

Feasibility of an AI-Simulated Patient Encounter Delivered in Group and Individual Learning Formats for First-Year Medical Students: A Two-Cohort Pilot Study

Shelley Tischkau

Southern Illinois University School of Medicine

James Lynch

Southern Illinois University School of Medicine

Sean McGinity

Southern Illinois University School of Medicine

Andrew Primous

Southern Illinois University School of Medicine

Elijah Meldrim

Southern Illinois University School of Medicine

Samantha Lynch

Southern Illinois University School of Medicine

Richard Selinfreund

`rselinfreund39@siu.edu`

Southern Illinois University School of Medicine

Research Article

Keywords: artificial intelligence, large language model, simulated patient, standardized patient, undergraduate medical education, problem-based learning, psychological safety, feasibility study, instrument development

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10627241/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. This goes under Competing interests in your Declarations section – BMC requires it there as a standalone heading, separate from the Intellectual Property discussion already in your text. Competing interests The AI-simulated patient platform evaluated in this study was developed at Southern Illinois University School of Medicine and served as the educational intervention under investigation. One author (R.S.) is a named inventor on a pending patent application covering elements of the underlying AI-simulation technology. The patent application is held by the Board of Trustees of Southern Illinois University. No author has received royalties, licensing income, or any other financial distribution related to this technology, and no author holds equity in any entity associated with it. To limit the potential influence of this intellectual-property interest on study conduct and interpretation, all statistical analyses and data tables were prepared independently by co-author S.M. from the original study dataset, and platform-derived transcript measures are reported descriptively and are not used to support any inferential claim regarding educational effectiveness. The remaining authors (S.T., J.L., S.M., A.P., E.M., S.L.) declare that they have no competing interests.

Abstract

Background

Generative artificial intelligence (AI) and large language model (LLM)–based simulated patients are being evaluated for problem-based learning and case-based learning in undergraduate medical education, but delivery formats are often under-specified and outcome measurement is inconsistently reported. We assessed the feasibility and acceptability of a single LLM-driven, AI-simulated patient encounter delivered in two formats—a facilitated whole-group session and a self-paced individual session—to first-year medical students (MS1), and we describe a provisional behavior-anchored survey instrument and an automated transcript-analysis pipeline intended to support future validation.

Methods

Two non-overlapping, non-randomized MS1 cohorts at one United States medical school completed the same case (Group, $n = 7$; Individual, $n = 12$) and the Behavioral Safety Language Scale–Provisional (BSLS-P) before and after the encounter. Pre-to-post change was tested with paired t or Wilcoxon signed-rank exact tests, with Benjamini–Hochberg false discovery rate (FDR) correction across subscale comparisons. An automated transcript-analysis framework intended to support future validation is described.

Results

Both cohorts showed directionally positive pre-to-post shifts in engagement and psychological safety. The largest effect—a psychological-safety gain in the Individual cohort (Wilcoxon $r = 0.62$)—was significant before correction ($P = .026$) but not after (adjusted $P = .20$). No between-cohort comparison reached significance. A single motivation item declined by one median point in both cohorts. The modal post-encounter preference was a hybrid “solo first, then group” model (7/12, 58%).

Conclusions

The encounter was feasible and acceptable in both delivery formats. Observed pre-to-post shifts are exploratory and require confirmation in an adequately powered, randomized sample. The provisional instrument and the analysis pipeline are described to support that work.

Background

Generative artificial intelligence (AI), particularly large language models (LLMs), is rapidly transforming undergraduate medical education through applications in problem-based learning (PBL), case-based learning (CBL), and AI-simulated patient encounters. Recent systematic and scoping reviews

consistently report improvements in learner satisfaction and knowledge acquisition while emphasizing that the current evidence base remains methodologically heterogeneous, with inconsistent outcome measures, variable study quality, and limited evaluation of instructional design. [1–4] Collectively, these reviews identify two persistent gaps addressed by the present study: instructional delivery format has rarely been examined as an independent educational variable, and AI-supported educational assessment methods have seldom been described with sufficient transparency to facilitate independent replication. [3, 4]

AI-simulated patients represent a distinct educational modality that differs from traditional human standardized patients by providing scalable, conversational clinical encounters that learners can access repeatedly with consistent case presentation. Throughout this manuscript, the term **AI-simulated patient** refers to an AI-driven virtual patient rather than a human standardized patient. Previous investigations have shown that learners view AI-simulated patients as valuable educational supplements while emphasizing that they complement rather than replace human standardized patient experiences. [5, 6] Our previous institutional work likewise demonstrated the successful implementation and learner acceptance of AI-assisted history-taking simulations, providing the foundation for the present investigation. [8]

In addition to evaluating AI itself, instructional delivery format may independently influence learner engagement, learner perceptions of a **safe learning environment**, and the overall educational experience. Recent studies suggest that students value self-paced, technology-supported learning and that conversational AI may enhance learner-centered educational experiences. [9–11] These observations support examining instructional format as a separate educational variable rather than focusing exclusively on the underlying technology.

