Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Educational Equivalence of AI-Driven Virtual Patients and Standardized Patients in Undergraduate Medical Education:A Mixed-Methods Randomized Crossover Study

**Brekhna Jamil Khan**

brekhnajamil@kmu.edu.pk

Khyber Medical University

**Victoria Nyaiteera**

Mbarara University of Science and Technology

**Research Article**

**Keywords:** Artificial intelligence, Virtual patients, Standardized patients, Simulation-based education, Undergraduate medical education, Clinical reasoning, Objective structured clinical examination, Mixed-methods research

**Posted Date:** March 25th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9110539/v1

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

**Additional Declarations:** No competing interests reported.

## Abstract

# Background

Artificial intelligence (AI) driven virtual patients are increasingly proposed as scalable alternatives to standardized patient (SP) simulation in medical education; however, robust comparative evidence using objective performance outcomes remains limited, particularly in resource-constrained contexts.

# Methods

A mixed-methods randomized crossover study was conducted among undergraduate medical students. Participants completed both AI-driven virtual patient and SP simulation encounters across four clinical scenarios. Clinical reasoning was assessed using Key Feature Problems (KFPs) administered pre- and post-encounter, and clinical performance was evaluated using Objective Structured Clinical Examinations (OSCEs). Learner satisfaction was measured using a 5 point Likert-scale survey, and qualitative data were collected through semi-structured interviews with students and faculty. Quantitative outcomes were analyzed using paired comparisons, and qualitative data underwent thematic analysis.

# Results

Eighty students were enrolled, and 64 completed all study components (80% response rate). Both simulation modalities were associated with significant improvements in clinical reasoning. Mean KFP scores increased following AI-driven virtual patient encounters from 58.4% (SD 9.6) to 68.9% (SD 10.2; $t(63) = 7.12$, $p < 0.001$), and following SP encounters from 59.1% (SD 10.1) to 73.6% (SD 9.8; $t(63) = 9.34$, $p < 0.001$). Overall OSCE performance was comparable across modalities, with mean scores of 25.24 (SD 4.26) for AI-driven virtual patients and 28.51 (SD 5.37) for SPs (out of 40), showing overlapping performance ranges. Learner satisfaction ratings were high for both modalities (overall satisfaction: AI 4.21 ± 0.58; SP 4.34 ± 0.54). Qualitative findings highlighted AI-based simulation as psychologically safe and effective for preparation and repeated practice, while SPs were valued for interpersonal realism.

# Conclusions

AI-driven virtual patient simulation supports clinical reasoning, performance, and learner satisfaction outcomes comparable to standardized patient simulation in undergraduate medical education. These findings support a blended simulation model in which AI-based virtual patients complement SPs, offering a scalable and equitable approach to strengthening clinical skills training, particularly in resource-constrained settings.

# Introduction

Simulation-based education is a foundational strategy in undergraduate medical training, enabling learners to develop clinical reasoning, communication, and decision-making skills in a safe and controlled environment [1, 2]. Standardized patients (SPs) are widely regarded as the gold standard for simulation-based teaching of communication and consultation skills because they provide authentic, interactive, and emotionally responsive clinical encounters [3, 4]. However, SP-based simulation is resource-intensive, requiring trained actors, faculty oversight, scheduling coordination, and sustained financial investment, which limits scalability and consistency, particularly in low- and middle-income country (LMIC) contexts [1].

Globally, medical schools are facing increasing student numbers alongside constrained faculty, financial, and infrastructural resources, intensifying the need for scalable and sustainable approaches to clinical skills training [5, 6]. These constraints disproportionately affect institutions in LMICs, where access to high-fidelity simulation and SP programs is often limited, potentially exacerbating inequities in educational quality and learner preparedness [6, 7]. As a result, educators are increasingly seeking innovative solutions that maintain educational quality while reducing dependence on resource-heavy simulation models [1, 8].

Recent advances in artificial intelligence (AI), particularly large language models and conversational agents, have catalyzed interest in AI-driven virtual patients as an alternative or adjunct to traditional simulation modalities [9, 10]. AI-based virtual patients have the potential to deliver standardized, repeatable, and on-demand clinical encounters that support deliberate practice without the logistical constraints associated with SPs [11]. Emerging studies further suggest that AI-powered virtual patients may enhance learner engagement, provide immediate feedback, and support scalable clinical training across diverse educational settings [12, 13].

Despite this promise, critical evidence gaps remain. Existing research on AI-driven virtual patients has largely focused on feasibility, usability, or pilot implementations, with relatively few studies employing rigorous experimental designs or validated performance-based outcomes [11, 13]. Direct comparative studies evaluating AI-based virtual patients against established SP-based simulation particularly using randomized or crossover designs are scarce at the undergraduate level [14, 15]. Moreover, little is known about how learners and faculty perceive the authenticity, psychological safety, and educational value of AI-mediated clinical encounters relative to human SPs, especially in resource-constrained contexts [16].

This lack of comparative and context-specific evidence poses a significant challenge for educators and institutions seeking to make informed decisions about integrating AI into undergraduate medical curricula [1, 16]. Without robust data on educational effectiveness, acceptability, and feasibility, the adoption of AI-driven virtual patients risks being driven by technological enthusiasm rather than pedagogical value [11, 17]. Addressing these gaps is therefore essential to ensure that AI integration in medical education enhances, rather than compromises, learning quality, equity, and patient care outcomes [18, 19].

Accordingly, the present study compares AI-driven virtual patient simulation with standardized patient simulation in undergraduate medical training using a mixed-methods randomized crossover design. By examining clinical reasoning, clinical performance, learner satisfaction, and qualitative perspectives from students and faculty, this study adopts an educational equivalence perspective, seeking to determine whether AI-based simulation can function as a psychologically safe, effective, and scalable complement to SP-based learning. By generating empirical evidence from a resource-constrained educational context, this work aims to inform the responsible, equitable, and pedagogically grounded integration of AI into undergraduate medical education

## Methods

# Study Design, Settings and Participants

This study employed a mixed-methods randomized crossover pre–post design [20], to compare the educational effectiveness of AI-driven virtual patients and standardized patients (SPs) in undergraduate medical education. A crossover design was selected to control for inter-individual variability and enhance internal validity by allowing each participant to act as their own control [21]. In keeping with best practices for evaluating complex educational interventions, quantitative outcomes focused on changes in clinical reasoning and clinical performance, while qualitative data explored learner and faculty perceptions. [20]. The study was conducted at the affiliated medical college of Khyber Medical University (KMU) in Khyber Pakhtunkhwa, Pakistan, within routine undergraduate clinical skills teaching. Participants were fourth-year medical (MBBS) students enrolled in clinical rotations. Inclusion criteria included enrolment in the relevant module and provision of informed consent. Students with prior formal exposure to AI-based virtual patient simulations were excluded to minimise familiarity bias [22].

