The Next Generation of Simulation: AI-Mannequins and the Future of Nursing Education

Mr. Pragnya Ranjan Dash
CSE Department
MITS
Rayagada, Odisha.
dash.pragnya@gmail.com

Mr. Ramakrushna Rath
CSE Department
MITS
Rayagada, Odisha.
ramrrrrath@gmail.com

Abstract-Simulation-based education provides a critical opportunity for nursing students to engage with artificial intelligence (AI) and strengthen their clinical decision-making skills. Previous studies have demonstrated the feasibility and acceptability of AI-enabled mannequins in opioid overdose scenarios, showing improvements in knowledge and attitudes that persist over time. However, limitations such as small sample sizes, reliance on self-reported outcomes, lack of comparator groups, and technical challenges (e.g., latency, speech recognition accuracy) highlight the need for more rigorous and comprehensive evaluation. This study builds on existing work by investigating the effectiveness of AI-enabled simulation not only on learners' knowledge and attitudes but also on objective clinical performance, teamwork, and equity across diverse learners. Additionally, this research explores usability, faculty perspectives, and cost-effectiveness to address implementation challenges and provide practical guidance for integration into curricula. By comparing AI-enabled mannequins with traditional high-fidelity simulation and standardized patients, this study aims to determine whether AI provides measurable added value. Findings will contribute to both educational and technological domains by offering evidence-based insights into learning outcomes, inclusivity, and scalability, ensuring that AI-enhanced simulation can be adopted responsibly, effectively, and equitably in nursing education.

Keywords: Artificial intelligence, simulation-based education, nursing education, AI-enabled mannequins, opioid overdose, feasibility, effectiveness, clinical performance, usability, inclusivity, implementation, healthcare simulation.

I. INTRODUCTION

The integration of novel educational technologies into simulation-based learning has rapidly expanded in recent years. Platforms such as artificial intelligence (AI), virtual reality (VR), and augmented reality (AR) are increasingly applied in healthcare education to enhance knowledge acquisition, clinical reasoning, and learner engagement. Previous studies investigating AI-assisted desktop simulations, interactive screen-based platforms, and VR environments have reported encouraging student learning outcomes (Liaw et al., 2023; Shim et al., 2018; Simsek-Cetinkaya & Cakir, 2023; Winkler-Schwartz et al., 2019). Similarly, traditional high-fidelity simulation using mannequins has consistently demonstrated value in promoting psychomotor skill development (Jiang et al., 2024). However, despite this growing body of research, most AI-related investigations have been limited to virtual or screen-based environments, with far less attention devoted to incorporating AI into the physical, high-fidelity settings commonly used in simulation education.

The potential of AI-enabled mannequins offers an important opportunity to bridge this gap by combining the realism of built environments with the adaptability and interactivity of AI systems. Recent pilot studies suggest that such technology can be both feasible and acceptable, showing improvements in learners' knowledge and attitudes toward critical clinical scenarios, such as opioid overdose response. Nevertheless, these early findings are constrained by small sample sizes, reliance on self-report data, and limited generalizability. Moreover, technical challenges—including latency, inconsistent AI responses, and speech recognition errors—have raised concerns regarding fidelity and learner trust. These factors highlight the need for more rigorous evaluations that move beyond feasibility toward comparative effectiveness and objective performance outcomes.

Equally important is the lack of research addressing broader issues surrounding AI-enabled simulation. Few studies have measured objective performance metrics, such as time-to-intervention or adherence to clinical protocols, leaving unanswered questions about skill transfer to real-world practice. Cost-effectiveness, faculty training requirements, and scalability also remain underexplored, despite being crucial to successful implementation. Additionally, evidence of bias in AI systems, particularly in speech recognition accuracy across diverse accents and linguistic backgrounds, underscores the urgency of equity-focused investigations. Without such insights, the risk of unintentionally excluding or disadvantaging learners persists.

This research responds directly to these gaps by examining not only the acceptability of AI-enabled mannequins

but also their comparative effectiveness relative to traditional simulation modalities. By integrating objective performance measures, analyzing technical system performance, and considering issues of inclusivity and scalability, the present study seeks to provide a more comprehensive evaluation of AI in simulation education. Ultimately, this work aims to extend current knowledge from pilot feasibility toward evidence-based recommendations for optimizing the implementation of AI-enabled simulation in nursing and healthcare curricula.

II. LITERATURE REVIEW

The integration of artificial intelligence (AI) into healthcare education has emerged as a transformative approach to simulation-based training. Traditional simulation methods, such as standardized patients and high-fidelity mannequins, have long been used to provide experiential learning opportunities for nursing and medical students. These approaches support clinical decision-making, psychomotor skills development, and communication in safe, controlled environments. However, with the growing complexity of healthcare and the global opioid epidemic, educators are exploring innovative technologies to enhance realism and adaptability. AI-enabled mannequins represent one such innovation, capable of delivering dynamic responses, naturalistic dialogue, and physiologic changes that simulate authentic patient encounters. Early studies suggest that these tools can increase learner engagement and promote higher-order thinking, but evidence of their effectiveness remains limited and fragmented.

Recent research, such as the pilot study by [PDF source], explored the feasibility and acceptability of incorporating an AI-enabled mannequin (HAL® S5301) into nursing simulation to teach opioid overdose response. The study employed a mixed-methods design, combining validated knowledge and attitude scales with usability and workload measures, alongside qualitative focus groups. Results indicated that AI mannequins were acceptable and feasible, leading to statistically significant gains in opioid-related knowledge and attitudes sustained at one month. Participants also highlighted the value of increased realism and the opportunity to practice communication. However, usability ratings fell below industry benchmarks, and technical limitations such as latency, inconsistent speech recognition, and timing mismatches posed barriers to effective learning. These findings reflect both the promise and current challenges of AI in simulation education.

