

A Practice Copilot for Medical Education Developing Clinical Reasoning, Patient Communication, and Empathy through Virtual Patient Encounters

Anonymous Author(s)

Abstract

Medical trainees need practice eliciting information, interpreting ambiguous patient language, and making management decisions. Expressive virtual patients could broaden that practice, but uncontrolled disclosure and unreliable scoring can make an encounter educationally misleading. We present a practice copilot for medical trainees, including a design and evaluation framework for controlled virtual patient encounters in Indian medical education. A versioned case state separates clinical truth, patient knowledge, disclosure, and speech realization. An evidence record links each presented cue to learner follow-up, patient clarification, and subsequent action. Independent services retrieve examination findings and investigation results; asynchronous observers support reviewable formative feedback. The learning design connects supported rehearsal to independent practice, spaced reassessment, and transfer. We specify matched comparisons for clinical consistency, cue availability, scoring agreement, and delayed learning. Live platform testing is reported as ongoing; this paper reports no participant outcomes or validated competency estimates. Indic-language and expressive-speech capabilities remain evaluation targets.

Keywords: virtual patients, clinical reasoning, medical education, Indic languages, formative assessment

Data and Code Availability This design paper uses published literature and constructed examples. The source package contains manuscript sources and a proposed evaluation protocol. No participant dataset or speech benchmark is supplied. A local deterministic demonstrator informed feasibility boundaries; it is not evidence of expressive speech or educational effectiveness.

Institutional Review Board (IRB) No human-participant analysis is reported here. The project team reports ongoing live testing, but its protocol,

consent and ethics determination are outside this v0 report. No approval or exemption for that activity is asserted. Any participant study reported in a revision requires its actual institutional determination and methods.

1. Educational problem and contribution

An encounter requires a trainee to decide what to ask, what to examine, what information changes the differential, and what to do next. Clinical reasoning instruction connects representations of a problem with organized clinical knowledge (Bowen, 2006). Management also involves choices under uncertainty and patient-specific tradeoffs (Haimovich et al., 2026). Correctly naming a diagnosis cannot by itself establish either skill.

The practice copilot targets repeated low-stakes rehearsal of consequential skills. Initial live testing is reported for medical students in India; the intended curriculum spans students, interns, residents and fellows. The present contribution is a specified simulation and evidence contract, with a falsifiable evaluation plan. We do not report results from the ongoing activity. A local two-case demonstrator uses fixed facts and surface forms, which illustrates a narrow control approach but does not establish natural conversation.

The proposed distinction is between an authored opportunity and a demonstrated learner response. A patient might say “pet mein gola ho raha hai,” an author-supplied example of colloquial gastrointestinal concern. Its meaning must be clarified, rather than mapped directly to diarrhea or a diagnosis. Similarly, “I am feeling okay” can warrant different exploration depending on context and delivery. The intended educational target is sensitive clarification, not guessing a concealed emotion. Hindi, Bengali, Odia and mixed registers are intended coverage, not a claim of completed multilingual validation.

2. Evidence informing the design

Prior virtual-patient research already includes empathic communication training (Foster et al., 2016). EQClinic studied automated nonverbal feedback in telehealth education (Liu et al., 2016). These precedents rule out a novelty claim based only on adding a patient avatar or communication feedback. A recent voice-VP agreement study found communication scoring less concordant with faculty ratings than information gathering (Gao et al., 2026). We therefore separate evidence extraction, scoring and learner feedback.

In broader education, structured AI tutoring has produced positive learning results in a college physics trial (Kestin et al., 2025), while a mathematics study distinguishes supported performance from learning after assistance is removed (Bastani et al., 2025). Neither establishes transfer to clinical encounters. A diagnostic-reasoning trial coauthored by Rodman found no significant benefit from giving physicians LLM access, despite stronger standalone model performance (Goh et al., 2024). Model competence and learner development need separate endpoints.

Onco-Shikshak describes a related educational architecture combining illness scripts, adaptive practice, spaced retrieval and feedback (Makani, 2026). We adopt its question of how an encounter informs later practice, while retaining its stated limitation that learner efficacy was not evaluated. Retrieval research with residents supports delayed knowledge assessment (Larsen et al., 2009); it does not establish that flashcard scheduling alone improves clinical communication. Our proposal pairs knowledge review with new encounters and reduced assistance.

3. Controlled encounter architecture

Let C be the approved case version, K_t the patient’s knowledge, D_t disclosure state, and H_t encounter history. An interpreter proposes a typed learner action a_t . A policy returns permitted content $F_t = \pi(C, K_t, D_t, H_t, a_t)$. The realizer receives F_t and persona constraints, rather than an unrestricted case or rubric. Clinical truth is not writable through conversation. A clinician reviews the case before curricular use; synthetic examples remain marked as requiring approval.

The coordinator commits one finalized action at a time. Examination and order services return only available case findings, with explicit unavailable or

Table 1: Proposed component boundaries. All assessment is formative pending validation.