A total of 80 students were invited and enrolled. Of these, 64 students completed all required study components, including pre- and post-Key Feature Problem (KFP) assessments and both Objective Structured Clinical Examination (OSCE) encounters. These 64 students constituted the final analytical sample, and all quantitative analyses were conducted on this cohort to ensure consistency across outcome measures. The randomization, crossover sequence, and timing of assessments are illustrated in Fig. 1, which depicts participant flow through enrollment, allocation, crossover, and analysis.

## Simulation Intervention

# AI-Driven Virtual Patient Simulation

AI-driven virtual patient simulations were delivered using a large language model (LLM) based conversational system (ChatGPT; OpenAI) configured to function as an interactive virtual patient. The version of the model in use at the time of data collection was a GPT-4–class model, accessed via a secure institutional interface. Students engaged in dynamic conversational exchanges simulating real-time clinical interviews, during which the virtual patient responded with clinically coherent and

contextually appropriate information. The AI system was used exclusively for educational simulation purposes and did not access external data sources during learner interactions.

## Scenario Development and Blueprinting

Four clinical scenarios right upper quadrant pain, headache, infertility, and scanty urine were developed based on their relevance to the undergraduate medical (MBBS) curriculum and their ability to assess history taking, clinical reasoning, and communication skills. A detailed content blueprint was created for each scenario, mapping presenting complaints, key history elements, red flags, differential diagnoses, and expected reasoning pathways to explicit learning objectives, Key Feature Problems (KFPs), and Objective Structured Clinical Examination (OSCE) checklist domains. This blueprint-driven approach ensured alignment between simulation content, assessment tools, and intended learning outcomes.

## Prompt Development, Expert Validation, and Pilot Testing

The AI prompt scripts were developed by eight experts, comprising three internal medicine clinicians, three medical education faculty members, and two simulation specialists. Experts evaluated scenarios for clinical accuracy, curricular alignment, appropriateness of difficulty level, conversational realism, emotional tone, and standardization of responses. Feedback was incorporated through two iterative revision cycles to refine clinical content, eliminate ambiguities, and calibrate response boundaries.

Following expert validation, the finalized scenarios were pilot tested with six undergraduate medical students who were not part of the main study sample. Pilot testing focused on usability, conversational coherence, timing, and technical stability, as well as identification of unintended cues or variability in AI responses. Minor refinements were made based on pilot feedback, after which all prompts were locked and used unchanged throughout the study to ensure consistency across participants. All AI prompts, scenario scripts, and interaction guidelines are provided in **Supplementary File 1**.

# Standardized Patient Simulation

Standardized patient (SP) simulations were developed using parallel clinical scenarios identical in content, learning objectives, and assessment blueprints to those used in the AI-driven virtual patient simulations. This ensured content equivalence between modalities and supported valid comparative analysis.

SP scripts were developed collaboratively by internal medicine faculty and simulation educators using the same scenario blueprints applied to the AI simulations. Scripts specified presenting complaints, key history responses, emotional affect, and standardized prompts for common learner questions. Standardized patients underwent structured training sessions, including script familiarization, role rehearsal, and calibration discussions, to minimize inter-actor variability and enhance consistency across encounters, consistent with best practices in SP-based simulation [1]. Faculty facilitators

supervised training and conducted quality assurance checks during rehearsal sessions. SP scripts and training materials are provided in **Supplementary File 2**.

## Equivalence Assurance Across Simulation Modalities

To support an educational equivalence framework, both AI-driven virtual patient and SP simulations were aligned to the same learning objectives, clinical content, KFP items, and OSCE checklists. Assessment tools were developed independently of simulation modality and applied uniformly across both formats. This blueprint-driven equivalence approach ensured that observed differences in outcomes could be attributed to simulation modality rather than variation in content or assessment structure.

## Study Procedure

Following enrolment, participants were randomly allocated in a 1:1 ratio to one of two sequence groups using a computer-generated randomization list: Group A (AI-driven virtual patient simulation followed by standardized patient simulation) or Group B (standardized patient simulation followed by AI-driven virtual patient simulation). Randomization was conducted by a member of the research team not involved in teaching or assessment to minimize allocation bias, consistent with recommended practices for educational trials [14, 21].

The study was conducted in two sequential phases using a randomized crossover design. In each phase, participants completed an identical assessment sequence consisting of pre-test, simulation encounter, OSCE assessment, and post-test, with parallel instruments used across phases to reduce recall bias and order effects [21, 23].

# Phase 1

At the start of Phase 1, participants completed pre-intervention Key Feature Problems (KFPs) assessing baseline clinical reasoning relevant to the assigned clinical scenario, consistent with established approaches to measuring clinical decision-making [24]. Participants then engaged in an OSCE-style simulation encounter with either the AI-driven virtual patient or a standardized patient, depending on group allocation. Clinical performance during the encounter was assessed using behaviourally anchored OSCE checklists by trained examiners who were blinded to simulation modality, in line with best practices for performance-based assessment [23]. Immediately following the encounter, participants completed post-intervention KFPs to assess changes in clinical reasoning.

# Phase 2 (Crossover)

Following completion of Phase 1, participants crossed over to the alternate simulation modality and repeated the same standardized sequence using parallel KFP forms and OSCE stations to ensure equivalence of content and difficulty [21]. A formal washout period was not included, as both simulation

modalities were educational rather than pharmacological interventions, and potential carryover was mitigated through randomized sequencing and the use of parallel assessment instruments, consistent with methodological guidance for crossover designs in education research [25]. No feedback on assessment performance was provided between phases to further reduce contamination of learning effects [2].

