# Anchoring Medical Chatbot Feedback to Human Rubrics: A Pilot Toward Reliable Oversight

Alina Shah*[1], Yann Hicke[1], Siena Shah[1], and Gianna Cox[2]

[1]Cornell University, Ithaca, NY, USA
[2]University of Illinois Urbana–Champaign, Urbana, IL, USA

October 2025

**Abstract**

Current chatbot feedback systems often sound helpful but lack anchoring to validated rubrics, making their evaluations easy to mislead and impossible to audit. This creates a fundamental oversight problem: confident-sounding but unreliable outputs mask failure modes and mislead users. We piloted a rubric-anchored evaluation pipeline that grounds model outputs in the Medical Interview Rating Scale (MIRS), a validated framework for communication skills. In a study with 21 medical learners, rubric anchoring improved clarity and actionability of feedback versus a generic GPT system, highlighting a key risk: participants sometimes preferred outputs that felt clearer and more actionable, even when those outputs were less accurate. Rubric anchoring shows how to shift chatbot feedback from isolated evaluations toward structured, auditable assessment linked to explicit criteria. While the case study is in medical education, the underlying method—rubric anchoring—offers a template for oversight of frontier models, where correctness is non-optional. This work represents an early prototype of scalable evaluation infrastructure, enforcing dual standards of (1) evidentiary reliability and (2) human-centered usability. It connects domain-specific pilots to the broader agenda of making model behavior accountable, auditable, and safe for deployment.

## 1 Introduction

Effective feedback is critical for developing communication and interviewing skills with peers and patients in medical education [2, 8]. Feedback helps students to improve their ability to connect with patients, ease their concerns, and organize their encounters. However, in practice, feedback is often delivered inconsistently, without clear grounding in evidence. As a result much feedback fails to support reliable skill development.

Recent advances in large language models (LLMs) have created interest in their use as conversational coaches, as they are now capable of generating adaptive, conversational feedback. Prior work has explored their application in clinical education, from simulating standardized patient encounters to providing automated coaching [4]. However, most AI-generated feedback is evaluated only by its surface plausibility or user preference. These signals are neither reproducible nor auditable against validated educational standards, raising the risk that learners may trust well-phrased but unreliable outputs.

This study investigates whether rubric anchoring can address that gap. We compare two feedback systems for simulated medical interviews on the MedSimAI platform: (1) a generic GPT-based chatbot and (2)

*Correspondence: alina.shah1022@gmail.com

a rubric-anchored chatbot explicitly grounded in the Medical Interview Rating Scale (MIRS), a validated framework for communication skills. Participants interacted with both systems and rated their feedback along dimensions such as clarity, actionability, and safety.

By grounding chatbot evaluation in a validated rubric, this work demonstrates how feedback systems can be made more structured, reproducible, and auditable. Although tested in medical education, this work points to a broader principle: in safety-critical domains, interactive AI systems must pair engagement with reliability to earn trust at scale.

## 2 Related Work

Reliable feedback is essential in training communication and interviewing skills in medical interviewing contexts. Ende's seminal work [2] framed feedback as essential for self-reflection and growth. Building on this work, Branch and Paranjape emphasized the importance of timely and learner-centered feedback. Empirical studies consistently report that students receive comments that are vague, delayed, or not actionable. This highlights the challenge of consistently producing feedback that learners can trust and apply. Recent advances in large language models (LLMs) have prompted new approaches to automated coaching. Liu et al. [4] presented GPT-based systems for simulated patient interactions, while Lee et al. presented AI-powered standardized patients for communication training. These systems show promise, but also demonstrate the same reliability issues faced by AI more broadly. That is, often outputs are generic, inconsistent across contexts, and difficult to audit. This mirrors the central concern in AI oversight: unconstrained models produce behavior that is hard to evaluate and even harder to trust at scale. One response, both in education and in AI safety research, is to constrain models through external standards. In medical education, validated rubrics such as the Medical Interview Rating Scale (MIRS) provide structured criteria for evaluating interviewing skills across empathy, organization, and eliciting concerns. Decades of research show that rubric-based ratings improve inter-rater reliability and learner outcomes. In AI safety, analogous arguments emphasize that model behavior must be anchored to external, human-defined benchmarks in order to be interpretable, auditable, and ultimately controllable. Despite this parallel, little prior work has tested whether LLM-based feedback can be systematically anchored to such validated rubrics. Existing studies focus primarily on free-form generation, leaving open the question of whether rubric integration can transform feedback into an auditable, reproducible process. This work addresses that gap. We compare two chatbot feedback systems: a generic GPT-based model and a rubric-anchored variant grounded in the MIRS framework. By embedding AI feedback within validated educational standards, we test whether rubric anchoring improves categories like specificity, actionability, and more . More broadly, this work investigates whether external rubrics can serve as control mechanisms—constraining generative models into evaluation channels that are transparent, reproducible, and aligned with human judgment.

