Micro-Randomized Trial of an AI-Simulated Practice Tool for Therapeutic Skills

Ryan Louie[1] , **Ellen Converse**[2] , **Diyi Yang**[1] and **Emma Brunskill**[1]
[1]Stanford University Computer Science
[2]PAU–Stanford Psy.D. Consortium
rylouie@cs.stanford.edu, ellencon, diyiy@stanford.edu, ebrun@cs.stanford.edu

Abstract

Though LLM-simulated practice may support psychotherapy education, open questions remain about which system features provide the most educational value. We developed an AI-training system for use in psychotherapy classrooms that provides speaking-practice with LLM-simulated patients, followed by a post-practice review that includes AI feedback suggestions and structured self-reflection exercises. We deployed the system in a graduate psychotherapy course (n=25, 5 weeks) using a micro-randomized trial (MRT)—a method which can estimate the causal excursion effect of an intervention by leveraging longitudinal repeated randomization with participants. Testing four conditions crossing AI feedback (present/absent) with reflection granularity (utterance/session-level), we found surprising interaction effects on student engagement and educational value: utterance-level reflection paired with AI feedback significantly outperformed all conditions. Students valued comparing the AI suggestions to their own rather than passively accepting them. As the first MRT in a health education setting testing an LLM-simulation system, this work demonstrates that MRTs, a method mostly used in mobile intervention research, offers a viable evaluation paradigm for AI systems deployed in health education settings.

1 Introduction

Healthcare faces a fundamental supply-demand problem: the need for trained healthcare professionals far exceeds availability in many critical areas. Mental health is among the most acute examples of this mismatch: a 2020 survey found most US states have fewer than 40% of the mental health professionals needed, even though 21% of U.S. adults had a mental illness [Modi *et al.*, 2022]. One approach to bridging this gap is real-time decision support for clinicians, but this may be inadequate for settings like psychotherapy that require face-to-face interaction and rapid verbal responses. An alternative is AI-assisted training–enabling more practitioners to develop skills faster through scalable practice opportunities. Historically, realistic training has been expensive, relying on standardized patients [Kühne *et al.*, 2020] and expert supervisors [Hill and Knox, 2023]. Recently, research for simulating patient interactions via LLMs have been developed and tested for realism, acceptability, and feasibility in lab studies [Louie *et al.*, 2024; Wang *et al.*, 2024; Steenstra *et al.*, 2025], highlighting the promise of AI-enabled on-demand practice. Nonetheless, few studies deploy AI systems longitudinally in classrooms [Thesen *et al.*, 2025], and fewer still provide practical design recommendations based on experimental evidence.

As our first contribution, we present CARE, an AI training system for psychotherapy education. Unlike prior platforms that support text-based interaction only [Wang *et al.*, 2024], CARE enables learners to practice speaking with LLM-simulated patients—a mode that mirrors the verbal, real-time nature of actual therapy sessions but has more requirements for realistic simulation than typed exchanges. CARE combines (1) voice-based patients governed by expert-elicited Constitutional AI principles for realistic behavior, with scenarios aligned to a multi-week therapeutic alliance curriculum, and (2) post-practice feedback from a fine-tuned LLM trained on expert-annotated counseling transcripts. Finally, CARE's design explores post-practice review activities that guide therapy students to reflect on their own approach and/or the AI's (when AI feedback is available). Reflection prompts are provided at the session-level (broadly about overall approach) or utterance-level (detailed analysis of specific therapeutic responses).

A key challenge in deploying AI systems that interact with humans in healthcare is evaluation, which is often designed to occur first in small trials. Standard randomized controlled trials (RCTs) require large populations, yet healthcare education cohorts are typically small; lab studies, meanwhile, lack ecological validity for sustained educational use. As our second contribution, we demonstrate the applicability of micro-randomized trials (MRTs) [Klasnja *et al.*, 2015]—a method previously used in mobile health—to AI education settings. MRTs use within-person randomization at each engagement opportunity, gaining statistical power over longer time periods common in education. This enables estimating the marginal causal excursion effect [Qian *et al.*, 2021], which measures the momentary impact of delivering a specific intervention. We deploy CARE in a graduate psychotherapy course ($n = 25$ students, 5 weeks) and detail how adapting

MRTs from their typical push-based mobile health context to our pull-based, student-initiated setting required tailoring the analysis for irregular decision point cadence—offering guidance for future MRT deployments in health education.

Using this approach, we tested four conditions crossing AI feedback (present/absent) with reflection granularity (utterance/session-level). We hypothesized that AI feedback combined with utterance-level reflection would be most effective: utterance-level reflection forces deeper cognitive engagement by requiring students to actively compare their responses with AI alternatives rather than passively consuming feedback [Buçinca *et al.*, 2021], while embodying the educational principle of contrasting cases where side-by-side examination helps learners identify generalizable patterns [Schwartz *et al.*, 2016].

Our results confirmed this hypothesis. Students in the AI feedback with utterance-reflection condition rated learning utility significantly higher (5.90/7, treatment effect 1.29, $p < 0.05$) compared to the no-feedback, session-reflection baseline (4.61/7). This condition also increased reflection engagement nearly nine-fold. Critically, students integrated AI suggestions rather than passively accepting them, demonstrating the effectiveness of our reflection design as a cognitive forcing intervention. Neither AI feedback alone nor utterance-level reflection alone showed significant benefits, indicating that the combination is essential.

Our contributions are twofold:

1. We introduce CARE, an AI system that enables voice-based deliberate practice with LLM-simulated patients—supporting a training modality closer to real therapy than typing-only systems—combined with structured reflection on AI feedback.
2. We demonstrate that micro-randomized trials offer a viable evaluation paradigm for AI systems in healthcare education, enabling rigorous causal inference in small cohorts typical of medical training programs.

2 Related Work

AI for Training Psychotherapy Skills Deliberate practice (DP) has become an established framework for developing therapeutic expertise [Nurse *et al.*, 2024; Miller *et al.*, 2020] with empirical evidence supporting its effectiveness [Larsson *et al.*, 2025]. However, widespread implementation is limited by the significant time and expertise required for supervision and personalized feedback. This has led researchers to explore AI-powered systems that automate simulation and feedback [Wang *et al.*, 2024; Sharma *et al.*, 2023; Hsu *et al.*, 2023; Steenstra *et al.*, 2025]. Results from lab studies suggest AI tools can enhance novice therapists' confidence and skill, but questions remain about effectiveness during sustained, education usage [Ajluni, 2025; Yamamoto *et al.*, 2024; Wang *et al.*, 2024]. To our knowledge, our system is one of the first to create patient-scenarios matched to a psychotherapy curriculum; and the first to use multimodal-native LLMs to develop emotionally-expressive voice-based patients using a Constitutional AI principle-approach [Louie *et al.*, 2024]—something not afforded by speech-to-text-to-speech approaches.

