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

AIHealth 2026 - Benchmark and Implementation

Yuqi Hu1, Qiwen Xiong1, Zhenzhen Qin1, Brandon Watanabe2, Yujing Wang1, Ilmi Yoon1*
1: Northeastern University
2: San Francisco State University
*Contact: i.yoon@northeastern.edu

1. Why simulate Root Cause Analysis (RCA)?

Hands-on RCA training is hard to scale

- Constraints: instructor time, standardized patient availability, logistical burden
- RCA skills need practice: interviewing, reconciling perspectives, system-level reasoning
- LLMs can enable role-consistent multi-turn dialogue + automated rubric-guided feedback

Paper focus (replication artifacts)

- Fully specified ICU case narrative
- Role schemas + system prompts for 5 avatars
- “States-of-mind” prompts to vary disclosure style
- Voice design targets for affect & immersion
- Rubrics + assessment prompts (formative + summative)

2. Contribution Type

BENCHMARK + IMPLEMENTATION

Implementation

What you can replicate:

- Unity-based 3D session flow
- Speech-to-speech interviewing (STT → LLM → TTS + animation)
- Persona control (role + case grounding + policy + reminders)
- Automated evaluation outputs (formative + summative)

Benchmark

Reusable baseline for comparison:

- Fixed ICU adverse-event scenario + “known facts”
- Standardized roles (5 stakeholders)
- Controlled interview variability (5 states-of-mind)
- Analytic rubrics for interview quality + RCA report scoring

Use it to benchmark persona fidelity, interview coaching, and RCA synthesis quality.

Key idea: separate FACT constraints (case + known facts) from STYLE control (state-of-mind) to vary behavior without changing the ground truth.

3. System Overview (Unity + LLM pipeline)

Session flow

1) Review case background → choose interviewee
2) Speech-to-speech interview (STT → LLM dialogue → emotional TTS + avatar animation)
3) Write structured RCA report
4) LLM assessment produces:
   - Formative interview feedback
   - Summative report scoring + narrative feedback

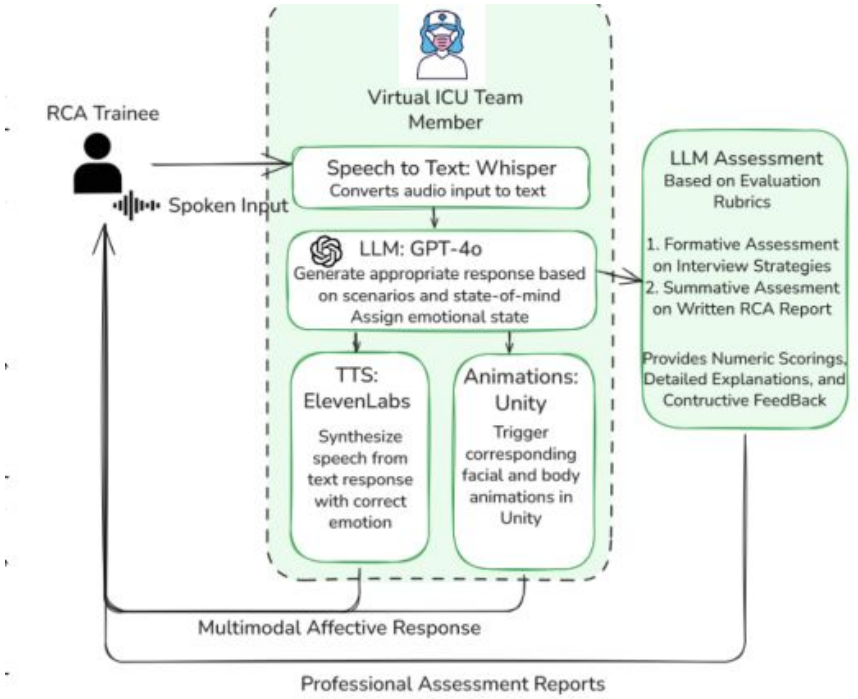

Figure 1. Architecture of the virtual RCA simulation system.

Figure: architecture of the virtual RCA simulation system

4. ICU Scenario (Adverse Event Case)

Core incident: wristband convention mismatch → resuscitation delay

Timeline highlights

- Intake: allergy noted; bands applied using conflicting conventions
- Infusion: anaphylaxis signs → infusion stopped → oxygen + code called
- During code: blue wristband interpreted as DNR → pause to verify status
- Delay: chart search; rhythm deteriorates (V-tach → asystole)
- Full-code confirmed; resuscitation resumes; patient pronounced dead

System contributors to probe

- Non-standardized wristband color systems
- Unclear ownership of wristband application
- Disorganized supplies / environment
- Communication breakdowns under time pressure
- Fatigue risk from excessive work hours