## 3 Methods

### 3.1 Participants

We recruited medical students from multiple universities (including the University of Illinois Chicago, Loyola University Chicago, and Lake Erie College of Osteopathic Medicine), residents, physician assistants, and physicians via email invitations and word of mouth. This study was approved by the Cornell University IRB (Protocol #607-255-6182). All participants provided informed consent prior to participation. A total of 21 learners completed the study.

### 3.2 Study Design

We used a within-subjects design in which each participant interacted with two feedback chatbots after completing a standardized patient interview with a simulated patient on the platform MedSimAI. One chatbot generated feedback solely based on its analysis of the interview trabscript, while the other produced rubric-anchored feedback based on dimensions from the Medical Interview Rating Scale (MIRS). The order of chatbot exposure was counterbalanced across participants to minimize order effects and increase internal validity.

### 3.3 Materials

**Interview Case.** We selected a standardized patient case featuring a 29-year-old with new-onset exertional shortness of breath. Dyspnea is a common and clinically important presenting complaint, and acute dyspnea cases have been validated in prior medical education research as effective for assessing clinical reasoning across learner levels.

### 3.4 Evaluation Rubric Justification

To evaluate the quality of AI-generated feedback, we derived our rubric dimensions from prior literature on feedback in medical education. Dimensions were adapted to the simulation context of MedSimAI, ensuring alignment with constructs empirically linked to learner uptake and skill transfer. Beyond education, these dimensions align with oversight priorities in AI safety—anchoring model outputs to criteria that make behavior interpretable, auditable, and safe.

- **Specificity.** High-quality feedback requires behavior-level comments; vague praise or criticism does not support change [6].
- **Actionability.** Feedback must contain concrete, feasible next steps to influence future performance.
- **Alignment with Learning Goals.** Anchoring feedback to the MIRS dimensions increases validity and reduces drift into unrelated advice.
- **Constructive Balance.** Balanced comments (strengths + growth areas) support motivation and self-efficacy, consistent with principles of effective feedback [3, 1].
- **Consistency and Focus.** LLMs often generate contradictory or scattered suggestions. Constraining outputs to focus on core rubric skills reduces cognitive load and increases uptake.
- **Transferability.** High-quality feedback extends beyond the immediate case to principles usable across encounters, a crucial mechanism for skill consolidation [9].
- **Cognitive Load / Clarity.** Feedback should be clear and concise, minimizing cognitive overload after demanding interviews [7].
- **Accuracy / Evidence Grounding.** Feedback must correctly reflect transcript events and, where possible, cite verbatim learner utterances. This reduces hallucinations and increases trust.
- **Clinical / Safety Appropriateness.** Feedback must not include unsafe or inappropriate medical advice. This was evaluated as a binary dimension (pass/flag).

In addition to these dimensions, the survey included three open-text items: (*Most useful aspect of the feedback*, *One way the feedback could be improved*, and *Goal for the next interview*). Participants also indicated which chatbot they preferred and explained their choice. These qualitative responses provided insight into how participants perceived the feedback, their reasons for their preference, and where user perception diverged from measured reliability.

### 3.5 Model Justification: GPT-4.1

We selected OpenAI's GPT-4.1 (April 2025 release) as the model for both chatbots.

**Stability and reproducibility.** Because our study required consistent rubric-anchored outputs, model predictability under fixed prompts and low-temperature settings was essential. GPT-4.1's improvements in instruction-following and context retention supported structured prompting and reproducibility.

