

Implementation of AI Characters for Simulation of Root Cause Analysis in the ICU Setting

Yuqi Hu¹, Qiwen Xiong¹, Zhenzhen Qin¹, Brandon Watanabe², Yujing Wang¹, Ilmi Yoon^{1*}

¹Northeastern University ²San Francisco State University

*Corresponding to: i.yoon@northeastern.edu

Abstract—Root Cause Analysis (RCA) is widely used to investigate adverse events in healthcare settings, yet hands-on RCA training is often constrained by instructor time, standardized patient availability, and logistical burden. This paper documents the implementation of an AI-based RCA simulation for an ICU adverse event case, with an emphasis on prompt engineering for large language model (LLM) components. Specifically, we present necessary details for implementation, including: (1) a fully specified ICU case narrative, (2) role schemas and system prompts for virtual avatars, (3) states-of-mind prompts that modulate interviewee behavior, (4) voice design specifications for virtual avatar, and (5) analytic rubrics and assessment prompts for formative and summative feedback. We also include examples of LLM-generated dialogue and assessment reports to illustrate expected outputs and formats. Lastly, we discuss limitations and future directions for the presented system.

Keywords—root cause analysis; AI; virtual avatars; healthcare education; ICU; formative assessment

I. INTRODUCTION

Adverse events remain a major patient safety challenge. Estimates suggest that more than 250,000 patients in the United States experience adverse events each year [1], in which medical treatment itself contributes to patient harm [2]. Root Cause Analysis (RCA) is widely used to investigate such events and reduce recurrence by focusing on system-level contributors rather than individual blame [3][4]. RCA is broadly adopted across healthcare organizations to examine failures involving workflows, communication, and policy gaps [5][6]. In large systems, RCAs occur routinely; for example, Veteran Affairs medical centers report multiple RCAs per year on average [7]. Although procedures vary, most RCAs follow a common structure: describing the event, identifying contributing factors, determining root causes, and proposing corrective actions [8][9].

Despite its value, RCA is often inconsistently executed, in part because the skills it requires are difficult to develop without repeated practice. Effective investigations depend on eliciting detailed accounts, reconciling conflicting perspectives, distinguishing proximate from systemic contributors, and translating findings into feasible corrective actions [10][11]. Yet hands-on RCA instruction is frequently constrained by instructor time and logistical burden [12]. Although RCA content appears in medical education, nursing programs, and continuing education, training depth and consistency vary widely [13]–[17]. As a result, learners may receive limited opportunities to practice the interpersonal and investigative

components of RCA, particularly interviewing stakeholders who may be defensive, uncertain, or emotionally affected by the event.

Recent advances in AI, especially large language models (LLMs), create opportunities for scalable, interactive practice. LLMs can generate coherent, context-sensitive dialogue and can be shaped through prompt engineering to produce role-consistent responses and controlled behavioral variability [18]. In healthcare education, LLM-enabled simulations have been explored to support clinical reasoning and communication skills through interactive virtual agents [19]–[21]. LLMs can also generate structured formative feedback and evaluate open-ended learner work when paired with explicit rubrics, improving the actionability of feedback while reducing educator workload [22]–[24].

In this paper, we document the implementation of an AI-based RCA simulation for an ICU adverse event case. The system builds on *Nurse Town*, a Unity-based 3D simulation game for nursing education [25], and adapts a textbook ICU failure scenario [26]. We focus on replication-oriented artifacts and design decisions, including: (1) a fully specified case narrative, (2) role schemas for five virtual ICU team member avatars, (3) states-of-mind prompts that modulate disclosure style, (4) voice design targets that support role-appropriate affect, (5) a structured RCA report template, and (6) analytic rubrics for formative interview feedback and summative report assessment.