Component	Access and output
Case compiler	Faculty-approved truth, disclosure and rubric; emits immutable version.
Disclosure policy	Case and encounter state; emits permitted facts, not new facts.
Speech realizer	Permitted content and persona; emits speech with delivery record.
Exam/orders	Typed request and approved findings; emits retrieval or clarification.
Observers	Case, rubric and evidence; emit attributed claims, not patient speech.
Feedback	Reviewed evidence and assistance history; emits practice suggestions.

clarification responses. Asking for an examination does not demonstrate physical execution. Ambiguous compound requests are confirmed before results appear. A statement about past medication is distinguished from a proposed prescription. Management scoring permits faculty-approved alternatives and treats unsupported interpretation as uncertainty.

Patient speech proceeds independently of assessment. Observers consume finalized events and propose evidence claims; a later scoring stage reconciles them. An observer failure leaves the event stream available for later review. A speech failure can fall back to text, but audio-dependent criteria become unassessable. Barge-in cancels queued speech, and the record distinguishes planned from delivered content. A fact in unplayed audio cannot be counted as heard.

3.1. Cue and evidence contract

For every assessable opportunity, record

$$o = (c, v, m, d, w, r, u),$$

where c identifies the authored cue, v its case version, m modality, d delivered event, w response window, r rubric criterion and u availability uncertainty. A learner response is linked to o and its source utterance; patient clarification and later management remain separate events. An acoustic intention is insufficient evidence that a cue was perceptible. Native-speaker review and playback checks must establish that boundary before a missed opportunity receives a score.

A diagnostic kernel is a small clue with plausible follow-up families and deeper information available after clarification. The rubric can distinguish noticing, characterization and use in management. A learner who asks about weight change but disregards the answer has demonstrated a different behaviour from one who updates the problem representation. Rojas and colleagues' expert/student comparison motivates examining information use, but its questionnaire case does not validate our conversational measure (Rojas et al., 2024).

4. Learning, scoring and clinical scope

Implementation details and constructed persona examples appear in Section 6. The inspected browser/Python demonstrator uses seeded text, JSON case lookups and logical event time; it does not establish expressive regional speech or measured latency. Faculty define anchors for history, problem representation, differential revision, examination planning, investigations, management, communication, documentation and closure. Observers cite events and abstain when evidence is insufficient. Deterministic rules check approved critical actions; semantic judgments require review against expert labels. No overall LLM grade is treated as a competency estimate. Score validity concerns interpretations and use, not a property conferred by a working interface (Cook and Beckman, 2006).

In teaching mode, the learner may request a broad prompt, a specific hint or a worked example. Assistance is logged and is not equivalent to independent performance. Feedback highlights a small number of consequential moments and supplies a next exercise. A subsequent case changes surface features and removes the hint. Spaced factual retrieval and encounter practice have separate histories. Repeated attempts do not increase the count of independent cases.

Figure 2 shows a design mock-up of the learner-facing loop for an OSCE-style station. In practice mode, cues the learner has not followed up appear during the encounter; in exam mode they are held for the debrief. The debrief sorts skills into strong, developing and next-to-work-on, and every item cites the turn it comes from. Weak skills become short drills that return at growing intervals until a full station without cues is passed. Appendix A shows the remaining screens. The mock-up uses synthetic data; it is not a measurement of learners, and its score

and pass line are placeholders that would need faculty setting and validation.

Advanced learners receive greater clinical ambiguity and management tradeoffs; beginners receive structure without arbitrary patient obstruction. Palliative-care and end-of-life communication are curriculum targets for later faculty-approved cases. A patient's agreement, emotional relief or compliance is not a required outcome for credit. Appropriate uncertainty, respecting preferences, and seeking supervision can be competent actions.

5. Oncology education and the Onco-Shikshak origin

Onco-Shikshak provides the oncology-specific starting point for the practice copilot Makani (2026). Its preprint describes clinical workflows including Morning Report, Tumor Board and Clinic Day, with illness-script modeling, adaptive scaffolding and spaced retrieval. The pedagogical link is between decisions made during a case and the learner's subsequent practice. The reported architecture is a precedent for design, while its stated absence of formal learner evaluation limits claims about educational benefit.

The project team reports demonstrations and discussions with oncology attending physicians and trainees at an Indian medical college about a prospective Onco-Shikshak pilot. The institution is withheld in this anonymous review draft. According to the team, these engagements prompted the idea of extending the oncology learning companion into the practice copilot, a broader medical-education learning copilot. This is an author-reported account of project origins, not a completed pilot or a qualitative study. We report no participant counts, interview findings, institutional endorsement or demonstrated learning outcomes from those engagements.

The oncology setting motivates proposed practice in explaining uncertainty, comparing management alternatives, discussing treatment burden and exploring patient and caregiver concerns. Tumor-board reasoning and patient-facing communication require different interactions. A trainee may justify a choice among specialists yet still need practice explaining it to a patient in an everyday regional register. the practice copilot would connect such reasoning with consultation, clarification, documentation and closure, including faculty-reviewed palliative and end-of-life scenarios appropriate to the learner's role.