**Relative to GPT-5.** Although GPT-5 introduced architectural advances (e.g., Mixture-of-Experts routing, multimodal extensions), it was newly released and had not been validated in constrained educational settings. To prioritize stability, auditability, and transparency over frontier novelty, we adopted GPT-4.1 for this pilot.

### 3.6 Chatbot Implementation

We developed two chatbots with shared scaffolding and minimal differences in system prompts. Both prompts emphasized a coaching style that moved through reflection $\rightarrow$ feedback $\rightarrow$ goal-setting.

- **Generic Feedback Chatbot.** Provided free-form feedback without explicit rubric alignment. The system prompt instructed the model to act as an *Expert Clinical Communication Coach*, offering practical suggestions in a reflective, supportive style.
- **Rubric-Anchored Chatbot.** Generated feedback explicitly tied to MIRS dimensions (e.g., agenda-setting, empathy, closing). We grouped MIRS items according to the Kalamazoo Essential Elements Communication Checklist [5] and required the model to map learner performance to rubric categories before generating feedback in order to decrease cognitive load on the model.

Both chatbots:

- Used identical transcripts for input (standardized MedSimAI cases),
- Were run with identical generation parameters (temperature 0.7, max tokens 650),
- Produced single-pass outputs with no cherry-picking; the first generation was retained for all raters.

All prompts and implementation code will be released in supplementary materials to ensure reproducibility.

### 3.7 Procedure

Participants first completed the standardized interview with the simulated patient. They then interacted with both chatbots in sequence, receiving feedback on their interview performance. After each chatbot, participants completed a survey evaluating the feedback quality, including Likert-scale ratings and open-ended questions.

### 3.8 Analysis

We analyzed survey data across quantitative ratings and qualitative responses.

**Preference comparisons.** Participants indicated which chatbot they preferred overall. Preferences were summarized with raw frequencies and binomial confidence intervals.

**Likert ratings.** Dimension-level ratings (specificity, actionability, alignment, clarity, etc.) were summarized with descriptive statistics (means, medians, standard deviations). We emphasized frequency distributions across the 7-point scale to highlight clustering.

**Exploratory tests.** We conducted paired non-parametric tests (Wilcoxon signed-rank) to compare chatbot conditions. Effect sizes (rank-biserial correlations) are reported to provide context to observed differences.

**Qualitative responses.** Open-text responses were used to illustrate quantitative findings and, critically, to identify cases where user preference diverged from reliability. This highlights a core oversight concern.

This mixed-methods approach provided both quantitative rigor and qualitative context, turning a small pilot into reproducible evidence for evaluation infrastructure.

### 3.9 Data Availability.

Study materials and code are available in a public GitHub repository GitHub Repository. Because this project is in preliminary stages, the repository is actively maintained and may undergo updates, but the analyses reported here reflect the version current as of October 4th 2025.

## References


[1] Edward L. Deci, Richard Koestner, and Richard M. Ryan. Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, 71(1):1–27, 2001.

[2] Jack Ende. Feedback in clinical medical education. *JAMA*, 250(6):777–781, 1983.

[3] Avraham N. Kluger and Angelo DeNisi. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2):254–284, 1996.

[4] Junnan Liu et al. Can large language models provide useful feedback on medical students' communication skills? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–12. Association for Computational Linguistics, 2023.

[5] Gregory Makoul. Essential elements of communication in medical encounters: The kalamazoo consensus statement. *Academic Medicine*, 76(4):390–393, 1999.

[6] Valerie J. Shute. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189, 2008.

[7] John Sweller, Jeroen J.G. van Merrienboer, and Fred G.W.C. Paas. Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296, 1998.

[8] Jeroen M. M. Van de Ridder, Kees M. Stokking, William C. McGaghie, and Olle T. J. Ten Cate. What is feedback in clinical education? *Medical Education*, 47(2):214–224, 2013.

[9] J. Jon Veloski, John Boex, Carolyn M. Grasberger, Gene E. Evans, and David R. Wolfson. Systematic review of the literature on assessment, feedback and physicians' clinical performance. *Medical Teacher*, 28(2):117–128, 2006.