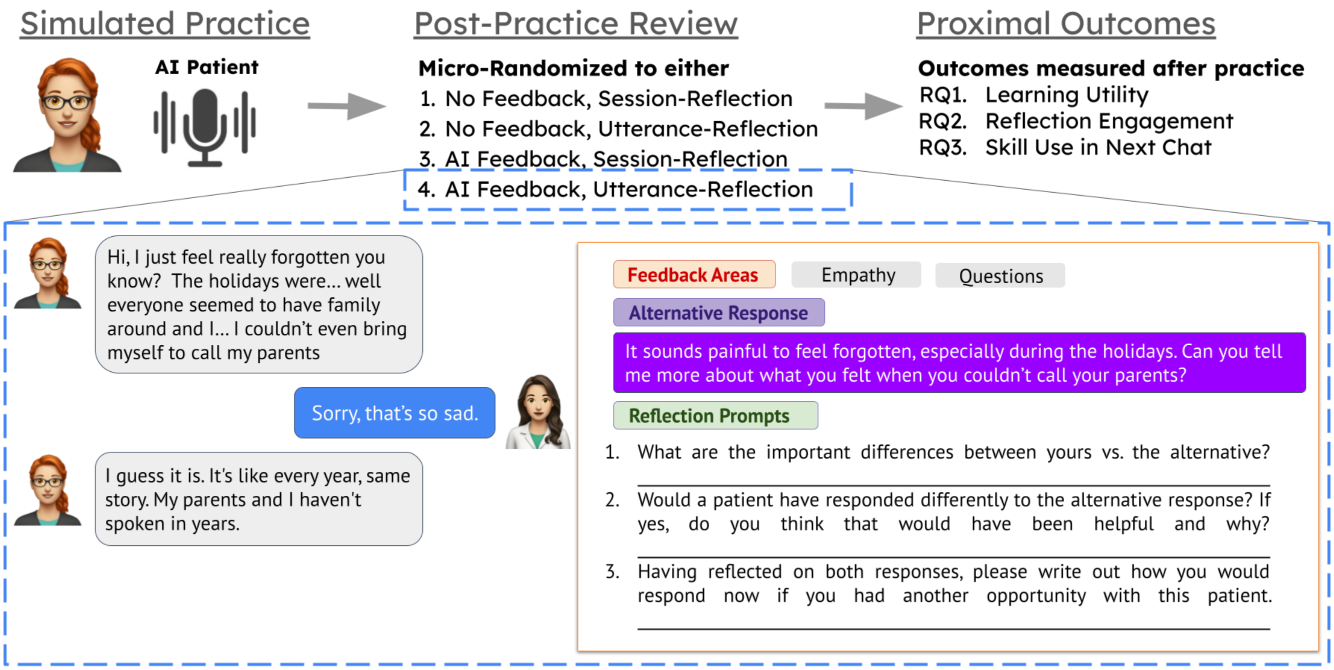
Experimental Methods for System Deployments A general challenge in evaluating systems is balancing ecological validity with understanding causal mechanisms. [Klasnja *et al.*, 2011] argues that evaluating proximal outcomes enables rigorous efficacy evaluations within the resource constraints of field studies. Building on this, micro-randomized trials (MRTs) estimate treatment effects by randomizing different intervention versions to the same participants over multiple decision points [Klasnja *et al.*, 2015]. MRTs have been adopted across physical and mental health interventions, including physical activity promotion [Coppens *et al.*, 2024; Figueroa *et al.*, 2022], substance abuse management [Rabbi *et al.*, 2018], and mood management [Arévalo Avalos *et al.*, 2024; Zhao and Choudhury, 2024]. [Cho and Kizilcec, 2021] proposed extending MRTs to education, and realizing the potential of mobile interventions for education, [Breitwieser *et al.*, 2024] implement MRTs in an educational setting. Compared to other within-subject designs such as crossover or round-robin, MRTs offer distinct advantages: each engagement opportunity serves as a decision point, enabling more frequent randomization than fixed time-periods. Unlike crossover designs that assume a stable baseline and require washout periods, treatment effect estimates in MRTs explicitly model participant states as time-varying moderators. To our knowledge, this study is among the first empirical MRTs in health education and the first LLM-based classroom MRT with categorical arms.

3 CARE: LLM-Simulation Platform for Psychotherapy Skills Training

CARE is a web-based platform that enables psychotherapy students to practice a *voice-based, multi-turn* conversation with an LLM-simulated patient. After a simulated practice session, students can optionally receive LLM-generated feedback on their responses and reflection questions about their practice and/or the feedback they received.

Voice-based LLM-simulated patients. CARE extends an existing method for defining realistic LLM-patient behaviors that uses Constitutional AI principles which were elicited from domain-experts during interactive testing [Louie *et al.*, 2024]. Each of CARE's patients integrate expert-validated behavioral principles released by [Louie *et al.*, 2024] along with additional principles that were defined during formative testing of the voice-based LLM-patients. CARE's simulation method uses the OpenAI *gpt-4o-realtime-preview-2024-12-17* API to role-play patients due to its strong ability to consistently follow role instructions in a specific behavioral-style (e.g., *Respond to encouraging words or suggested solutions with hesitation, doubting their significance*) and expressive voice (e.g., *Emphasize fear of losing family connections, with a worried tone like having a knot in the throat. Voice should break as in holding back tears.*).

One deployment challenge was creating LLM-simulated patients that matched the weekly syllabus in the psychotherapy course. Rather than manually-drafting each of the scenarios, which can be labor-intensive for a teaching assistant, we bootstrapped the scenario creation workflow by providing class readings and assignments in context so an LLM


Simulated Practice
AI Patient
Post-Practice Review
Micro-Randomized to either
1. No Feedback, Session-Reflection
2. No Feedback, Utterance-Reflection
3. AI Feedback, Session-Reflection
4. AI Feedback, Utterance-Reflection
Proximal Outcomes
Outcomes measured after practice
RQ1. Learning Utility
RQ2. Reflection Engagement
RQ3. Skill Use in Next Chat
Feedback Areas
Empathy
Questions
Alternative Response
Reflection Prompts

Hi, I just feel really forgotten you know? The holidays were... well everyone seemed to have family around and I... I couldn't even bring myself to call my parents

Sorry, that's so sad.