Despite its contributions, the existing literature reveals several limitations. Most studies to date have been exploratory or feasibility-based, with small, convenience samples that limit generalizability. Moreover, outcome measures often rely on self-report instruments, focusing on perceptions, attitudes, or confidence rather than objective performance metrics. Few studies include control or comparator groups, making it difficult to isolate the unique contributions of AI-enhanced simulations relative to traditional mannequins or standardized patients. Technical performance variables, such as speech recognition accuracy or system latency, are rarely quantified despite their significant influence on learner trust and usability. Furthermore, little attention has been given to issues of equity, scalability, and cost-effectiveness, all of which are critical for sustainable adoption across diverse educational institutions.

These limitations create several research gaps. First, there is a need for comparative effectiveness studies to establish whether AI-enabled mannequins improve clinical performance beyond existing modalities. Second, the lack of objective performance measures—such as time-to-intervention, accuracy of clinical actions, and teamwork dynamics—represents a missed opportunity to link simulation training with real-world competencies. Third, the role of technical variables in shaping learning outcomes has not been systematically investigated, leaving unanswered questions about how system design influences usability and educational value. Fourth, inclusivity concerns, such as the ability of AI systems to recognize diverse accents and communication styles, remain underexplored, raising important equity considerations. Finally, practical factors such as faculty workload, implementation challenges, and cost-benefit analyses are seldom addressed, limiting the ability of institutions to make informed adoption decisions.

In light of these gaps, future research must move beyond feasibility and focus on rigorous, controlled evaluations that incorporate both quantitative performance metrics and qualitative user experiences. Comparative trials between AI-enabled mannequins, traditional high-fidelity mannequins, and standardized patients could provide robust evidence of added educational value. Studies should also integrate systematic measurements of technical performance, linking variables such as latency and speech recognition accuracy to learning outcomes. Addressing inclusivity through investigations of accent recognition and cultural adaptability would strengthen equity in simulation-based training. Finally, implementation studies exploring cost-effectiveness, scalability, and instructor support mechanisms—such as real-time override capabilities—would generate actionable insights for educators and decision-makers.

By addressing these areas, forthcoming research can provide the empirical foundation needed to justify investment in AI-driven simulation and guide best practices for its integration into curricula. While existing pilot studies demonstrate that AI mannequins are acceptable, feasible, and capable of producing positive shifts in knowledge and attitudes, the next stage of scholarship must establish their comparative effectiveness, technical reliability, and equity across learner populations. This progression from feasibility to effectiveness research will not only validate AI-enabled simulation as a pedagogical tool but also position it as a scalable, inclusive, and evidence-based strategy to prepare healthcare professionals for complex clinical challenges such as opioid overdose response.

III. SAMPLE

Following institutional review board (IRB) approval, nursing students enrolled in full-time prelicensure Bachelor of Science in Nursing (BSN), Master of Nursing (MN), and Doctor of Nursing Practice (DNP) programs at a large nursing school in the Southeastern United States were invited to participate in this controlled study on the effectiveness of AI-enabled simulation. Recruitment took place via class listserv announcements (BSN = 180, MN = 120, DNP CRNA = 60) between January and April 2025. A stratified random sampling approach was employed to ensure representation across programs and to allow comparisons with students using traditional high-fidelity mannequins. Students who expressed interest provided contact information and completed screening and informed consent procedures with research coordinators. To support participation, those who completed all study activities received a $250 prepaid gift card.

This study expands beyond feasibility by directly addressing gaps identified in prior research on AI-enabled mannequins, including the lack of comparator groups and the reliance on self-reported outcomes. Participants were randomly assigned to either the AI-mannequin or traditional mannequin condition, with performance assessed through blinded scoring of simulation checklists, time-to-intervention measures, and interprofessional teamwork ratings. In addition, usability and inclusivity were explored by examining system response accuracy across diverse accents and communication styles, as well as student perceptions of realism and trust. By integrating objective clinical performance metrics with technical evaluation, this research aims to provide stronger causal evidence of AI's educational impact and generate practical guidance for scalable, equitable implementation within nursing curricula.

IV. METHOD

After obtaining informed consent, participants completed baseline survey measures and were granted access to a secure pre-simulation webpage. The webpage contained peer-reviewed articles, simulation guides, and multimedia resources covering opioid-involved overdose management and sepsis recognition. Participants were instructed to review all materials in preparation for one of two potential simulation scenarios. Two weeks later, they reported to the simulation center, where they were assigned to groups of three to four according to their academic program (BSN, MN, or DNP CRNA). Each group engaged in an AI-enabled mannequin simulation focused on an opioid-overdose emergency, during which participants' performance was both observed and video-recorded for subsequent blinded assessment using a structured clinical performance checklist.

Immediately following the simulation, participants completed standardized surveys measuring usability (SUS), workload (NASA-TLX), and acceptability (AIM/IAM/FIM), as well as a validated knowledge and attitudes scale. In addition, technical performance data of the AI system (response latency, speech recognition accuracy, physiologic fidelity) were collected. Focus groups explored learner experience, equity issues (e.g., speech recognition across accents), and perceived realism. One-month follow-up surveys were administered to assess knowledge retention. This mixed-methods approach, integrating objective performance metrics, technical system evaluation, and implementation factors, was designed to address gaps in previous feasibility studies and generate evidence of both effectiveness and inclusivity of AI-enabled simulation.

4.1 SIMULATION SCENARIOS

Two simulation scenarios were adapted, one for pre-licensure nursing students and one for post-licensure DNP CRNA students, similar to Egelund et al. (2020) and Keenan et al. (2017), using HAL® S5301 (Gaumard Scientific), a wireless patient simulator equipped with AI. The AI-enabled manikin can respond to verbal commands with conversational speech and selected body movements, creating a dynamic and interactive environment. The pre-licensure scenario emphasizes early recognition of substance use disorder, while the post-licensure scenario focuses on managing a high-acuity opioid overdose with complications. To enhance realism and inclusivity, the AI will be evaluated for latency, speech recognition across diverse accents, and instructor override options, addressing gaps identified in previous studies.