II. RELATED WORK

A. Root Cause Analysis and Patient Safety Training

RCA gained prominence in the 1990s quality improvement movement as healthcare organizations increasingly recognized that harm often stems from system failures rather than isolated individual mistakes [27]. In the United States, RCA was promoted as a key method for investigating adverse events and preventing recurrence [28], and it became institutionalized through accreditation and safety oversight (e.g., sentinel event analyses) [29][30]. Guidance from major safety organizations has further supported broad adoption in both clinical and administrative domains [31], and structured RCA programs have been used in large healthcare systems to operationalize patient safety initiatives [32]. RCA's emphasis on learning-oriented, system-focused improvement can support a culture of safety, shifting attention away from blame and toward actionable redesign [33][34].

However, RCA effectiveness depends heavily on execution quality. Teams must gather accurate accounts, navigate interpersonal dynamics during interviews, and synthesize evidence into root causes and measurable corrective actions [3][9]. Training gaps are common, and inconsistent preparation can limit the rigor and impact of RCA findings [10][11]. While RCA concepts are taught in professional programs, scalable hands-on practice remains difficult; simulation-based or case-based interventions can improve realism but are often constrained by staffing and logistics [12][13][15].

B. LLM-Enabled Simulation and Automated Feedback

LLMs can sustain multi-turn, context-sensitive dialogue, enabling conversational agents that simulate realistic interviews for educational purposes [18]. With prompt engineering and role constraints, LLM-driven agents can maintain persona consistency, selectively disclose information, and express interview stances that mirror real-world dynamics (e.g., defensiveness, uncertainty) that are central to RCA practice [35]. In healthcare education, such agents have been explored to support clinical reasoning and communication skill development through interactive practice [19]–[21]. LLMs can also support assessment and feedback on open-ended learner work; prior studies suggest they can generate formative coaching and evaluate text outputs, especially when guided by analytic rubrics that clarify expectations and improve scoring consistency [22]–[24]. These findings motivate our use of rubric-guided assessment and prompt-controlled behavioral variability within an RCA simulation context.

III. SIMULATION SYSTEM IMPLEMENTATION

A. System Overview and Session Flow

The RCA training experience is delivered as a Unity-based 3D simulation in which learners investigate an ICU adverse event by interviewing five virtual ICU team members. Figure 1 illustrates the overall architecture. A typical session begins with the learner reviewing the case background and RCA instructions in an in-game help panel, then selecting an interviewee (e.g., primary nurse, ICU nurse, physician, respiratory therapist, or medical student). Interviews are conducted as speech-to-speech conversations: the learner speaks a question, the system captures audio and transcribes it via speech-to-text (STT), and the transcript is appended to the ongoing interview record. The transcript is then sent to an LLM-driven dialogue component that generates a role-consistent response conditioned on (i) the case facts, (ii) the character schema, and (iii) a sampled state-of-mind prompt that controls the interviewee’s disclosure style (e.g., defensive vs. self-reflective). Next, the generated response is rendered through emotional text-to-speech (TTS), while the avatar’s facial expression, lip sync, and body gestures are animated to align with the intended affect.

After completing all interviews, the learner transitions to a synthesis phase in which they complete a structured RCA report using the provided template. The report is uploaded within the system and evaluated by a separate LLM-based assessment module guided by analytic rubrics. The module produces two