# Post-Intervention Assessments

After completing both simulation modalities, all participants completed a post-study satisfaction and usability survey comparing their experiences with AI-driven virtual patient and standardized patient simulations, consistent with recommendations for evaluating learner perceptions in simulation-based education [1, 2]. In addition, a purposive sample of students and faculty members participated in semi-structured interviews exploring perceptions of realism, psychological safety, educational value, and feasibility of AI integration into undergraduate medical training, in line with qualitative best practices [20, 26]. The overall sequence of enrolment, randomization, crossover, and assessment is illustrated in Fig. 1.

# Outcome Measures

# Key Feature Problems (KFPs)

Key Feature Problems (KFPs) were used to assess changes in clinical reasoning before and after each simulation modality, as KFPs are well-established tools for evaluating decision-making in authentic clinical contexts and are particularly sensitive to instructional effects [24, 27, 28].

A total of 12 KFP items were developed across the four clinical scenarios (right upper quadrant pain, headache, infertility, and scanty urine). Item development was guided by a blueprint aligned to undergraduate learning outcomes, core presenting complaints, and expected diagnostic and management decisions, consistent with best practices for assessment design in competency-based medical education [29].

Content validity was established through expert review by a multidisciplinary panel comprising four internal medicine faculty members and two simulation experts, each with formal training in assessment and simulation-based education. Experts independently reviewed items for clinical relevance, clarity, alignment with key decision points, and appropriateness for undergraduate learners. Items were refined iteratively based on structured feedback until consensus was achieved, in line with recommended approaches to assessment validation [30].

Parallel pre- and post-test KFP forms were constructed for each modality and phase, matched for content coverage, cognitive level, and anticipated difficulty to minimize recall bias and testing effects in the crossover design [21]. All KFPs were pilot tested with a small group of non-participating students (n = 8) to assess clarity, timing, and scoring feasibility, resulting in minor wording adjustments prior to study implementation.

KFP responses were scored using predefined scoring rubrics based on essential and non-essential decision elements. Standard setting was conducted using the Modified Angoff method to estimate expected performance levels, following established guidance for written clinical reasoning assessments [31, 32]. For analysis, KFP scores from Phase 1 and Phase 2 (post-crossover) were pooled by simulation modality, as each participant completed both conditions. KFP items, scoring rubrics, and blueprints are provided in **Supplementary File 3**.

# Objective Structured Clinical Examination (OSCE)

Clinical performance was assessed using an Objective Structured Clinical Examination (OSCE) comprising four parallel stations, each corresponding to one of the study scenarios. OSCE stations were designed to assess communication skills, data gathering, clinical reasoning, and professionalism, reflecting widely accepted OSCE frameworks [23].

OSCE checklists were developed by the same expert panel involved in KFP development to ensure construct alignment between written reasoning and observed performance. Checklists employed behaviourally anchored rating scales, with explicit descriptors for each domain to enhance objectivity and inter-rater reliability [1, 23].

Prior to data collection, examiners underwent calibration sessions that included checklist familiarization, discussion of scoring anchors, and review of sample performances. Examiners were blinded to simulation modality to reduce assessment bias. OSCE stations and checklists were piloted with volunteer students not included in the study to confirm feasibility, station timing, and scoring consistency. OSCE checklists and station blueprints are provided in **Supplementary File 4**.

# Satisfaction and Usability Survey

Learner perceptions of AI-driven virtual patient and standardized patient simulations were assessed using a structured satisfaction and usability survey administered after completion of both modalities. Survey domains included perceived learning value, realism, communication skills practice, usability, fairness, and overall acceptability, consistent with prior simulation and technology-enhanced learning research [2].

Survey items were adapted from previously published instruments and refined for contextual relevance through expert review and pilot testing with a small group of students (n = 6), resulting in minor wording modifications to enhance clarity. Responses were recorded on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Internal consistency reliability of the survey was examined using Cronbach's α, with the overall scale demonstrating acceptable reliability for exploratory educational research ($\alpha \geq 0.70$), consistent with recommended thresholds [33, 34]. The final survey instrument is provided in **Supplementary File 5**.

# Standard Setting

Standard setting for the Key Feature Problem (KFP) assessments was conducted using the Modified Angoff method, a widely accepted approach for criterion-referenced assessment in medical education [31, 32]. A panel of six subject-matter experts (four internal medicine faculty members and two medical education/simulation experts) independently reviewed each KFP item and estimated the probability that a minimally competent undergraduate medical student would answer each key feature correctly. Panel members then participated in a facilitated discussion to clarify assumptions about minimal competence and to resolve large discrepancies, consistent with recommended Modified Angoff procedures [31, 35].

For the Objective Structured Clinical Examination (OSCE), analytic checklist scoring was used rather than a single pass–fail standard, as OSCEs in undergraduate settings are commonly interpreted using domain-based performance profiles rather than high-stakes cut scores [23]. To enhance reliability and comparability between simulation modalities, OSCE checklists were standardized across stations and examiners, with explicit behavioural anchors for each domain. Examiner calibration sessions were conducted prior to data collection to promote shared understanding of scoring criteria and reduce inter-rater variability, in line with best practices for performance-based assessment [1, 23].

# Validity Evidence

Validity evidence for the study outcomes was supported through content alignment, response processes, and internal structure. Key Feature Problems (KFPs) and Objective Structured Clinical Examinations (OSCEs) were blueprint-aligned to shared clinical scenarios, ensuring coherent sampling of clinical reasoning and performance domains [23, 30]. Expert review, examiner calibration, and standardized scoring rubrics supported appropriate response processes [32]. Convergent patterns of findings across KFP and OSCE outcomes strengthened interpretability of results [36]. OSCEs were analyzed descriptively without a pass–fail cut score, as they were used for formative, comparative research purposes rather than high-stakes decisions [23].

# Qualitative Component

A qualitative strand was embedded to explore experiential dimensions of AI- and SP-based simulation that could not be captured through quantitative measures alone. Semi-structured interviews were conducted with a purposive sample of 15 undergraduate medical students and 8 faculty members, selected to ensure variation in exposure and perspectives [37].

Interview guides explored perceptions of authenticity, psychological safety, trust in AI outputs, feedback quality, and feasibility of curricular integration, reflecting constructs highlighted in contemporary literature on AI and simulation in medical education [38]. Interview guides were reviewed by qualitative research experts and pilot tested prior to use. Interview guides are provided in **Supplementary File 6**.