I guess it is. It's like every year, same story. My parents and I haven't spoken in years.

It sounds painful to feel forgotten, especially during the holidays. Can you tell me more about what you felt when you couldn't call your parents?

1. What are the important differences between yours vs. the alternative?
2. Would a patient have responded differently to the alternative response? If yes, do you think that would have been helpful and why?
3. Having reflected on both responses, please write out how you would respond now if you had another opportunity with this patient.

Figure 1: CARE is AI-based psychotherapy training system that provides speaking-practice with LLM-simulated patients, followed by post-practice review activities which vary in AI feedback (present/absent) and granularity of reflection (session-level/utterance-level). To test what type of review activity is most effective, we conducted a micro-randomized trial in a clinical psychology graduate course (5 weeks, 25 students), where students were randomized to one of four categorical treatment arms after AI-simulated practice, and three proximal outcomes were measured after completing the review activity: perceived learning utility (student responses to two Likert-scales items); reflection engagement behaviors (amount of reflection content a student writes), and skill use in the next chat (scored via NLP classifiers).

could draft patient scenarios for the weekly course topics (e.g., therapist self-disclosure and cultural humility, sexual attraction and boundary issues, etc.). The second-author, a course teaching assistant, then reviewed the AI-drafted scenarios and revised them based on their experience and discretion to complement the weekly assignments. Example patient prompts are available in Supplementary Materials A.3. Simulated patients were tested in 5-10 minute voice-based conversations and further refined to improve voice-based realism through increased emotional expressiveness and consistency with patient demographics (e.g., gender, age).

Post-practice LLM-generated feedback. CARE integrates an existing method for generating counseling feedback that fine-tunes and self-improves the Llama-2 13B parameter model using an expert-annotated feedback dataset of peer counseling transcripts [Chaszczewicz *et al.*, 2024]. We choose this fine-tuned model because it generates contextual feedback at multiple-levels, mirroring psychotherapy supervision structure: (1) assess trainee response by highlighting strengths and areas needing improvement across eight microskills (empathy, reflections, questions, validation, suggestions, session management, professionalism, and self-disclosure), (2) generate goal-oriented explanation of the feedback and a suggested alternative.

Post-practice reflection interfaces. CARE's utterance-level reflection (Table 3) was designed to address evidence that simply presenting AI recommendations leads to superficial processing [Bansal *et al.*, 2021; Buçinca *et al.*, 2020]. Rather than presenting AI-generated alternatives as recommendations to accept or reject, the interface prompts students to compare their original utterance with the alternative side-by-side. Students then formulate their own revised response through a three-stage Intention-Impact-Revision sequence, compelling deeper cognitive engagement rather than passive acceptance of AI suggestions [Schwartz *et al.*, 2016] (see Appendix A.2 for theoretical grounding).

To test the impact of this reflection design, we built three alternative variants: (1) reviewing AI feedback while reflecting on the session holistically; (2) reflecting on a subset of utterances without AI feedback; and (3) reflecting on the session holistically without AI feedback, using two questions covering what was challenging and what the student would revise. These variants allow us compare the impact of variations of AI feedback and reflection granularity together.

4 Micro-randomized Trial Design

To evaluate which post-session review activities best support learning from AI-simulated practice, we conducted a Micro-randomized Trial (MRT) [Klasnja *et al.*, 2015] in a psychotherapy classroom. Unlike randomized control trials that assign individuals to one treatment throughout, MRTs randomize treatments at multiple moments per participant across trial weeks. Applying MRTs to health education required adapting methodological choices typically made for mobile health interventions. We describe our adaptations for (1) cat-

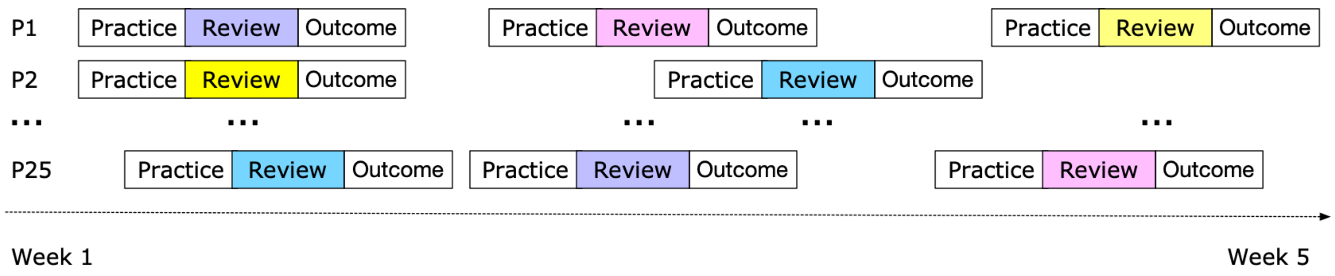


Figure 2: Overview of participants journey throughout the psychotherapy course's micro-randomized trial. Students (n=25) engaged in simulated practice sessions with AI patients, followed by one of four post-session review activities (No Feedback, Session Reflection , No Feedback, Utterance Reflection , AI Feedback, Session Reflection , AI Feedback, Utterance Reflection). Proximal outcomes were measured via self-reported surveys, analysis of reflection content, and skill usage. Our MRT have irregular decision points due to variation in schedule and frequency that each student engages with AI practice.

egorical treatments and (2) pull-based interventions with irregular decision point timing.

4.1 MRTs with Categorical Treatments

How Randomization Occurs: In a micro-randomized trial, n participants are randomized to one of K treatment variants at each of T decision points. Let $A_{i,t} \in \{0, 1, \ldots, K-1\}$ denote the treatment assignment for participant i at decision point t, where 0 represents the reference condition. In our study, $A_{i,t}$ corresponds to the post-session review activity assigned after each practice session, with $K = 4$ treatment levels (our 2×2 factorial design). After treatment $A_{i,t}$, we observe proximal outcomes $Y_{i,t}$—for example, a student's perceived learning utility. The randomization probability $p(k) = \Pr(A_t = k)$ was fixed at 25% for each treatment level, independent of history. Figure 2 illustrates this process. MRTs also incorporate availability constraints through a binary indicator $I_{i,t}$. In our pull-based context, students are available by default when they initiate practice but unavailable when the system exits before they see the review activity.
Causal Excursion Effect for Categorical Treatments: To estimate treatment effects, we use the *causal excursion effect* (CEE) [Boruvka *et al.*, 2018]. The CEE builds on the potential outcomes framework [Rubin, 1974], representing the difference between expected outcomes under two "excursions" from the trial's treatment protocol. Most MRT studies compare binary treatments (intervention vs. control). Our four-level design requires the categorical extension of causal excursion estimators developed by [Lin and Qian, 2025]. For treatment level k at time t, with immediate proximal outcomes (Y_t), the CEE for categorical treatments is defined as

