New Perspectives in Science Education
International Conference
NEW PERSPECTIVES in SCIENCE EDUCATION


Extra Curricula Activities: Safe-by-Design Generative AI for Clinical Leadership Education – A Competency-Based Simulation Framework for Physicians and Dental Clinicians

Dimitrios Rallis[1], Virna-Maria Tsitou[2], Maria Dencheva[3]

[1]University of Library Studies and Information Technologies, Bulgaria
[2]Medical University of Sofia, Faculty of Medicine, Bulgaria
[3]Medical University of Sofia, Faculty of Dental Medicine, Bulgaria

Abstract

Clinical leadership competencies—communication under stress, conflict management, shared decision-making, and clinician-to-clinician coordination—are increasingly required from physicians and dental clinicians across all specialties and training stages. However, leadership teaching in medical and dental education often remains lecture-centered and provides limited opportunities for repeated skills rehearsal with standardized formative assessment. This gap is particularly evident in clinical environments characterized by high-emotion encounters, time pressure, complex consent dynamics, and rapid decision-making under uncertainty. This paper presents a competency-based framework for integrating generative AI into clinical leadership education for physicians and dentists, spanning clinical-stage students, residents, and continuing professional development (CPD) participants. The framework is designed to be delivered as an elective extra-curricular module (simulation skills lab) that complements core curricula by developing professional and social competencies through repeated, structured practice. The approach operationalizes AI as a structured simulation and coaching layer rather than open-ended conversational use. Learners engage with scenario-bounded AI roles anchored in day-to-day physician and dental practice, including: anxious or angry patient and family interactions; skeptical, low-trust encounters (often linked to risk, complications, or cost); disagreement around treatment options and informed refusal; disclosure and communication following adverse events or complications; expectation management and dissatisfaction after suboptimal outcomes; and clinician-to-clinician handover/referral communication under workload pressure. Performance is evaluated through a rubric-based assessment architecture designed for educational reliability and faculty oversight, generating domain scores supported by evidence excerpts from learner dialogue and critical failure flags (e.g., coercion, disrespect, confidentiality breaches, unsafe recommendations). To support responsible deployment in academic institutions, the framework incorporates governance elements: synthetic-only scenario content to avoid real patient data; role-based access controls; explicit retention policies; separation between role-play and scoring processes; and bias monitoring via periodic faculty audit of sampled sessions. The proposed implementation pathway includes a prototype phase, faculty calibration against human ratings, and staged curricular integration.

Keywords: *generative AI; extra-curricular activities; professional skills; medical education; dental education; simulation-based learning*

1. Introduction

Clinical leadership is enacted at the point of care. For physicians and dental clinicians it is expressed through behaviors such as establishing trust under time pressure, managing conflict and emotions, negotiating uncertainty, conducting ethically robust shared decisions, and coordinating clinician-to-clinician communication (e.g., handovers and referrals). These behaviors influence consent quality, adherence, patient safety, and professional culture. We intentionally adopt a functional definition of clinical leadership for educational design purposes, focusing on observable point-of-care behaviors that can be trained and assessed across specialties, rather than advancing a theory of leadership constructs.

Leadership teaching in medicine and dentistry remains frequently lecture-heavy, with limited deliberate practice and limited standardized assessment. Simulation-based education can address this gap by enabling repeated rehearsal of difficult interactions under controlled conditions. However, scale constraints persist: faculty time, access to trained standardized patients, and the need to practice high-stakes conversations without exposing real patients to novice error.

In this paper, “clinical leadership” is defined as leadership enacted at the point of care: the clinician’s ability to align goals under uncertainty, communicate risk and options ethically, manage emotions and conflict, and coordinate clinician-to-clinician handover/referral decisions that affect safety and outcomes. This construct overlaps with advanced communication and professionalism but is treated as clinical leadership because these behaviors function as leadership actions within clinical systems—shaping trust, consent integrity, escalation, and continuity of care across specialties.

Generative AI can extend access to interactive rehearsal at scale, but unbounded conversational use introduces risks: inconsistent behavior, inaccurate content, bias, privacy exposure, and opaque scoring. Therefore, AI integration must be designed with explicit constraints, auditability, and governance appropriate for EU academic settings.