Accordingly, this investigation was designed as a feasibility and instrument-development pilot rather than a comparative efficacy study. The objectives were to determine whether an AI-simulated patient encounter could be successfully implemented in facilitated group and self-paced individual learning formats, to explore learner engagement and the **Psychological Safety** domain of the Behavioral Safety Language Scale–Provisional (BSLS-P), and to describe a reproducible automated transcript-analysis framework intended to support future validation. Because the study enrolled small, non-randomized cohorts, all quantitative findings are interpreted as exploratory and hypothesis-generating rather than confirmatory.

Methods

Study Design and Setting

This study was designed as a two-cohort feasibility and instrument-development pilot conducted at Southern Illinois University School of Medicine during the first-year (MS1) gastrointestinal pathology curriculum. Two naturally occurring, non-overlapping student cohorts participated in identical AI-simulated patient encounters delivered in different instructional formats. One cohort participated in a

faculty-facilitated group session, whereas the second cohort completed the encounter individually in a self-paced format. Because instructional delivery format was confounded with cohort membership and several study-design variables, comparisons between cohorts were considered descriptive and exploratory rather than confirmatory.

The primary objectives were to evaluate the practical implementation of AI-simulated patient encounters across two instructional delivery formats, to explore changes in learner engagement and the **Psychological Safety** domain of the Behavioral Safety Language Scale–Provisional (BSLS-P), and to describe an automated transcript-analysis framework intended to support future validation. Consistent with the exploratory nature of this feasibility pilot, the study was not powered to detect between-group differences, and no claims of comparative educational effectiveness are made.

The facilitated group session was conducted on April 22, 2026, and the self-paced individual sessions on May 6, 2026. In the group format, seven students shared a single encounter in a classroom, with the AI-simulated patient projected on an 85-inch display with external speakers; peer observation was therefore present throughout, and the session lasted approximately 35 minutes. In the individual format, twelve students each completed one encounter on a personal device through an individualized link after relocating to a private space, without peer observation; encounters lasted a median of 31 minutes (range 17–83), with most between 30 and 40 minutes. Both cohorts were recruited through a class email circulated one week in advance, supplemented in the group cohort by a verbal invitation at a resource session and a reminder email, and in the individual cohort by an enrollment link circulated by the class president.

The group cohort completed version 1 of the Behavioral Safety Language Scale–Provisional (BSLS-P; items 1–10, subscales A–C) on platform version 1.1.1 and provided qualitative data through a verbal group debriefing held immediately after the session (approximately 10 statements). The individual cohort completed version 2 of the BSLS-P (items 1–12, subscales A–D plus a single-select format-preference item) on platform version 1.1.2 and provided qualitative data through private written responses to three optional prompts. Paired pre- and post-encounter BSLS-P responses were the quantitative data source in both cohorts. Platform-derived performance data were available only for the individual cohort, because the two-agent automated transcript-analysis pipeline (Behavioral, Free-Text Evaluation, Action, and Time components) was not deployed in platform version 1.1.1.

Both cohorts completed the same clinical case on the same platform. The cohorts differ simultaneously on several of the dimensions described above, and delivery format is therefore fully confounded with instrument version, session date, recruitment procedure, debriefing modality, setting, peer observation, sample size, and platform version. Between-cohort contrasts are accordingly reported as descriptive only.

Participants and Recruitment

Participants were first-year medical students enrolled in the gastrointestinal pathology curriculum at Southern Illinois University School of Medicine. Participation was voluntary and did not affect course grades or academic standing. Recruitment occurred through a general invitation extended to the entire class rather than by individual faculty solicitation, thereby minimizing potential power differentials between investigators and students.

Two non-overlapping cohorts participated. The facilitated group cohort included seven students, and the self-paced individual cohort included twelve students. Because participation was voluntary, no attempt was made to balance cohort sizes or participant characteristics. Demographic information, including age, sex, race, ethnicity, and prior experience with generative artificial intelligence or standardized patients, was not collected. Accordingly, no demographic comparisons between cohorts were performed.