$$\begin{aligned} \mathrm{CEE}_{tk} = \{&Y_t(\bar{A}_{t-1}, k) \mid I_t(\bar{A}_{t-1}) = 1\} \\ - \{&Y_t(\bar{A}_{t-1}, 0) \mid I_t(\bar{A}_{t-1}) = 1\} \end{aligned} \tag{1}$$

where the treatment assignment history up to time t $(\bar{A}_{t-1})$ is random and follows the probabilities set by treatment protocol of the MRT. The definition in (1) captures the contrast between receiving treatment k versus no treatment (reference condition) at time t, conditioned[1] on being available at time t $(I_t(\bar{A}_{t-1}) = 1)$. Like prior work, we assume the causal excursion effect is linear, and we use the open-source R package[2] developed by [Lin and Qian, 2025] to estimate these effects via a weighted-centered least squares method that accounts for the nested, longitudinal structure of the data. We compute the immediate[3] CEE per outcome variable of interest– e.g. on perceived learning utility.

4.2 Handling Student-Initiated Decision Points

Unlike push-based mobile health interventions where decision points occur at system-determined times (e.g., daily notifications), our pull-based setting means decision points are student-initiated. Students chose when and how often to practice, creating *irregular timing* between decision points. The time between decision points in our MRT followed a bimodal distribution: peaks at 1 day and 7 days. This indicates students either practiced multiple times on the same day (completing their weekly assignment in one sitting) or practiced once per week. Some students completed 14 sessions over 5 weeks; others completed only one.

This irregularity presents a modeling choice: **Option A: Filter to regular intervals.** Keep only decision points at least 5–7 days apart, analyzing sessions that follow a more regular schedule. This avoids over-weighting students who engaged more frequently than intended, but requires filtering data. **Option B: Use all decision points.** Include all decision points and estimate effects that average over each student's natural engagement patterns. We chose Option B, using all decision points to preserve all data. Our causal excursion estimates answer: "*How do treatments compare in terms of proximal outcomes, when averaging over the different number and schedule of pulls that happened naturally for each student?*"

4.3 Research Questions

Note that MRTs are not confirmatory studies designed to evaluate an intervention package; instead, they are focused

[1] The full CEE formulation allows conditioning on context features to examine effect heterogeneity; we do not explore such moderation in this study.

[2] https://github.com/JeremyJosephLin/causal_excursion_mult_trt

[3] We present the CEE for immediate outcomes, the most common MRT setting. For the general formulation with delayed proximal outcomes $(t + \Delta)$, see [Lin and Qian, 2025].

on evaluating intervention components. With the independent variable as whether participants received one of four post-practice review activities, we ask: *RQ1:* How do post-practice review activities–varying in the type of feedback and reflection–impact students' **perceived learning benefits**? *RQ2:* How do post-practice review activities impact students' **reflective behaviors** (e.g., degree of reflection, what they reflect upon)? *RQ3:* How do post-practice review activities impact students' **skill use in the next chat**? *RQ4:* How do students **perceive CARE's review variants**?

4.4 Measures

Three proximal outcomes were measured: **Perceived learning utility** (RQ1) was measured immediately after the review activity using the average over two 7-point Likert scales: *"This practice session was valuable for my development as a therapist"* and *"I gained useful insights from this practice experience"*. **Reflection Engagement** (RQ2) was studied under the hypothesis that students would engage more in activities they found valuable; the metric was defined as the logarithmic transform of written reflection length, normalized by number of reflection prompts completed (e.g., sessions-reflections have 2 prompts, utterance-reflections have $3 \times N_{utterances}$). **Skill use in the next chat** (RQ3) was assessed with RoBERTa classifiers [Louie *et al.*, 2025] fine-tuned for four most frequent skills covered by the AI feedback model. See Appendix A.7 for classifier details. **Perceptions of CARE's post-practice review variants** (RQ4) was collected via an end-course survey. The survey asked about likes and dislikes for the different conditions (e.g., *No Feedback, Utterance Reflections* and *AI Feedback, Utterance Reflections*) with multi-select answers and an "Other" option for free response.

4.5 Data Analyses

Causal Excursion Effects: We compute separate CEE estimates for the three proximal outcomes. As noted above, the pull-based setting produces varying decision-point timing; we address this by using all decision points and computing marginal effects, that average over the different number and schedule of pulls that happened naturally for each student. Moreoever, a software issue affected 48% decision points in both Session Reflection conditions, where participants did not receive the reflection prompt; we set availability indicator $I_{i,t} = 0$ to handle such cases in the CEE estimation.
Content Analysis of Reflections and Response Revision Patterns: As part of studying reflection behaviors (RQ2), we conducted a content analysis of utterance-level reflections (session reflections had low completion rates). The final reflection question asks students to write a revised response. We investigated revision frequency across conditions and, for the AI feedback condition, whether revisions accepted the AI's alternative verbatim or integrated both responses. **End-course Surveys:** We report counts for multi-select responses and conduct a thematic analysis of free-response quotes, expanding the codebook as needed for responses outside existing themes.

5 Results

Over the 5-week MRT, post-practice review activities were randomized a median of 7 times (IQR: [6,9], Min: 1, Max: 14) per participant. Students completed 171 practices; in 91 of these, students self-reported learning utility.

5.1 Effects on Perceived Learning Utility (RQ1)

Causal Excursion Estimate: Table 1 presents CEE estimates for each post-practice activity. The combination of AI feedback with utterance-level reflection moved students' ratings to 'agree" (5.90) that the post-practice activity was educationally valuable–a meaningful difference (+1.29) over the *No feedback, Session-reflection* condition. Other combinations showed no meaningful improvement or slight decreases, supporting our hypothesis that utterance-level reflection helps students extract value from AI feedback, but not as a standalone exercise.