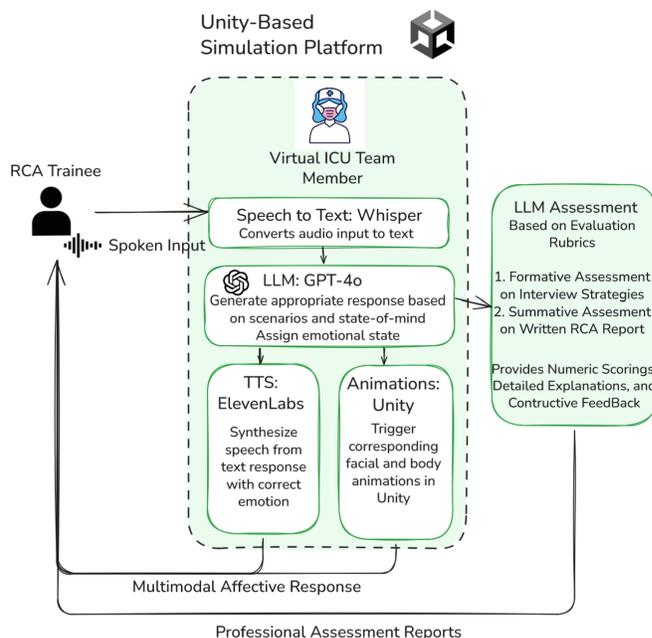


Figure 1. Architecture of the virtual RCA simulation system.

deliverables: (1) formative feedback on interviewing behaviors (e.g., depth of inquiry, follow-up quality, and theme coverage) based on the interview transcripts, and (2) summative scoring and narrative feedback on the written RCA report (e.g., clarity of problem statement, identification of direct and root causes, feasibility of corrective actions, and use of interview evidence). Feedback is returned to the learner in a structured report view that can be revisited to support iterative practice and improvement. An example of learner-virtual character dialogue and feedback report can be found in the Appendix.

B. Simulation Scenario Overview

The simulation centers on an ICU adverse event in which inconsistent wristband conventions and communication breakdowns during an emergency contribute to a fatal outcome. A patient admitted for IV antibiotics for pneumonia was initially stable; during intake, the primary nurse identified a penicillin allergy and applied a red wristband for “no blood draw” and a blue wristband mistakenly believed to indicate “allergy” based on another hospital’s convention, even though in this hospital blue signifies “Do Not Resuscitate” (DNR). Shortly after the infusion began, the patient developed signs of anaphylaxis; the nurse stopped the infusion, administered oxygen, and called a code when the patient became unresponsive. During resuscitation, an ICU nurse noticed the blue wristband and raised concern that the patient might be DNR, prompting the team to pause and verify code status while the primary nurse searched for the chart; during this delay, the rhythm deteriorated from ventricular tachycardia to asystole, and despite resuming efforts after confirming full-code status, the patient was pronounced dead. The case surfaces multiple

system contributors relevant to RCA training, including non-standardized wristband color systems, unclear ownership of wristband application, disorganized supplies, communication breakdowns under time pressure, and fatigue-related risk from excessive work hours.

C. Virtual Characters in the Simulation

The simulation includes five interviewees representing key ICU stakeholders involved in the event. Each character is instantiated as a unique 3D avatar with an LLM-driven persona, specified with (i) professional responsibilities and perspective, (ii) the subset of case facts the character is expected to know, (iii) likely concerns or biases during an RCA interview, and (iv) communication tendencies (e.g., willingness to elaborate, formality, and stance toward accountability). In addition to language, each character is assigned a voice profile to support immersion and to cue typical interpersonal dynamics encountered in real RCA interviews (e.g., fatigue, hesitancy, urgency). Table I summarizes each character's role and key attributes/actions, along with the corresponding voice design targets used to realize their speech.

Pedagogically, the *Primary Nurse* serves as the central witness and is positioned to reveal frontline workflow pressures (e.g., workload and fatigue) as well as local conditions that enabled the wristband mix-up. The *Code Team Medical Student* represents a junior member who can surface hierarchy and speaking-up dynamics during emergencies, creating opportunities for learners to practice eliciting information from hesitant informants. The *Code Team ICU Nurse* is designed as a more assertive safety advocate who identifies the wristband ambiguity during the code, helping learners probe system-level safeguards and workaround behaviors. The *Code Team Doctor* reflects leadership priorities and time-critical decision-making under pressure, allowing learners to explore how urgency and institutional metrics can shape actions during a code. Finally, the *Respiratory Therapist/Anesthesiologist* focuses on airway management and visual status indicators, prompting learners to examine whether safety signals were sufficiently clear and redundant for high-acuity settings.