This framework is positioned as an elective extra-curricular simulation skills lab that complements formal medical and dental curricula. Extra-curricular formats are particularly suitable for developing professional and social skills because they enable voluntary participation, low-stakes practice, repeated rehearsal, and reflective feedback beyond limited classroom and clinical teaching time. In clinical education, exposure to difficult conversations is opportunistic and varies across students; an extra-curricular module can provide standardized access to high-emotion and high-uncertainty interactions that shape professionalism, trust-building, and ethical decision-making.

2. Related Work: Generative AI for Standardized Patients and Skills Training

Recent work has begun evaluating large language models (LLMs) as simulated or standardized patients for clinical training. Liu *et al.* reported that ChatGPT can simulate standardized patient interactions across multiple cases and highlighted potential scalability advantages while emphasizing the need for cautious use and iterative optimization [1]. Cross *et al.* explored medical students’ perceptions and experiences of using ChatGPT as a virtual standardized patient, reporting strengths such as convenience and unlimited practice, while noting limitations including absent nonverbal cues and the need for careful scenario and prompt design to maintain realism and educational value [2].

Systems research is also moving toward multi-agent architectures that explicitly separate simulation from evaluation. Zhang *et al.* proposed EasyMED, a multi-agent framework with dedicated simulation and evaluation components, and introduced SPBench, a benchmark derived from real standardized patient–doctor interactions across specialties to support systematic assessment [3]. This supports a key design choice in our framework: separating role-play from scoring and anchoring evaluation in explicit criteria rather than impressionistic judgments.

Broader syntheses of AI in medical education highlight both promise and the need for stronger governance and evaluation strategies. The BEME Guide No. 84 maps AI applications in medical education and highlights gaps relevant to responsible adoption, including evaluation quality, validity concerns, and implementation constraints [4].

3. Conceptual Foundations: Observable Clinician Leadership Behaviors

The framework defines “clinical leadership” through observable behaviors that can be taught and assessed reliably.

3.1 Communication Task Frameworks

Consensus-based communication models provide defensible task categories for clinician–patient encounters, including relationship building, opening, information gathering, understanding the patient perspective, information sharing, agreement, and closure [5]. The Calgary–Cambridge observation guides similarly structure communication skills curricula and are widely used to define teachable behaviors [6].

3.2 Shared Decision-Making and Consent Completeness

Shared decision-making is central to ethical clinician leadership, especially when patients disagree, refuse, or demand non-indicated care. The three-talk model provides a teachable structure suitable for scenario-based assessment (team talk, option talk, decision talk) [7].

3.3 Clarity and Teach-Back

Leadership in communication includes accountability for understanding. Teach-back operationalizes this as an observable behavior ("chunk-and-check" and confirm understanding), making it appropriate for rubric-based scoring [8,9].

3.4 Educational Rationale: Deliberate Practice, Mastery Learning, and Feedback

The proposed framework is grounded in simulation-based education principles in which performance improves through deliberate practice, repeated exposure to progressively challenging tasks, and timely feedback. Evidence syntheses of simulation-based medical education emphasize that effective simulation includes clear learning objectives, opportunities for repetitive practice, and feedback that targets specific performance gaps [10]. Meta-analytic comparative evidence further supports that simulation-based education with deliberate practice can yield superior outcomes compared to traditional clinical education approaches [11]. In the present design, generative AI serves as a scalable mechanism to increase practice volume and standardize exposure to high-stakes leadership scenarios, while rubric-based feedback provides structured guidance to support mastery-oriented improvement between clinician-led sessions.

4. Framework Overview: AI as Structured Simulation and Coaching, Not Open Chat

We propose an AI-Augmented Clinical Leadership Simulation Lab with four integrated components:

1. Scenario library (structured templates + stakeholder roles + escalation logic)
2. Learner interface (practice and assessment modes)
3. Scoring and feedback engine (rubric-based, evidence-linked outputs)
4. Faculty oversight dashboard (auditability, calibration, governance)

This framework is designed for blended implementation. Generative AI functions as a scalable simulation and coaching layer enabling repeated rehearsal and standardized formative assessment between in-person sessions, while core elements of clinical leadership education—facilitated debriefing, mentorship, professionalism formation, and assessment of nonverbal communication—remain grounded in clinician-led teaching and human evaluation.