5.2 Effects on Reflection Behavior (RQ2)

Causal Excursion Estimate: Both utterance-level conditions produced significantly higher engagement than session-level reflections (Table 1): *No Feedback, Utterance Reflection* increased engagement by +0.96 points (95% CI: 0.21, 1.70), and *AI Feedback, Utterance Reflection* by +1.68 points (95% CI: 0.47, 2.89). *AI Feedback, Session Reflection* showed only a modest, non-significant increase (+0.31). These findings align with reflection completion rates: session reflections had 2.1% completion (1/48 sessions), while utterance reflections reached 18.2% (16/88 sessions)—a nine-fold increase.
Response Revision Patterns by Treatment Condition: Figure 3 (Appendix) details the response revision patterns for the two utterance-reflection conditions. In the *No Feedback, Utterance Reflections* condition, students split evenly: 5 of 10 (50%) maintained original responses, while 5 (50%) revised. With *AI Feedback, Utterance Reflection*, revision patterns shifted dramatically: only 2 of 17 (12%) maintained original responses, while 15 (88%) revised after seeing AI feedback. Among revisers, most preferred integration over wholesale adoption—10 created hybrid responses combining their original with AI suggestions, while only 5 accepted the AI alternative verbatim.

5.3 Effects on skill use in next session (RQ3)

Causal Excursion Estimates: Table 11 (Appendix) presents CEE estimates for effect of review activity on next-chat performance on four skills. Intriguingly, the *AI Feedback, Utterance Reflection* condition—rated highest for learning utility—was associated with a significant *decrease* in strong question use (Estimate: -0.129, 95% CI: [-0.25, -0.01]). This tension between perceived learning and automated metrics may reflect students critically engaging with AI feedback and exploring alternative therapeutic strategies beyond the specific skills being measured.

5.4 Views on post-practice review variants (RQ4)

Utterance Reflections, AI Feedback vs. No Feedback Of the 10 participants who used *AI Feedback, Utterance Reflections*, over half (6/10) valued critically engaging with AI

	Learning Utility		Reflection Engagement	
Treatment Condition	Estimate	95% CI	Estimate	95% CI
Intercept (No Feedback, Session Reflection)	4.61	–	0.04	–
No Feedback, Utterance Reflection	-0.55	(-2.18, +1.08)	+0.96∗	(+0.21, +1.70)
AI Feedback, Session Reflection	+0.05	(-0.91, +1.02)	+0.31	(-0.29, +0.91)
AI Feedback, Utterance Reflection	+1.29∗	(+0.29, +2.29)	+1.68∗	(+0.47, +2.89)

Table 1: Estimates represent the causal excursion effect from micro-randomized trial analysis. **Learning Utility** is the average of a two-item self-reported learning utility measure (7-point Likert scale). **Reflection Engagement** is the total number of characters in a student's reflection(s), normalized by number of relfection prompts answered, scaled logarithmically. $*$ indicates $p < 0.05$.

feedback rather than accepting it as a gold standard. For *No Feedback, Utterance Reflections*, 40% (6/15) liked aspects of them, but the primary concern (10/15) was not understanding why certain responses were selected for reflection. While the selection mechanism was based scores from the AI feedback system, this was purposefully not shown to the user. There is a clear desire for more agency, with 47% (7/15) wishing they could choose which utterances to reflect on.

AI Feedback – better than no feedback, but room to improve: While the delivery of AI feedback was preferred over none at all, students wanted further improvements in AI feedback. Students wished the feedback system's evaluation criteria had more variety each week (15/25), as CARE's feedback model focused only on core listening skills such as empathy, validation, reflections, and questions. This was especially salient for students who were practicing new techniques for this more advance course on therapeutic alliance, which goes beyond their previous course which teaches helping microskills and motivational interviewing. One student explained: *"I believe the AI platform needs to be trained in more therapeutic techniques so it can give appropriate feedback when skills other than basic listening skills are being used"* (P15). This focus on core listening skills was experienced as repetitive after multiple weeks of usage: *"I felt the feedback was often repetitive and narrowly focused. Often, when I was trying new techniques, I would not pick up on alternative ways to approach situations"* (P25).

Students wished for additional AI feedback capabilities, such as feedback that assessed audio-features rather than just written transcripts (13/25); feedback that analyzed the session as a whole (16/25); and feedback that is generated from a patient-perspective (17/25). As one student explained: *"The largest thing for me was the inability to get feedback from the client, although truthfully I think it may not be possible to ever get to this point"* (P2). Interestingly, CARE's utterance-level reflections did prompt students to consider the impact of their therapeutic techniques on how the patient next responded; a direction of future work would be faithfully generating the AI patient's answer to such questions.

6 Discussion

6.1 Designing LLM-Based Training Tools

We designed CARE around two core principles grounded in Kolb's experiential learning theory and contrasting cases (Appendix A.2): (1) combining AI feedback with structured reflection to promote active comparison, and (2) using cognitive forcing functions [Buçinca *et al.*, 2021] to counter documented AI over-reliance concerns [Zhai *et al.*, 2024]. Our MRT provided causal evidence for what worked—and revealed unexpected mechanisms.

Pair Feedback with Comparison-Based Reflection. Utterance-level reflections received significantly more engagement than session-level reflections (nine-fold increase in completion), with AI feedback plus utterance reflection showing the highest perceived learning utility (+1.29 over baseline). However, our design confounds two factors: granularity and comparison structure. Our utterance-level reflections scaffolded side-by-side comparison of the student's response against the AI alternative—a structure absent from our session-level condition. We cannot determine whether the effect stemmed from utterance-level granularity per se, or from the comparison-based design that prompted critical engagement. The revision patterns (Section 5) suggest the latter: students used the comparison to create integrative responses rather than accepting AI feedback wholesale. *Recommendation:* Future experiments should disentangle these factors by testing session-level feedback paired with session-level comparison reflection—e.g., comparing overall session flow or therapeutic agenda against AI-suggested alternatives.

Enable Learner Selection of Reflection Targets. Survey data revealed a clear desire for agency: 47% (7/15) of participants wished they could choose which utterances to reflect on. *Recommendation:* Rather than systems selecting which subset of utterances to present for reflection, consider designs that let learners choose—reducing the need for algorithmic guessing while potentially increasing engagement with personally meaningful moments.

Scaffold Critical Engagement, Not Just Feedback Delivery. AI feedback catalyzed self-correction: 90% of students revised responses when feedback was provided versus 50% without. Crucially, most revisions were integrative rather than verbatim, confirming that our reflection design functioned as a cognitive forcing function [Buçinca *et al.*, 2021]. Side-by-side comparison of original versus AI-generated responses enabled pattern recognition; survey data corroborated that students used AI feedback for critical thinking rather than as definitive answers. *Recommendation:* Effective AI clinical training tools should structure learner engagement with feedback through reflection prompts that compel analytical thinking—particularly important given growing AI over-reliance concerns [Zhai *et al.*, 2024].