Generalization should preserve specialty-specific content and expectations. Oncology fellowship milestones cannot serve unchanged as a general undergraduate rubric, and a treatment decision cannot be reduced to one universal answer. The proposed common layer consists of controlled patient knowledge, evidence-linked formative feedback, visible assistance and repeated independent practice. Faculty retain responsibility for clinical facts, acceptable management alternatives and stage-appropriate expectations. Whether that common layer supports learning across specialties requires evaluation beyond the reported oncology demonstrations and discussions.

6. Implementation and practical encounter design

6.1. Inspected prototype and development boundary

The inspected local demonstrator has a browser interface served by a Python standard-library HTTP server. JSON case files supply facts, response variants, disclosure triggers, examinations and investigations. Requests create an encounter or record a message, examination, order, differential, management statement, interruption or final note. Encounters are stored in process memory, with an exportable event sequence. The implementation selects seeded surface forms and matches configured terms; it is not an unrestricted conversational model.

The browser uses available Web Speech recognition and synthesis. It submits finalized recognition text and selects a voice by language tag where available. Such selection does not establish dialect fidelity or controllable emotional expression. The event clock in this demonstrator advances logically with actions, so it cannot support measured audio latency. Authentication, durable storage, faculty approval, expressive speech control and longitudinal competency estimation are not established by this local implementation. The author-reported live-testing platform must be documented separately before equating it with this snapshot.

6.2. Faculty authoring and independent persona dimensions

The proposed authoring form separates clinical truth from age, gender, occupation, residence, family context, material constraints, language, register, liter-

acy and health literacy. Faculty should also specify the patient’s own understanding, remembered facts, concerns and disclosure boundaries. These are independent fields: a rural worker may be articulate and medically knowledgeable, while an urban professional may be unsure of medication names. Disfluency must not be assigned automatically from literacy, income or occupation.

Figure 1 presents constructed profiles, not a representative sample of Indian patients. The respiratory and glycemic scenarios correspond broadly to the two local demonstration case themes; the third is a proposed extension. All require clinician and language review before curricular use. Faculty must distinguish a presenting complaint from a working syndrome and an underlying diagnosis. A vague presenting phrase does not conceal a single “correct chief complaint” that the student must guess.

Socioeconomic context should enter through actionable circumstances such as travel, wage loss, medication access, caregiving duties or preferred follow-up arrangements when specified in a case. A learner can ask about those constraints and adapt an appropriate plan. The rubric should not presume poverty, nonadherence, limited understanding or a particular family role from an occupation or address. Sensitive information can be disclosed after an appropriate invitation without requiring a culturally narrow style of rapport.

6.3. Indic speech and symptom expression

AMIE Video provides an adjacent architectural precedent: a fast Talker receives asynchronous Planner and Perception updates Nagda et al. (2026). Its AI plays the clinician, whereas our patient must never receive the examiner’s desired diagnosis or next question. We adopt concurrency as a candidate design while reversing clinical authority. Technical delays should be distinguishable from authored hesitation, and stale observations must retain source times. The preprint’s discrete-turn limitation also means interruption support cannot be assumed to provide full-duplex conversation. Keep approved semantic content separate from a reviewed set of local expressions, possible meanings and clarification paths. Preserve the original utterance alongside any normalized transcript. Hindi, Haryanvi, Khari Boli/Western Hindi registers, Bengali, Odia and code-mixing are evaluation targets; a standard language code does not validate a regional variety. A native speaker should assess register and pragmatics, and a clinician should check that varia-

tion has not changed the facts or made a diagnosis obvious.

The intended speech adapter can use a streaming cascade or a native realtime provider after benchmarking. No provider is selected here. Both alternatives must preserve the disclosure boundary, interruption semantics and a record of played speech. Measure turn-end detection, transcription, response planning, first audio and cancellation separately. Recognition errors involving negation, duration, drug names or examination requests should trigger correction or clarification rather than unearned credit or a silent penalty.

Fatigue, hesitation, breath grouping and occasional cough can be independently authored expressive features. They are hypotheses about simulation design, not validated signatures of fever or respiratory disease. Keep cough events separate from clinical content where possible, constrain their frequency, and test whether they leak the diagnosis or obscure the student's speech. The interface should expose a degraded-audio state and suspend affected criteria. A patient saying "I am okay" with uncertain delivery should invite respectful exploration, not a compulsory emotion label.

6.4. Examination, orders and reasoning evidence

The student can state a proposed maneuver, for example listening to both lung fields. An interpreter should normalize that request and retrieve the authored finding. Assessment mode returns available findings without steering toward a better maneuver. Teaching mode may offer a graduated hint, labelled and recorded as assistance. Neither mode establishes the learner's ability to execute a hands-on examination. Broad requests require a declared granularity policy, so one learner is not given an entire examination for a phrase that yields a single finding for another.