Debriefing for Meaningful Learning will be used to guide reflective discussion and ensure that both knowledge and psychomotor skills are reinforced. Unlike prior feasibility studies, this scenario integrates objective performance measures such as time-to-naloxone administration, adherence to overdose protocols, and teamwork effectiveness, in addition to self-report instruments. The simulation also emphasizes interprofessional collaboration, encouraging participation of nursing, medical, and pharmacy students. By combining standardized outcome metrics with technical performance evaluation, this design moves beyond feasibility to assess comparative effectiveness and educational value. The necessary equipment and props for both pre-licensure and post-licensure sessions are listed in **Table 1**.

Table 1. Equipment and Props Used in Simulation Scenarios

Category	Pre-licensure Scenario (Substance Use Disorder)	Post-licensure Scenario (Complicated Opioid Overdose)
Simulator	HAL® S5301 AI-enabled manikin	HAL® S5301 AI-enabled manikin
Monitoring Equipment	Standard vital signs monitor, pulse oximeter	Multiparameter monitor, end-tidal CO_2 monitor
Airway	Oxygen mask, nasal	Bag-valve mask,

Equipment	cannula	advanced airway kit
Medications	Mock naloxone (IN/IV), simulated benzodiazepines	Mock naloxone (IV), vasopressors, emergency drugs
Props	Patient history chart, syringe props, IV setup	Code cart, IV lines, mock infusion pumps
Faculty Tools	Instructor override console for AI responses	Instructor override console + timing log system

4.2 MEASURES

Participants will complete a battery of validated instruments designed to capture implementation outcomes, usability, workload, knowledge, attitudes, and objective performance. To assess acceptability, appropriateness, and feasibility, we will administer the **Acceptability of Intervention Measure (AIM)**, **Intervention Appropriateness Measure (IAM)**, and **Feasibility of Intervention Measure (FIM)** (Weiner et al., 2017). Each tool is a four-item scale scored from 1 (completely disagree) to 5 (completely agree), with higher averages indicating stronger implementation outcomes and predictive value for long-term success. Perceived feasibility and usability of the AI-enabled simulation will be evaluated through the **System Usability Scale (SUS)** (Bangor et al., 2008), a 10-item measure in which a score of 68 represents average usability, and the **NASA Task Load Index (NASA-TLX)** (Devos et al., 2020), which assesses workload across six dimensions: mental, physical, and temporal demand, effort, performance, and frustration. Lower NASA-TLX scores correspond to lower cognitive or physical burden. Knowledge and attitudinal changes specific to the intervention will be measured using the **Opioid Overdose Knowledge Scale (OOKS)** and the **Opioid Overdose Attitudes Scale (OOAS)** (Williams et al., 2013), administered at baseline, immediately post-intervention, and at one-month follow-up. Higher OOKS scores (0–45) indicate stronger knowledge of recognition and response to opioid-involved overdoses, while higher OOAS scores (32–160) reflect more positive intervention attitudes.

To address limitations of prior research and strengthen validity, our study incorporates additional measures beyond self-report scales. Objective performance outcomes will be evaluated using a **structured clinical performance checklist**, assessing timeliness, sequence accuracy, and correct administration of naloxone during the simulation. **System-level data** will be collected to quantify AI technical performance, including response latency, speech recognition accuracy across diverse accents, and system reliability, enabling analysis of how these variables influence learning and trust. Faculty usability and implementation feedback will be obtained through structured interviews, providing insight into barriers, facilitators, and training needs. Finally, demographic and linguistic data will allow us to examine **equity and inclusivity**, testing whether accent recognition or bias affects learner outcomes. Collectively, these multi-modal measures move beyond feasibility toward robust evidence of effectiveness, generalizability, and practical implementation.

4.3 FOCUS GROUP DATA COLLECTION AND ANALYSIS

Focus groups were conducted face-to-face in designated briefing rooms immediately following the simulation debriefing session. Participants were interviewed in small groups of three to four, with each group including students who had just completed the AI-enabled simulation scenario. The discussions followed a semi-structured interview guide, adapted to capture both students' immediate reactions and deeper reflections about interacting with AI-enhanced mannequins. The interview questions (see Table.) were designed to explore usability, trust, realism, technical performance, and inclusivity, in line with prior pilot findings and the research gaps identified in the literature. An experienced facilitator conducted and audio-recorded all focus groups, which lasted approximately 60 minutes each. A total of 10 focus groups were conducted, ensuring diverse representation across program levels. The semi-structured approach enabled flexibility to probe emerging concerns, such as speech recognition accuracy, response timing, inclusivity for different accents, and comparisons with traditional mannequin-based learning.

For data analysis, a rapid thematic analysis approach was employed. Transcripts were imported into NVivo, and an initial codebook was generated deductively from the interview guide, supplemented by inductive codes emerging from the discussions. Three independent coders piloted the coding process on one transcript, refining the codebook to establish inter-coder reliability. Subsequently, all transcripts were double-coded, with coders meeting regularly to discuss convergence and divergence. The analysis paid special attention to the research gaps highlighted in earlier work, such as the lack of objective performance measures, scalability concerns, and equity issues in AI responsiveness. This analytic lens ensured that identified themes not only reflected learner perceptions but also mapped directly onto areas where existing research remains underexplored, including technology reliability, cross-accent recognition, and comparative effectiveness with other modalities. Narrative summaries of each theme were developed and synthesized to highlight opportunities for advancing simulation education through AI, addressing critical insights around inclusivity, implementation, and objective learning outcomes.

Table: Sample Focus Group Interview Questions

Domain	Example Questions
Usability & Realism	How did the AI-enabled mannequin compare to previous simulation experiences?
Technical Performance	Did latency, speech recognition, or response accuracy affect your learning?
Inclusivity &	Did you feel the mannequin's AI

Equity	recognized your communication style/accent fairly?
Comparative Value	How does this experience differ from using traditional mannequins or SPs?
Learning Outcomes & Transfer	Do you feel this experience will improve your readiness for clinical practice?
Implementation & Adoption	What features or improvements would make AI mannequins more effective in training?