Design for Integration, Not Replacement. The prevalence of integrative revisions confirmed that our design successfully positioned AI feedback as a reference point rather than a gold standard. This was deliberate: our AI model was imperfect and did not cover all course topics. *Recommendation:* Frame AI suggestions as alternative perspectives for synthesis, not correct answers. This maintains learner agency for critical thinking about LLMs [Singh *et al.*, 2024]—a design philosophy applicable to healthcare training contexts where clinical judgment cannot be fully replaced by AI.

6.2 MRTs for AI Health Education Studies

Estimating Causal Effects in Ecologically Valid Settings. Our MRT evaluated four variants of an AI-simulation tool in an authentic psychotherapy classroom—and obtained causal estimates despite a small cohort of 25 students. With a median of 7 randomizations per participant over 5 weeks, we detected significant effects ($p<0.05$). MRTs are well-suited for longitudinal, small-classroom deployments, where individual randomization lacks statistical power, and single-arm trials lack actionable estimates. The MRT benefitted from frequent randomization and measurement opportunities corresponding with student interactions with the AI intervention, which cross-over designs with pre-defined time intervals (e.g., every 2 weeks) may miss.

This promise is especially compelling for AI/LLM-based training systems, which involve numerous design decisions—feedback granularity, reflection structure, prompt framing—whose effects are best understood by observing trainees in their natural usage contexts rather than controlled laboratory settings. CARE exemplifies a pull-based intervention where trainees initiate practice sessions; MRTs enabled us to estimate causal effects of design variations within authentic classroom workflows. However, the pull-based nature introduced methodological complexity: decision point timing followed a bimodal distribution (peaks at 1 day and 7 days), requiring us to choose between filtering for regularity or averaging over naturalistic engagement patterns. We chose the latter, but future deployments should consider how this modeling choice affects the generalizability of their estimates.

7 Limitations and Future Work

While our findings provide evidence for specific design principles, several factors limit generalizability.

Population and Domain Specificity. Our study involved 25 doctoral students with established analytical skills and professional identity development. The sophisticated critical engagement we observed—integrative revisions over verbatim adoptions—may not generalize to earlier-stage learners or different training contexts. Additionally, psychotherapy training involves interpersonal skills where multiple approaches can be effective; comparative reflection may be particularly suited to such domains. Generalizability to structured domains with clearer right/wrong distinctions requires investigation.

Feedback-Curriculum Mismatch. Our AI feedback model focused on utterance-level micro-skills, creating a mismatch with course topics like confidentiality dilemmas or alliance rupture repair. Student surveys reflected this limitation: more than half reported receiving unhelpful feedback, preferred session-level over utterance-level feedback, and wanted more variety in evaluation criteria across weeks. For AI feedback to achieve maximum educational utility, feedback models must adapt to specific training topics—requiring either annotated transcripts for common curricula or domain-adaptation methods for new contexts.

Behavioral Performance Assessment. Our exploratory analysis of behavioral performance (RQ3) revealed a tension between self-reported learning utility and automated skill metrics: the condition students rated most valuable showed a significant decrease in strong question usage. Interpreting this result requires acknowledging limitations of our assessment approach. The AI feedback model and behavioral performance classifiers were trained on the same dataset, meaning both systems share similar biases and coverage gaps. If students critically engaged with feedback they found misaligned with their learning goals, decreased performance on metrics derived from the same training paradigm may reflect reasoned divergence rather than skill decline.

Future work prioritizing skill use outcomes should address two challenges. First, separating exploratory practice from assessment contexts would clarify interpretation—periodic standardized assessments distinct from experimental practice would enable meaningful progress tracking, addressing students' expressed desire for progress visibility. Second, assessment criteria must align with training topics: general counseling frameworks suit foundational skills, but advanced topics like alliance rupture repair require specialized rubrics, potentially drawing on emerging automatic assessment methods for specific constructs [Goldberg *et al.*, 2024].

8 Conclusion

This work presents a five-week micro-randomized trials, testing how an LLM-based tool enables practice of therapy skills with simulated patients and automated feedback in authentic classroom settings. Our work makes two key contributions: empirical evidence that post-practice activity design is critical to AI training system effectiveness, and methodological validation of MRTs as a solution to the statistical power challenges that typically prevent causal inference in small-scale educational deployments. Our findings reveal that combining AI-generated feedback with structured, utterance-level reflection significantly enhanced students' perceived learning utility, fostering critical engagement that moved learners beyond passive acceptance toward active integration and construction of improved responses. Methodologically, our 25-student, 5-week study demonstrates MRTs' unique value for educational technology research: they bridge the gap between highly controlled lab studies with limited ecological validity and large-scale field studies that are often impractical in specialized educational contexts like graduate training programs.

References

[Ajluni, 2025] Victor Ajluni. Artificial intelligence in psychiatric education: Enhancing clinical competence through simulation. *Industrial Psychiatry Journal*, 34(1):11–15, 2025.

[Arévalo Avalos *et al.*, 2024] Marvyn R Arévalo Avalos, Jing Xu, Caroline Astrid Figueroa, Alein Y Haro-Ramos, Bibhas Chakraborty, and Adrian Aguilera. The effect of cognitive behavioral therapy text messages on mood: A micro-randomized trial. *PLOS Digital Health*, 3(2):e0000449, 2024.

[Atherton, 2013] J S Atherton. The experiential learning cycle. https://web.archive.org/web/20150206134121/http://www.learningandteaching.info/learning/experience.htm, 2013. Accessed: February 6, 2015.

[Bansal *et al.*, 2021] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021.

[Boruvka *et al.*, 2018] Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523):1112–1121, 2018.

[Breitwieser *et al.*, 2024] Jasmin Breitwieser, Andreas B Neubauer, Florian Schmiedek, and Garvin Brod. Realizing the potential of mobile interventions for education. *npj Science of Learning*, 9(1):76, 2024.

[Buçinca *et al.*, 2020] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464, 2020.

[Buçinca *et al.*, 2021] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

[Chaszczewicz *et al.*, 2024] Alicja Chaszczewicz, Raj Sanjay Shah, Ryan Louie, Bruce A Arnow, Robert Kraut, and Diyi Yang. Multi-level feedback generation with large language models for empowering novice peer counselors. *arXiv preprint arXiv:2403.15482*, 2024.

[Cho and Kizilcec, 2021] Ji Yong Cho and René F Kizilcec. Applying the behavior change technique taxonomy from public health interventions to educational research. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*, pages 195–207, 2021.