Laboratory, imaging, referral and management requests need typed actions with clarification of uncertainty. Distinguish what was proposed, ordered, returned and interpreted. Capture differential revision and the final clinical note separately from elicited history. A note can contain a correct fact that was never obtained, and a requested test can be appropriate without being interpreted correctly. Diagnostic-kernel assessment should link the initial clue, relevant follow-up and use of the answer, rather than count keywords.

6.5. Closure and longitudinal feedback

The proposed closure rubric records summary, understandable explanation, uncertainty, next steps, checking understanding, invitations for questions, safety-netting and follow-up. Cases may include access barriers relevant to completing the plan. Scoring an explanation does not establish actual adherence or reduced loss to follow-up; those are downstream outcomes requiring separate evidence.

The student view should show a small set of priority observations with replayable evidence and concrete practice suggestions. Faculty should see sampling coverage, assistance, uncertainty and recurring patterns, with a route from cohort to encounter evidence. Candidate visualizations include a case-by-skill matrix and a trajectory with observation counts and uncertainty; they should be evaluated for interpretation rather than adopted for appearance. Repetitions, supported attempts and independent cases remain distinct. Avoid a single radar chart or scalar mastery score that hides sparse evidence or unequal difficulty.

7. Evaluation protocol and failure criteria

Evaluation separates simulation consistency, clinical correctness, measurement quality and learning. First, test paraphrased questions, withheld information, incompatible orders and injection attempts against approved cases. Count fact contradictions, unauthorized disclosure and unsupported findings. A boundary violation blocks use of the affected case/version. Such tests establish behaviour only for tested inputs, not universal security.

Second, compare matched expressive audio, neutral audio and text conditions. Preserve relevant clinical facts, while reviewing whether a wording or voice change alters diagnostic cueing. Include neutral controls to measure false emotional attribution. Use human recordings as a reference when consent and resources permit. Clinical-concept transcription errors and delivery failures are reported separately from learner mistakes. TTS naturalness does not establish cultural fidelity.

Third, clinicians and native speakers annotate opportunities and acceptable response families before observing automated scores. Independently trained raters assess held-out encounters with patient/caregiver perspectives included in rubric devel-