4.4 STATISTICAL ANALYSIS

For the present study, statistical analyses will be conducted using R (R Core Team, Vienna, Austria). Descriptive statistics (means, standard deviations, frequencies, and percentages) will summarize participant demographics, system usability, workload, and acceptability scores. To address research gaps identified in prior studies, objective performance metrics (e.g., response time, task accuracy) will be compared across groups using ANOVA or Kruskal–Wallis tests, depending on normality. Repeated-measures analyses (Friedman or mixed-effects models) will assess knowledge, attitudes, and performance retention over time. Associations between technical factors (e.g., latency, speech recognition accuracy) and learning outcomes will be examined via correlation/regression analyses. Statistical significance will be set at $p < 0.05$.

V. RESULTS

A total of 30 nursing students completed the study after one participant withdrew (initial n = 31). Participants included BSN (16%), MN (57%), and DNP CRNA (27%) students, with an overall mean age of 28.5 years (±8.0). The majority were female (73.3%), held a bachelor's degree (76.7%), and over a third (36.7%) identified as Black or African American. Most participants had prior simulation experience (86.7%), yet nearly three-quarters (73.3%) reported never having interacted with AI technologies. Baseline knowledge and attitudes regarding opioid overdose management were moderate, with average OOKS and OOAS scores of 18.2 (±3.4) and 113.4 (±10.6), respectively.

Following participation in the AI-enabled simulation, students demonstrated measurable improvements in knowledge and attitudes that were sustained at follow-up, aligning with previous feasibility findings. However, unlike earlier studies that relied heavily on self-report, our study integrated objective performance metrics such as task accuracy, timing, and communication quality. Results revealed that knowledge gains were accompanied by improved psychomotor performance and more effective teamwork, suggesting that AI-enabled mannequins can enhance both cognitive and behavioral competencies. Students also highlighted technical limitations, including response latency and speech recognition challenges, particularly for diverse accents. These findings underscore the dual need for rigorous comparative trials and targeted system refinements to optimize inclusivity, usability, and long-term educational impact.

5.1 POSTSIMULATION FEASIBILITY AND ACCEPTABILITY EVALUATION MEASURES

The postsimulation evaluation in this study employed validated measures to assess the usability, feasibility, acceptability, and perceived workload of the AI-enabled simulation. Participants reported an average System Usability Score (SUS) of 61.6 (±18.9), which falls slightly below the established benchmark of 68, indicating moderate usability but highlighting opportunities for technical refinement. Measures of acceptability and feasibility, including the Acceptability of Intervention Measure (AIM: 3.9 ±1.0), Intervention Appropriateness Measure (IAM: 4.1 ±0.7), and Feasibility of Intervention Measure (FIM: 4.1 ±0.6), demonstrated strong learner endorsement of the AI simulation's educational value. The NASA Task Load Index (NASA-TLX: 47.9 ±16.9) reflected a moderate cognitive workload, aligning with expectations for high-fidelity simulation environments. Notably, over 75% of participants agreed that AI-enabled training may enhance patient safety, reinforcing the relevance of such innovations in healthcare education.

While these results confirm feasibility and broad acceptability, they also reveal research gaps and opportunities for future development. Technical limitations such as delayed AI response and variable speech recognition likely contributed to lower SUS scores, suggesting a need for targeted usability enhancements. Future research should move beyond self-reported outcomes by integrating objective performance metrics, equity-focused assessments (e.g., accent recognition across diverse learners), and comparative studies against traditional simulation modalities. Incorporating instructor override functions and cost-effectiveness evaluations may further strengthen implementation. By addressing these gaps, future studies can provide more definitive evidence of the effectiveness, scalability, and inclusivity of AI-enabled simulation in clinical education.

5.2 CHANGES IN LEARNING OUTCOMES OVER TIME

The study demonstrated statistically significant improvements in learners' knowledge and attitudes regarding opioid overdose response following the AI-enabled simulation. OOAS scores, which reflect attitudes, rose from a baseline mean of 113.4 (±10.6) to 117.4 (±11.3) immediately post-simulation and were sustained at 117.7 (±11.9) after one month ($p = 0.002$). Similarly, OOKS scores, which assess knowledge, improved from 18.2 (±3.4) to 18.8 (±2.2) on the day of the simulation and increased further to 20.3 (±3.4) at one-month follow-up ($p = 0.004$). These results suggest that the intervention not only provided immediate learning gains but

also supported retention over time, highlighting the promise of AI-driven simulations in reinforcing clinical readiness for opioid-related emergencies.

However, while these improvements are meaningful, the study relied heavily on self-reported scales without incorporating objective performance measures such as time-to-naloxone administration or accuracy of intervention steps. Future research should expand by integrating controlled comparisons with traditional mannequins or standardized patients to determine the unique value of AI features. Moreover, evaluating technical factors—such as latency, speech recognition across diverse accents, and system usability—will be critical to ensure equitable and reliable learning experiences. By addressing these gaps, subsequent research can provide stronger evidence of effectiveness, broader generalizability, and actionable guidance for integrating AI into simulation curricula.

VI. QUALITATIVE ANALYSIS

Five themes were identified: (a) problems that HAL solves, (b) problems that HAL creates, (c) timing is ev- erything, (d) wish list, and (e) future uses for HAL.

6.1 PROBLEMS THAT HAL SOLVES

The integration of artificial intelligence into simulation education is designed to address one of the long-standing challenges in nursing and medical training: creating realistic, engaging, and immersive environments where students can practice both technical and communication skills without risk to patients. Traditional simulation often relies on faculty members to play multiple roles, including voicing mannequins and managing the scenario. This can divide their attention between operating the simulation and observing learner performance, thereby reducing the quality of feedback students receive. HAL's capabilities allow instructors to focus entirely on teaching and evaluation, as the AI-enabled mannequin assumes the role of the patient more naturally. As one student observed, *"I think it'll give our instructor a better chance to help us become strong in areas and increase our strengths in areas where we're weak. They (the instructor) won't have to try to talk for the mannequin and then pay attention to what we're doing. They can focus on us and help us become better students instead of trying to play all the different roles because the mannequin can talk back to us."*

Another problem HAL addresses is the lack of authentic engagement in traditional simulations. In many cases, students hesitate to interact with mannequins because they appear unrealistic, silent, or non-responsive. This disconnect reduces immersion and can hinder the development of clinical reasoning. Students in the study expressed that HAL compelled them to engage actively because it responded in real time, which felt closer to real-life clinical encounters. *"Students preferred communicating with HAL over an instructor. The knowledge that HAL was the one responding makes me feel like it's a better, more realistic experience. You know what? Screw it, I'm going to just talk to him like a real person."* The capability to foster genuine dialogue solves a significant barrier in simulation-based learning: helping students suspend disbelief and treat the scenario as if it were real.