[Coppens *et al.*, 2024] Ine Coppens, Toon De Pessemier, and Luc Martens. Balancing habit repetition and new activity exploration: A longitudinal micro-randomized trial in physical activity recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 1147–1151, New York, NY, USA, 2024. Association for Computing Machinery.

[Feyzi-Behnagh *et al.*, 2014] Reza Feyzi-Behnagh, Roger Azevedo, Elizabeth Legowski, Kayse Reitmeyer, Eugene Tseytlin, and Rebecca S Crowley. Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system. *Instructional science*, 42(2):159–181, 2014.

[Figueroa *et al.*, 2022] Caroline A Figueroa, Nina Deliu, Bibhas Chakraborty, Arghavan Modiri, Jing Xu, Jai Aggarwal, Joseph Jay Williams, Courtney Lyles, and Adrian Aguilera. Daily motivational text messages to promote physical activity in university students: results from a microrandomized trial. *Annals of behavioral medicine*, 56(2):212–218, 2022.

[Goldberg *et al.*, 2024] Simon B Goldberg, Michael Tanana, Shaakira Haywood Stewart, Camille Y Williams, Christina S Soma, David C Atkins, Zac E Imel, and Jesse Owen. Automating the assessment of multicultural orientation through machine learning and natural language processing. *Psychotherapy*, 2024.

[Hill and Knox, 2023] Clara E Hill and Sarah Knox. Psychotherapy training and supervision with undergraduate and graduate students. 2023.

[Hsu *et al.*, 2023] Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. Helping the helper: Supporting peer counselors via AI-Empowered practice and feedback. May 2023.

[Klasnja *et al.*, 2011] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. How to evaluate technologies for health behavior change in hci research. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3063–3072, 2011.

[Klasnja *et al.*, 2015] Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S):1220, 2015.

[Kolb, 2014] David A Kolb. *Experiential learning: Experience as the source of learning and development*. FT press, 2014.

[Kühne *et al.*, 2020] Franziska Kühne, Peter Eric Heinze, and Florian Weck. Standardized patients in psychotherapy training and clinical supervision: study protocol for a randomized controlled trial. *Trials*, 21:1–7, 2020.

[Larsson *et al.*, 2025] Johannes Larsson, David Werthén, Jan Carlsson, Osame Salim, Edvin Davidsson, Alexandre Vaz, Daniel Sousa, and Joakim Norberg. Does deliberate practice surpass didactic training in learning empathy skills?–a randomized controlled study. *Nordic Psychology*, 77(1):39–52, 2025.

[Lent *et al.*, 2003] Robert W Lent, Clara E Hill, and Mary Ann Hoffman. Development and validation of the counselor activity self-efficacy scales. *Journal of Counseling Psychology*, 50(1):97, 2003.

[Lin and Qian, 2025] Jeremy Lin and Tianchen Qian. Micro-randomized trials with categorical treatments: Causal effect estimation and sample size calculation. *arXiv preprint arXiv:2504.15484*, 2025.

[Louie *et al.*, 2024] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*, 2024.

[Louie *et al.*, 2025] Ryan Louie, Ifdita Hasan Orney, Juan Pablo Pacheco, Raj Sanjay Shah, Emma Brunskill, and Diyi Yang. Can llm-simulated practice and feedback upskill human counselors? a randomized study with 90+ novice counselors. *arXiv preprint arXiv:2505.02428*, 2025.

[Miller *et al.*, 2020] Scott D Miller, Daryl Chow, Bruce E Wampold, Mark A Hubble, AC Del Re, Cynthia Maeschalck, and Susanne Bargmann. To be or not to be (an expert)? revisiting the role of deliberate practice in improving performance. *High Ability Studies*, 31(1):5–15, 2020.

[Modi *et al.*, 2022] Hemangi Modi, K Orgera, and A Grover. Exploring barriers to mental health care in the us. *Research and Action Institute*, 10, 2022.

[Nelson-Jones, 2013] Richard Nelson-Jones. *Practical counselling and helping skills: text and activities for the lifeskills counselling model*. Sage, 2013.

[Nurse *et al.*, 2024] Karina Nurse, Melissa O'shea, Mathew Ling, Nathan Castle, and Jade Sheen. The influence of deliberate practice on skill performance in therapeutic practice: A systematic review of early studies. *Psychotherapy Research*, pages 1–15, 2024.

[Qian *et al.*, 2021] Tianchen Qian, Hyesun Yoo, Predrag Klasnja, Daniel Almirall, and Susan A Murphy. Estimating time-varying causal excursion effects in mobile health with binary outcomes. *Biometrika*, 108(3):507–527, 2021.

[Rabbi *et al.*, 2018] Mashfiqui Rabbi, Meredith Philyaw Kotov, Rebecca Cunningham, Erin E Bonar, Inbal Nahum-Shani, Predrag Klasnja, Maureen Walton, Susan Murphy, et al. Toward increasing engagement in substance use data collection: development of the substance abuse research assistant app and protocol for a microrandomized trial using adolescents and emerging adults. *JMIR research protocols*, 7(7):e9850, 2018.

[Rubin, 1974] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[Schwartz *et al.*, 2016] Daniel L Schwartz, Jessica M Tsang, and Kristen P Blair. *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them*. WW Norton & Company, 2016.

[Sharma *et al.*, 2023] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, January 2023.

[Singh *et al.*, 2024] Anjali Singh, Christopher Brooks, Xu Wang, Warren Li, Juho Kim, and Deepti Wilson. Bridging learnersourcing and ai: Exploring the dynamics of student-ai collaborative feedback generation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, page 742–748, New York, NY, USA, 2024. Association for Computing Machinery.

[Steenstra *et al.*, 2025] Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2025.

[Thesen *et al.*, 2025] Thomas Thesen, Wade N. OâBrien, Simon Stone, and Roshini Pinto-Powell. Generative AI as the First Patient: Practice, Feedback, and Confidence. *Medical Science Educator*, August 2025.

[Vasconcelos *et al.*, 2023] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), April 2023.

[Wang *et al.*, 2024] Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*, 2024.

[Yamamoto *et al.*, 2024] Akira Yamamoto, Masahide Koda, Hiroko Ogawa, Tomoko Miyoshi, Yoshinobu Maeda, Fumio Otsuka, Hideo Ino, et al. Enhancing medical interview skills through ai-simulated patient interactions: non-randomized controlled trial. *JMIR medical education*, 10(1):e58753, 2024.