To ensure that each LLM-driven avatar reliably maintains its assigned persona, we use a layered prompt structure with explicit constraints. First, a *role system prompt* encodes the character schema (profession, responsibilities, viewpoint, and communication style) and establishes hard boundaries on what the avatar can claim to have seen or known. Second, a *case grounding block* provides the incident narrative and a character-specific "known facts" list, reducing hallucination by constraining answers to an approved knowledge set. Third, a short *behavioral policy* enforces interaction norms (e.g., answer strictly as the character; do not provide omniscient summaries; if uncertain, state uncertainty and ask for clarification; avoid inventing documentation that was not reviewed; keep answers within a reasonable length; and remain consistent with prior turns). In practice, these constraints are complemented by lightweight runtime checks (e.g., maintaining a rolling memory

of prior claims and re-injecting a compact persona reminder) to reduce drift over longer interviews.

Beyond textual persona control, we assign each character a voice target that is consistent with their role and typical affect during an RCA interview. These voice profiles function as an additional modality-level constraint that helps preserve characterization (e.g., a fatigued nurse sounds strained rather than exuberant; a junior learner sounds hesitant; a physician leader sounds steady and commanding). Operationally, we treat voice design as part of the persona specification: each role schema includes a short voice descriptor (tone, pace, intensity, and emotional range), which is used to select or configure the TTS voice and to maintain consistent delivery across turns.

D. Assigning States-of-Mind to Virtual Avatars

To reduce "scripted" interviews and train adaptive questioning, each avatar adopts one of five states-of-mind at the start of an interview. The state-of-mind does not change the underlying case facts the character may disclose; instead, it modulates *how* information is presented (e.g., level of detail, willingness to accept responsibility, emotional tone, and tendency to elaborate). This design encourages learners to practice follow-up strategies, clarification requests, and rapport-building techniques when encountering reluctance, uncertainty, or frustration.

We implement states-of-mind using a dedicated *style controller prompt* that is injected alongside the fixed character schema. The prompt specifies (i) the target interview stance (e.g., defensive vs. self-reflective), (ii) the preferred linguistic markers (e.g., hedging, minimization, apology, formality), and (iii) disclosure guidelines (e.g., "do not volunteer extra details unless asked" for detached; "redirect blame unless pressed" for defensive). Importantly, we separate *content constraints* from *style constraints*: the character schema and case-grounding block define what the avatar can know and say, while the state-of-mind prompt determines how forthcoming, emotional, or detailed the delivery should be. This separation reduces the risk that style modulation changes factual content.

To keep behavior consistent across long interviews, the state-of-mind is reiterated in a short reminder prefix for each turn (e.g., "Maintain a frustrated tone; emphasize system issues; avoid admitting fault unless directly asked."). We also include guardrails that prevent the avatar from breaking character (e.g., "do not speak as an AI"; "do not narrate the simulation") and from introducing facts outside the approved "known facts" list. When the learner asks ambiguous questions, the prompt instructs the avatar to request clarification rather than guessing, which helps preserve both persona fidelity and factual consistency.

E. Automated Evaluation of Simulation Sessions

The simulation includes an automated evaluation component that assesses learner performance across two complementary products of an RCA investigation: (i) the learner's interview process and (ii) the learner's final written RCA report. At a high level, the system treats each session as a structured record composed of interview transcripts (time-stamped learner

TABLE I. SUMMARY OF VIRTUAL CHARACTERS IN THE ICU FAILURE SIMULATION, INCLUDING ROLE ATTRIBUTES/ACTIONS AND VOICE DESIGN TARGETS.