5. Virtual ICU Team (5 interviewees)

Each avatar has (1) role schema, (2) known-facts boundary, and (3) voice target

Character	Key attributes / actions	Voice design target
Primary Nurse	Fatigued/overworked; confused wristband codes; delayed resuscitation while searching chart.	Weary, slightly breathy; professional but strained; may convey frustration or defensiveness.
Medical Student	Noticed blue wristband but unsure; hesitant to speak up; relied on others for direction.	Soft-spoken, hesitant; pauses indicating uncertainty; youthful/tentative.
ICU Nurse	Raised concern immediately; frustrated by unclear DNR indicators; suggested signage above beds.	Direct and experienced; clear articulation; slightly impatient at times but professional.
Doctor (Code Lead)	Led code team; focused on rapid defibrillation; frustrated by delay; emphasized protocols/teamwork.	Calm and commanding; steady pacing; urgency with control.
Resp. Therapist	Managed airway; questioned reliance on wristbands alone; recommended better visual indicators.	Pragmatic and concise; task-focused; occasional disbelief/concern when noting system flaws.

Pedagogical intent: cover frontline workflow pressure, hierarchy/speaking-up, safety advocacy, and time-critical leadership perspectives.

6. Persona Control via Layered Prompting

Goal: role-consistent answers without “omniscient summaries” or hallucinated facts

1) Role system prompt (schema)
Profession, responsibilities, viewpoint, communication style + hard boundaries

2) Case grounding block
Incident narrative + character-specific “known facts” list to constrain content

3) Behavioral policy
Answer strictly as character; if uncertain, ask; avoid inventing documentation; stay consistent

4) Runtime checks (lightweight)
Rolling memory of claims + compact persona reminders to reduce drift

Design principle
Separate:
- CONTENT constraints (schema + known facts)
- STYLE modulation (state-of-mind)

This keeps behavioral variability from changing the ground truth.

Guardrails examples
- “Do not speak as an AI”
- “If uncertain, ask for clarification”
- “Do not introduce facts outside the known-facts list”

7. “States-of-Mind” to Modulate Interview Dynamics

Five styles are sampled per interview; facts stay fixed, disclosure style changes

Defensive
Deflects blame; minimizes responsibility; vague unless pressed.

Self-reflective / Honest
Acknowledges mistakes; detailed chronology; regret; volunteers contributing factors.

Confused / Uncertain
Hedging; inconsistent recall; needs clarification and timeline reconstruction.

Overly Professional / Detached
Formal and narrow; avoids interpretation; minimal context volunteered.

Frustrated
Emphasizes systemic problems; vents about processes/resources; downplays personal mistakes.

Example (Frustrated): “This wouldn’t have happened if we had standardized wristband colors…”

8. Automated Evaluation (Rubric-guided LLM Assessment)

Two complementary products: interview process + final RCA report

Formative (Interview transcripts)

Rubric emphasizes:

- Depth of inquiry (open-ended + targeted follow-ups)
- Comprehensiveness (actions, decisions, comms, system contributors)
- Active listening / adaptability
- Key theme pursuit (wristband meaning, role ownership, fatigue…)
- Professionalism and clarity

Summative (Written RCA report)

Rubric focuses on:

- Clear problem statement + event description
- Correct direct causes + contributing factors
- Depth of systemic analysis (workflow, training/protocol gaps, comms structures)
- Use of interview evidence (triangulation + contradictions)
- Corrective actions: specific, feasible, measurable

Outputs are delivered as a structured report with scores + rationale + actionable suggestions.

9. Example Outputs (Dialogue + Feedback Format)

Interview transcript excerpt (Primary Nurse – "Frustrated")

Learner: Why did the code team pause resuscitation?
Nurse: It was the wristband confusion. At my other job, blue means allergy. Here it means DNR.

Learner: What would prevent this from happening again?
Nurse: Standardize wristband colors… make it crystal clear who is responsible for applying bands.

Nurse (adds): Fatigue… I've been working nonstop. That clouded my judgment.

Formative feedback (format)

- Depth of Inquiry: 8/10
- Comprehensiveness: 7/10
- Active Listening: 9/10

Each includes strengths + targeted improvement suggestions.

Summative report feedback (format)

- Problem Statement: 9/10
- Causes: Immediate 8/10; Contributing 9/10
- Solutions: 9/10

Scoring + rationale + concrete next steps for stronger RCA synthesis.

10. Limitations & Future Directions

What the authors call out (and what to build next)

Limitations / open issues

- Expand scenario library (more adverse-event types)
- Refine persona control to reduce drift over long interviews
- Evaluate how automated feedback configurations affect learning outcomes and user trust

Reusability & adoption

- Deploy artifacts beyond Unity (paper-based, OSCE-style, small groups)
- Use states-of-mind for controlled interview variability
- Use templates + rubrics to standardize expectations and reduce faculty workload

Takeaways

- Implementation package: Unity session flow + speech/LLM components + assessment outputs
- Benchmarkable kit: fixed ICU scenario + roles + states-of-mind + analytic rubrics
- Design trick: separate “what happened” (facts) from “how it’s told” (style)
- Practical goal: scalable RCA interview practice + faster feedback with less faculty load

Contribution type: Benchmark + Implementation