Verbal informed consent was obtained from all participants before study enrollment. Participants were informed that the project was conducted for educational research purposes, that participation was voluntary, and that they could withdraw at any time without academic penalty.

AI-Simulated Patient Encounter

Both cohorts completed the same AI-simulated patient encounter, which was designed to support history taking, clinical reasoning, differential diagnosis development, and patient-centered communication. The clinical scenario, patient history, and expected learning objectives were identical for both instructional formats. The AI-simulated patient responded conversationally to learner questions while maintaining a consistent clinical presentation throughout each encounter.

In the facilitated group format, students collaboratively interviewed the AI-simulated patient under faculty guidance, with discussion occurring throughout the encounter. In the individual format, students independently completed the same encounter at their own pace without real-time faculty facilitation. Following completion of the encounter, participants in both cohorts completed the post-intervention Behavioral Safety Language Scale–Provisional (BSLS-P).

The AI-simulated patient was developed using a commercially available large language model integrated with conversational voice and avatar technologies. The same patient configuration, underlying clinical case, and response framework were used for both instructional formats to minimize variability attributable to the simulation itself.

Behavioral Safety Language Scale–Provisional (BSLS-P)

The Behavioral Safety Language Scale–Provisional (BSLS-P) was developed as an exploratory research instrument to assess learner perceptions following AI-simulated patient encounters. The instrument was designed to evaluate educational constructs considered relevant to early clinical skills development, including learner engagement, **Psychological Safety**, motivation to prepare, and perceptions of the overall learning experience. [11]

Participants completed the BSLS-P immediately before and immediately after the AI-simulated patient encounter. [12] Responses were recorded using Likert-type scales, with higher scores indicating greater agreement with positively worded statements. Because the BSLS-P is a newly developed instrument, the present study was intended to provide preliminary exploratory evidence regarding its use rather than establish definitive psychometric properties.

Accordingly, BSLS-P findings should be interpreted as hypothesis-generating. Formal psychometric validation, including evaluation of reliability, construct validity, and measurement performance across larger and more diverse learner populations, remains an important objective for future research. In this instrument, Psychological Safety refers to the established educational and organizational-behavior construct describing a learner's perceived ability to ask questions, express uncertainty, and make errors without fear of embarrassment or penalty; it is a property of the learning climate and is not a measure of participant mental health or any other health outcome.

Automated Transcript Analysis

All AI-simulated patient encounters generated complete conversational transcripts. These transcripts served as the basis for an automated transcript-analysis framework developed to support exploratory educational research and future validation studies. The framework was designed to identify predefined communication behaviors using standardized operational definitions applied consistently across participant transcripts.

The automated transcript-analysis framework was intended to provide reproducible, scalable measures of learner communication behaviors while reducing the manual effort associated with traditional transcript coding. In the present study, transcript-derived measures are described to demonstrate the practical implementation of the analytical approach rather than to establish independent measures of learner performance.

Because the automated transcript-analysis framework remains under development, findings derived from transcript analyses should be interpreted as exploratory. Independent validation against expert human ratings will be required before these methods can be considered appropriate for summative educational assessment.

Reporting Standards

The quantitative component is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, and the qualitative component in accordance with the Standards for Reporting Qualitative Research (SRQR); completed checklists are provided as Additional Files.

Statistical Analysis

Statistical analyses were performed to characterize within-cohort changes following the AI-simulated patient encounter and to provide descriptive comparisons between instructional formats. Because of the exploratory nature of this feasibility pilot, analyses were intended to estimate the magnitude and direction of observed effects rather than establish comparative effectiveness.

Pre- and post-encounter BSLS-P scores were evaluated using paired *t* tests for normally distributed variables and Wilcoxon signed-rank tests for non-normally distributed variables, as appropriate. To account for multiple comparisons across BSLS-P subscales, *P* values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) procedure. Effect sizes were calculated to facilitate interpretation of observed changes.

Between-cohort comparisons were considered descriptive because instructional delivery format was confounded with cohort membership and other study-design variables. Consequently, no causal inferences regarding instructional format were made.

Given the modest sample size, all statistical findings should be interpreted as exploratory and hypothesis-generating. The analyses were intended to inform the design of future research rather than provide definitive evidence of educational effectiveness.

Use of Generative Artificial Intelligence

During the preparation of this manuscript, the authors used Claude Opus 5 (Anthropic, accessed 5 August 2026) for generative editorial assistance, including drafting and revising manuscript text. All study data, statistical analyses, and data tables were prepared, verified, and approved by co-author Sean McGinity from the raw study data, without the use of generative AI. The authors reviewed, verified, and edited all AI-assisted text and take full responsibility for the content of this manuscript.