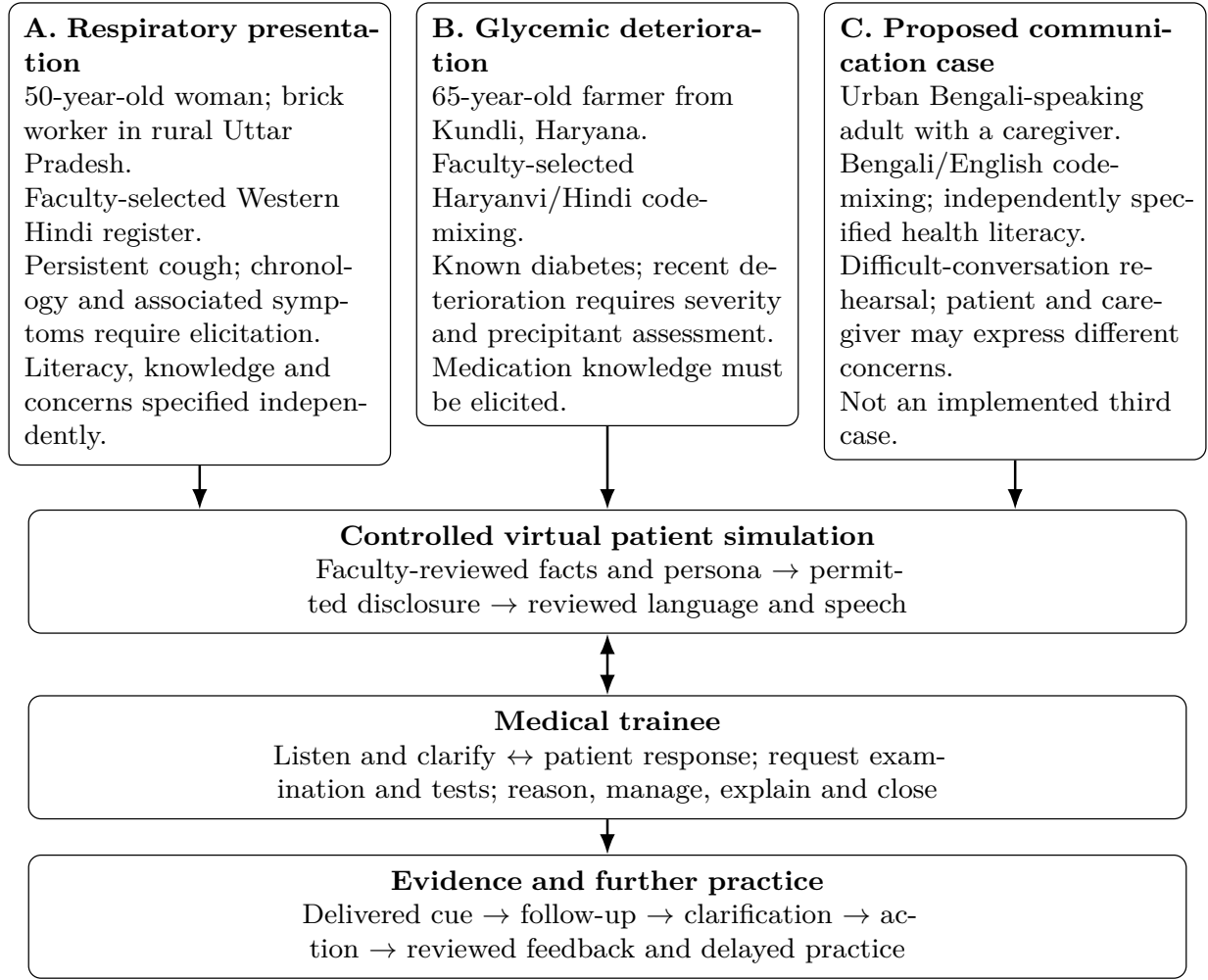


Figure 1: Constructed Indic-language persona examples interacting with the practice copilot. The first two reflect demonstration themes; the third is proposed. Profiles are illustrative, not population estimates or stereotypes. Medical content and language require clinician and native-speaker approval.

opment. Report evidence precision and recall, false credit, missed credit, and safety-flag sensitivity. Use absolute-agreement ICC for continuous domain ratings when assumptions fit, weighted kappa for ordinal judgments, and intervals with the learner/case clustering preserved. Agreement with one examiner does not establish validity.

Fourth, compare the controlled system with a strong patient-prompt baseline and a deterministic scripted baseline under matched cases and resource budgets. A proposed learning study compares usual case practice

with practice using the copilot, matched for time and feedback access. The primary endpoint is blinded rating of a novel, unassisted encounter after a prespecified delay; immediate supported scores and perceived realism are secondary. Sample size must follow a meaningful effect and pilot variance, not an arbitrary convenience number. Two cases can test mechanisms but cannot support general competency claims.

Failure criteria include unavailable cues being scored as missed, hallucinated evidence producing remediation, systematic language-specific penalties, or im-