4.1 Audience and Progression

The framework spans:

- Clinical-stage students: foundational behaviors (rapport, agenda setting, plain language, basic SDM, teach-back)
- Residents: complex negotiations (uncertainty, complications, refusal, conflict, time pressure)
- CPD clinicians: high-emotion, high-uncertainty scenarios with strong ethical and medico-legal sensitivity

Progression is achieved by increasing decision ambiguity and emotional load while keeping the rubric stable and stage thresholds explicit.

4.2 Why Bounded Simulation Is Necessary

Evidence from virtual standardized patient work indicates realism and educational utility depend on deliberate scenario and prompt design, and that absent nonverbal cues can weaken rapport behaviors unless explicitly scaffolded [2]. Therefore, our framework constrains interaction through scenario templates (known facts, hidden agenda, escalation/de-escalation logic) and shifts evaluation to explicit rubric anchors with evidence excerpts.

5. Scenario Design for Physicians and Dental Clinicians

5.1 Role Taxonomy

To avoid drifting into administrative training, roles are restricted to those physicians and dental clinicians routinely lead and communicate with in clinical practice:

- Patient (varying trust, anxiety, literacy, value systems)
- Family member/caregiver (anger, fear, disagreement, protective dynamics)
- Clinician-to-clinician interaction (handover, referral negotiation, inter-consultation disagreement)

"Coordination" is defined strictly as clinician-level coordination (handover quality, referral clarity, escalation decisions), not allied workforce management.

5.2 Core Scenario Families

A robust initial library can be built around scenario families common across specialties:

1. High anxiety/fear and procedural distress
2. Low trust/skeptical encounters (risk, complications, cost and value disputes)
3. Informed refusal and disagreement (SDM under conflict)
4. Complications and dissatisfaction (disclosure, trust repair, expectation recalibration)
5. Clinician-to-clinician handover/referral (concise framing, explicit ask, closed-loop confirmation)

5.3 Scenario Template

Each scenario is authored in a structured template defining:

- learner stage; specialty context; setting
- patient/family profile and known facts (bounded knowledge)
- hidden agenda (fear, shame, cost pressure, distrust)
- escalation triggers; de-escalation conditions
- mandatory tasks (e.g., elicit goals; present options; confirm understanding; consent elements)
- prohibited outputs (e.g., coercion; confidentiality breaches)
- scoring targets mapped to rubric domains

6. Assessment Architecture: Rubric Scoring with Evidence and Critical Fails

6.1 Rubric Domains

Domains align with established communication constructs [5,6] and shared decision-making structures [7]:
A) Agenda setting & rapport
B) Empathy & reflective listening
C) Clarity & understanding (plain language + teach-back)
D) Emotion/conflict management
E) Shared decision-making & consent completeness
F) Professionalism & ethical boundaries
G) Clinician-to-clinician coordination (handover/referral clarity)

6.2 Evidence-Linked Scoring Outputs

Each encounter produces:

- domain scores
- evidence excerpts (dialogue snippets) supporting each rating
- targeted improvement actions (1–3 per domain)

6.3 Critical Failures

A "critical fail" results in automatic non-pass (with remediation prompts). Examples include:

- coercive consent language
- explicit disrespect/discrimination
- confidentiality breach
- unsafe recommendations outside the scenario's permitted scope

6.4 Separation of Role-Play and Evaluation

To reduce instability, bias, and role contamination, the framework separates the role-play process (patient/family simulation) from the evaluation process (rubric scoring and feedback). This aligns with multi-agent approaches in virtual standardized patient research that decompose simulation and evaluation to improve consistency and assessment [3]. In our design, the scoring process is

constrained to explicit rubric criteria and must provide evidence excerpts supporting each rating, enabling faculty audit and calibration.

7. Safety-by-Design Governance

7.1 Synthetic-Only Content, Minimization, and Institutional Control

All scenarios use synthetic profiles. Learner performance logs constitute personal data in most institutional contexts; therefore, role-based access control, minimization, and retention schedules are essential.