Ethical Considerations

This study was reviewed by the Southern Illinois University School of Medicine Institutional Review Board (IRB) and received an exemption determination (IRB #24–525; reference number 036792) under applicable federal regulations governing exempt educational research. Protocol modifications covering the work reported here were approved before data collection.

Participation was voluntary, uncompensated, and had no effect on course grades or academic standing. Participants received an information sheet describing the study, and verbal informed consent was obtained in accordance with the IRB determination that written documentation of consent was not required. All study data were de-identified before analysis to protect participant confidentiality.

The study was conducted in accordance with institutional policies governing human subjects research and with the ethical principles outlined in the Declaration of Helsinki.

Results

Participant Characteristics and Group Cohort Outcomes

Nineteen first-year medical students participated in the study. Seven students completed the AI-simulated patient encounter in the faculty-facilitated group format, and twelve students completed the encounter individually in a self-paced format. Because participation was voluntary and demographic information was not collected, participant characteristics were not compared between cohorts.

All participants completed the AI-simulated patient encounter and the pre- and post-encounter Behavioral Safety Language Scale–Provisional (BSLS-P). Complete survey data were available for all analyses. No adverse events or technical issues requiring participant exclusion occurred during the study.

Within the group cohort, BSLS-P scores demonstrated directionally positive changes in learner engagement and the **Psychological Safety** subscale following the AI-simulated patient encounter, although no within-cohort comparison remained statistically significant after false discovery rate correction (Table 1). These findings support the practical implementation of the instructional approach while remaining consistent with the exploratory nature of this feasibility pilot.

Table 1

Group-cohort BSLS-P (version 1) subscale summaries, paired pre versus post (reverse-scored; n = 7).

Subscale	Pre M (SD)	Post M (SD)	Mean Δ	Hedges g (95% CI)	P	P-adj
A – Enjoyment / Engagement	4.10 (0.32)	4.19 (0.74)	+ 0.09	0.13 (-0.53, 0.77)	.7	> .9
B – Psychological Safety	3.81 (0.94)	4.10 (0.88)	+ 0.29	0.47 (-0.24, 1.2)	.2	.8
C – Motivation to Prepare	4.14 (0.81)	3.89 (0.67)	-0.25	-0.30 (-0.95, 0.38)	.4	.8
Composite (10 items)	4.03 (0.63)	4.04 (0.66)	+ 0.01	0.03 (-0.62, 0.67)	> .9	> .9

Paired-samples t test; effect size is the paired Hedges g. P-adj is the Benjamini–Hochberg FDR-adjusted P value. Reverse-scoring applied to items 3, 6, 9. Ninety-five percent CIs for effect sizes are shown in the table and in Additional File 4.

Individual Cohort Outcomes

Within the individual cohort, BSLS-P scores also demonstrated directionally positive changes following the AI-simulated patient encounter (Table 2). The largest observed effect occurred in the **Psychological Safety** subscale (Wilcoxon $r = 0.62$). Although this comparison reached statistical significance before adjustment for multiple comparisons ($P = .026$), statistical significance was not maintained following

Benjamini–Hochberg false discovery rate correction (adjusted $P = .20$). No other within-cohort comparison reached statistical significance after correction.

Motivation to Prepare demonstrated a modest decline following the encounter. Examination of the item-level responses suggested that this finding was primarily attributable to a single survey item rather than representing a generalized decline across the motivation domain. Because of the exploratory study design and modest sample size, all observed changes should be interpreted as hypothesis-generating rather than confirmatory.

Table 2

Individual-cohort BSLS-P (version 2) subscale summaries, paired pre versus post (reverse-scored; $n = 12$).

Subscale	Pre	Post	Δ	ES (95% CI)	P	P-adj
A – Enjoyment / Engagement ☒	3.67 (3.50, 4.33)	4.33 (3.83, 4.33)	+ 0.67	$r = 0.41$ (0.02, 0.83)	.2	.4
B – Psychological Safety ☒	4.00 (4.00, 4.00)	4.50 (3.83, 4.67)	+ 0.50	$r = 0.62$ (0.16, 0.89)	.026	.2
C – Motivation to Prepare ☒	3.78 (0.54)	3.36 (0.74)	-0.42	$g = -0.52$ (-1.1, 0.06)	.077	.2
D – Delivery-Format Preference ☒	3.36 (0.44)	3.50 (0.50)	+ 0.14	$g = 0.23$ (-0.31, 0.76)	.4	.5
Composite v1 (Q1–Q9) ☒	3.78 (3.72, 4.00)	3.89 (3.78, 4.33)	+ 0.11	$r = 0.22$ (0.01, 0.72)	.5	.5
Composite v2 (Q1–Q12) ☒	3.74 (0.23)	3.83 (0.40)	+ 0.10	$g = 0.24$ (-0.30, 0.77)	.4	.5