Character	Key attributes and actions	Voice design target
Primary Nurse	Fatigued and overworked; confused wristband codes due to experience at another hospital; delayed resuscitation while searching for the chart.	Weary and slightly breathy tone reflecting fatigue; professional but strained; may convey frustration or defensiveness.
Code Team Medical Student	Noticed the blue wristband but unsure of its meaning; hesitant to speak up; relied on others for direction.	Soft-spoken and hesitant with pauses indicating uncertainty; youthful and tentative.
Code Team ICU Nurse	Immediately raised concern about the wristband; frustrated by lack of standard DNR indicators; suggested DNR signage above beds.	Direct and experienced; clear articulation; slightly impatient at times but professional.
Code Team Doctor	Led the code team; focused on rapid defibrillation; frustrated by the delay; emphasized teamwork and clear protocols.	Calm and commanding; steady pacing; urgency with control.
Respiratory Therapist / Anesthesiologist	Managed airway; questioned reliance on wristbands alone; recommended better visual indicators for patient status.	Pragmatic and task-focused; concise delivery; occasional disbelief/concern when noting system flaws.

questions and avatar responses) and a completed RCA report template submitted at the end of the scenario. These artifacts are then analyzed by an LLM-based assessment module guided by analytic rubrics, producing both formative feedback (to improve interviewing technique) and summative feedback (to evaluate the quality of RCA synthesis and recommendations).

After completing interviews with the five virtual team members, learners transition to a synthesis phase and complete an RCA report using a structured template. The template first captures basic event metadata, including the event title, date of the event, dates the RCA was initiated and completed, and the names and roles of the facilitator and team members. It then guides learners through a stepwise RCA workflow. Learners begin by selecting and describing the event, articulating why the incident warrants investigation. They next document the RCA charter and team formation, then build a detailed timeline to establish a shared narrative of what occurred. The template prompts learners to identify contributing factors linked to specific steps in the timeline, then iteratively identify root causes using structured “why” reasoning. Finally, learners propose corrective actions that target identified root causes and specify how success will be measured, encouraging a shift from description to prevention-oriented planning.

The assessment module uses analytic rubrics to ensure feedback is interpretable and actionable. Each rubric dimension is scored and accompanied by a brief rationale, a description of strengths and weaknesses, and concrete suggestions for improvement. For *formative* evaluation of interview performance, the rubric emphasizes the learner’s depth of inquiry (e.g., use of open-ended questions and targeted follow-ups), comprehensiveness of investigation (e.g., coverage of actions, decisions, communication breakdowns, and system-level contributors such

as fatigue or unclear protocols), and active listening/adaptability (e.g., probing inconsistencies and adjusting questioning strategy in response to new information). The formative rubric also evaluates whether the learner identifies and pursues key themes relevant to the case (e.g., wristband meaning, role ownership, protocol gaps, workload pressures, emergency communication) and whether the learner maintains professionalism and clarity throughout the interview.

For *summative* evaluation of the written RCA report, the rubric focuses on the clarity of the problem statement and event description, the correctness and completeness of identified direct causes and contributing factors, and the depth of systemic analysis (e.g., workflow design, protocol and training gaps, and communication structures). It further evaluates how effectively the report integrates interview evidence—including triangulation across perspectives and acknowledgement of contradictions or uncertainty—and whether proposed corrective actions are specific, feasible, and appropriately targeted to root causes rather than superficial fixes. Finally, the summative rubric considers overall organization, coherence, and writing quality, reflecting the expectation that an RCA report should be clear enough to support implementation and accountability.

IV. DISCUSSION AND CONCLUSION

This paper documents an implementation-oriented RCA simulation package for an ICU patient safety case, emphasizing scenario specification, character and persona design, behavioral variability via states-of-mind, learner deliverables, and rubric-guided assessment scaffolds. A key goal of this work is reusability: the artifacts are structured so they can be adopted in multiple instructional formats and with varying levels of technical infrastructure.

TABLE II. STATES-OF-MIND USED TO MODULATE AVATAR DISCLOSURE STYLE AND TONE.