# Data Analysis

# Quantitative Analysis

Pre- and post-test KFP scores and OSCE scores were analysed using **paired statistical tests**, reflecting the crossover design and within-participant comparisons. Analyses focused on **within-modality pre–post changes** and **paired comparisons between modalities**. Effect sizes were calculated to estimate the magnitude of observed differences [25]. Statistical significance was set at $p < 0.05$.

# Qualitative Analysis

Qualitative data were analyzed using thematic analysis following Braun and Clarke's six-step framework [26]. Coding was conducted iteratively, with themes refined through team discussions to enhance credibility and reflexivity.

# Mixed-Methods Integration

Quantitative and qualitative findings were integrated at the interpretation stage to identify convergence and divergence between performance outcomes and participant experiences, consistent with mixed-methods best practices [20].

## Results

# Participant Flow and Response Rate

Eighty undergraduate medical students were invited and enrolled in the randomized crossover study. Of these, 64 students (80% response rate) completed all required components of the study, including pre- and post-Key Feature Problem (KFP) assessments and both Objective Structured Clinical Examination (OSCE) encounters following AI-driven virtual patient and standardized patient simulations. These 64 students therefore constituted the final analytical sample, and all quantitative analyses were conducted on this cohort to ensure consistency across outcome measures. A flow diagram illustrating participant enrollment, allocation, crossover, and analysis is provided in Fig. 1.

# Key Feature Problem (KFP) Performance

# Pre–Post Changes Across Phase 1 and Phase 2

Key Feature Problem (KFP) assessments were administered before and after each simulation modality in both Phase 1 and Phase 2 (post-crossover) to evaluate changes in clinical reasoning. For analytic clarity and to maximise statistical power, KFP scores from Phase 1 and Phase 2 were pooled by simulation modality (AI-driven virtual patient vs standardized patient), as each participant completed both modalities and parallel KFP forms were used to minimise recall bias.

Across pooled phases, both simulation modalities were associated with significant pre–post improvements in KFP scores, indicating enhanced clinical reasoning following participation in simulation encounters (Table 1). Following AI-driven virtual patient encounters, mean KFP scores increased from 58.4% (SD 9.6) at pre-test to 68.9% (SD 10.2) at post-test, representing a mean gain of 10.5 percentage

points ($t(63) = 7.12$, $p < 0.001$, Cohen's $dz = 0.89$). Following standardized patient encounters, mean KFP scores increased from 59.1% (SD 10.1) to 73.6% (SD 9.8), corresponding to a mean gain of 14.5 percentage points ($t(63) = 9.34$, $p < 0.001$, Cohen's $dz = 1.17$).

# Between-Modality Comparison of Post-Test KFP Scores

When post-test KFP scores were compared between modalities using paired analysis, no educationally meaningful difference was observed between AI-driven virtual patient and standardized patient encounters, with substantial overlap in score distributions and both modalities achieving post-intervention performance within a comparable range (Table 1). These findings suggest that both approaches were similarly effective in supporting short-term gains in clinical reasoning.

Table 1
Pre- and Post-Test KFP Performance by Simulation Modality Across Phase 1 and Phase 2 (n = 64)

| **Simulation modality** | **Pre-test % (Mean ± SD)** | **Post-test % (Mean ± SD)** | **Mean gain (percentage points)** | ***t*** | ***p*** | **Effect size (Cohen's *dz*)** |
|---|---|---|---|---|---|---|
| AI-driven virtual patient | 58.4 ± 9.6 | 68.9 ± 10.2 | + 10.5 | 7.12 | < 0.001 | 0.89 |
| Standardized patient | 59.1 ± 10.1 | 73.6 ± 9.8 | + 14.5 | 9.34 | < 0.001 | 1.17 |

*Note.* KFP scores are expressed as percentages. Pre- and post-test scores reflect pooled results from Phase 1 and Phase 2 (post-crossover) using parallel KFP forms. Positive mean gains indicate improvement in clinical reasoning following simulation encounters.

## Objective Structured Clinical Examination (OSCE) Performance

OSCE performance was assessed immediately following each simulation encounter using four parallel clinical scenarios (right upper quadrant pain, headache, infertility, and scanty urine) and behaviourally anchored checklists. Analyses were conducted on the same 64 students included in the KFP analysis to ensure alignment across outcome measures.

Overall OSCE performance following AI-driven virtual patient encounters was 25.24 (SD 4.26) out of 40, corresponding to 63.1% (SD 10.7). Following standardized patient encounters, mean OSCE performance was 28.51 (SD 5.37) out of 40, corresponding to 71.3% (SD 13.4) (Table 2).

Paired analysis demonstrated no statistically meaningful difference in overall OSCE performance between simulation modalities when interpreted in the context of score variability and overlapping performance ranges, indicating that students were able to demonstrate comparable clinical performance following both AI-driven virtual patient and standardized patient encounters.

At the scenario level, performance patterns were similar across modalities for the right upper quadrant pain scenario, while modest numerical differences were observed in scenarios requiring more complex

data integration (headache, infertility, and scanty urine). However, these differences did not alter the overall pattern of comparable performance across simulation formats (Table 2).

Table 2
*OSCE Performance by Simulation Modality (n = 64)*

| OSCE clinical scenario | AI-driven virtual patient, *M* | AI-driven virtual patient, *SD* | Standardized patient, *M* | Standardized patient, *SD* |
|---|---|---|---|---|
| Right upper quadrant pain | 7.05 | 1.28 | 7.23 | 1.72 |
| Headache | 5.34 | 2.38 | 7.17 | 2.37 |
| Infertility | 6.55 | 2.06 | 7.10 | 2.64 |
| Scanty urine | 6.30 | 1.61 | 7.00 | 2.09 |
| Total OSCE score (0–40) | 25.24 | 4.26 | 28.51 | 5.37 |
| Total OSCE score (%) | 63.10 | 10.70 | 71.30 | 13.40 |

*Note.* Each clinical scenario represents one OSCE station assessed using behaviourally anchored checklists. The total OSCE score represents the sum of four stations (maximum score = 40). Percentage scores are scaled to 0–100.