Subscale B (Psychological Safety) is the only within-cohort effect significant before correction; it did not survive FDR correction. ☒ Non-parametric: medians (Q1, Q3), Wilcoxon signed-rank exact test, effect size (ES) is the paired Wilcoxon r (used where Shapiro–Wilk rejected normality, $P < .05$). ☒ Parametric: means (SD), paired t test, Hedges g . P-adj is the FDR-adjusted P value. Ninety-five percent CIs are shown in the table and in Additional File 4.

Motivation to Prepare

A modest decline in the Motivation to Prepare domain was observed in both cohorts. Inspection of the individual questionnaire responses indicated that this change was largely attributable to a single survey item rather than reflecting a consistent pattern across the entire domain. Because the present study was designed as a feasibility pilot, these findings should be interpreted cautiously and viewed as hypothesis-generating.

One possible explanation is that participants recalibrated their expectations after completing the AI-simulated patient encounter, resulting in different interpretations of the survey item during the post-encounter assessment. An alternative explanation is that the encounter genuinely altered perceptions of preparedness for subsequent discussion. The present study was not designed to distinguish between these possibilities. Future research should evaluate revised wording for this item and formally examine its measurement characteristics during psychometric validation of the BSLS-P.

Between-Cohort Comparisons

Post-encounter comparisons between the facilitated group and self-paced individual cohorts demonstrated no statistically significant differences across the BSLS-P subscales after adjustment for multiple comparisons (Table 3). Directional differences were observed for the **Psychological Safety** and Motivation to Prepare domains; however, because instructional delivery format was confounded with cohort membership and multiple study-design variables, these findings should be interpreted as descriptive rather than evidence of comparative educational effectiveness.

Accordingly, the between-cohort analyses are presented to characterize observed patterns within this feasibility pilot and to inform the design of adequately powered randomized studies rather than to support causal conclusions regarding instructional format.

Table 3
Post-test summaries by cohort on identically worded items (Q1–Q9).

Subscale (post-test)	Individual M (SD)	Group M (SD)	Diff (G – I)	Hedges g (95% CI)	P	P-adj
Composite (Q1–Q9)	3.94 (0.49)	4.00 (0.69)	+ 0.06	0.09 (-0.80, 0.98)	> .9	> .9
A – Enjoyment / Engagement	4.11 (0.66)	4.19 (0.74)	+ 0.08	0.11 (-0.78, 1.0)	.8	> .9
B – Psychological Safety	4.36 (0.50)	4.10 (0.88)	-0.26	-0.39 (-1.3, 0.52)	.7	> .9
C – Motivation to Prepare (Q7–Q9)	3.36 (0.74)	3.71 (0.76)	+ 0.35	0.45 (-0.46, 1.3)	.7	> .9

Wilcoxon rank-sum exact test; effect size is Hedges g with 95% CI; P-adj is the FDR-adjusted P value. Diff (G – I) = Group mean minus Individual mean. No comparison reached significance; directional patterns are hypothesis-generating only and are not attributable to delivery format given the confounds described in the text.

Learner Preferences

Following completion of the AI-simulated patient encounter, participants in the individual cohort reported their preferred instructional format (Table 4). The most frequently selected preference was a hybrid

educational model in which students first completed an individual AI-simulated patient encounter followed by a faculty-facilitated group discussion. This preference was selected by seven of the twelve participants (58%).

These findings suggest that learners valued both independent interaction with the AI-simulated patient and subsequent faculty-guided discussion. The observed preference for a hybrid instructional model supports further investigation of educational approaches that combine self-directed AI-supported learning with collaborative reflection and clinical reasoning

Table 4
Single-choice format-preference distribution (Individual cohort, n = 12).

Preferred format	Pre n (%)	Post n (%)	Change	Notes
Solo first, then group	5 (42%)	7 (58%)	+ 2	Modal preference, post
Small Group	5 (42%)	4 (33%)	-1	
Solo only	1 (8.3%)	1 (8.3%)	0	
No preference	1 (8.3%)	0 (0%)	-1	Migrated away
Large Group	0 (0%)	0 (0%)	0	Never selected

Values are counts (percentages) of the n = 12 Individual cohort before and after the solo encounter.