State-of-mind	Behavioral signature in interviews	Example dialogue snippet
Defensive	Deflects blame; minimizes responsibility; may redirect to policies, other roles, or external constraints; can be vague unless prompted with specifics.	“Placing wristbands isn’t really my job. I assumed the ED had handled it. The cabinet was a mess anyway.”
Self-Reflective / Honest	Acknowledges mistakes; offers detailed chronology; expresses regret; more likely to volunteer contributing factors and opportunities for improvement.	“I should’ve double-checked the wristband color. I just assumed, and that was a serious mistake on my part.”
Confused / Uncertain	Hesitant recall; inconsistent details; uses hedging language; requires clarification and timeline reconstruction.	“I think... it was the blue band that made everyone pause? Or maybe someone said something about DNR... I’m not totally sure.”
Overly Professional / Detached	Formal and emotionally flat; answers narrowly and literally; avoids interpretation or speculation; limited volunteering of context.	“At the time of arrival, the patient was in V-tach. I began airway management. I cannot comment on code status decisions.”
Frustrated	Emphasizes systemic problems; expresses dissatisfaction with processes and resources; may downplay personal mistakes; tends to “vent” about organizational issues.	“This wouldn’t have happened if we had standardized wristband colors. It’s ridiculous how disorganized things are here.”

First, the case narrative and character role schemas can be used beyond the Unity simulation itself. Instructors may deploy them as paper-based RCA exercises, as interview-based assessments (e.g., OSCE-style stations), or as guided small-group activities in which learners practice eliciting perspectives and reconstructing a timeline. Second, the states-of-mind mechanism provides a lightweight way to introduce controlled variability in interview dynamics. By shifting disclosure style and emotional tone without altering the underlying facts, the system encourages learners to practice adaptive questioning strategies, clarification, and rapport-building in the face of defensiveness, uncertainty, or frustration. Third, the structured RCA template and analytic rubrics support consistent expectations across instructors and cohorts. They can be used with automated assessment to reduce faculty workload and provide rapid feedback, or they can be applied in instructor-led settings to standardize grading and feedback.

Overall, these materials are intended to lower barriers to adopting simulation-based RCA training and to support replication and adaptation across institutions and curricula. Future extensions include expanding the scenario library to cover a wider range of adverse events, refining persona control to reduce drift over long interviews, and evaluating how different configurations of automated feedback influence learning outcomes and user trust.

APPENDIX

A. Sample Interview Transcript (Excerpt)

Below is an excerpt demonstrating a “Frustrated” primary nurse interview style.

- Learner: Hello, I’m here to talk about the ICU incident earlier.
- Primary Nurse (Frustrated): Hi. I’m here to discuss that, too. It’s been... quite overwhelming, to say the least. Where would you like to start?
- Learner: Why did the code team pause resuscitation?
- Primary Nurse (Frustrated): It was the wristband confusion. At my other job, blue means allergy. Here it means DNR. Everyone hesitated because of that.
- Learner: Have you ever seen this confusion before?
- Primary Nurse (Frustrated): Not here, but that’s exactly the problem—different hospitals use different systems. It’s frustrating.
- Learner: What would prevent this from happening again?
- Primary Nurse (Frustrated): Standardize wristband colors, and organize supplies. Also, make it crystal clear who is responsible for applying bands.
- Learner: Anything else contribute?
- Primary Nurse (Frustrated): Fatigue. I’ve been working nonstop. That absolutely clouded my judgment.

B. Sample Assessment Output Format (Formative)

- **Depth of Inquiry: 8/10.** Strong open-ended questions and relevant follow-ups; add targeted probes about decisions during the code.
- **Comprehensiveness: 7/10.** Covered communication/protocol issues; further explore staffing and training gaps.
- **Active Listening: 9/10.** Adapted questions to interviewee comments; probe inconsistencies more explicitly.