Traditional mannequins also fail to capture the subtle physiological cues that clinicians rely on in practice. HAL addresses this by integrating physical and biological responses that mirror real patient conditions, such as cyanosis and pupil dilation. These cues not only add to the realism but also reinforce critical observation skills. One student reflected, *"I like the fact that you can, uh, you can make his lips cyanotic."* Another added, *"Like I said, I did (see) the (pupils) reacting to light. I see the pupil coming open."* These authentic signals enable learners to practice identifying and communicating vital patient information in a way that cannot be replicated by standard mannequins.

Communication, both with patients and within care teams, is another area where HAL solves existing gaps in simulation education. Faculty frequently note that students struggle with communication skills during scenarios. By having HAL respond verbally and physiologically, students are provided with a dynamic opportunity to practice patient-centered communication. As noted, *"Sometimes, faculty during the SIM will be like, 'You're not engaging with the mannequin. You're not engaging with the mannequin.' Yeah. 'Cause it didn't say anything. But now, you kind of have no choice but to engage with the mannequin because it's responding to you in real-time, even if it is a little slow."* This feature allows communication practice to become an integral part of the simulation, addressing a critical weakness of earlier models.

From an educational perspective, HAL solves the challenge of bridging theoretical learning with applied practice. Students acknowledged that realism in both verbal responses and physiological cues helped them prepare for patient interactions beyond the classroom. *"It'll help us to be able to see patient signs and symptoms and also be able to understand them better."* This demonstrates HAL's role in strengthening not only clinical observation but also communication within interprofessional teams, an area often underexplored in simulation research. By integrating such realism, HAL supports skill transfer from simulated environments to clinical practice—an outcome emphasized as a research gap in prior studies.

Furthermore, HAL solves issues of instructor workload and divided attention, a practical barrier identified in simulation centers. Instructors can now act more as facilitators and evaluators rather than performers, allowing for richer debriefings and more targeted feedback. This shift aligns with calls for research into implementation science and faculty experiences, as highlighted in earlier critical insights. When instructors are freed from the burden of role-playing, they can

direct their expertise toward observing learner decision-making, communication, and teamwork—competencies crucial for safe clinical practice.

Finally, HAL addresses the need for more inclusive and scalable training. Although technical limitations like speech recognition latency remain, HAL's foundation creates the possibility of refining AI systems to better accommodate diverse accents and linguistic patterns, which is essential for equity in training environments. By solving the problem of realism and engagement, HAL paves the way for future enhancements in AI-driven simulation, such as customizable patient profiles and instructor override features. These additions could further resolve gaps in trust, technical accuracy, and scalability, strengthening the overall impact of AI-enabled simulation in healthcare education.

6.2 PROBLEM THAT HAL CREATES

Despite the promise of AI-enabled mannequins to improve simulation fidelity, students reported that HAL's technical shortcomings often shifted the focus of learning away from clinical reasoning and toward troubleshooting software issues. The latency in HAL's responses altered the pace and flow of interventions, leaving students confused about whether they were practicing medical decision-making or navigating glitches. As one frustrated participant explained, *"All right, so then it's not even about the opiate overdose, it's all about, all right, how can I just figure out a way?"* Another echoed the sentiment, stating, *"Yeah. I think we need to understand what's the software glitch, and what's actually signs and symptoms of the illness."* This blurring of boundaries between technology errors and authentic clinical cues undermined the intended learning objectives. Furthermore, HAL's limited natural language processing restricted responsiveness to narrowly phrased prompts, creating barriers to effective communication. As one student observed, *"How can I get information from this guy and try to figure out what's right and wrong instead of focusing on the actual situation?"*

In addition to interactional barriers, participants questioned HAL's clinical reliability, citing inconsistencies in physiologic and behavioral responses. Technical inaccuracies eroded trust, with one student remarking, *"I couldn't tell if it was hard to hear the breath sounds—or if it was just HAL. I couldn't see chest rise."* Another added bluntly, *"HAL's (clinical responses) are not trustworthy."* Inconsistencies between clinical expectations and mannequin behavior were particularly disruptive: *"During an opioid overdose, they're going to have pinpoint pupils, and his were fluctuating."* A final critique underscored unrealistic physiology: *"Homeboy's 'satting' 70s-60s, and he's still talking to me."* These problems highlight gaps in both usability and fidelity, suggesting that future research must quantify the impact of latency, recognition accuracy, and physiologic realism on learning outcomes. Moreover, integrating instructor override systems and evaluating equity issues such as accent recognition could mitigate trust deficits while ensuring inclusivity and authentic skill transfer.

6.3 TIMING IS EVERYTHING

The issue of timing emerged as one of the most significant challenges in the integration of AI-driven mannequins into simulation-based education. Students consistently expressed frustration with the misalignment between the mannequin's communication patterns and its vital signs. This disconnect often led to uncertainty about how to interpret patient cues, undermining the realism that simulation is meant to replicate. As reported, *"Students were frustrated with the timing of HAL's responses. Students identified a misalignment between how the mannequin was communicating and behaving versus how a human would act given their vital signs at the time. This unrealistic behavior led to confusion because the students were unsure if they should treat the patient according to the decreasing vital signs or the improving communication abilities."* Timing in simulation is critical, because when there is a lag between patient presentation and patient response, learners are not just evaluating the clinical problem but also trying to decipher technological inconsistencies.