Qualitative Findings

Participant feedback complemented the quantitative findings by providing additional insight into learner experiences with the AI-simulated patient encounter. Across both instructional formats, participants described the simulation as engaging and valuable for practicing clinical reasoning, history taking, and patient communication within a **safe learning environment**.

Participants in the individual cohort frequently expressed a preference for completing the AI-simulated patient encounter independently before participating in a faculty-guided discussion. This sequence was perceived as providing an opportunity to develop diagnostic reasoning and clinical confidence before comparing approaches with peers. Participants also identified the value of structured faculty feedback to reinforce key learning objectives and address misconceptions.

Several participants suggested opportunities to improve the educational experience, including refinements to the user interface, navigation, and AI interaction. These comments provide practical guidance for continued platform development while reinforcing the overall feasibility and acceptability of AI-simulated patient encounters within undergraduate medical education.

Discussion

This feasibility pilot supports the **practical implementation** of AI-simulated patient encounters delivered in both faculty-facilitated group and self-paced individual learning formats for first-year medical students. Although the study was not designed to compare instructional effectiveness between formats, both cohorts demonstrated directionally favorable changes in learner engagement and the **Psychological Safety** subscale of the Behavioral Safety Language Scale–Provisional (BSLS-P). These findings suggest that AI-simulated patient encounters can be successfully incorporated into undergraduate medical education while providing learners with a **safe learning environment** for practicing clinical reasoning and patient communication.

This study contributes to the growing body of literature supporting the educational value of AI-simulated patients. Previous investigations have reported high learner acceptance, improved engagement, and opportunities for repeated deliberate practice. [1, 5, 6] The present findings extend this literature by examining two instructional delivery formats while emphasizing **practical implementation** rather than comparative effectiveness. The observed learner preference for a hybrid instructional model further suggests that independent AI-simulated patient encounters may be most effective when integrated with subsequent faculty-guided discussion and reflection.

Comparison With Prior Work

The present findings are consistent with a growing body of literature demonstrating that AI-simulated patients can enhance learner engagement and provide valuable opportunities for deliberate practice in undergraduate medical education. Recent systematic and scoping reviews have reported favorable learner acceptance while emphasizing the need for more rigorous study designs and standardized educational outcome measures. [1–4] The present study contributes to this literature by focusing on instructional delivery format and by transparently describing both a provisional learner-perception instrument and an automated transcript-analysis framework.

Although the observed improvements in learner engagement and the **Psychological Safety** domain were exploratory, they align with previous reports suggesting that conversational AI can support learner-centered educational experiences. [6, 9, 11] Unlike many previous investigations, this study emphasizes methodological transparency by describing the analytical framework in sufficient detail to support replication and future validation.

The observed preference for a hybrid instructional model further supports integrating AI-simulated patient encounters with faculty-guided discussion rather than viewing AI as a replacement for traditional instructional approaches. Together, these findings reinforce the potential role of AI-simulated patients as scalable educational tools while highlighting the importance of continued evaluation through adequately powered studies.

Intellectual Property and Potential Conflicts

The AI-simulated patient platform evaluated in this study was developed at Southern Illinois University School of Medicine and served as the educational intervention investigated in this research. One author (R.S.) is a named inventor on a pending patent application covering elements of the underlying AI-simulation technology. The patent application is held by the Board of Trustees of Southern Illinois University, and no royalties or other financial distributions related to this technology have been received by any author.

To reduce the potential influence of this intellectual-property interest on study interpretation, statistical analyses and preparation of all data tables were performed independently by co-author Sean McGinity using the original study dataset. Platform-derived transcript analyses are reported descriptively and are not used to support inferential claims regarding educational effectiveness. The analytical framework has been described transparently to facilitate independent evaluation and future validation.

Future research should incorporate independent statistical analysis, prospective registration of the planned randomized trial, and external validation of transcript-derived outcome measures to further strengthen methodological rigor and minimize potential sources of bias.

Limitations

As a feasibility pilot, this study has several limitations that should be considered when interpreting the findings. First, the study included a small sample of first-year medical students from a single institution, limiting statistical power and the generalizability of the findings. Second, participants were not randomized to instructional format. Because cohort membership and instructional delivery format were confounded with several study-design variables, between-cohort comparisons should be interpreted as descriptive rather than causal.

Third, demographic information, including age, sex, race, ethnicity, prior standardized patient experience, and previous experience with generative artificial intelligence, was not collected. Consequently, the potential influence of participant characteristics on learner perceptions could not be evaluated. Recruitment through a general invitation to the entire class helped minimize investigator–student power differentials but does not eliminate the possibility of volunteer bias.