Summative report feedback (example formatting):

- **Problem Statement: 9/10.** Clear description of wristband misinterpretation and resuscitation delay.
- **Causes: Immediate 8/10; Contributing 9/10.** Strong analysis of fatigue and environment; add detail on hierarchy/speaking-up dynamics.
- **Solutions: 9/10.** Practical and targeted; include accountability and measurement plans.

REFERENCES

- [1] J. G. Anderson and K. Abrahamson, "Your health care may kill you: Medical errors", in *Building Capacity for Health Informatics in the Future*, IOS Press, 2017, pp. 13–17.
- [2] M. Jason Boulanger and C. Keohane, "Role of patient safety organizations in improving patient safety", *Patient Safety in Obstetrics and Gynecology, An Issue of Obstetrics and Gynecology Clinics*, vol. 46, no. 2, p. 257, 2019.
- [3] J. J. Rooney and L. N. Heuvel, "Root cause analysis for beginners", *Quality progress*, vol. 37, pp. 45–56, 2004.
- [4] I. Tjia et al., "Wake up safe and root cause analysis: Quality improvement in pediatric anesthesia", *Anesthesia & Analgesia*, vol. 119, no. 1, pp. 122–136, 2014.
- [5] E. Flink et al., "Lessons learned from the evolution of mandatory adverse event reporting systems", in *Advances in Patient Safety: From Research to Implementation (Volume 3: Implementation Issues)*. Agency for Healthcare Research and Quality, 2005.
- [6] K. M. Kellogg et al., "Our current approach to root cause analysis: Is it contributing to our failure to improve patient safety?", *BMJ Quality & Safety*, vol. 26, no. 5, pp. 381–387, 2017.
- [7] K. B. Percarpio, B. V. Watts, and W. B. Weeks, "The effectiveness of root cause analysis: What does the literature tell us?", *The Joint Commission Journal on Quality and Patient Safety*, vol. 34, no. 7, pp. 391–398, 2008.
- [8] P. M. Williams, "Techniques for root cause analysis", in *Baylor University Medical Center Proceedings*, vol. 14, Taylor & Francis, 2001, pp. 154–157.
- [9] P. Gangidi, "A systematic approach to root cause analysis using 3×5 whys technique", *International Journal of Lean Six Sigma*, vol. 10, no. 1, pp. 295–310, 2019.
- [10] A. W. Wu, A. K. Lipshutz, and P. J. Pronovost, "Effectiveness and efficiency of root cause analysis in medicine", *Jama*, vol. 299, no. 6, pp. 685–687, 2008.
- [11] R. J. Latino, "How is the effectiveness of root cause analysis measured in healthcare?", *Journal of Healthcare Risk Management*, vol. 35, no. 2, pp. 21–30, 2015.
- [12] M. Aboumrad, J. Neily, and B. V. Watts, "Teaching root cause analysis using simulation: Curriculum and outcomes", *Journal of Medical Education and Curricular Development*, vol. 6, p. 2382120519894270, 2019.
- [13] L. W. Hall et al., "Effectiveness of patient safety training in equipping medical students to recognise safety hazards and propose robust interventions", *BMJ Quality & Safety*, vol. 19, no. 1, pp. 3–8, 2010.
- [14] M. A. Dolansky, K. Druschel, M. Helba, and K. Courtney, "Nursing student medication errors: A case study using root cause analysis", *Journal of Professional Nursing*, vol. 29, no. 2, pp. 102–108, 2013.
- [15] J. Lambton and L. Mahlmeister, "Conducting root cause analysis with nursing students: Best practice in nursing education", *Journal of Nursing Education*, vol. 49, no. 8, pp. 444–448, 2010.
- [16] P. Bowie, J. Skinner, and C. de Wet, "Training health care professionals in root cause analysis: A cross-sectional study of post-training experiences, benefits and attitudes", *BMC Health Services Research*, vol. 13, pp. 1–10, 2013.