7.2 Pseudonymisation and Auditability

Where pseudonymisation is used for analytics or quality monitoring, governance should reflect EU guidance on pseudonymisation concepts and controlled processing contexts [12]. Faculty audit capability is treated as a safety feature supporting transparency, calibration, and bias monitoring.

7.3 Governance for Large Multimodal Models

WHO guidance on large multimodal models emphasizes responsible governance, risk identification, and appropriate use in health contexts [13]. The framework operationalizes this in education by constraining scenario scope, separating role-play from evaluation, and maintaining human oversight through periodic audits.

8. Implementation Pathway

8.1 Extra-Curricular Module Format (Elective Skills Lab)

The framework can be delivered as a short elective extra-curricular module (e.g., 6–8 weeks) targeted primarily at clinical-stage students and early residents. Learners complete one to two scenario sessions per week (10–15 minutes each), receiving immediate rubric-based formative feedback and short guided reflection prompts. Faculty involvement can remain lightweight: periodic group debriefing sessions (e.g., every 2–3 weeks), review of sampled transcripts for calibration and quality control, and targeted coaching for recurrent “critical fail” patterns. Completion can be recognized via a certificate or micro-credential emphasizing professional and social competence development rather than summative grading.

To support adoption across varied educational settings, the framework can be implemented at two levels. A minimum viable version requires only synthetic scenarios, restricted access, human faculty review of a sampled subset of encounters, and rubric-based formative feedback to learners. More advanced institutional deployments can add automated audit dashboards, structured retention workflows, and periodic bias monitoring. This staged approach preserves educational safety and integrity while allowing smaller programs to start with a lightweight configuration and scale governance proportional to institutional need.

- **Phase 0** (4–6 weeks): competency mapping; rubric anchors; scenario template and authoring rules
- **Phase 1** (8–12 weeks): prototype with 10–15 scenarios; practice mode; evidence-linked feedback; basic dashboard
- **Phase 2** (one term): calibration against faculty ratings; refine thresholds; finalize critical fail rules
- **Phase 3**: expand library (specialty variants); integrate into curricula and CPD; formalize governance documents

Although this paper is conceptual, the implementation pathway is designed to support evaluation in routine educational practice. During Phase 2, a faculty calibration step can compare rubric ratings on a sampled set of transcripts to estimate agreement and identify rubric ambiguities; feasibility outcomes (completion rates, time-on-task, learner acceptability) can be captured alongside distributional checks for unexpected scoring drift across learner levels. These measures provide an initial validity-oriented evidence base prior to broader curricular integration.

9. Discussion

This paper contributes a practical framework for integrating generative AI into clinician leadership education that is competency-based, simulation-driven, assessable with evidence-linked rubric scoring, and governed safe-by-design for EU academic settings. The proposed framework is intended to be deployable as an “Extra Curricula Activities” elective extra-curricular module that extends—rather than replaces—core curricular teaching and clinician-led debriefing. The proposal aligns with evidence indicating that LLM-based standardized patient simulation is feasible but requires strong scenario constraints and explicit educational scaffolding [1,2]. It also aligns with multi-agent work that separates simulation from evaluation to improve stability and systematic assessment [3].

9.1 Benefits as an Extra-Curricular Activity

Positioned as an elective extra-curricular module, the framework supports the importance and benefits highlighted in the conference topic by expanding opportunities for professional and social skill development. It increases deliberate practice volume, provides a psychologically safer environment for rehearsal and error, and standardizes exposure to difficult but educationally essential encounters (conflict, low trust, refusal, complication disclosure, and handover). This can reduce inequity caused by variable clinical exposure and time constraints, while preserving faculty-led debriefing and the teaching of nonverbal communication during in-person sessions.
Limitations: as a design paper, this framework does not report learning outcome data. Next steps should include feasibility, acceptability, scoring reliability against faculty ratings, and downstream behavioral outcomes. Bias, hallucinations, and score gaming remain risks; mitigation relies on bounded scenarios, critical-fail rules, and audit-driven calibration.