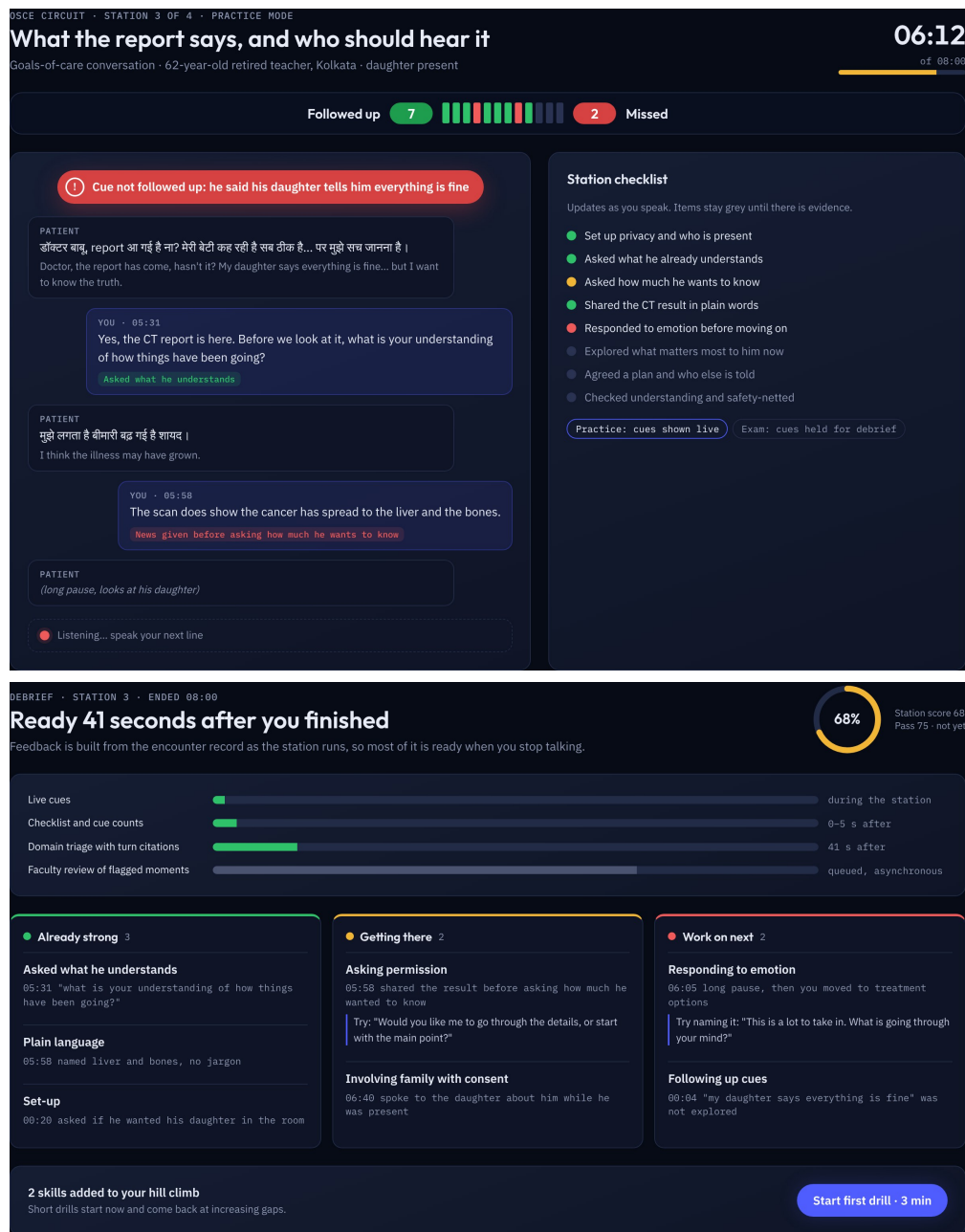


Figure 2: Design mock-up of the learner-facing feedback loop, with synthetic data for one invented learner. Top: a live goals-of-care station in practice mode, with a cue tracker, a prompt for an unexplored cue and a checklist that fills as evidence appears. Bottom: the debrief, with the timing of each feedback layer and skills sorted by evidence, each citing its turn. Timings, scores and the pass line are illustrative. No learner outcomes are reported.

provement confined to rehearsed cases. Preserve a successful control design even if expressive audio fails to add benefit. Remove an automated scoring dimension if expert agreement or evidence attribution is inadequate.

8. Limitations and conclusion

The design draws on a targeted narrative review, not systematic screening. No new learner outcomes, provider benchmark, native-speaker validation or calibrated mastery threshold is reported. Ongoing live testing must be documented before empirical claims are added. Authenticity, regional coverage and transfer remain open. Student audio and performance require consent, restricted access, retention limits and an appeal path; practice records should not silently become summative records.

The proposed practice copilot uses controlled clinical facts and reviewable evidence as the basis for formative practice. Its testable claim is that preserving the link between presented opportunity, learner action and feedback makes simulation more interpretable and assessable. Whether that structure improves independent clinical reasoning and communication requires prospective evidence.

AI assistance OpenAI Codex assisted source discovery, synthesis, drafting and document preparation. No participant observations or educational outcomes were generated as research data. Authors remain responsible for checking citations, claims and the final submission.

References

- Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakci, and Rei Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025. doi: 10.1073/pnas.2422633122.
- Judith L. Bowen. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355:2217–2225, 2006. doi: 10.1056/NEJMr054782.
- David A. Cook and Thomas J. Beckman. Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine*, 119(2):166.e7–166.e16, 2006. doi: 10.1016/j.amjmed.2005.10.036.
- Adriana Foster, Neelam Chaudhary, Thomas Kim, Jennifer L. Waller, Joyce Wong, Michael Borish, and Andrew Cordar. Using virtual patients to teach empathy: A randomized controlled study to enhance medical students’ empathic communication. *Simulation in Healthcare*, 11(3):181–189, 2016. doi: 10.1097/SIH.0000000000000142.
- Xiaoxing Gao, Xiaoming Huang, Rongrong Hu, et al. Evaluating large language model-based automated scoring in a voice-based virtual standardized patient platform for medical students: A cross-sectional agreement study. *JMIR Medical Education*, 2026. doi: 10.2196/95578.
- Ethan Goh, Robert Gallo, Jason Hom, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969, 2024. doi: 10.1001/jamanetworkopen.2024.40969.
- Adrian D. Haimovich, Alexander T. Janke, Keith E. Kocher, Courtney W. Mangus, Andrew S. Parsons, Liam McCoy, Richard Andrew Taylor, Adam Rodman, and Martin Pusic. Managing clinical uncertainty: Formalizing management reasoning in emergency care delivery. *Annals of Emergency Medicine*, 87(5):578–585, 2026. doi: 10.1016/j.annemergmed.2025.09.007.
- Greg Kestin, Kelly Miller, Anna Kales, Timothy Milbourne, and Gregorio Ponti. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15:17458, 2025. doi: 10.1038/s41598-025-97652-6.
- Douglas P. Larsen, Andrew C. Butler, and Henry L. Roediger. Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. *Medical Education*, 43(12):1174–1181, 2009. doi: 10.1111/j.1365-2923.2009.03518.x.
- Chunfeng Liu, Renee L. Lim, Kathryn L. McCabe, Silas Taylor, and Rafael A. Calvo. A web-based telehealth training platform incorporating automated nonverbal behavior feedback for teaching communication skills to medical students: A randomized crossover study. *Journal of Medical Internet Research*, 18(9):e246, 2016. doi: 10.2196/jmir.6299.