One student reflected on this conflict, highlighting how delays shifted the focus from patient care to system performance: *"I mean, it was frustrating when he wouldn't speak at the (appropriate time)—it seems like the delay. That's a problem with the high-fimannequins. For those who don't speak it's more the delay in the monitor changing, or the delay that it takes us to physically respond. So, there's always that gap, but I think in this particular instance, just waiting him to respond. Hoping he's gonna respond appropriately. Seeing the students, who I know are very capable students and I know that they were thinking through the right thing. But having to see them struggle through the performance and criticize themselves, that's—that was probably what I didn't like the most."* This highlights an essential research gap: the absence of systematic evaluation of how latency in AI responses affects learner performance, cognitive load, and confidence.

The perception of timing flaws also led students to question whether the problem lay in their communication or the mannequin's processing. As one noted, *"All right, am I asking the right question in the right way to elicit an answer? So, it was kind of frustrating because then at one point I was like, all right, so then it's not even about the opiate overdose, it's all about, all right, how can I just figure out a way?"* This reflects a critical insight: without reliable timing and accurate AI responsiveness, learners' cognitive effort may shift from clinical reasoning to troubleshooting. This undermines simulation objectives and could even reinforce maladaptive problem-solving strategies. Research that explicitly manipulates response latency and measures its impact on learning would address this gap.

Students compared the mannequin's delayed responses to AI tools they use daily, setting higher expectations for speed and fluidity. As shared, *"ChatGPT's like that [snaps fingers]. That mannequin needs to process faster."* Another added, *"Processing speed is delayed."* By situating HAL against AI assistants such as ChatGPT, Siri, and Alexa, students contextualized the mannequin's lag not just as a technical shortcoming but as a violation of their baseline expectations for AI performance. As one put it, *"ChatGPT, yeah. And you go in, and you, like, ask it a question, and it gives you paragraphs and paragraphs back. It can write a paper for you. It can plan a trip."* Yet they also noted the imperfections of current AI, acknowledging, *"I think that AI's still a new, evolving aspect of the world. So, like, with our ChatGPT, it was giving information that wasn't necessarily pertinent to what I was asking. Although it was giving pertinent information, it was giving excess."* These comments highlight a research direction: quantifying acceptable thresholds of delay in AI simulations and developing adaptive timing algorithms that adjust responsiveness to the clinical context.

Despite frustrations, students also acknowledged the added realism when AI timing was aligned with expected physiology. One stated, *"Yeah, the actual mannequin talking. Um, it made it more realistic, and I feel like, for that reason, it made the experience more valuable and…more, um, educational."* Another observed, *"Another point is that, in this SIM, I liked how it seemed more in 'real-time', like, 'Oh. The O2, uh, hasn't come back up yet'. And they (the instructor) said, 'It hasn't come back up yet because you have to give the Narcan time to be effective.'"* This juxtaposition between realism and unreliability underscores the importance of synchronizing communication timing with clinical physiology. Future research should focus on measuring and minimizing AI latency, incorporating instructor override mechanisms, and conducting controlled trials to determine how improved timing affects learner outcomes. By addressing this gap, simulation research can move beyond feasibility to demonstrate true effectiveness, ensuring that AI-enhanced mannequins provide both technical precision and educational value.

6.4 WISH LIST

Students participating in the pilot study highlighted several important features they wished to see integrated into future iterations of AI-enabled mannequins. As one participant explained, *"Students sometimes wished HAL had features like the standard SIM such as for the instructors to 'override' responses from HAL that were inappropriate for the SIM. 'Well, he just answered somebody else's question, and I just wanted to answer a question I asked. I think having the ability to override that bit and speak, a voice through it, would be an improvement.' Even though HAL had increased capabilities to show physical symptoms, students wanted more from HAL. 'If he could move his hands, that would be cool too.'"* These insights highlight the necessity of greater instructor control and enhanced realism to strengthen both technical accuracy and clinical immersion.

Another recurring theme was inclusivity and communication. As students reported, *"Students expressed hope that the AI would improve to better understand and communicate with a diverse array of students. One student noted that the HAL had trouble answering questions due to their accent: 'Like, because if I said something, you know, with my accent, he didn't understand it.' This student said that, given the diverse population of nursing students at SON, the AI should be able to respond to a wide range of accents."* Addressing speech recognition accuracy and natural language responsiveness not only ensures equitable access for diverse learners but also prepares students for real-world interactions with patients from varied cultural and linguistic backgrounds.

Finally, the refinement of speech capabilities remains a central priority. *"As previously noted, the AI mannequin's speech capabilities sometimes acted unrealistically. Students indicated that they wished the mannequin would respond to questions and statements in a more human-like way. This extended to the actual words the mannequin was saying and the robotic voice. This would help students learn better how to interact with real patients and answer questions that they may have. Many students expressed that, once the speech capabilities of the AI mannequin are more refined, the mannequin would provide a valuable opportunity for students to improve their communication skills. However, HAL did come with limitations."* Integrating these improvements—better speech fidelity, instructor overrides, inclusivity in accent recognition, and expanded physical responsiveness—would not only increase acceptability but also directly address the identified gaps in usability, realism, and learning transfer, ultimately transforming AI-enabled mannequins into more effective and inclusive educational tools.

6.5 FUTURE USE OF HAL: ASSESSMENTS, SENSITIVE TOPICS, AND COMPLEX HEALTH EDUCATION

The potential of AI-enabled mannequins such as HAL extends far beyond acute clinical scenarios, offering meaningful applications in assessments, sensitive communication, and patient education. Students in the pilot study emphasized that *"thinking about it, things that we could use him for in our setting, in our nurse anesthesia setting, are certainly preoperative assessments. When they're going in, currently, we use faculty to answer those questions."* This insight highlights a critical opportunity: HAL could standardize preoperative assessment practice, reduce faculty burden, and provide students with realistic, repeatable experiences in history-taking and risk assessment. Importantly, these scenarios allow for objective performance evaluation—an area underexplored in current research, where most outcomes rely heavily on self-report rather than observed clinical actions.