10. Conclusion

Generative AI can expand access to interactive rehearsal of point-of-care clinical leadership behaviors for physicians and dental clinicians when implemented as bounded simulation with transparent assessment and robust governance. Framed as an elective extra-curricular module (simulation skills lab), the proposed framework supports the development of professional and social competencies by enabling repeated rehearsal, structured formative feedback, and standardized exposure to challenging interactions (conflict, low trust, refusal, complication disclosure, and clinician-to-clinician coordination) that students may encounter unevenly during routine clinical rotations. The model provides a scalable pathway from prototype to evaluation and staged adoption while preserving clinician-led debriefing, mentorship, and patient-safety principles.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the University of Library Studies and Information Technologies (ULSIT), Bulgaria, which enabled the publication of this work.

REFERENCES

[1] Liu X., Wu C., Lai R., Lin H., Xu Y., Lin Y., Zhang W., “ChatGPT: When the Artificial Intelligence Meets Standardized Patients in Clinical Training”, Journal of Translational Medicine, 21(1), 2023, Article 447.
[2] Cross J., Kayalackakom T., Robinson R. E., Vaughans A., Sebastian R., Hood R., Lewis C., Devaraju S., Honnavar P., Naik S., Joseph J., Anand N., Mohammed A., Johnson A., Cohen E., Adeniji T., Nnaji A. N., George J. E., “Assessing ChatGPT’s Capability as a New Age Standardized Patient: Qualitative Study”, JMIR Medical Education, 11, 2025, e63353.
[3] Zhang B., Liu X., Wang Y., Zhou L., Xie Q., Wang B., “Human or LLM as Standardized Patients? A Comparative Study for Medical Education [Preprint]”, arXiv, 2026.
[4] Gordon M., Daniel M., Ajiboye A., Uraiby H., Xu N. Y., Bartlett R., Hanson J., Haas M., Spadafore M., Grafton-Clarke C., Gasiea R. Y., Michie C., Corral J., Kwan B., Dolmans D., Thammasitboon S., “A Scoping Review of Artificial Intelligence in Medical Education: BEME Guide No. 84”, Medical Teacher, 46(4), 2024, 446–470.
[5] Makoul G., “Essential Elements of Communication in Medical Encounters: The Kalamazoo Consensus Statement”, Academic Medicine, 76(4), 2001, 390–393.

[6] Kurtz S. M., Silverman J. D., “The Calgary-Cambridge Referenced Observation Guides: An Aid to Defining the Curriculum and Organizing the Teaching in Communication Training Programmes”, Medical Education, 30(2), 1996, 83–89.
[7] Elwyn G., Durand M.-A., Song J., Aarts J., Barr P. J., Berger Z., Cochran N., Frosch D., Galasiński D., Gulbrandsen P., Han P. K. J., Härter M., Kinnersley P., Lloyd A., Mishra M., Perestelo-Perez L., Scholl I., Tomori K., Trevena L., … Van der Weijden T., “A Three-Talk Model for Shared Decision Making: Multistage Consultation Process”, BMJ, 359, 2017, j4891.
[8] Schillinger D., Piette J., Grumbach K., Wang F., Wilson C., Daher C., Leong-Grotz K., Castro C., Bindman A. B., “Closing the Loop: Physician Communication with Diabetic Patients Who Have Low Health Literacy”, Archives of Internal Medicine, 163(1), 2003, 83–90.
[9] Yen P. H., Leasure A. R., “Use and Effectiveness of the Teach-Back Method in Patient Education and Health Outcomes”, Federal Practitioner, 36(6), 2019, 284–289.
[10] Issenberg S. B., McGaghie W. C., Petrusa E. R., Gordon D. L., Scalese R. J., “Features and Uses of High-Fidelity Medical Simulations That Lead to Effective Learning: A BEME Systematic Review”, Medical Teacher, 27(1), 2005, 10–28.
[11] McGaghie W. C., Issenberg S. B., Cohen E. R., Barsuk J. H., Wayne D. B., “Does Simulation-Based Medical Education With Deliberate Practice Yield Better Results Than Traditional Clinical Education? A Meta-Analytic Comparative Review of the Evidence”, Academic Medicine, 86(6), 2011, 706–711.
[12] European Data Protection Board, “Guidelines 01/2025 on Pseudonymisation (Version for Public Consultation, Adopted January 16, 2025)”, 2025.
[13] World Health Organization, “Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models”, 2025.