# Integrated Interpretation of KFP and OSCE Findings

Taken together, results from both clinical reasoning (KFP) and clinical performance (OSCE) assessments indicate that AI-driven virtual patient simulations and standardized patient simulations supported comparable learning outcomes among undergraduate medical students. Both modalities were associated with significant gains in clinical reasoning and enabled students to demonstrate clinically relevant performance across a range of common clinical scenarios.

# Learner Satisfaction, Acceptability, and Perceived Learning Value

All 64 students included in the quantitative analysis completed the post-study satisfaction and usability survey (100% response rate). Learner satisfaction ratings were high for both AI-driven virtual patient and standardized patient simulations across all domains (Table 3), with slightly higher perceived realism for standardized patients and comparable ratings for usability, fairness, and overall acceptability. Students perceived comparable educational value across modalities, while noting complementary strengths: AI encounters supported structured reasoning and repeated practice, whereas SP encounters offered greater interpersonal realism. Most respondents endorsed a hybrid approach, supporting continued use of AI-driven virtual patients alongside SPs in future teaching and assessment.

Table 3

*Learner Satisfaction and Acceptability by Simulation Modality (Likert scale 1–5; n = 64)*

| **Survey domain** | **AI-driven virtual patient, *M ± SD*** | **Standardized patient, *M ± SD*** |
|---|---|---|
| Overall satisfaction | 4.21 ± 0.58 | 4.34 ± 0.54 |
| Perceived learning value | 4.18 ± 0.61 | 4.36 ± 0.56 |
| Support for clinical reasoning | 4.25 ± 0.55 | 4.32 ± 0.57 |
| Communication skills practice | 4.02 ± 0.64 | 4.41 ± 0.52 |
| Realism/authenticity | 3.89 ± 0.70 | 4.48 ± 0.49 |
| Usability and ease of interaction | 4.33 ± 0.51 | 4.29 ± 0.55 |
| Perceived fairness of assessment | 4.26 ± 0.57 | 4.30 ± 0.53 |
| Overall acceptability for future use | 4.28 ± 0.56 | 4.35 ± 0.54 |

Note. *Items were rated on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Values represent mean item scores averaged across respondents. Results are presented descriptively to support interpretation of learning outcomes rather than for inferential comparison.*

## Qualitative Findings

Semi-structured interviews were conducted with 15 undergraduate medical students and 8 faculty members to explore experiences with AI-driven virtual patient simulation. Thematic analysis identified four overarching themes, each reflecting perspectives from both groups and highlighting perceived benefits, limitations, and conditions for effective integration. As summarized in Table 4, both students and faculty viewed AI-driven virtual patient simulation as acceptable and educationally valuable, while emphasizing the need for faculty oversight and complementary use alongside standardized patients.

Table 4

**Qualitative Themes on Acceptability and Educational Value of AI-Driven Virtual Patient Simulation**

| Theme | Student perspectives (summary and illustrative quotes) | Faculty perspectives (summary and illustrative quotes) |
|---|---|---|
| **Perceived realism and authenticity** | Students perceived AI virtual patients as clinically coherent and "real enough" to support structured history taking and reasoning, though some noted limited emotional responsiveness compared with human patients.<br><br>*"It didn't feel fake. I could ask questions logically and get proper answers, which helped me think through the case."* (Student)<br><br>*"Sometimes the emotional response was missing, like the patient didn't react the way a real person might."* (Student) | Faculty viewed AI realism as adequate for undergraduate learning objectives and consistency, but insufficient for teaching advanced empathy and nuanced communication.<br><br>*"For structured history taking and reasoning, the AI was quite realistic and consistent."* (Faculty)<br><br>*"It lacks the subtle emotional cues that are important for teaching empathy."* (Faculty) |
| **Psychological safety and learning comfort** | Students reported reduced anxiety and greater comfort interacting with AI, allowing exploration of questioning strategies and mistakes without fear of judgment; some felt this reduced the seriousness of the encounter.<br><br>*"I was more relaxed and could try different questions without worrying about being judged."* (Student)<br><br>*"Because it felt safe, I didn't take it as seriously as a real OSCE sometimes."* (Student) | Faculty recognized psychological safety as a strength for novice learners, while cautioning that AI encounters should not fully replace exposure to real-world performance pressure.<br><br>*"Students seemed more willing to explore their thinking when anxiety was reduced."* (Faculty)<br><br>*"They still need real pressure at some point; AI alone cannot provide that."* (Faculty) |
| **Trust, accuracy, and need for oversight** | Students generally trusted AI-generated clinical information but emphasized verifying responses with faculty or learning resources, expressing caution about blind reliance.<br><br>*"Most of the information matched what we were taught."* (Student)<br><br>*"I wouldn't rely on it fully without checking books or teachers."* (Student) | Faculty expressed conditional trust in AI outputs, emphasizing the importance of faculty oversight, prompt refinement, and curricular alignment to prevent inaccuracies. *"With good prompts and supervision, AI can be a reliable teaching aid."* (Faculty)<br><br>*"Without oversight, subtle inaccuracies could influence learning."* (Faculty) |
| **Complementarity versus replacement** | Students valued AI simulation for preparation, repetition, and confidence-building, but preferred standardized patients for | Faculty strongly supported AI as a complementary tool that can expand simulation capacity and reduce workload, while affirming that standardized patients remain essential |

| Theme | Student perspectives (summary and illustrative quotes) | Faculty perspectives (summary and illustrative quotes) |
|---|---|---|
| | emotionally complex and high-stakes interactions.<br><br>*"AI helped me practice before facing a real standardized patient."* (Student) *"For emotional conversations, I still prefer a real person."* (Student) | for assessment of professionalism and empathy.<br><br>*"AI can expand access to simulation and reduce workload."* (Faculty)<br><br>*"It should complement SPs, not replace them, especially for professionalism."* (Faculty) |

# Discussion

This mixed-methods randomized crossover study examined whether AI-driven virtual patient simulation can support learning outcomes comparable to standardized patient (SP) simulation in undergraduate medical education, addressing a recognized evidence gap in the evaluation of AI-enabled educational tools [14, 22]. Across clinical reasoning (KFPs) and clinical performance (OSCEs), students demonstrated meaningful learning gains following participation in both simulation modalities, reinforcing the educational value of simulation-based learning independent of delivery format [2, 23, 39]. When interpreted through a within-participant crossover framework, the pattern of findings supported educational equivalence rather than modality superiority, consistent with recommended analytic approaches for comparative educational research [21, 25].