Fourth, the Behavioral Safety Language Scale–Provisional (BSLS-P) is an exploratory research instrument that has **not yet undergone formal psychometric validation**. Although the instrument demonstrated utility for this feasibility investigation, future research should establish its reliability, construct validity, responsiveness, and measurement performance across larger and more diverse learner populations before it is considered for broader educational use.

Finally, the automated transcript-analysis framework remains under development. Although the framework provides a transparent and reproducible approach for analyzing learner–patient conversations, transcript-derived measures have not yet been independently validated against expert human ratings. Accordingly, these measures should not be interpreted as validated indicators of learner

performance or used for high-stakes educational assessment until additional validation studies have been completed.

Recommendations

The findings from this feasibility pilot support several priorities for **future research**. First, multicenter studies with larger, randomized cohorts are needed to evaluate the educational effectiveness of AI-simulated patient encounters across different instructional settings. [7, 10] Second, the Behavioral Safety Language Scale–Provisional (BSLS-P) should undergo formal psychometric evaluation, including assessment of reliability, construct validity, responsiveness, and measurement performance in diverse learner populations.

Additional work should focus on validating the automated transcript-analysis framework through direct comparison with expert human ratings. Refinement of the user interface, learner workflow, and AI interaction based on participant feedback may further improve the educational experience. Finally, future investigations should examine knowledge retention, clinical reasoning, communication skills, and learner confidence over longer follow-up periods to better define the educational value of AI-simulated patient encounters.

Conclusions

This feasibility pilot supports the **practical implementation** of AI-simulated patient encounters in both faculty-facilitated group and self-paced individual learning environments for first-year medical students. Across both instructional formats, learners demonstrated directionally favorable changes in engagement and the **Psychological Safety** domain of the Behavioral Safety Language Scale–Provisional (BSLS-P), while qualitative feedback indicated strong acceptance of a hybrid educational model combining independent AI-simulated patient encounters with subsequent faculty-guided discussion.

The findings should be interpreted as exploratory and hypothesis-generating rather than evidence of comparative educational effectiveness. In addition to evaluating instructional delivery formats, this study introduces two methodological contributions: the Behavioral Safety Language Scale–Provisional (BSLS-P) as an exploratory research instrument and an automated transcript-analysis framework designed to support reproducible educational research.

Future research should include multicenter randomized studies, formal psychometric validation of the BSLS-P, independent validation of transcript-derived outcome measures against expert human ratings, and evaluation of long-term educational outcomes. Collectively, these efforts will help define the role of AI-simulated patients as scalable, evidence-informed tools for undergraduate medical education.

Abbreviations

AI

artificial intelligence
API
application programming interface
BSLS-P
Behavioral Safety Language Scale–Provisional
CBL
case-based learning
CI
confidence interval
ED
emergency department
FDR
false discovery rate
IRB
Institutional Review Board
LLM
large language model
MS1
first-year medical student
PBL
problem-based learning
Q1, Q3
first and third quartiles
SD
standard deviation
SIU-SOM
Southern Illinois University School of Medicine
SP
simulated patient
SRQR
Standards for Reporting Qualitative Research
STROBE
Strengthening the Reporting of Observational Studies in Epidemiology
UME
undergraduate medical education.

Declarations

Ethics approval and consent to participate

The study received an exemption determination from the Southern Illinois University School of Medicine Institutional Review Board (IRB #24-525; reference number 036792) under applicable federal regulations governing exempt educational research, and protocol modifications covering the work reported here were approved before data collection. All participants were adults (18 years of age or older). Participation was voluntary, uncompensated, and had no effect on course grades or academic standing. Participants received an information sheet describing the study, verbal informed consent was obtained from all participants, and the Institutional Review Board determined that written documentation of consent was not required. All methods were performed in accordance with the Declaration of Helsinki and with institutional policies governing human subjects research.

Consent for publication

Not applicable.

Availability of data and materials

The de-identified item-level dataset supporting the conclusions of this article — paired pre- and post-encounter BSLS-P responses for both cohorts — is included as Additional File 6. Verbatim encounter transcripts and free-text responses are not publicly available because they were generated by small cohorts of students at a single institution and may permit re-identification; de-identified excerpts are available from the corresponding author on reasonable request, subject to approval by the SIU School of Medicine IRB. The locked system prompts and case configuration governing the AI-simulated patient are provided in Additional File. The AI-simulation platform itself is not publicly available, as it is the subject of a pending patent application held by the Board of Trustees of Southern Illinois University.

