for teaching consultation skills: multicenter randomized crossover study. JMIR Form Res. Nov 20, 2025;9:e71667. [FREE Full text] [doi: 10.2196/71667] [Medline: 41264856]
11. Suárez-García RX, Chavez-Castañeda Q, Orrico-Pérez R, Valencia-Marin S, Castañeda-Ramírez AE, Quiñones-Lara E, et al. DIALOGUE: a generative AI-based pre-post simulation study to enhance diagnostic communication in medical students through virtual type 2 diabetes scenarios. Eur J Investig Health Psychol Educ. Aug 07, 2025;15(8):152. [FREE Full text] [doi: 10.3390/ejihpe15080152] [Medline: 40863274]
12. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ. Sep 23, 2024;10:e58753. [FREE Full text] [doi: 10.2196/58753] [Medline: 39312284]
13. Gilbert V, Philip P, Llorca PM, Samalin L. Use of artificial intelligence-enhanced virtual patients in educational approaches to medical interview training: a systematic review. BMC Med Educ. Mar 04, 2026;26(1):588. [FREE Full text] [doi: 10.1186/s12909-026-08804-9] [Medline: 41781962]
14. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res. Jun 19, 2023;25:e43311. [FREE Full text] [doi: 10.2196/43311] [Medline: 37335593]
15. Antila AK, Lindblom S, Louhiala P, Pyörälä E. Creating a safe space: medical students' perspectives on using actor simulations for learning communication skills. BMC Med Educ. Oct 28, 2024;24(1):1225. [FREE Full text] [doi: 10.1186/s12909-024-06184-6] [Medline: 39468604]
16. Nojima K, Yoshida R, Sato S, Fukuda T, Kakinuma K. Simulation-based education and nursing student learning motivations: a scoping review. Cureus. May 18, 2025;17(5):e84375. [doi: 10.7759/cureus.84375] [Medline: 40539177]
17. Donovan KO, Connolly M, Browne F. Within healthcare education how do simulated patients enable (or not) the transfer of learning into clinical practice: a realist review. Nurse Educ Today. Mar 2026;158:106927. [doi: 10.1016/j.nedt.2025.106927] [Medline: 41317425]
18. Yu J, Lee S, Kim M, Lee J, Jung YJ. Evaluating medical students' engagement and confidence across three simulation-based education methods: standardized patient, high fidelity simulator, and virtual reality. BMC Med Educ. Jan 23, 2026;26(1):283. [FREE Full text] [doi: 10.1186/s12909-026-08634-9] [Medline: 41578259]
19. Andrade R. An exploration of virtual standardized patients and their effect on clinical readiness in pharmacy education. Curr Pharm Teach Learn. May 2026;18(5):102599. [FREE Full text] [doi: 10.1016/j.cptl.2026.102599] [Medline: 41713008]
20. SimPatient. URL: https://simpatient.co.uk/ [accessed 2026-10-08]
21. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA. SAGE Publications; 2017.
22. RCT and trial registration requirements. JMIR Publications. URL: https://support.jmir.org/hc/en-us/articles/115001389307-RCT-and-Trial-Registration-Requirements [accessed 2026-09-02]

XSL•FO
RenderX

23. L. Clinical trials. International Committee of Medical Journal Editors. URL: https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html [accessed 2026-09-02]
24. Ng DT, Wu W, Leung JK, Chiu TK, Chu SK. Design and validation of the AI literacy questionnaire: the affective, behavioural, cognitive and ethical approach. Br J Educ Technol. Dec 13, 2023;55(3):1082-1104. [FREE Full text] [doi: 10.1111/bjet.13411]
25. Grassini S. Development and validation of the AI Attitude Scale (AIAS-4): a brief measure of general attitude toward artificial intelligence. Front Psychol. Jul 24, 2023;14:1191628. [FREE Full text] [doi: 10.3389/fpsyg.2023.1191628] [Medline: 37554139]
26. Delone WH, McLean ER. The DeLone and McLean model of information systems success: a ten-year update. J Manag Inf Syst. Dec 23, 2014;19(4):9-30. [doi: 10.1080/07421222.2003.11045748]
27. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media. May 16, 2014;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
28. Suresh K. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. J Hum Reprod Sci. Jan 2011;4(1):8-11. [FREE Full text] [doi: 10.4103/0974-1208.82352] [Medline: 21772732]
29. Rosseel Y. lavaan: an R package for structural equation modeling. J Stat Softw. 2012;48(2):1-36. [doi: 10.18637/jss.v048.i02]
30. Gilligan C, Powell M, Lynagh MC, Ward BM, Lonsdale C, Harvey P, et al. Interventions for improving medical students' interpersonal communication in medical consultations. Cochrane Database Syst Rev. Feb 08, 2021;2(2):CD012418. [FREE Full text] [doi: 10.1002/14651858.CD012418.pub2] [Medline: 33559127]
31. Williams GC, Saizow RB, Ryan RM. The importance of self-determination theory for medical education. Acad Med. Sep 1999;74(9):992-995. [doi: 10.1097/00001888-199909000-00010] [Medline: 10498090]
32. Jadoon M, Naushad K, Aman F, Khan Y, Durrani M, Ali S. Virtual patients for communication skills training: a mixed methods evaluation. Cureus. Aug 26, 2025;17(8):e91005. [doi: 10.7759/cureus.91005] [Medline: 41018381]
33. Gottlieb M, Chan TM, Zaver F, Ellaway R. Confidence-competence alignment and the role of self-confidence in medical education: a conceptual review. Med Educ. Jan 2022;56(1):37-47. [doi: 10.1111/medu.14592] [Medline: 34176144]

Abbreviations

DID: difference-in-differences
GenAI: generative AI
ICMJE: International Committee of Medical Journal Editors
LCSM: latent change score model
LLM: large language model
OSCE: objective structured clinical examination
RQ: research question
VADER: Valence Aware Dictionary and Sentiment Reasoner
VSP: virtual standardized patient

Edited by I Steenstra; submitted 21.Jan.2026; peer-reviewed by B Rubinstein, M Dehnert, C Thompson; comments to author 13.May.2026; revised version received 15.Sep.2026; accepted 24.Sep.2026; published 05.Nov.2026

Please cite as:
Wang X, Duggal S, Hughes A, Murad S, O'Malley A
Evaluating an AI-Powered Patient Simulator for Communication Skills Training: One-Year Randomized Controlled Trial
J Med Internet Res 2026;28:e91886
URL: https://www.jmir.org/2026/1/e91886
doi: 10.2196/91886
PMID:

©Xining Wang, Sandhya Duggal, Alun Hughes, Sayed Murad, Andrew O'Malley. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.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 the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX