

Scaling multimodal agentic AI in medical education: a multi-site cross-sectional study of simulation effectiveness in primary care

Chris Jacobs, Hans Johnson, Kirsty Brownlie, Richard Joiner, Trevor Thompson

Submitted to: JMIR Formative Research
on: December 17, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 24

 Figures 25

 Figure 1..... 26

 Figure 2..... 27

 Figure 3..... 28

 Figure 4..... 29

 Figure 5..... 30

CONSORT (or other) checklists..... 31

 CONSORT (or other) checklist 0..... 32

TOC/Feature image for homepages 33

 TOC/Feature image for homepage 0..... 34

Scaling multimodal agentic AI in medical education: a multi-site cross-sectional study of simulation effectiveness in primary care

Chris Jacobs^{1, 2} MBBCHIR, MRES, MD; Hans Johnson³ MBChB, MRES; Kirsty Brownlie³ MBChB; Richard Joiner¹ MSc, PhD; Trevor Thompson³ MBBS, MSc, PhD

¹Department of Psychology University of Bath Bath GB

²Faculty of Life Sciences & Medicine King's College London London GB

³Medical School University of Bristol Bristol GB

Corresponding Author:

Chris Jacobs MBBCHIR, MRES, MD
Department of Psychology
University of Bath
Claverton Down
Bath
GB

Abstract

Background: Conversational artificial intelligence (AI) systems offer potential solutions to traditional constraints in medical consultation skills training, including high costs, scheduling difficulties, and varied standardisation.

Objective: There is limited evidence evaluating medical professionals' perceptions of AI generated patient interactions across multiple fidelity dimensions and assess the educational value of conversational AI for consultation skills training.

Methods: Cross-sectional evaluation study at a UK medical school involving 179 participants (medical students and general practitioners) at 70 sites. Participants completed standardised clinical scenarios using SimFlow.ai conversational AI system, followed by structured questionnaires evaluating AI Realism, Medical Content, Educational Value, Feedback, and Usability domains. Data were analysed using descriptive statistics and non-parametric tests to assess domain performance and participant characteristics.

Results: Medical Content received highest ratings (median 4.0, IQR 1.0 [4.0-5.0]), with 97.9% rating clinical plausibility highly. Educational Value was rated positively (IQR 1.0-2.0 [3.0-4.0]) though AI Realism received moderate scores (median 3.0, IQR 2.0 [2.0-4.0]). Participants with prior AI experience gave significantly higher ratings for AI Realism than those without prior experience (mean 3.81 vs 3.07, $p < 0.05$). Qualitative analysis revealed four themes: clinical authenticity, interactional limitations, educational potential, and implementation considerations.

Conclusions: Conversational AI demonstrates strong capabilities in functional fidelity (clinical accuracy) despite limitations in conversational fidelity (realism). The technology shows promise for supplementary clinical skills training rather than higher-stakes assessment, with future development needed in dialogue naturalness and feedback capabilities. Clinical Trial: <https://doi.org/10.17605/OSF.IO/NUGBD>
Not a clinical trial

(JMIR Preprints 17/12/2025:88905)

DOI: <https://doi.org/10.2196/preprints.88905>

Preprint Settings

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

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

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

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

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

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#), my accepted manuscript PDF will be available to all.

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in [JMIR Publications](#), my accepted manuscript PDF will be available to all.



Original Manuscript

Scaling multimodal agentic AI in medical education: a multi-site cross-sectional study of simulation effectiveness in primary care

Abstract

Introduction: Conversational artificial intelligence (AI) systems offer potential solutions to traditional constraints in medical consultation skills training, including high costs, scheduling difficulties, and varied standardisation. There is limited evidence evaluating medical professionals' perceptions of AI generated patient interactions across multiple fidelity dimensions and assess the educational value of conversational AI for consultation skills training.

Methods: Cross-sectional evaluation study at a UK medical school involving 179 participants (medical students and general practitioners) at 70 sites. Participants completed standardised clinical scenarios using SimFlow.ai conversational AI system, followed by structured questionnaires evaluating AI Realism, Medical Content, Educational Value, Feedback, and Usability domains. Data were analysed using descriptive statistics and non-parametric tests to assess domain performance and participant characteristics.

Results: Medical Content received highest ratings (median 4.0, IQR 1.0 [4.0-5.0]), with 97.9% rating clinical plausibility highly. Educational Value was rated positively (IQR 1.0-2.0 [3.0-4.0]) though AI Realism received moderate scores (median 3.0, IQR 2.0 [2.0-4.0]). Participants with prior AI experience gave significantly higher ratings for AI Realism than those without prior experience (mean 3.81 vs 3.07, $p < 0.05$). Qualitative analysis revealed four themes: clinical authenticity, interactional limitations, educational potential, and implementation considerations.

Conclusion: Conversational AI demonstrates strong capabilities in functional fidelity (clinical accuracy) despite limitations in conversational fidelity (realism). The technology shows promise for supplementary clinical skills training rather than higher-stakes assessment, with future development needed in dialogue naturalness and feedback capabilities.

Introduction

Artificial intelligence (AI) is widely regarded as the suite of technologies that allow computers to perform tasks that resemble human abilities, including pattern recognition, decision-making, and content generation [1]. It has shown considerable promise in healthcare applications over recent years, with prospects for continued development in medical training contexts [2]. The integration of AI into medical education represents a significant development in healthcare training approaches, as traditional consultation skills training has long relied on human standardized patients and role-playing exercises. While these conventional methods remain valuable, they face notable constraints including high costs, scheduling difficulties, and limited standardization across encounters [3].

Large language models (LLMs) and conversational AI systems have demonstrated notable capabilities across healthcare applications. These systems have shown the ability to pass standardized medical examinations such as the USMLE with performance levels approaching human benchmarks [4]. More sophisticated domain-specific models like Med-PaLM2 have achieved expert-level performance on complex medical questions, indicating their potential for specialized healthcare applications [5,6]. These capabilities translate effectively into virtual patient simulations, where LLM sourced avatars can provide authentic and professional patient-scenario dialogue flows that enable realistic clinical interactions through large data resourced models [7]. Multimodal technologies have quickly updated from basic text generation to systems capable of processing voice, text, and visual inputs.

Conversational AI utilising multimodal agentic systems designed for medical training offer several advantages over traditional methods of simulating patient interactions. They provide continuous availability, enabling students to practice consultation skills without the logistical constraints of coordinating with human actors. These systems can deliver standardized case presentations, ensuring learners encounter consistent scenarios and receive uniform training experiences. The scalability allows multiple simultaneous interactions, addressing resource limitations that have historically restricted access to hands-on practice opportunities [8]. Recent studies have demonstrated the practical application of these technologies, with research showing that medical students can effectively use readily available AI tools like ChatGPT-4o Advanced Voice Mode for challenging communication skills practice, finding the technology both useful and acceptable for supplementing traditional training methods [9]. Models can be developed for feedback that provides insights into clinical reasoning, language use, and response effectiveness [10]. Furthermore, AI created feedback offers potential advantages over human supervisor feedback by providing detailed, criterion-based assessments with multiple verbatim quotations from transcribed clinical interactions, enabling personalised learning conversations [11].

Simulation fidelity encompasses multiple dimensions beyond visual and auditory realism, including psychological fidelity and functional fidelity that support specific learning objectives [12]. Understanding these different dimensions of fidelity is important for developing effective AI based training systems, as research suggests that functional fidelity, the degree to which a simulation evokes targeted cognitive processes, may be more significant for learning outcomes than perfect physical or conversational realism [13]. The conceptual framework for how users interact with immersive technologies encompasses multiple interconnected domains including engagement, system usability, and cognitive processes that together influence the overall user experience and educational effectiveness [14].

Recent work has explored how AI can benefit simulation-based training through advanced assessment capabilities and adaptive learning approaches [15,16]. Current research on AI in medical education has primarily focused on knowledge assessment and content generation rather than interpersonal skills training through experiential learning approaches, though systematic reviews indicate growing evidence of AI's broader educational impact across health professions [17,18].

While studies have examined medical students' interactions with AI tutoring systems and assessment tools, there is growing recognition of AI's potential for supporting communication skills training in healthcare education, though limited research has evaluated how different dimensions of simulation fidelity impact learning outcomes in conversational AI applications for consultation skills development [19]. Training effectiveness and optimising learning outcomes are becoming apparent, though questions remain about their readiness for widespread implementation [20].

The specific aims of this research were to:

1. Evaluate how medical professionals perceive the realism and clinical authenticity of AI generated patient interactions across multiple fidelity dimensions, distinguishing between functional fidelity (clinical accuracy and medical content) and conversational fidelity (realism and flow)
2. Assess the educational value of conversational AI for experiential learning in consultation skills training
3. Identify factors that influence user acceptance and perceived simulation effectiveness, particularly regarding the relationship between technical realism and functional educational outcomes
4. Determine current readiness and implementation considerations for AI-constructed patient simulation in medical education curricula

Methods

Study design and participants

We conducted a cross-sectional evaluation study of a conversational AI system designed for patient consultation training. Participants were recruited from a single UK university medical school incorporating 70 general practice sites. All medical students in their 3rd year of study and general practitioners within these settings were invited to participate via email invitation during March 2025. A retrospective protocol for this study, including the full methodology, questionnaire, and analysis plan, was registered on the Open Science Framework prior to journal submission.

From the total eligible population (n=305), included grouped and individual participant response (n=51), the estimated total individual participation was 179 (adjusted response rate of 61.3%).

Procedure

Participants completed a standardised clinical scenario with the conversational artificial intelligence system during a scheduled primary care training day. The simulation was delivered through the SimFlow.ai web-hosted platform (figure 1), which operates behind institutional firewall protection and uses a single-stage authentication process. The system is powered by a large language model based on OpenAI technology, enabling naturalistic conversational exchanges throughout the consultation. A synthetic voice engine generated a realistic, age-appropriate female voice matched to the demographic profile of the virtual patient. The AI functioned in a one-way configuration in which user input was neither stored nor transmitted externally, ensuring that interactions remained contained within the protected hosting environment.

Figure 1 Screenshot of SimFlow.ai web-hosted platform. Patient avatar image AI generated.

3

Simulation



Patient Info: Jane Cohen, a 65 year old woman, has booked an appointment at Riverside Health Centre, her local GP practice in Bristol. She lives at 34 Pembroke Crescent, Clifton, Bristol, BS8 4QH and was born on 12th January 1960.

Task: You should gather a focused history about the patient's symptoms. After 6 minutes, the tutor will provide you with the examination findings and urinalysis. You should then explain to the patient what you have found and the possible reasons for their symptoms, and provide options for next steps in order to come to a shared management plan.

[Begin Simulation](#)

More than seventy participating teaching practices accessed the simulation simultaneously during the same morning session. Each site undertook the same scenario, which involved a general practice consultation with a woman in her sixties presenting with a first episode of visible haematuria. Learners were instructed to conduct the consultation as they would in routine clinical practice, including eliciting the history, identifying relevant red-flag symptoms and discussing appropriate investigation and management options.

The prompt design underlying the AI consultation was developed from a summarised version of the case-based learning materials used in this teaching session. These materials were mapped closely to the intended learning outcomes, ensuring that the virtual patient's history, symptom cues and conversational structure aligned with established clinical reasoning expectations and supported the educational aims of the activity.

Prior to deployment, the scenario underwent extensive face-validity testing involving practising clinicians, medical educators and individuals experienced in simulation design. Feedback from this process led to iterative improvements to the underlying prompt, producing clearer symptom cues, greater clinical accuracy and more coherent conversational flow. Supporting materials for students and facilitators, including case notes, learning points and guidance for debriefing, were also reviewed to ensure accuracy and alignment with the curriculum.

Members of the academic team were present across participating sites during the session to provide support and oversee implementation. No safety concerns were reported, and there were no observed instances of inappropriate or clinically unsafe hallucinations from the AI system.

Data collection

Following the consultation, participants completed a structured questionnaire evaluating the system across five domains (online supplementary file 1): AI Realism (6 items), Medical Content (2 items), Educational Value (3 items), Feedback (2 items), and Usability (2 items). Each item was rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree). Three open-ended questions captured qualitative feedback on the system's most realistic aspects, areas for improvement, and suitability for assessment purposes such as OSCEs. Domains relating to engagement and system usability abridged from immersive technology in medical education measure (ITEM) [21]. Furthermore, working group pilot tested for construct validity check. The overall questionnaire demonstrated excellent internal consistency (Cronbach's $\alpha = 0.92$). Factor analysis supported the five-domain structure, with internal consistency ranging from good (User Experience: $\alpha = 0.753$) to excellent (Feedback: $\alpha = 0.954$, Learning Value: $\alpha = 0.935$), confirming the validity of the theoretical framework for evaluating conversational AI acceptability in medical education. Survey methods are reported using the CROSS checklist [22] (supplementary file).

Demographic information collected included professional role, years of clinical experience, and prior experience with conversational AI systems for medical training.

Data analysis plan

Quantitative data analysis was performed on 47 unique questionnaire responses, representing a mix of individual submissions and group consensus submissions. Group responses were treated as single data points representing the consensus of the training site. Data cleaning and statistical analysis were conducted using R statistical software. Internal consistency of the questionnaire domains was assessed using Cronbach alpha. One-sample Wilcoxon signed-rank tests were used to determine if domain scores differed significantly from a neutral median of 3.0. Mann-Whitney U tests were used to evaluate subgroup differences based on professional role and prior AI experience. All statistical tests were two-tailed, and a P value less than .05 was considered statistically significant.

Qualitative analysis of open-ended survey responses was conducted using reflexive thematic analysis following the framework of Braun and Clarke [23]. Two researchers independently familiarised themselves with the data and generated initial codes. These codes were iteratively reviewed and collated into overarching themes regarding clinical authenticity, interactional limitations, and educational potential. Qualitative findings were triangulated with quantitative results to provide context for statistical trends, specifically regarding the divergence between medical content accuracy and conversational realism.

Patient and public involvement

Healthcare professionals and students were interviewed in the design process of ITEM survey [14,24]. During the evaluation patients' in one primary care practice were asked on their view on this training method. This revealed no objections to using agentic AI to simulate conversations. Medical actors (n=34) were consulted about using AI in simulation with 88% positive to be involved in future researching of agentic AI.

Ethical considerations

This study was approved by Swindon Academy Medical Education and Ethics Committee IRB (CJ062023). All participants provided voluntary consent to take part after receiving a comprehensive explanation of the study. No compensation was provided for participating in the study

Results

Participant characteristics

From the eligible population of 305 healthcare professionals (245 medical students, 60 general practitioners), the adjusted response rate was 61.3% (179/305). The final sample included evaluations from medical students (57.4%), general practitioners (38.3%), and specialist doctors (1/47, 2.1%). Most participants had clinical experience of either 1-2 years (36.2%) or 11+ years (29.8%), with fewer in intermediate experience categories (3-5 years: 21.3%; 5-10 years 12.8%). Only 12.8% reported prior experience with conversational AI for medical training.

Quantitative findings

Domain-level evaluation

Table 1 presents descriptive statistics for the five evaluation domains (additionally in figure 2). Medical Content received the highest ratings (median 4.0, IQR 1.0 [4.0-5.0]), with nearly all participants (97.9%) rating the medical plausibility of the AI system highly. Educational Value was also rated positively (median 4.0, IQR 1.0-2.0 [2.0-4.0]), though with greater variability in responses. AI Realism received moderate ratings (median 3.0, IQR 1.0-2.0 [2.0-4.0]), with voice quality (median 4.0) scoring higher than other realism aspects (median 3.0). Feedback capabilities received the lowest consistent ratings (median 3.0, IQR 0.0-1.0 [2.0-3.0]).

Based on individual-level concordance analysis, 46.8% of participants demonstrated high or very high agreement with the group findings (≥ 3 exact domain rank matches), with an additional 25.5% showing moderate concordance (2 exact matches), and an average Spearman correlation of 0.685 (moderate to strong agreement), see figure 3.

Table 1 Domain level evaluation scores

Domain	Median	IQR	Range
Medical Content	4.0	1.0 [4.0-5.0]	3-5
Learning Value	4.0	1.0 [3.0-4.0]	1-5
User Experience (usability)	4.0	2.0 [3.0-5.0]	2-5
AI Realism	3.0	2.0 [2.0-4.0]	1-5
Feedback	3.0	1.0 [2.0-3.0]	1-5

Figure 2 Domain level question responses with median and interquartile ranges

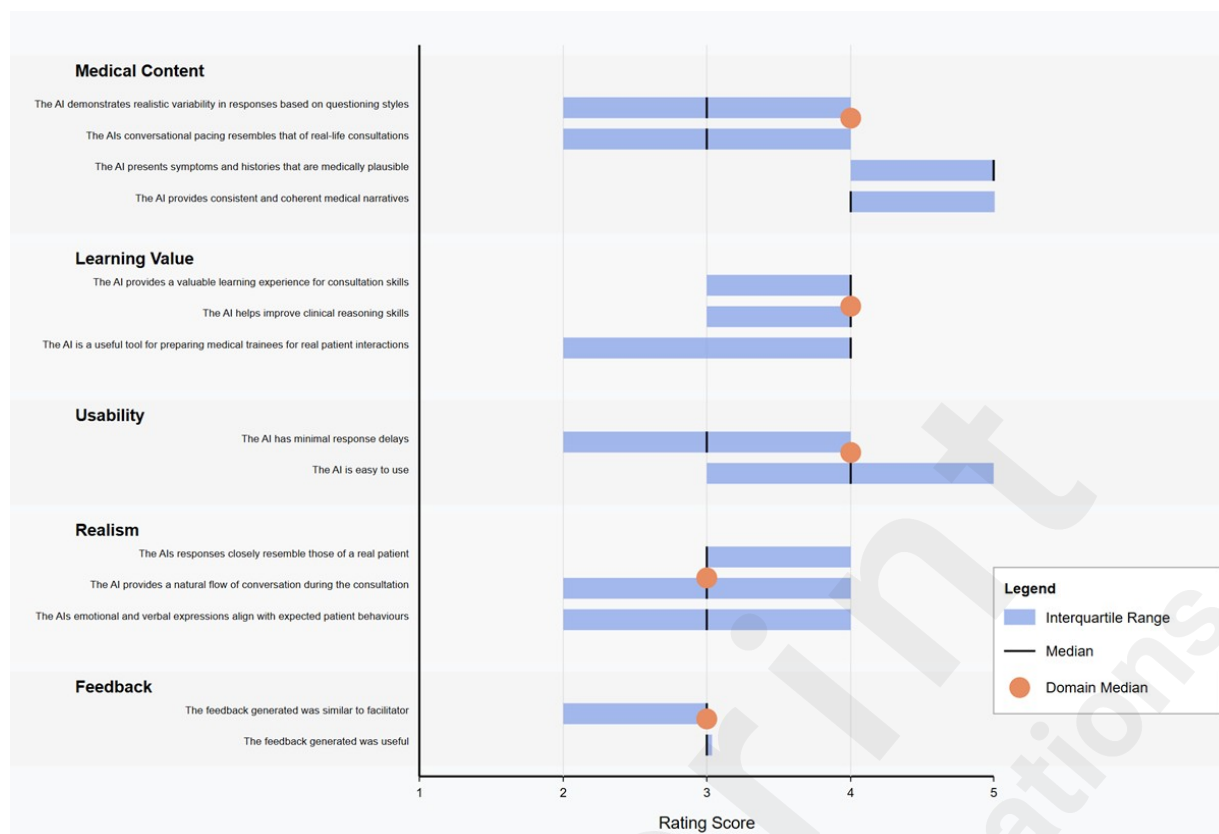
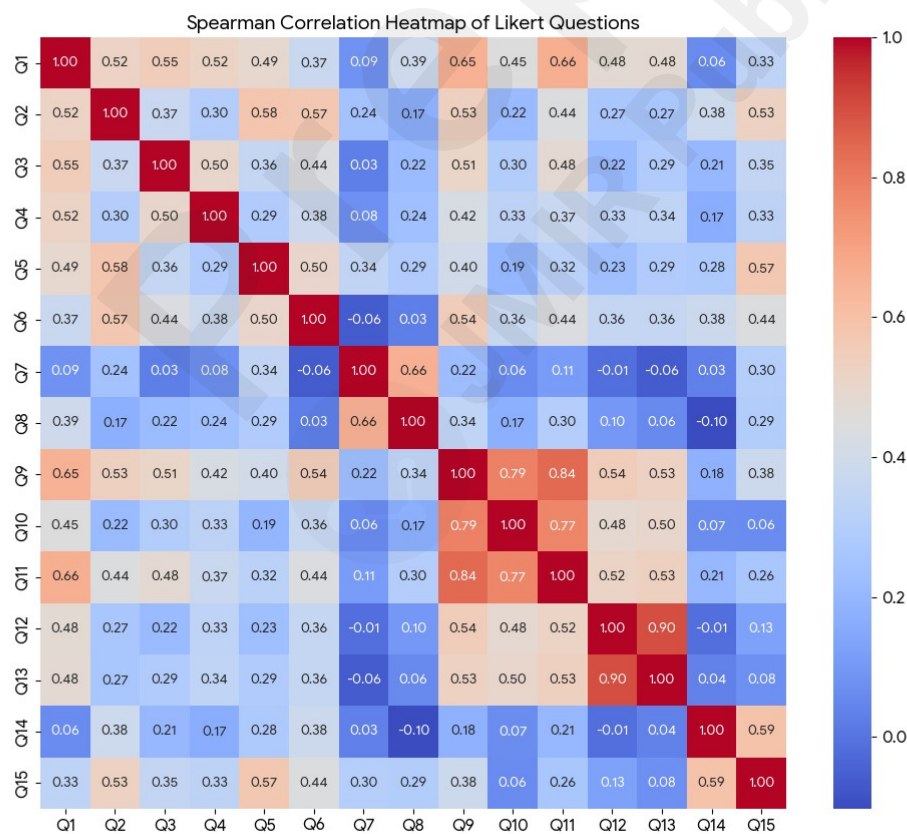


Figure 3 Heat map indicating the correlation between questions



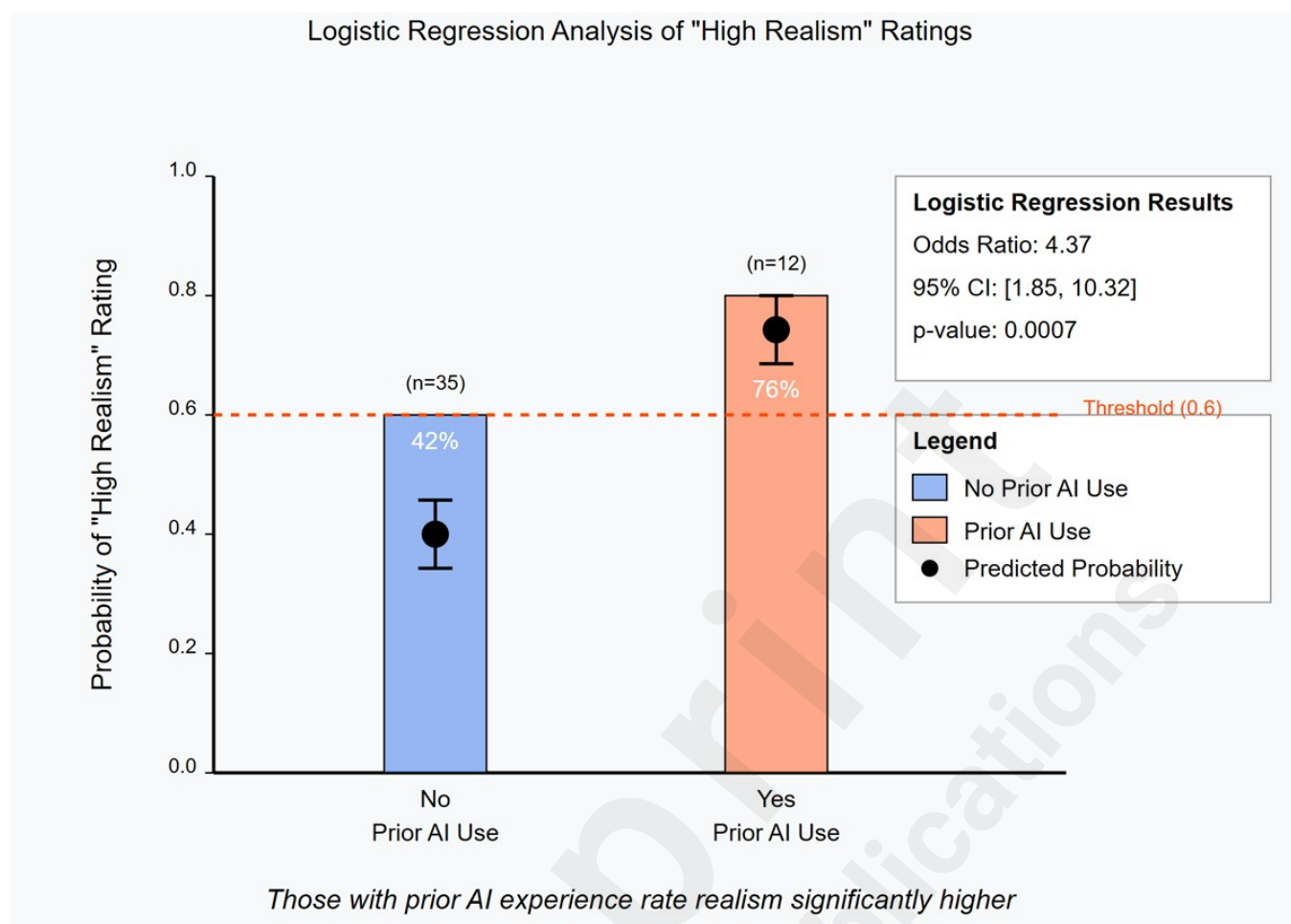
Effect of participant characteristics

Prior AI experience “expectation effect” was associated with higher ratings for AI Realism when analysing mean scores (with experience: 3.81, without experience: 3.07, difference: 0.74, $t=2.84$, $p<0.05$) as seen in Figure 4. In the odds ratio analysis, participants with prior AI experience were more likely to give high ratings (≥ 4) for AI Realism (OR 3.60, 95% CI 0.52-25.00) and User Experience (OR 2.56, 95% CI 0.42-15.55), though these confidence intervals included 1.0 (Table 2). Clinical experience of ≥ 5 years was associated with higher ratings for AI Realism (OR 4.17, 95% CI 0.72-24.23) and Feedback quality (OR 2.67, 95% CI 0.55-12.82), but these associations were not statistically significant. Group versus individual questionnaire completion showed no significant impact on domain scores (all $p>0.05$), with the largest difference observed for Educational Value (group mean 3.33, individual mean 3.69, difference -0.36, $t=1.11$).

Table 2 Impact of prior AI experience on domain scores

Domain	High ratings with AI experience	High ratings without AI experience	Odds ratio (95% CI)
AI Realism	2/6 (33.3%)	5/41 (12.2%)	3.60 (0.52-25.00)
User Experience	4/6 (66.7%)	18/41 (43.9%)	2.56 (0.42-15.55)
Educational Value	3/6 (50.0%)	20/41 (48.8%)	1.05 (0.19-5.83)
Feedback	1/6 (16.7%)	7/41 (17.1%)	0.97 (0.10-9.65)
Medical Content	6/6 (100.0%)	40/41 (97.6%)	0.30 (0.00-Infinity)

Figure 4 Logistic regression of higher realism associated with prior AI use



Question-level analysis

The highest-rated individual item was "The AI presents symptoms and histories that are medically plausible" (median 5.0, IQR 1.0 [4.0-5.0]), followed by "The AI provides consistent and coherent medical narratives" (median 4.0, IQR 1.0 [4.0-5.0]). "Is easy to use" also received high ratings (median 4.0, IQR 2.0 [3.0-5.0]). The lowest-rated items were "The feedback generated was similar to facilitator" (median 3.0, IQR 1.0 [2.0-3.0]) and "has minimal response delays" (median 3.0, IQR 2.0 [2.0-4.0]). Table 3 and figure 2 summarises these item results against a neutral baseline.

Table 3 Participant ratings of conversational AI fidelity and utility (N=47) *Statistical significance was determined using a one-sample Wilcoxon signed-rank test against a neutral median of 3.0.*

Domain / Item	Median (IQR)	P Value
Medical Content (Functional Fidelity)		
Medically plausible symptoms (Q7)	5.0 (4.0-5.0)	<.001
Consistent/coherent narratives (Q8)	4.0 (4.0-5.0)	<.001
AI Realism (Conversational Fidelity)		

Voice sounds natural (Q5)	4.0 (3.0-4.0)	.001
Responses resemble real patient (Q1)	3.0 (3.0-4.0)	.43
Natural flow of conversation (Q2)	3.0 (2.0-4.0)	.76
Emotional/verbal alignment (Q3)	3.0 (2.0-4.0)	.58
Realistic response variability (Q4)	3.0 (2.5-4.0)	.33
Conversational pacing (Q6)	3.0 (2.0-4.0)	.89
Educational Value		
Valuable learning experience (Q9)	4.0 (3.0-4.0)	.01
Improves clinical reasoning (Q10)	4.0 (3.0-4.0)	.02
Useful preparation tool (Q11)	4.0 (2.0-4.0)	.11
Feedback & Usability		
Easy to use (Q15)	4.0 (3.5-5.0)	<.001
Feedback similar to facilitator (Q12)	3.0 (2.25-3.0)	.17
Feedback was useful (Q13)	3.0 (3.0-3.0)	.21

Qualitative findings

Thematic analysis of open-ended responses revealed four key themes: clinical authenticity, interactional limitations, educational potential, and implementation considerations. Figure 5 summarises the key themes and subthemes related to response.

Clinical authenticity

This clinical reliability was frequently noted as the system's greatest strength, even when other aspects were criticised.

“Mannerisms sounded human, plausible responses, given good history. Responding well to questions asked. Patient sounded appreciative and positive.” (P18)

Many participants observed that the authenticity of clinical content partially compensated for conversational limitations, suggesting that strong medical accuracy might be more important than perfect conversational flow for educational applications.

Interactional limitations

Participants frequently identified unnatural conversational patterns that diminished the realism of the consultation experience.

"The AI answered almost every question with a question which disrupted the flow of the consultation. Almost every answer was followed by 'is that what you'd expect'/ 'does that rule anything out' which became a bit repetitive and did not feel natural." (P1)

Educational potential

Despite limitations, participants recognised significant training value, particularly for early-stage learners who need practice with basic consultation skills and clinical reasoning before engaging with real patients. The system's ability to provide consistent scenarios with reliable clinical presentations was seen as valuable for skill development.

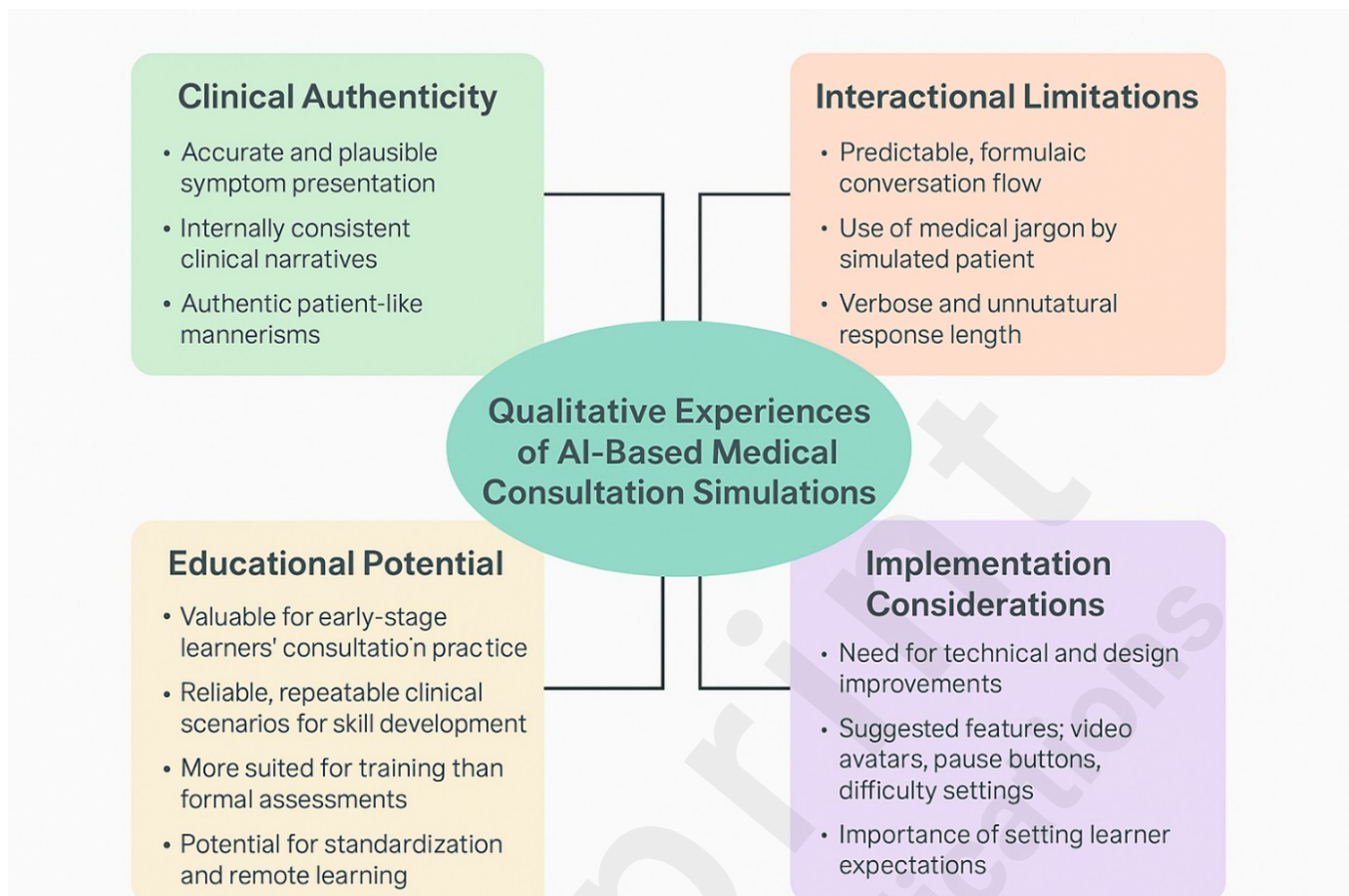
" Consultation skills - her answers were life like, as if talking to a real patient." (P25)

Implementation considerations

Participants identified specific technical and design improvements needed before wider deployment, focusing on both technological limitations and pedagogical design elements that would enhance the educational experience.

" 1. Monologue 2. Information given too readily 3. Knew what question would be 4. Less prompt based and more response based." (P22)

Figure 5 Thematic map



Triangulation of findings

Quantitative and qualitative findings showed strong convergence. The high ratings for medical content aligned with qualitative praise for clinical plausibility. Similarly, the moderate ratings for conversational aspects corresponded with qualitative feedback about unnatural dialogue patterns. The statistical trend showing that participants with prior AI experience gave higher realism ratings aligns with qualitative comments suggesting expectations influence perception of the system's capabilities.

Discussion

Principal findings

The system demonstrated high functional fidelity, with "Medical Plausibility" scoring significantly above the neutral baseline (Median 5.0, $P < .001$). This confirms that current LLM-based simulations can effectively replicate the "medical logic" of a patient encounter. However, conversational fidelity (Realism) remained statistically neutral (Median 3.0). Participants did not actively reject the realism but perceived it as average, supporting the hypothesis that while the voice was excellent (Median 4.0), the emotional and conversational flow lagged behind.

Qualitative analysis identified four key themes: clinical authenticity, interactional limitations, educational potential, and implementation considerations. Participants consistently praised the system's medical knowledge base and clinically accurate presentations, which partially compensated for conversational limitations.

Notably, prior experience with AI technology significantly influenced perception of system realism ($p < 0.05$), with experienced users rating realism higher (mean 3.81) than those without prior exposure (mean 3.07), highlighting a potential expectation effect.

Comparison with other studies

Our findings align with previous research on AI virtual patient simulation in healthcare education. Systematic reviews of large language model applications in medical training have similarly found that clinical accuracy and educational utility were rated more favourably than conversational ability with relationship to human interaction [25]. The challenge of creating natural dialogue patterns in AI systems has been identified as a technical hurdle despite advances in foundation models [5,26].

The high ratings for medical content accuracy in our study correspond with findings from performance analyses of AI in adopting large data sets, which have documented capabilities in representing detailed information while highlighting limitations in naturalistic conversation [27]. Our findings regarding the high ratings for clinical content accuracy despite lower ratings for conversational realism reinforce important principles about simulation fidelity in healthcare education. Effective simulation doesn't require faithful replication of reality, but rather accurate representation of essential cues and stimuli [12]. In our study, participants clearly distinguished between clinical plausibility and conversational authenticity, yet still found significant educational value in the AI simulation. This supports that learning outcomes can be achieved without perfect fidelity in all dimensions.[28,29] Aspects of fidelity significantly hinge on the learners' perceived realism of the context of the learning episode as opposed to any one particular element.

LLM can perpetuate and amplify existing biases in datasets, which may explain our participants' observations of the AI's tendency toward medically precise but conversationally unnatural responses. Research shows that LLMs are primarily optimised for generating accurate informational content rather than simulating natural human conversation patterns [30]. This optimisation bias helps explain our qualitative findings where participants noted the AI system would often answer questions with questions and provide verbosity in responses when simpler answers would be more natural. The inclination of AI to provide excessively detailed information reflects what describes as a fundamental characteristic of current LLM as they provide comprehensive information delivery over conversational naturalness. For medical education specifically, our participants' tolerance of these conversational limitations suggests that the educational value of accurate clinical content may outweigh the need for perfect conversational simulation. In a pilot study similar conversational limitations in their VP system were noted, including unusually formal language, repetitive answers, excessively long responses, and consistently polite demeanor. Latency is an ongoing limitation with current AI technology that incorporates human like voice [31], which is due to multiple systems working in tandem to provide the necessary computational steps from text transcription to voice creation [32,33].

There has been significant progress in LLM ability to accurately assess a wide range of clinical material [34-36], furthermore, the computational capacity exists to analyse the conversation transcription against a chosen framework or rubric [37]. Our findings on conversational AI feedback capabilities align with research on AI generated educational feedback. In a study using a custom prompt for GPT-4 demonstrated how LLMs can be specifically tailored to educational objectives when guided by established pedagogical frameworks [38]. Similar to our results showing moderate feedback ratings (median 3.0, IQR 1.0 [2.0-3.0]), their study revealed variability in AI performance across different feedback dimensions. This suggests that while current AI systems can successfully incorporate elements of effective feedback principles, but they still require refinement in consistently providing high-quality information about learner performance.

The "expectation effect" we observed, where prior AI experience influenced system perception, which can be interpreted as technological readiness [39]. This phenomena may have influenced the perception of the conversation AI. In particular, our findings suggest improved realism with prior experience. Users with prior AI exposure develop more realistic expectations about system capabilities and limitations, leading them to evaluate performance relative to known technological constraints rather than human interaction standards [29].

Implications for training

Our findings suggest several important implications for medical education. Conversational AI systems demonstrate particular promise as supplementary tools for early-stage communication skills training, offering consistent, accessible practice opportunities both prior and as an adjunct to patient contact. The high ratings for medical content and educational value indicate that current technology can provide clinically valid and effective simulations despite conversational limitations. For medical educators, these systems may address challenges in standardising training experiences and providing equitable access to practice opportunities, particularly in settings with limited standardised patient resources or during disruptions to in-person teaching [3]. The ability to deploy multiple simultaneous interfaces (>70 in our study) demonstrates potential for efficient large-group teaching applications, addressing scalability issues identified in educational implementation research [40].

Content may compensate for realism despite the neutral ratings for conversational flow, the educational value remained significantly positive (Median 4.0). This supports the theory that functional fidelity (the accuracy of the case) is more critical for clinical reasoning training than conversational fidelity (perfect chat flow). Learners appear willing to forgive some "robotic" conversational traits provided the medical signs, symptoms, and history are robust.

However, participants' distinction between the system's utility for practice versus assessment (noted in qualitative feedback) suggests caution in higher stakes implementation. Recent commentary on AI assessment tools in healthcare education emphasise that technology-enhanced learning tools should complement rather than replace human instruction and assessment [41]. Transparency and disclosing the fundamental nature of AI interactions, avoids misrepresentations of technological capabilities and to cultivate an ethical approach and awareness of system biases [42].

Strengths and limitations of this study

Strengths

The study's unique methodology allowed for simultaneous testing with over 70 AI interfaces in a single time window, demonstrating the system's scalability for large group teaching sessions.

The study also achieved a relatively high adjusted response rate (61.3%), suggesting good engagement with the evaluation process [43]. The systematic evaluation across multiple domains using a structured questionnaire built upon validated measures provides a reliable framework for future studies in this area.

Limitations and future research

Several limitations warrant consideration. First, the study was conducted at a single institution that potentially limits generalisability. The uneven distribution of clinical experience among participants (bimodal distribution toward either 1-2 years or 11+ years) may have influenced overall ratings. Additionally, the reliance on self-reported evaluations rather than objective performance metrics presents potential for response bias. The cross-sectional design provides only a snapshot evaluation rather than longitudinal assessment of the technology's impact on learning outcomes or skill

development.

The design and evaluation of AI driven virtual platforms that focus on communication skills would benefit from greater stakeholder participation throughout the development process. Including input from diverse user groups, subject matter experts, educators, and industry professionals can provide valuable perspectives that enhance the relevance and effectiveness of these systems. Educational effectiveness can be explored with controlled studies that expand from co-designed learning approach. This collaboration ensures that the resulting platforms address actual communication challenges while incorporating real-world contexts and expectations that might otherwise be overlooked in purely technical development environments [8].

Conclusion

This evaluation demonstrates that conversational AI systems for patient consultation skills training show promising capabilities in providing clinically accurate scenarios with educational value, despite limitations in conversational authenticity. The technology appears particularly suited for supplementary training in clinical skills development rather than high-stakes assessment. Future refinements for larger scale randomised study should focus on enhancing dialogue naturalness and feedback capabilities while maintaining the strong medical content accuracy that participants valued most highly.

Data availability statement

Anonymised data is in supplementary file.

Ethics statements

This study was approved by Swindon Academy Medical Education Committee IRB (CJ062023). Study participants provided a single anonymised survey response. All participants were consented.

Conflict of interest statement

Authors have no financial relationship with SimFlow.ai, but received non-financial support in the form of platform access; the developers did not have access to any data collected and were not involved in analysis.

Sources of funding

There were no sources of funding for this study

Acknowledgements

We would like to thank Dr Jon Turvey CEO of SimFlow.ai for providing platform use during this study

1. Zhong JY, Fischer NL: Commentary: The desire of medical students to integrate artificial intelligence into medical education: An opinion article. *Frontiers in Digital Health*. 2023, Volume 5 - 2023.
2. Davenport T, Kalakota R: The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019, 6:94-98. 10.7861/futurehosp.6-2-94
3. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C: Cost-effectiveness

- of peer role play and standardized patients in undergraduate communication training. *BMC Medical Education*. 2015, 15:183. 10.1186/s12909-015-0468-1
4. Kung TH, Cheatham M, Medenilla A, et al.: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*. 2023, 2:e0000198. 10.1371/journal.pdig.0000198
5. Tu T, Schaekermann M, Palepu A, et al.: Towards conversational diagnostic artificial intelligence. *Nature*. 2025. 10.1038/s41586-025-08866-7
6. Singhal K, Tu T, Gottweis J, et al.: Toward expert-level medical question answering with large language models. *Nature Medicine*. 2025, 31:943-950. 10.1038/s41591-024-03423-7
7. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L: Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:240413066*. 2024.
8. 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 Journal of Educational Technology*. 2024, 40:39-57. 10.14742/ajet.9307
9. Mukadam A, Suresh S, Jacobs C: Beyond Traditional Simulation: An Exploratory Study on the Effectiveness and Acceptability of ChatGPT-4o Advanced Voice Mode for Communication Skills Practice Among Medical Students. *Cureus*. 2025, 17:e84381. 10.7759/cureus.84381
10. Marshall NJ, Loades ME, Jacobs C, Biddle L, Lambert JD: Integrating Artificial Intelligence in Youth Mental Health Care: Advances, Challenges, and Future Directions. *Current Treatment Options in Psychiatry*. 2025, 12:11. 10.1007/s40501-025-00348-x
11. Chan KS, Zary N: Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. *JMIR Med Educ*. 2019, 5:e13930. 10.2196/13930
12. Tun JK, Alinier G, Tang J, Kneebone RL: Redefining Simulation Fidelity for Healthcare Education. *Simulation & Gaming*. 2015, 46:159-174. 10.1177/1046878115576103
13. Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA: Reconsidering Fidelity in Simulation-Based Training. *Academic Medicine*. 2014, 89.
14. Jacobs C, Wheeler J, Williams M, Joiner R: Cognitive interviewing as a method to inform questionnaire design and validity - Immersive Technology Evaluation Measure (ITEM) for healthcare education. *Computers & Education: X Reality*. 2023, 2:100027. <https://doi.org/10.1016/j.cexr.2023.100027>
15. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al.: Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. *Journal of Surgical Education*. 2019, 76:1681-1690. <https://doi.org/10.1016/j.jsurg.2019.05.015>
16. Komasaawa N, Yokohira M: Simulation-Based Education in the Artificial Intelligence Era. *Cureus*. 2023, 15:e40940. 10.7759/cureus.40940
17. Mishra V, Lurie Y, Mark S: Accuracy of LLMs in medical education: evidence from a concordance test with medical teacher. *BMC Medical Education*. 2025, 25:443. 10.1186/s12909-025-07009-w
18. Feigerlova E, Hani H, Hothersall-Davies E: A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Medical Education*. 2025, 25:129. 10.1186/s12909-025-06719-5

19. Stamer T, Steinhäuser J, Flügel K: Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. *Journal of Medical Internet Research*. 2023, 25:e43311.
20. Liaw SY, Tan JZ, Bin Rusli KD, et al.: Artificial Intelligence Versus Human-Controlled Doctor in Virtual Reality Simulation for Sepsis Team Training: Randomized Controlled Study. *J Med Internet Res*. 2023, 25:e47748. 10.2196/47748
21. Jacobs C, Vaidya K, Medwell L, Old T, Joiner R: Case study of virtual reality sepsis management- instructional design and ITEM outcomes. *Journal of Visual Communication in Medicine*. 2023, 46:168-177. 10.1080/17453054.2023.2280611
22. Sharma A, Minh Duc NT, Luu Lam Thang T, et al.: A Consensus-Based Checklist for Reporting of Survey Studies (CROSS). *J Gen Intern Med*. 2021, 36:3179-3187. 10.1007/s11606-021-06737-1
23. Braun V, Clarke V, Boulton E, Davey L, McEvoy C: The online survey as a qualitative research tool. *International Journal of Social Research Methodology*. 2020:1-14. 10.1080/13645579.2020.1805550
24. Jacobs C, Foote G, Williams M: Evaluating user experience with immersive technology in simulation-based education: A modified Delphi study with qualitative analysis. *PLOS ONE*. 2023, 18:e0275766. 10.1371/journal.pone.0275766
25. Lucas HC, Upperman JS, Robinson JR: A systematic review of large language models and their implications in medical education. *Medical Education*. 2024, 58:1276-1285. <https://doi.org/10.1111/medu.15402>
26. Chan KY, Yuen TH, Co M: Using ChatGPT for medical education: the technical perspective. *BMC Medical Education*. 2025, 25:201. 10.1186/s12909-025-06785-9
27. Lewandowski T, Kučević E, Leible S, Poser M, Böhmman T: Enhancing conversational agents for successful operation: A multi-perspective evaluation approach for continuous improvement. *Electronic Markets*. 2023, 33:39. 10.1007/s12525-023-00662-3
28. Massoth C, Röder H, Ohlenburg H, et al.: High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students. *BMC Medical Education*. 2019, 19:29. 10.1186/s12909-019-1464-7
29. Jicol C, Clarke C, Tor E, et al.: Realism and Field of View Affect Presence in VR but Not the Way You Think. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Hamburg, Germany; 2023:Article 399.
30. Kasneci E, Sessler K, Küchemann S, et al.: ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*. 2023, 103:102274. <https://doi.org/10.1016/j.lindif.2023.102274>
31. Fareez F, Parikh T, Wavell C, et al.: A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*. 2022, 9:313. 10.1038/s41597-022-01423-1
32. Maicher KR, Adam S, Marisa S, et al.: Artificial intelligence in virtual standardized patients: Combining natural language understanding and rule based dialogue management to improve conversational fidelity. *Medical Teacher*. 2023, 45:279-285. 10.1080/0142159X.2022.2130216