Competing interests

The AI-simulated patient platform evaluated in this study was developed at Southern Illinois University School of Medicine. One author (R.S.) is a named inventor on a pending patent application covering elements of the underlying AI-simulation technology; the application is held by the Board of Trustees of Southern Illinois University, and no royalties or other financial distributions related to this technology have been received by any author. To reduce the potential influence of this interest on study interpretation, all statistical analyses and data tables were prepared independently by S.M. from the original study dataset. The remaining authors declare that they have no competing interests.

Funding

This study was supported by Southern Illinois University School of Medicine and by a generous gift from the Southern Illinois University Foundation. Neither funder had any role in the design of the study; in the

collection, analysis, or interpretation of the data; or in the writing of the manuscript.

Authors' contributions

S.T. and R.S. conceived and designed the study. R.S. led development of the AI-simulated patient platform and the automated transcript-analysis framework. S.L., J.L., and R.S. delivered the AI-simulated patient encounters and coordinated data collection. S.M. performed the statistical analyses and prepared all data tables from the raw study data. A.P., E.M., and S.L. contributed to case development, platform testing, and acquisition of study data. S.T. and R.S. drafted the manuscript. All authors critically revised the manuscript for important intellectual content and read and approved the final version.

Acknowledgements

The authors thank the first-year medical students at Southern Illinois University School of Medicine who volunteered their time to participate in this study.

References

1. Wei H, Dai Y, Yuan K, Li KY, Hung KF, Hu EM, et al. AI-powered problem- and case-based learning in medical and dental education: a systematic review and meta-analysis. *Int Dent J*. 2025;75(6):100858. 10.1016/j.identj.2025.100858.
2. Abidi SH, Almazan J, Fabiyi O, Zehra F, Tariq M. AI-supported case-based learning in medical education: a comprehensive scoping review. *Front Med (Lausanne)*. 2026;13:1798097. 10.3389/fmed.2026.1798097.
3. Pham TD, Karunaratne N, Exintaris B, Liu D, Lay T, Yuriev E, et al. The impact of generative AI on health professional education: a systematic review in the context of student learning. *Med Educ*. 2025;59(12):1280–9. 10.1111/medu.15746.
4. Simoni J, Urtubia-Fernandez J, Mengual E, Simoni DA, Royo M, Egaña-Yin D, et al. Artificial intelligence in undergraduate medical education: an updated scoping review. *BMC Med Educ*. 2025;25(1):1609. 10.1186/s12909-025-08188-2.
5. Cross J, Kayalackakom T, Robinson RE, Vaughans A, Sebastian R, Hood R, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. *JMIR Med Educ*. 2025;11:e63353. 10.2196/63353.
6. Tong X, Hu Y, Long Y, Zhang Q, Yang Y, Yuan J, et al. The application of problem-based learning (PBL) guided by ChatGPT in clinical education in the Department of Nephrology. *BMC Med Educ*. 2025;25:1048. 10.1186/s12909-025-07427-w.
7. Çiçek FE, Ülker M, Özer M, Kiyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. *Postgrad Med J*. 2025;101(1195):458–63. 10.1093/postmj/qgae170.

8. Mool A, Schmid J, Johnston T, et al. Using generative AI to simulate patient history-taking in a problem-based learning tutorial: a mixed-methods study. *Technol Knowl Learn*. 2026. 10.1007/s10758-025-09929-4.
9. Kassab SE, Rathan R, Taylor DC, et al. Understanding how medical students learn in the era of artificial intelligence: a mixed methods study. *BMC Med Educ*. 2025;25:1521. 10.1186/s12909-025-08145-z.
10. Kämmer JE, Hautz WE, Krummrey G, Sauter TC, Penders D, Birrenbach T, et al. Effects of interacting with a large language model compared with a human coach on the clinical diagnostic process and outcomes among fourth-year medical students: study protocol for a prospective, randomised experiment using patient vignettes. *BMJ Open*. 2024;14(7):e087469. 10.1136/bmjopen-2024-087469.
11. Luo J, Peng Q, Cao H, et al. Medical undergraduate students' readiness and anxiety toward artificial intelligence: a systematic review and meta-analysis. *BMC Med Educ*. 2026;26:44. 10.1186/s12909-025-08388-w.
12. Huang Y, Xu BB, Wang XY, Luo YC, Teng MM, Weng X. Implementation and evaluation of an optimized surgical clerkship teaching model utilizing ChatGPT. *BMC Med Educ*. 2024;24(1):1540. 10.1186/s12909-024-06575-9.