- [17] M. Murphy et al., "Implementation of a mock root cause analysis to provide simulated patient safety training", *BMJ Open Quality*, vol. 6, no. 2, e000096, 2017.
- [18] H. Naveed et al., "A comprehensive overview of large language models", *arXiv preprint arXiv:2307.06435*, 2023.
- [19] D. P. Mohapatra et al., "Leveraging large language models (llm) for the plastic surgery resident training: Do they have a role?", *Indian Journal of Plastic Surgery*, vol. 56, no. 5, pp. 413–420, 2023.
- [20] C. Levin, M. Suliman, E. Naimi, and M. Saban, "Augmenting intensive care unit nursing practice with generative ai: A formative study of diagnostic synergies using simulation-based clinical cases", *Journal of Clinical Nursing*, 2024.
- [21] N. Kapadia et al., "Evaluation of large language model generated dialogues for an ai based vr nurse training simulator", in *International Conference on Human-Computer Interaction*, Springer, 2024, pp. 200–212.
- [22] T. Ruwe and E. Mayweg-Paus, "Embracing llm feedback: The role of feedback providers and provider information for feedback effectiveness", in *Frontiers in Education*, Frontiers Media SA, vol. 9, 2024, p. 1461362.
- [23] J. Schneider, B. Schenk, and C. Niklaus, "Towards llm-based autograding for short textual answers", *arXiv preprint arXiv:2309.11508*, 2023.
- [24] W. Xie, J. Niu, C. J. Xue, and N. Guan, "Grade like a human: Rethinking automated assessment with large language models", *arXiv preprint arXiv:2405.19694*, 2024.
- [25] Y. Hu, Q. Xiong, L. Yi, and I. Yoon, "Nurse town: An llm-powered simulation game for nursing education", in *2025 IEEE Conference on Artificial Intelligence (CAI)*, IEEE, 2025, pp. 215–222.
- [26] S. H. Campbell and K. Daley, Eds., *Simulation scenarios for nursing educators: making it real*. Springer Publishing Company, 2017.
- [27] L. Emanuel et al., "What exactly is patient safety?", *Journal of Medical Regulation*, vol. 95, no. 1, pp. 13–24, 2009.
- [28] D. H. Havens and L. Boroughs, "'to err is human': A report from the institute of medicine", *Journal of Pediatric Health Care*, vol. 14, no. 2, pp. 77–80, 2000.
- [29] J. Sherwin, "Contemporary topics in health care: Root cause analysis", *PT Motion*, vol. 3, pp. 28–31, 2011.
- [30] The Joint Commission, *Sentinel event policy and procedures*, <https://www.jointcommission.org/resources/sentinel-event/>, Accessed: April 3, 2025, 2025.
- [31] I. for Healthcare Improvement, *Patient safety 104: Root cause and systems analysis summary sheet*, [https://www.ihl.org/sites/default/files/lms/legacy/education/IHIOpenSchool/Courses/Documents/SummaryDocuments/PS % 20104 % 20SummaryFINAL.pdf](https://www.ihl.org/sites/default/files/lms/legacy/education/IHIOpenSchool/Courses/Documents/SummaryDocuments/PS%20104%20SummaryFINAL.pdf), Accessed: April 3, 2025, 2015.
- [32] J. P. Bagian et al., "Developing and deploying a patient safety program in a large health care delivery system: You can't fix what you don't know about", *The Joint Commission Journal on Quality Improvement*, vol. 27, no. 10, pp. 522–532, 2001.
- [33] R. A. M. Iedema et al., "Turning the medical gaze in upon itself: Root cause analysis and the investigation of clinical error", *Social Science & Medicine*, vol. 62, no. 7, pp. 1605–1615, 2006.
- [34] C. Vincent, "Understanding and responding to adverse events", *New England Journal of Medicine*, vol. 348, no. 11, pp. 1051–1056, 2003.
- [35] C. Grévisse, "Raspatient pi: A low-cost customizable llm-based virtual standardized patient simulator", in *International Conference on Applied Informatics*, 2024, pp. 125–137.