33. Nivetha S, Ponnappalli PS, Nivetha S: ConverseLearn: A Multimodal AI Educational Assistant For Dynamic Learning And Practicing The Communication Skills And Real-Time Support. *IOSR Journal of Computer Engineering*. 2024, 26:10-20. 10.9790/0661-2606031020
34. Law AKK, So J, Lui CT, et al.: AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Medical Education*. 2025, 25:208. 10.1186/s12909-025-06796-6
35. Zarei M, Eftekhari Mamaghani H, Abbasi A, Hosseini M-S: Application of artificial intelligence in medical education: A review of benefits, challenges, and solutions. *Medicina Clínica Práctica*. 2024, 7:100422. <https://doi.org/10.1016/j.mcpsp.2023.100422>
36. Sadeq MA, Ghorab RMF, Ashry MH, et al.: AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. *Scientific Reports*. 2024, 14:18859. 10.1038/s41598-024-68996-2
37. Zou B, Du Y, Wang Z, Chen J, Zhang W: An Investigation Into Artificial Intelligence Speech Evaluation Programs With Automatic Feedback for Developing EFL Learners' Speaking Skills. *SAGE Open*. 2023, 13:21582440231193818. 10.1177/21582440231193818
38. Venter J, A. CS, and Schmulian A: Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assessment & Evaluation in Higher Education*. 2024:1-21. 10.1080/02602938.2024.2415649
39. Kelly S, Kaye S-A, Oviedo-Trespalacios O: What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*. 2023, 77:101925. <https://doi.org/10.1016/j.tele.2022.101925>
40. Clarke J, Dede C, Ketelhut DJ, Nelson B: A design-based research strategy to promote scalability for educational innovations. *Educational Technology*. 2006, 46:27-36.
41. Sezgin E: Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digit Health*. 2023, 9:20552076231186520. 10.1177/20552076231186520
42. Franco D'Souza R, Mary M, Vedprakash M, and Surapaneni KM: Twelve tips for addressing ethical concerns in the implementation of artificial intelligence in medical education. *Medical Education Online*. 2024, 29:2330250. 10.1080/10872981.2024.2330250
43. Fincham JE: Response rates and responsiveness for surveys, standards, and the Journal. *Am J Pharm Educ*. 2008, 72:43. 10.5688/aj720243

Supplementary Files

Figures

Screenshot of SimFlow.ai web-hosted platform. Patient avatar image AI generated.

3

Simulation

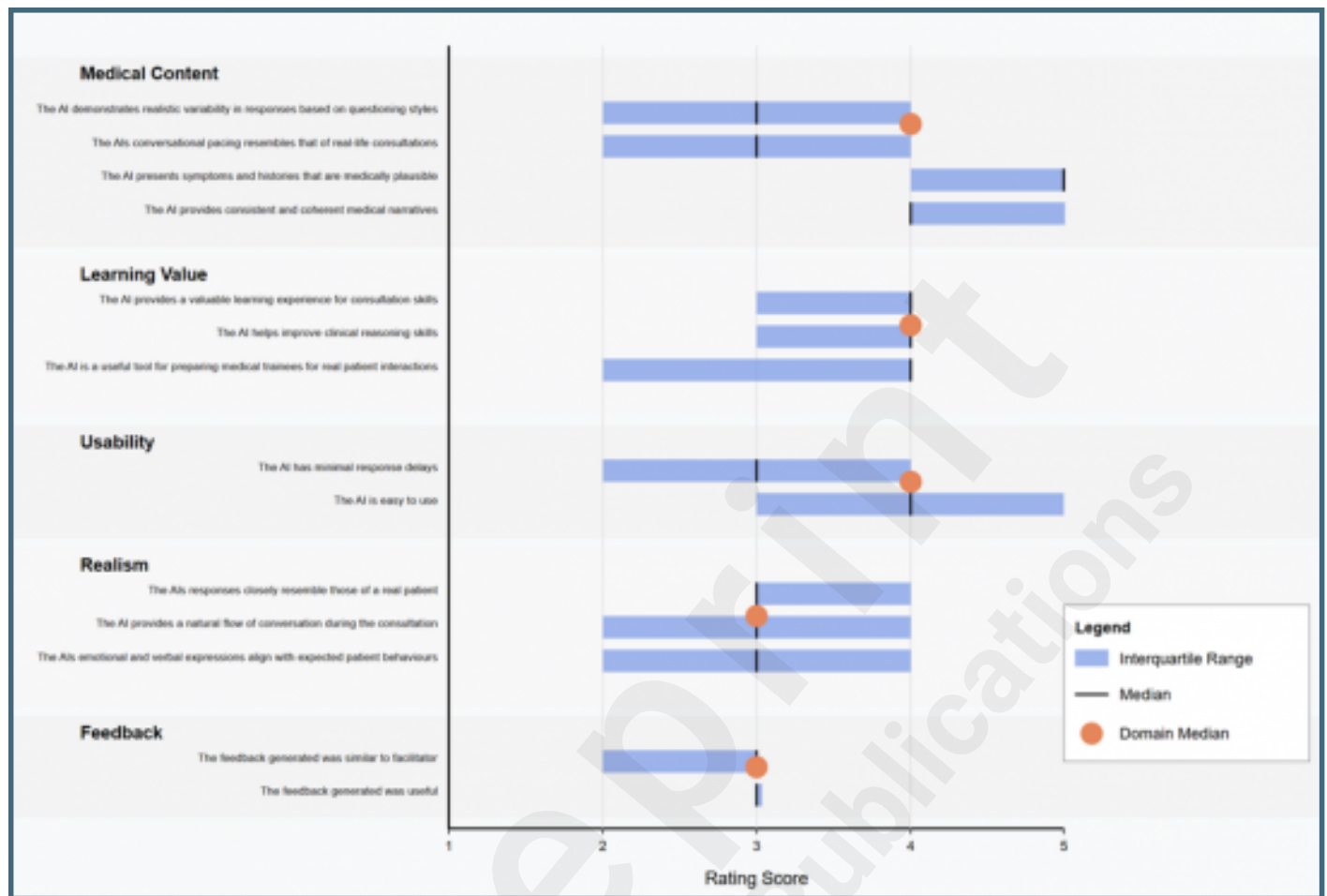


Patient Info: Jane Cohen, a 65 year old woman, has booked an appointment at Riverside Health Centre, her local GP practice in Bristol. She lives at 34 Pembroke Crescent, Clifton, Bristol, BS8 4QH and was born on 12th January 1960.