Ashish Makani. Onco-shikshak: An ai-native adaptive learning ecosystem for medical oncology education. *medRxiv*, 2026. doi: 10.64898/2026.02.23.26346944. Preprint, version 1.

Mahvish Nagda, Jihyeon Lee, Matthew Thompson, et al. Towards expert-level medical AI for real-time video consultations, 2026. Preprint, version 1.

Marcos Rojas, Argenta Price, Candice Jeehae Kim, Sharon F. Chen, Kathleen Gutierrez, Carl Wieman, and Shima Salehi. Exploring differences in clinical decisions between medical students and expert clinicians. *Advances in Medical Education and Practice*, 15:1285–1297, 2024. doi: 10.2147/AMEP.S492302.

Appendix A. Feedback interface mock-up

This appendix shows the full set of screens from the mock-up in Figure 2. All learner data are synthetic. The case titles and the nine feedback domains come from the implemented system; the attempts, scores, dates and timings do not. The interface itself is a design proposal, not the current implementation.

Three design rules shape the screens. First, feedback arrives in layers: cue prompts during practice-mode encounters, checklist counts at the end of the station, a domain-level debrief shortly after, and faculty review of flagged moments asynchronously. Second, rewards follow what the learner did, such as cues followed up, safety-netting and passing without hints. They never follow whether the simulated patient or family agreed, calmed down or accepted a plan. There is no leaderboard, and streaks count days of practice rather than days of high scores. Third, each weak skill becomes a short drill that returns at growing intervals, and the climb ends with a full station without cues. Whether this loop improves unassisted performance is a hypothesis for the delayed-transfer study in Section 7, not a result.