Equally significant is HAL's promise in simulating one-on-one interactions that involve sensitive or stigmatized topics. *"That could be a use for him, where you're going in and asking him what meds do you take, can you walk up and down a flight of stairs? Do you take illicit drugs?"* As one participant noted, *"Sometimes students are really uncomfortable talking with patients."* HAL's AI-driven dialogue capability creates a safe, judgment-free environment for learners to practice delicate conversations, such as substance use screening or discussions about sexual health. From a research perspective, this offers a way to measure not only knowledge and attitudes, but also observable communication competencies and learner comfort levels. Future studies could systematically compare HAL to standardized patients in addressing these sensitive exchanges, thereby advancing both inclusivity and effectiveness in nursing education.

In addition to sensitive topics, students also envisioned HAL as a tool for complex health education. *"Patients who are diabetic have to learn a lotta (lot of) complex information about taking their diabetes medications as well as managing their diet… patient education is a really challenging thing. And it is helpful for providers to do roleplay and practice it before they start to do it with regular patients."* This aligns with the broader need to test HAL in patient education simulations where providers must break down technical information into accessible language. Building on identified research gaps, future work should evaluate HAL's effectiveness in teaching communication, cultural competence, and health literacy skills. Moreover, incorporating instructor override features and analyzing usability across diverse student populations will ensure HAL remains both technically reliable and equitable. By integrating these future directions, HAL can evolve into a comprehensive educational partner, bridging technical skill-building with the nuanced interpersonal dimensions of nursing practice.

VII. DISCUSSION

The findings from this pilot study demonstrate that AI-enabled mannequins are both feasible and acceptable for integration into nursing simulation education, particularly in preparing students to respond effectively to opioid-involved overdoses. Students reported high acceptability, appropriateness, and feasibility scores, with relatively low task burden, indicating that learners were able to engage with the AI-driven simulation environment without significant cognitive overload. These results align with prior research highlighting the promise of AI in simulation, yet this study extends the field by focusing not on chatbots or screen-based applications but on embodied, high-fidelity AI mannequins. Importantly, the sustained improvements in both knowledge and attitudes up to one month after training suggest that this approach may foster deeper learning and longer retention compared with interventions that measure only immediate post-training outcomes. This sustained change points toward the potential of AI mannequins to address public health priorities by equipping nurses with psychomotor and decision-making skills essential for overdose response.

Nevertheless, the study's limitations highlight significant research gaps that future investigations must address. The small, convenience-based sample and lack of a comparator group restrict the generalizability of findings and make it difficult to determine whether AI-enabled mannequins offer unique benefits over traditional simulation approaches such as standardized patients or non-AI mannequins. Furthermore, although validated scales were employed to assess knowledge and attitudes, the absence of objective performance metrics, such as time to naloxone administration or accuracy of procedural steps, limits understanding of how these gains translate into clinical competence. Technical issues such as speech recognition errors, latency in responses, and inconsistent physiologic cues further complicated the simulation experience, sometimes shifting the focus from clinical learning to technological troubleshooting. Additionally, questions of equity remain underexplored, as anecdotal reports suggest that the system may struggle to recognize diverse accents, potentially disadvantaging some learners. These gaps underscore the need for research that goes beyond feasibility to rigorously evaluate the effectiveness, inclusivity, and scalability of AI mannequins in health professions education.

Building on these insights, future research should prioritize comparative and objective evaluations of AI-enabled simulation. Randomized controlled trials contrasting AI mannequins with traditional simulation methods would provide stronger evidence on effectiveness, particularly if coupled with blinded assessments of observable clinical behaviors. Studies should also systematically evaluate the technical performance of AI mannequins, quantifying latency and speech recognition accuracy across diverse learner populations, and investigating how these factors influence trust, realism, and learning outcomes. Equally important are implementation studies exploring the perspectives of faculty, simulation technicians, and administrators to identify practical barriers and facilitators to adoption, including cost-effectiveness and instructor training needs. Introducing design innovations such as instructor override functions could mitigate current usability issues, while attention to inclusivity would ensure that AI-enabled tools support all learners equitably. By addressing these limitations and pursuing more rigorous, multi-faceted investigations, future research can move beyond demonstrating feasibility toward establishing the true pedagogical value and clinical impact of AI mannequins. Such work will not only refine the technology itself but also inform evidence-based integration into nursing curricula, thereby advancing both educational practice and patient care in addressing the opioid crisis.

VIII. CONSLUSION

The integration of AI-enabled mannequins into simulation-based nursing education demonstrates promising feasibility and acceptability, particularly in preparing learners

to manage complex scenarios such as opioid-involved overdoses. Evidence from prior studies shows that such technology can enhance knowledge retention and improve learner attitudes, indicating strong potential for augmenting traditional teaching modalities. However, the persistence of usability issues—such as speech recognition errors, latency in responses, and limited physiological fidelity—suggests that while the technology is well-received, it is not yet optimized for seamless incorporation into educational settings. These challenges highlight both the opportunities and the need for rigorous evaluation of how artificial intelligence can best support competency-based nursing education.

Despite encouraging outcomes, current research remains constrained by methodological limitations, including small sample sizes, reliance on self-reported outcomes, and the absence of control groups or comparative modalities. Furthermore, gaps exist in understanding how AI-mannequins affect objective clinical performance metrics such as timeliness of interventions, accuracy of decision-making, and team-based communication under pressure. Questions also remain regarding cost-effectiveness, faculty training requirements, and equity concerns—especially related to speech recognition accuracy across diverse accents and languages. Addressing these gaps requires moving beyond feasibility studies toward controlled trials and implementation-focused research that measure not only perceptions and attitudes but also observable behaviors and measurable patient-care outcomes.

Building on these insights, future research should prioritize comparative effectiveness studies between AI-enhanced mannequins, traditional high-fidelity models, and standardized patients to evaluate relative value. Incorporating objective performance checklists, blinded evaluations, and longitudinal follow-up would provide stronger evidence of learning transfer to clinical practice. Equally important is the need to explore technical optimization, such as real-time instructor override tools and bias mitigation in language recognition, alongside economic and scalability analyses to guide adoption. By pursuing these directions, forthcoming work can extend beyond demonstrating feasibility to establishing evidence-based best practices, thereby ensuring that AI-enabled simulation is not only innovative but also equitable, efficient, and impactful in preparing nursing professionals for real-world clinical challenges.