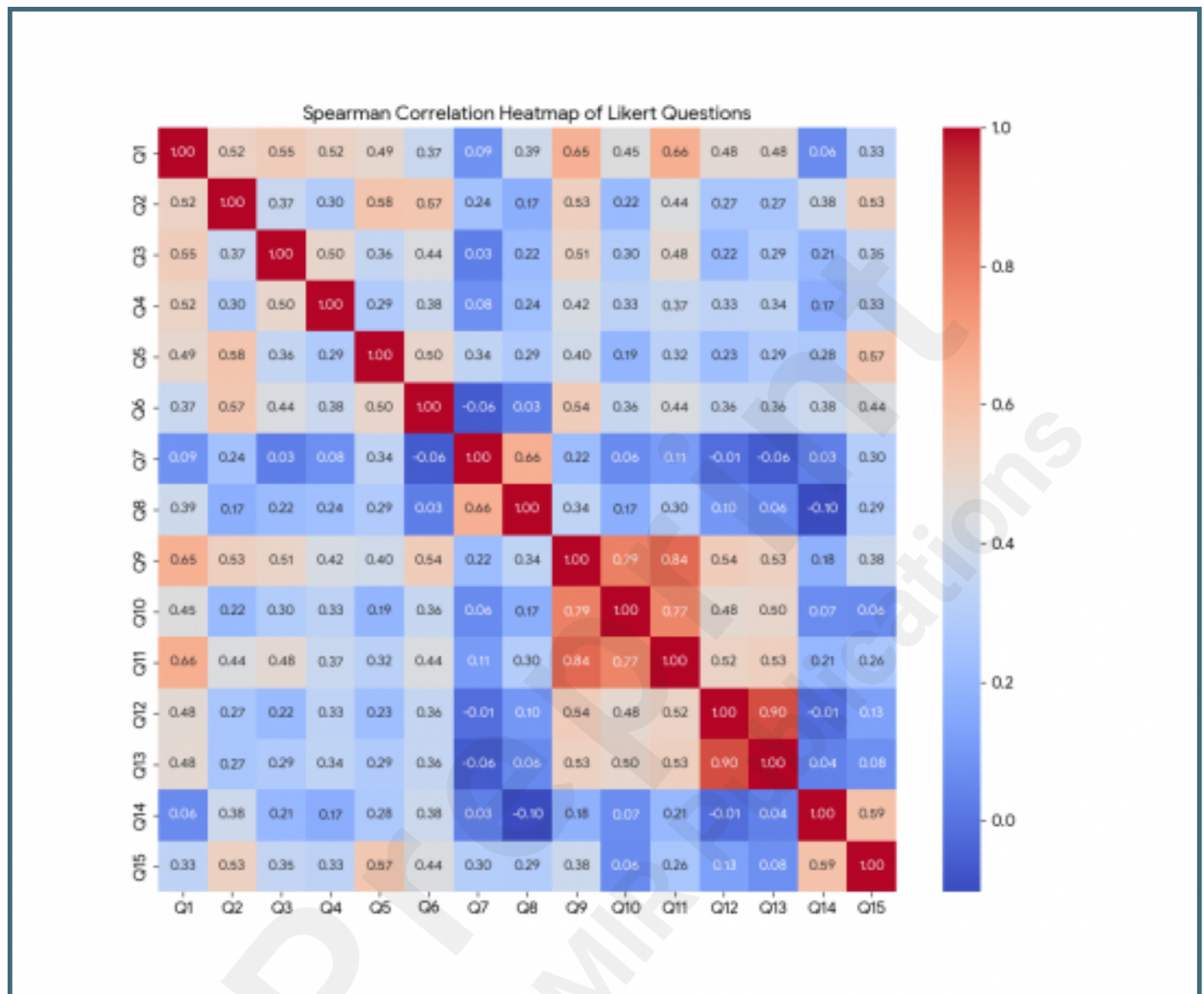
Task: You should gather a focused history about the patient's symptoms. After 6 minutes, the tutor will provide you with the examination findings and urinalysis. You should then explain to the patient what you have found and the possible reasons for their symptoms, and provide options for next steps in order to come to a shared management plan.

Begin Simulation

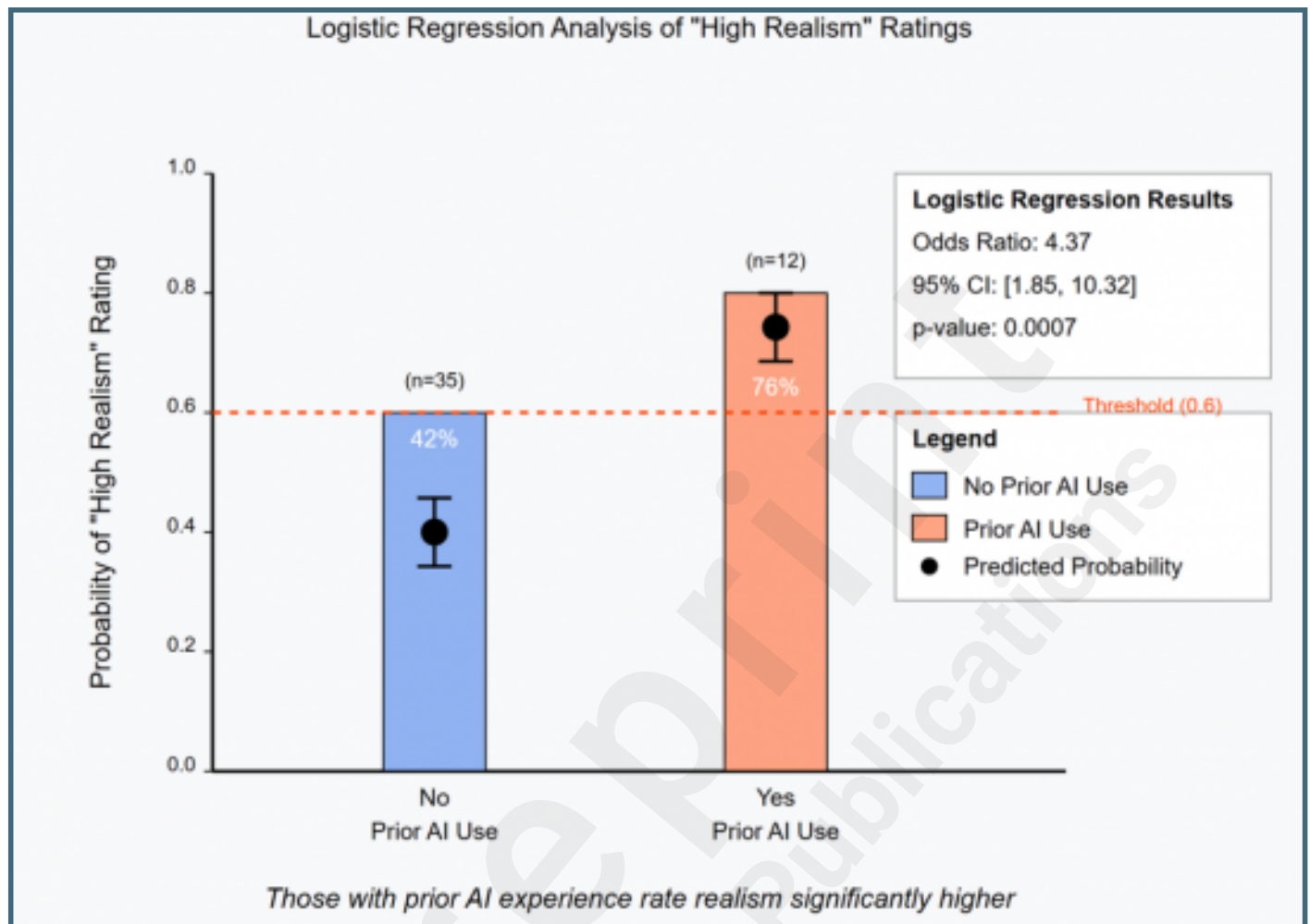
Domain level question responses with median and interquartile ranges.



Heat map indicating the correlation between questions.



Logistic regression of higher realism associated with prior AI use.



Thematic map.



CONSORT (or other) checklists

Untitled.

URL: <http://asset.jmir.pub/assets/f008919342b20a040e4c6d88ad732372.pdf>



TOC/Feature image for homepages

Chat GPT 5.2 image created with prompt- image that captures multimodal agentic AI, using computer avatars for consultation simulation in medical training.