Both AI-driven virtual patient encounters and SP encounters were associated with significant pre–post improvements in KFP scores, indicating enhanced clinical reasoning following simulation exposure [24, 40]. Although SP encounters yielded numerically higher post-test scores, the magnitude of difference was modest and post-intervention score distributions overlapped substantially, a pattern commonly observed in crossover educational trials [14, 21]. These findings suggest that AI-based virtual patients can effectively engage learners in key decision-making processes, particularly when scenarios are carefully aligned with curricular objectives and assessment blueprints [22, 41].

Similarly, OSCE outcomes demonstrated that students were able to perform core clinical tasks across both simulation formats, including communication, data gathering, reasoning, and professionalism, reflecting the robustness of simulation-based assessment [1, 23]. While SP encounters showed slightly higher mean OSCE scores in some scenarios requiring nuanced contextual interpretation, overall performance ranges overlapped across modalities, supporting a comparability interpretation rather than a deficit model for AI simulation [14, 42]. This pattern aligns with emerging evidence that AI-driven virtual patients can approximate key elements of performance-based assessment when evaluated using structured, behaviourally anchored tools [13, 43].

The concurrent use of KFPs and OSCEs enabled examination of learning outcomes across complementary domains; clinical reasoning and observable performance recommended for

comprehensive evaluation of complex educational interventions [23]. Convergence of findings across these outcome measures strengthens confidence in internal validity and supports the interpretation of educational equivalence between modalities [20, 25].

Learner satisfaction findings further contextualized performance outcomes, with high ratings for both modalities across domains of usability, fairness, and perceived learning value, and slightly higher perceived realism for SPs [1, 2]. Qualitative data provided explanatory depth, indicating that psychological safety, consistency, and opportunities for repeated practice were valued strengths of AI-based simulation, while emotional realism and interpersonal nuance remained strengths of SP encounters [22, 44]. Triangulation of qualitative insights with KFP, OSCE, and satisfaction results therefore positions AI-driven virtual patients as a psychologically safe and effective preparatory modality that complements, rather than replaces, standardized patient-based learning [14, 45, 46].

From an educational and practical perspective, these findings suggest that AI-driven virtual patient simulation can serve as a viable complementary modality alongside SP simulation, particularly in resource-constrained settings [16, 22]. Institutions in low- and middle-income contexts often face financial, logistical, and workforce limitations that restrict access to high-fidelity SP programs, making scalable AI-based solutions especially attractive [1, 8, 11]. By enabling standardized, repeatable, and accessible simulation experiences, AI-driven virtual patients may contribute to greater equity in access to clinical skills training [22, 47].

Importantly, our findings do not support replacement of SP simulation but instead align with a blended or tiered simulation model, in which AI-based tools are integrated strategically to support preparation, formative practice, and early skills acquisition [1, 2]. This approach reflects contemporary perspectives on technology-enhanced learning that emphasize augmentation of human-centered education rather than substitution [2, 22].

Methodological strengths include the randomized crossover design, which minimizes between-participant variability and enhances internal validity in educational trials [14, 21]. The use of parallel KFP forms and OSCE stations, examiner blinding, and behaviourally anchored scoring further strengthens rigor and credibility [23, 24]. Integration of quantitative and qualitative strands is consistent with best practices for evaluating complex interventions in health professions education [20].

# Implications and Recommendations

AI-driven virtual patients can be effectively integrated into undergraduate medical education as a complementary simulation modality alongside standardized patients. They are particularly useful for early clinical reasoning practice, structured history taking, and repeated formative learning in psychologically safe environments. A blended simulation approach, where AI-based practice precedes SP encounters, may enhance learner preparation and confidence.

Medical schools, especially in resource-constrained settings, should consider incorporating AI-driven simulation to expand access to clinical skills training. Successful implementation requires curricular alignment, expert-designed scenarios, and faculty oversight, along with faculty development to support responsible and effective use of AI in teaching.

Further studies should explore long-term learning outcomes, cost-effectiveness, and multi-institutional implementation of AI-based simulation, as well as the potential of adaptive AI feedback to enhance clinical reasoning and communication skills.

## Conclusion

AI-driven virtual patient simulation produced clinical reasoning, performance, and learner satisfaction outcomes comparable to standardized patient simulation in undergraduate medical education. When evaluated using a rigorous randomized crossover design, both modalities supported meaningful learning gains without educationally meaningful differences. Students and faculty perceived AI-based simulation as acceptable, psychologically safe, and particularly valuable for preparation and formative practice. These findings support a blended simulation approach, in which AI-driven virtual patients complement rather than replace standardized patients, offering a scalable and equitable strategy for strengthening clinical skills training, especially in resource-constrained settings.

## Declarations

### Ethics approval and consent to participate

Ethical approval for this mixed-methods needs assessment study was obtained from the Institute of Health Professions Education & Research, Khyber Medical University Ethics Board (Ref No: 1–13/IHPER/KMU/25–134. Dated: 20-08-2025). The study involved undergraduate medical students and faculty members participating and did not involve real patients or clinical interventions. Participation was voluntary, and confidentiality was maintained throughout the study. Participants were informed about the use of AI in educational simulation, including its capabilities and limitations, in line with emerging ethical guidance for AI in medical education.All procedures were performed in accordance with the ethical principles of the Declaration of Helsinki.

### Consent for publication

Not applicable. This manuscript does not contain any individual person’s data in an identifiable form.

## Competing interests

The authors declare that they have no competing interests.

# Use of Large Language Models and AI Tools

ChatGPT (OpenAI) was used as a supportive tool during manuscript preparation to assist with language refinement, clarity, and organization of text. All content generated with the assistance of the LLM was critically reviewed, edited, and validated by the authors, who retain full responsibility for the accuracy, originality, and integrity of the work.

# Funding

None

# Author Contribution

**BJK** conceptualised the study, collected and analyse data , and drafted the manuscript. **VN** mentored this project and contributed to study design, and analysis. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

# Acknowledgements

The authors would like to acknowledge the students, faculty, and programme leadership who contributed their time and insights to this study.

# Data Availability

The datasets generated and/or analysed during the current study are not publicly available due to institutional policies and the educational context of the data but are available from the corresponding author on reasonable request.

## References

1. Bevis Z, Nestel D, Kumar A, Gibson S, Kavanagh M, Rosado C, Chianáin LN, Battista A. Instruction and guidance in healthcare simulation: a scoping review. 2025.
2. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hamstra SJ. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA. 2011;306(9):978–88.
3. Cleland JA, Abe K, Rethans J-J. The use of simulated patients in medical education: AMEE Guide 42. Med Teach. 2009;31(6):477–86.

4. Cleland J. Simulation-based education: what's it all about? Perspect Med Educ. 2018;7(Suppl 1):30–3.
5. Puri L, Das J, Pai M, Agrawal P, Fitzgerald JE, Kelley E, Kesler S, Mate K, Mohanan M, Okrainec A. Enhancing quality of medical care in low income and middle income countries through simulation-based initiatives: recommendations of the Simnovate Global Health Domain Group. BMJ Simul Technol Enhanced Learn. 2017;3:S15.
6. Irfanullah EA, Chandra A, Solaiman RH, Siems C, Chethan S, Belani K, Harmon J, Siems CB, Harmon JV Jr. Simulation training in a lower middle-income country: Supporting a new center and developing low-cost models for critical skill acquisition. Cureus 2023, 15(6).
7. Bajwa M, Najeeb F, Alnazzawi H, Ayub A, Bell JG, Sadiq F. A Scoping Review of Pakistani Healthcare Simulation: Insights for Lower-Middle-Income Countries. Cureus 2024, 16(12).
8. Khan MA, Asim N, Jamil B, Iqbal M. Perceptions of benefits, challenges, and utility of simulation-based education among medical faculty of Khyber Pakhtunkhwa, Pakistan. Pakistan J Med Sci. 2025;41(11):3201.
9. Alblooshi AS, AlMarzooqi FM, Almansoori TM, Ahmed G, Al-Shamsi S, AlRadini FA. A bibliometric analysis of simulation-based learning in medical education: trends, gaps, and future directions. Front Med. 2026;12:1692991.
10. Khalifa A, Tahhan O, Albazooni M, Saeed M, Hamdi R, Stanners M, Malik A, Malik A. Automated and Artificial Intelligence (AI)-Derived Performance Assessment in Surgical Simulation: A Systematic Review. Cureus 2025, 17(12).
11. Zary A, Zary N. Artificial Intelligence in Technical and Vocational Education and Training: Empirical Evidence, Implementation Challenges, and Future Directions. 2025.
12. Kim T, Pankhania Y. AI-Powered Simulation: Leveraging Technology to Advance Nursing Communication Skills with AI SimBot. J Teach Learn Technol 2025, 14(1).
13. Bowers P, Graydon K, Ryan T, Lau JH, Tomlin D. Artificial intelligence-driven virtual patients for communication skill development in healthcare students: A scoping review. Australasian J Educational Technol 2024.
14. Tyrrell EG, Sandhu SK, Berry K, Ghannam SF, Lewis SA, Crowfoot D, Sahota GS, Carson J, Wilson EE, Taggar J. Web-Based AI-Driven Virtual Patient Simulator Versus Actor-Based Simulation for Teaching Consultation Skills: Multicenter Randomized Crossover Study. JMIR Formative Res. 2025;9:e71667.
15. Harder N, Ali F, Turner S, Workum K, Gillman L. Comparing artificial intelligence-enhanced virtual reality and simulated patient simulations in undergraduate nursing education. Clin Simul Nurs. 2025;105:101780.
16. Meng L, Liu X, Ni J, Shen P, Jiao F. An investigation for the efficacy of teaching model of combining virtual simulation and real experiment for clinical microbiology examination. Front Med. 2024;11:1255088.

17. Palaganas J, Bajwa M, Patel A, Morton A, Danaher-Garcia N, Gross I. The State of Artificial Intelligence in Healthcare Simulation: A 2024 International Survey. Cureus J Comput Sci. 2025;2(1):3491.
18. Domrös-Zoungrana D, Rajaeean N, Boie S, Fröling E, Lenz C. Medical education: considerations for a successful integration of learning with and learning about AI. J Med Educ Curric Dev. 2024;11:23821205241284719.
19. Mansourzadeh A, Rasouli S. The Future of Medical Education: A Review of the Opportunities and Challenges of Artificial Intelligence Integration. Med Educ Bull. 2024;5(2):973–82.
20. Creswell JW, Plano Clark V. Revisiting mixed methods research designs twenty years later. Handb mixed methods Res designs. 2023;1(1):21–36.
21. Senn SS. Cross-over trials in clinical research. Wiley; 2002.
22. Masters K, MacNeil H, Benjamin J, Carver T, Nemethy K, Valanci-Aroesty S, Taylor DC, Thoma B, Thesen T. Artificial intelligence in health professions education assessment: AMEE guide 178. Med Teach 2025:1–15.
23. Harden RM, Lilley P, Patricio M. The definitive guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment. Elsevier Health Sciences; 2015.
24. Hrynchak P, Glover Takahashi S, Nayer M. Key-feature questions for assessment of clinical reasoning: a literature review. Med Educ. 2014;48(9):870–83.
25. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. J graduate Med Educ. 2012;4(3):279–82.
26. Braun V, Clarke V. Toward good practice in thematic analysis: Avoiding common problems and be (com) ing a knowing researcher. Int J transgender health. 2023;24(1):1–6.
27. Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. Acad Med. 1995;70(3):194–201.
28. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 2005;39(12):1188–94.
29. Frank JR, Karpinski J, Sherbino J, Snell LS, Atkinson A, Oswald A, Hall AK, Cooke L, Dojeiji S, Richardson D. Competence By Design: a transformational national model of time-variable competency-based postgraduate medical education. Perspect Med Educ. 2024;13(1):201.
30. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–7.
31. Cizek GJ, Bunch MB. Standard setting: A guide to establishing and evaluating performance standards on tests. SAGE Publications Ltd; 2007.
32. Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, Hays R, Palacios Mackay MF, Roberts T, Swanson D. 2018 Consensus framework for good assessment. Med Teach. 2018;40(11):1102–9.
33. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53.

34. DeVellis RF, Thorpe CT. Scale development: Theory and applications. Sage; 2021.
35. Sridharan K, Sivaramakrishnan G. Angoff methods in standard setting in health professional education: a systematic review and meta-analysis. BMC Med Educ. 2025;25(1):1727.
36. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50(9):741.
37. Nyimbili F, Nyimbili L. Types of purposive sampling techniques with their examples and application in qualitative research studies. 2024.
38. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res. 2016;26(13):1753–60.
39. Harden RM, Lilley P. The eight roles of the medical teacher: the purpose and function of a teacher in the healthcare professions. Elsevier Health Sciences; 2018.
40. Nayer M, Glover Takahashi S, Hrynchak P. Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning. Médical teacher. 2018;40(11):1116–22.
41. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: A meta-analysis. Rev Educ Res. 2020;90(4):499–541.
42. Liaw SY, Tan JZ, Bin Rusli KD, Ratan R, Zhou W, Lim S, Lau TC, Seah B, Chua WL. Artificial intelligence versus human-controlled doctor in virtual reality simulation for sepsis team training: randomized controlled study. J Med Internet Res. 2023;25:e47748.
43. Zambrano-Serrano B, Cueto-Galán R, Escamilla-Sánchez A, del Rocio Ramirez-Rodríguez M, Sanchez-Corrales N, Fontalba-Navas A. AI-Driven Virtual Patient Simulations for Communication Training in Medical Students: A Pilot Study. 2025.
44. Zhu Y. Revolutionizing simulation-based clinical training with AI: Integrating FASSLING for enhanced emotional intelligence and therapeutic competency in clinical psychology education. J Clin Technol Theory|. Vol 2025;2:39.
45. Hamilton A. Artificial intelligence and healthcare simulation: the shifting landscape of medical education. Cureus 2024, 16(5).
46. Mateusz M, Jakub M, Anna Ż, Ameen N, Tomasz T. Artificial intelligence in medical education: A narrative review. AI. 2025;6(12):322.
47. Mohsin S, Jamil B, Khan KI, Virk S, Ahmad AMR. Learning effectiveness of simulation based teaching in Gynaecology and Obstetrics among medical students: a mixed-methods study. Front Med. 2025;12:1652105.

# Figures

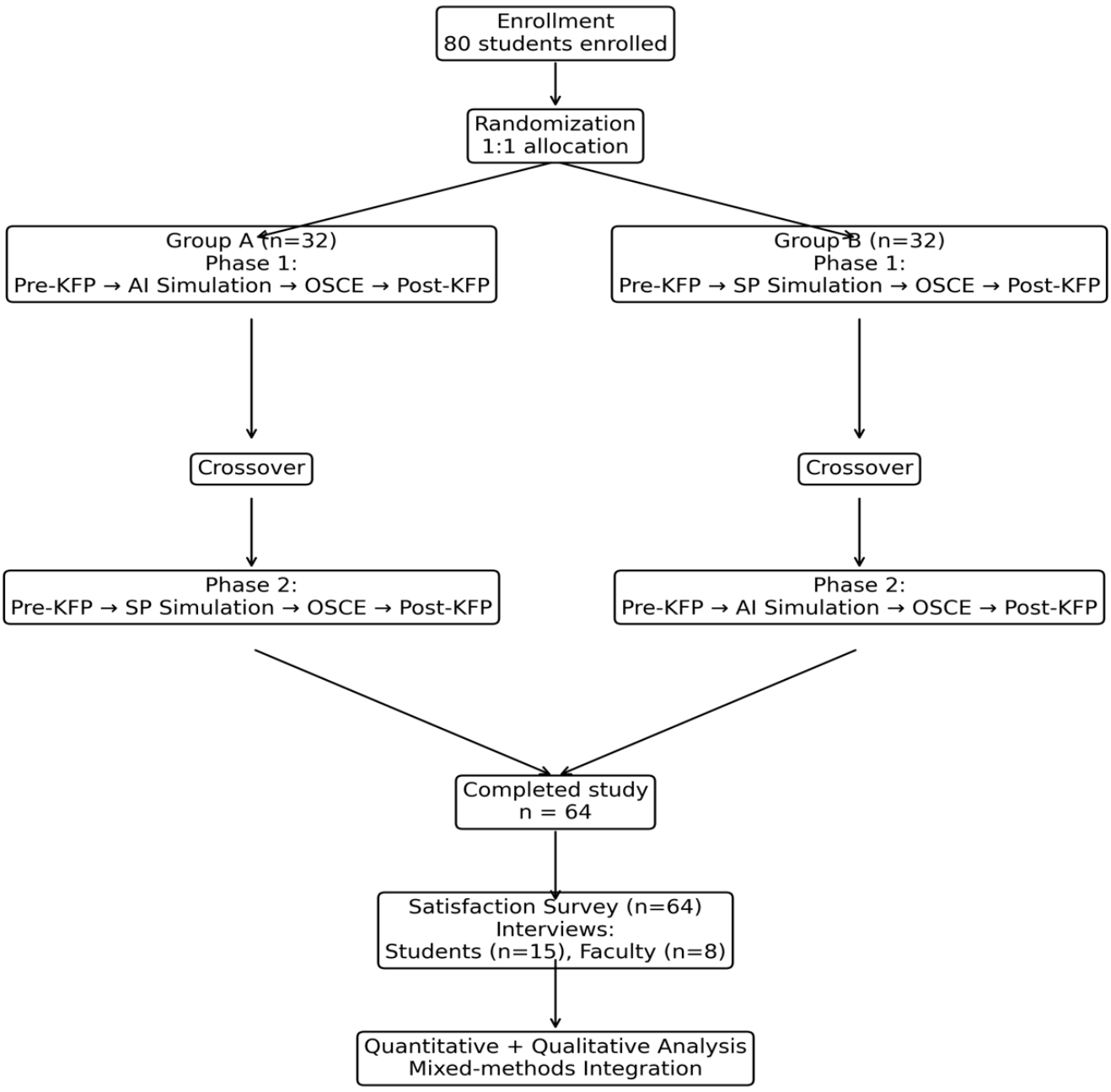



**Figure 1**

**Study design and participant flow**

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.docx
- SupplementaryFile2.docx
- SupplementaryFile3.docx
- SupplementaryFile4.docx
- SupplmenetaryFile5.docx
- SupplementaryFile6.docx