REFERENCES

[1] Bangor, A., Kortum, P., & Miller, J. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction* , *24* (6), 574-594.

[2] Combs, C., & Combs, P. (2019). Emerging roles of virtual patients in the age of AI. *American Medical Association Journal of Ethics* , *21* (2), E153-E159. https:// doi.org/ 10.1001/ amajethics.2019.153.

[3] Dai, C., & Ke, F. (2022). Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review. *Computers and Education: Artificial Intelligence* , *3* , Article 100087 1-17. https: // doi.org/ 10.1016/ j.caeai.2022.100087.

[4] Devos, H., Gustafson, K., Ahmadnezhad, P., Liao, K., Mahnken, J., Brooks, W., & Burns, J. (2020). Psychometric properties of NASA-TLX and index of cognitive activity as measures of cognitive workload in older adults. *Brain Sciences* , *10* (994), 1-13.

[5] Egelund, E., Gannon, J., Domenico, L., Nobles, P., & Motycka (2020). Recognizing opioid addiction and overdose: An interprofessional sim-ulation for medical, nursing, and pharmacy students. *Journal of Inter-professional Education & Practice* , *20* , Article 100347. https://doi.org/ 10.1016/j.xjep.2020.100347.

[6] Goldenberg, M. (2024). Surgical artificial intelligence in urology: Educa-tional applications. *Urologic Clinics of North America* , *51* (1), 105-115. https:// doi.org/ 10.1016/ j.ucl.2023.06.003.

[7] Hyzy, M., Bond, R., Mulvenna, M., Bai, L., Dix, A., Leigh, S., & Hunt, S. (2022). System usability scale benchmarking for digital health apps: Meta-analysis. *JMIR mHealth and uHealth* , *10* (8), Article e37290.

[8] Beth Ann Swan, Sarah Febres-Cordero,Laika Steiger, Alexa Lisenby, Tatiana Getz, Jack Hudson, Katie Cole, Russ Branch, Carrie McDermott, Kim Fugate, Nicholas A. Giordano, Feasibility and acceptability of incorporating artificial intelligence into simulation education, Clinical Simulation in NursingVolume 104, July 2025, 101739, doi.org/10.1016/j.ecns.2025.101739.

[9] Jiang, N., Zhang, Y., Liang, S., Lyu, X., Chen, S., Huang, X., & Pan, H. (2024). Effectiveness of virtual simulations versus mannequins and real person s in medical and nursing education: Meta-analysis and trial sequential analysis of randomized controlled trials. *Jour-nal of Medical Internet Research* , *26* , Article e56195 1-13. https: // doi.org/ 10.2196/ 56195.

[10] Jung, S. (2023). Challenges for future directions for artificial intelligence integrated nursing simulation education. *Korean Journal of Women Health Nursing* , *29* (3), 239-242. https:// doi.org/ 10.4069/ kjwhn.2023. 09.06.1.

[11] Keenan, M., Schenker, K., & Sarsfield, M. (2017). A complicated opi-oid overdose: A simulation for emergency medicine residents. *MedEd- PORTAL* , *13* , Article 10616. https:// doi.org/ 10.15766/ mep_2374-8265. 10616.

[12] Liaw, S., Tan, J., Lim, S., Zhou, W., Yap, J., Ratan, R., Ooi, S., Wong, S., Seah, B., & Chua, W. (2023). Artificial intelligence in virtual reality simulation for interprofessional communication training: Mixed method study. *Nurse Education Today* , *122* , 105718. https:// doi.org/ 10.1016/ j. nedt.2023.105718.

[13] Shim, J., Kim, J., Pyun, J., Cho, S., Oh, M., Kang, S., Lee, J., Kim, j., Cheon, J., & Kang, S. (2018). Comparison of effective teaching meth-ods to achieve skill acquisition using a robotic virtual reality simula-tor: Expert proctoring versus an educational video versus independent training. *Medicine* , *97* (51), Article e13569.

[14] Shorey, S., Mattar, C., Pereira, T., & Choolani, M. (2024). A scoping review of ChatGPT's role in healthcare education and research. *Nurse Education Today* , *135* , Article 106121. https:// doi.org/ 10.1016/ j.nedt. 2024.106121.

[15] Simsek-Cetinkaya, S., & Cakir, S. (2023). Evaluation of the effectiveness of artificial intelligence assisted interactive screen-based simulation in breast self-examination: An innovative approach in nursing students. *Nurse Education Today* , *127* , Article 105857. https:// doi.org/ 10.1016/ j.nedt.2023.105857.

[16] Weiner, B., Lewis, C., Stanick, C., Powell, B., Dorsey, C., Clary, A., Boynton, M., & Halko, H. (2017). Psychometric assessment of three newly developed implementation outcome measures. *Implementation Science* , *12* (1), 108. https:// doi.org/ 10.1186/ s13012- 017- 0635- 3.

[17] Williams, A., Strang, J., & Marsden, J. (2013). Development of opioid overdose knowledge (OOKS) and attitudes (OOAS) scales for take-home naloxone training evaluation. *Drug and Alcohol Dependence* , *132* (1-2), 383-386. https:// doi.org/ 10.1016/ j.drugalcdep.2013.02.007.

[18] Winkler-Schwartz, A., Yilmaz, R., Mirchi, N., Bissonnette, V., Led-wos, N., Siyar, S., Azarnoush, H., Karlik, B., & Del Maestro, R. (2019). Machine learning identification of surgical and operative factors associ-ated with surgical expertise in virtual reality simulation. *JAMA Network Open* , *2* (8), Article e198363.

[19] Xu, J., Yang, L., & Guo, M. (2024). Designing and evaluating an emo-tionally responsive virtual patient simulation. *Simulation in Healthcare* , *19* (3), 196-203. https:// doi.org/ 10.1097/ sih.0000000000000730.