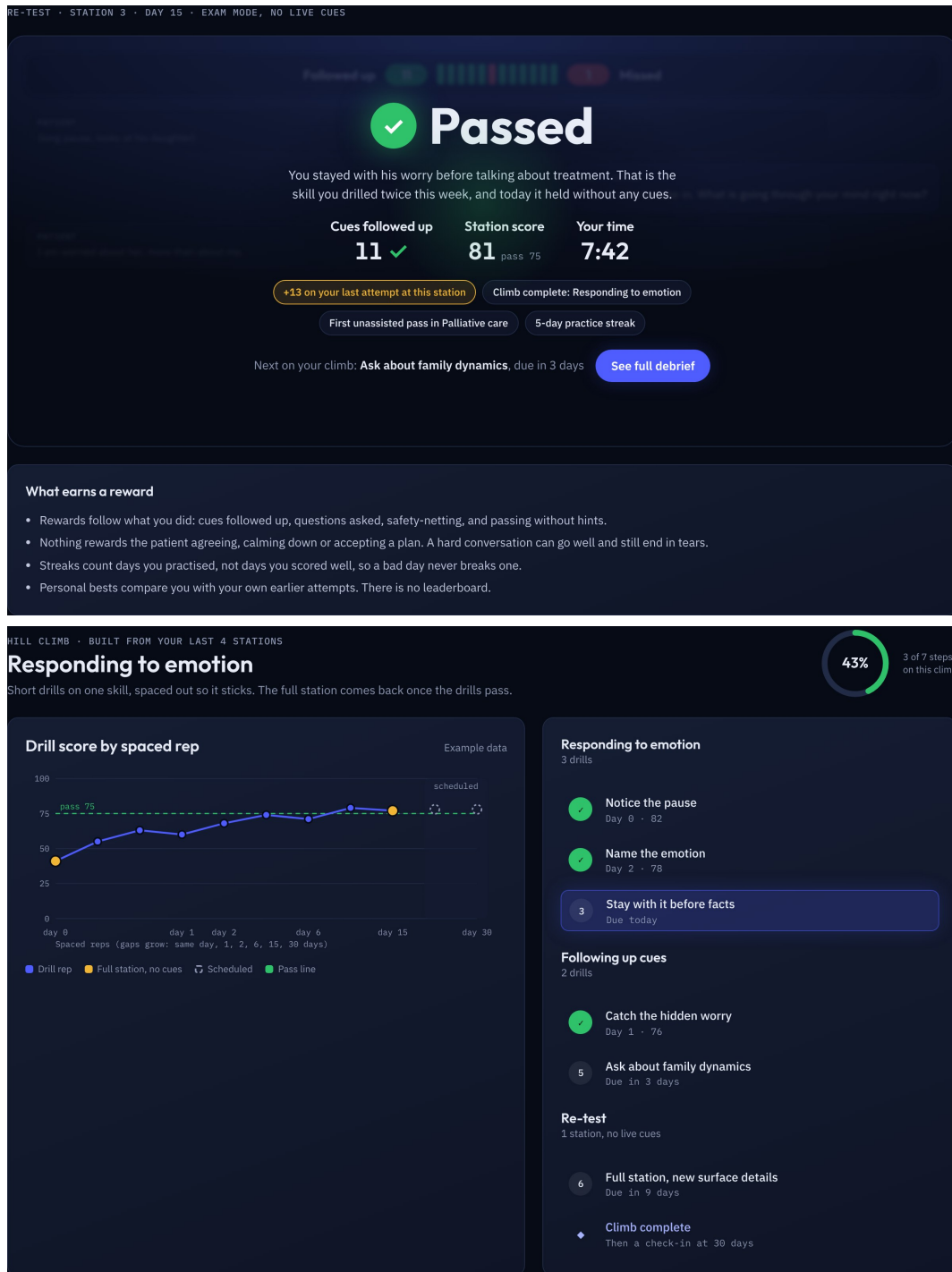


Figure 3: Top: the result screen after a passed re-test in exam mode, with feedback naming the behaviour that changed and rewards tied to learner actions. Bottom: a hill climb on one weak skill, with drill scores across spaced repetitions and the remaining schedule. Synthetic data.

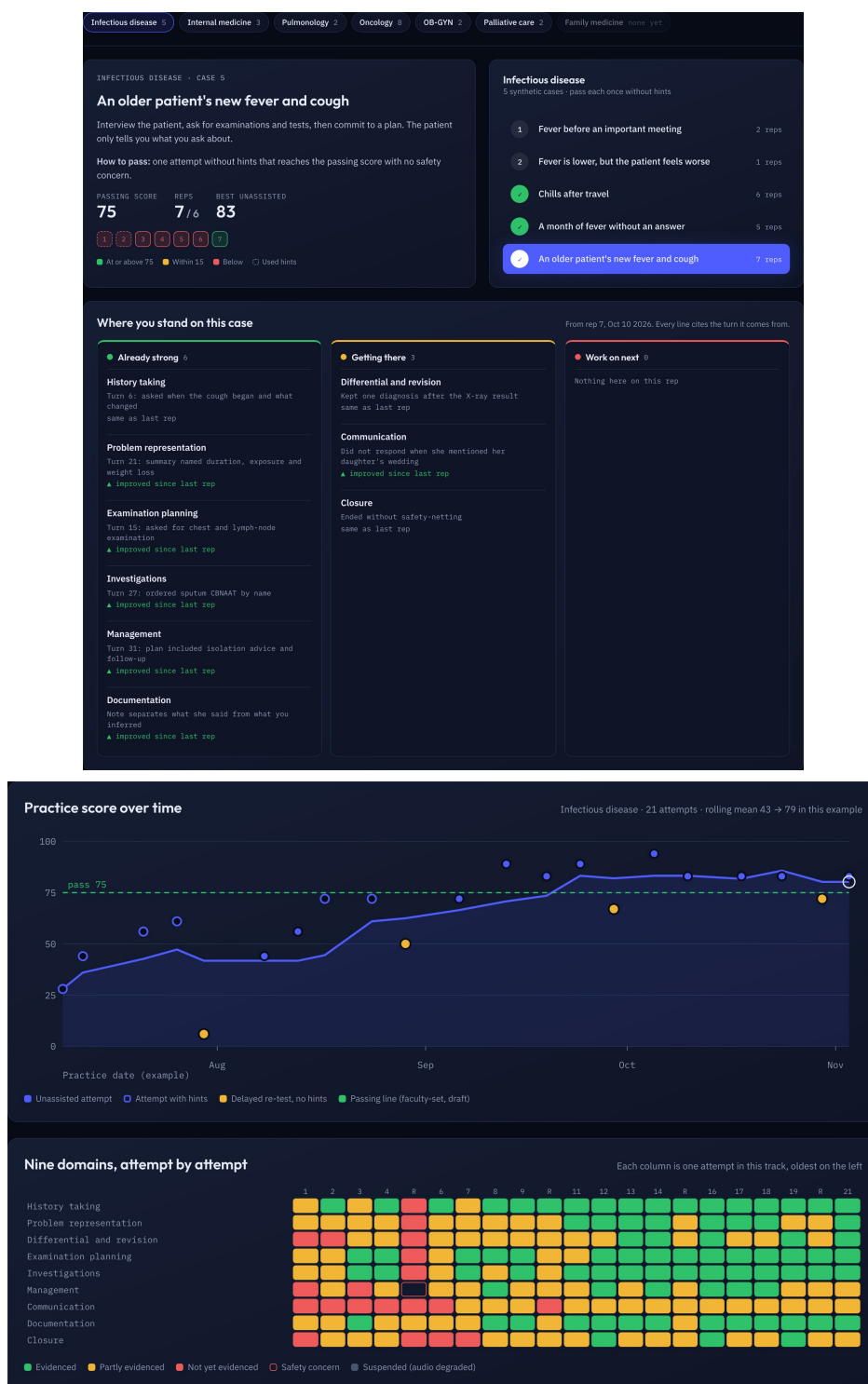


Figure 4: Top: a specialty track with reps per case against a placeholder pass score and the latest debrief for the selected case. Bottom: practice score over time with hinted, unassisted and delayed re-test attempts marked, and the nine feedback domains attempt by attempt. Specialty case counts come from the case library; all attempts are synthetic.