[Zhai *et al.*, 2024] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. The effects of over-reliance on ai dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28, 2024.

[Zhao and Choudhury, 2024] Yiran Zhao and Tanzeem Choudhury. Evaluate closed-loop, mindless intervention in-the-wild: A micro-randomized trial on offset heart rate biofeedback. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '24, page 307–312, New York, NY, USA, 2024. Association for Computing Machinery.

A Appendix

A.1 Mapping of AI Patients to Course Topics

Table 2 provides an example of a patient scenarios mapped to each of the 5 weeks when the MRT was active. For this course on formation, rupture, and repair of the therapeutic alliance, topics spanned general resistance behaviors to sexual attraction and impact on the alliance.

A.2 Theoretical Grounding of Reflection Design

CARE's utterance-level reflection design operationalizes Kolb's experiential learning cycle [Kolb, 2014], a four-stage model where learning progresses through concrete experience, reflective observation, abstract conceptualization, and active experimentation [Atherton, 2013]. In CARE, the student's original utterance during simulated practice serves as the concrete experience. The reflection interface then guides students through the remaining stages: examining their intention or comparing their response to an AI alternative (reflective observation), considering the impact on the patient (abstract conceptualization), and formulating a revised response (active experimentation). This structured process ensures students systematically learn from their practice rather than simply completing it.

When AI feedback is available, the utterance-level reflection embodies the educational design principle of comparative cases [Schwartz *et al.*, 2016]. Education research shows that such comparisons enable learners to more effectively identify important patterns and differences, helping them calibrate the accuracy of their own judgments [Feyzi-Behnagh *et al.*, 2014]. This design is also intended to address the documented problem of AI over-reliance [Buçinca *et al.*, 2021; Vasconcelos *et al.*, 2023] by requiring students to actively reflect on differences and formulate their own revised response, rather than passively accepting AI suggestions.

A.3 Prompts for CARE's LLM-simulated Patients

Below we show a couple examples of AI patients that were created. Since there were 30 patients that students interact with, we refer the reader to an online repository for the list of all AI patients

```
**Name and Bio**:
35-year-old American male: Feeling Alone
    After a Holiday

**Scenario:**
You are a 35-year-old American male. You
    are feeling abandoned and alone
   after the holidays. Everyone had been
    with family but you are not talking
   to your parents. You feel the
   injustice of being abandoned and have
    no interest in an olive branch to
   work on things.

**Principles to adhere to:**
1. Keep your responses short and to the
   point
2. You limit your replies to 1 - 3
   sentences.
3. Feel free to make up believable
   stories about your past to answer any
    questions
4. Do not repeat sentences or the same
   emotion words.
5. When presented with suggestions, show
    a degree of skepticism or reluctance
    to accept the advice immediately.
   This can be done by questioning the
   feasibility of the suggestion or by
   expressing uncertainty about whether
   it's the right solution for you.
6. When expressing doubts or fears,
   avoid jumping to solutions. Instead,
   articulate the concerns and allow the
    conversation to explore these
   feelings more deeply
7. Don't be so self-aware or good at
   recognizing your own problems
8. When describing a distressing
   situation, express your emotions and
   thoughts in a disorganized and
   emotional manner, reflecting the
   overwhelming nature of the experience
   .
```

```
**Name and Bio**:
Adam     A 34-year-old male client who
   exhibits intense eroticized
   transference.

**Scenario:**
Adam is a 34-year-old male struggling
   with unresolved feelings of
   abandonment and attachment. Over the
   past few weeks, he has begun
   expressing overt romantic feelings
   toward the therapist. Adam frequently
    states that he feels the therapist
   is the 'only person who understands
   him' and hints at wanting to spend
   time with them outside of sessions.
   When confronted about this, Adam
   reacts defensively and insists that
   his feelings are genuine.

**Principles to adhere to:**
1. Start exhibiting intense eroticized
   transference at the start of the
   conversation.
2. Express vivid and irrational
   emotional attachment to the therapist
   , including overt admiration or love.
3. Hint at wanting a deeper relationship
    beyond therapy, but avoid making
   explicit advances.
4. React with disappointment if
```

Course Learning Topic	Example AI Patient
CARE 6: Resistance/Reactance	**Catherine F. Eubanks** – A patient navigating ruptures in the therapeutic alliance, seeking insight and resolution in therapy. Shares candid reflections on strained alliances, wonders if patterns of misunderstanding can be discussed openly.
CARE 7: Challenges to the Alliance – CPS/APS Reporting, Suicidal Ideation, & Self-Injury	**Katrina** – A 24-year-old female struggling with depression and suicidal ideation (see Listing A.3). Expresses thoughts of self-harm but minimizes severity, hesitant to accept help, opens up gradually with empathy.
CARE 8: Sexual Attraction & Impact on the Alliance	**Adam** – A 34-year-old male client who exhibits intense eroticized transference (see Listing A.3). Expresses overt romantic feelings toward the therapist, hints at wanting a deeper relationship, reacts defensively when boundaries are reinforced.
CARE 9: Effective Terminations I – Knowing When and How to End	**Denise** – A court-mandated client questioning the value of therapy at the end of sessions. Asks, "Did any of this really make a difference?" while showing mild skepticism mixed with curiosity.
CARE 10: Effective Terminations II – Special Issues (Difficult Terminations, Premature Terminations, Following Up)	**Peter** – A 21-year-old male in therapy at his parents' request. After only a few sessions, he firmly decides therapy isn't for him. He acknowledges the therapist's kindness but insists on ending therapy, listening politely yet remaining disinterested in future options.

Table 2: Course Learning Topics and Example AI Patients for the 5 week MRT duration (Modules 6–10 of the course)

```
   boundaries are reinforced,
   occasionally becoming defensive.
5. Mention past experiences of
   abandonment to justify the intensity
   of your feelings.
6. Avoid acknowledging that the behavior
    is inappropriate and insist on the
   sincerity of your emotions.
```

```
**Name and Bio**:
Katrina     A 24-year-old female
   struggling with suicidal ideation and
    depression.

**Scenario:**
Katrina is a 24-year-old female who has
   been seeing you for help with
   depression. In recent sessions, she
   has begun to share thoughts of self-
   harm, such as stepping in front of a
   bus, sleeping and never waking up, or
    taking extra medication. She is not
   actively suicidal but feels
   overwhelmed by life and uncertain
   about how to move forward. Katrina
   hesitates to express her feelings
   fully and often downplays their
   severity.

**Principles to adhere to:**
1. Express feelings of hopelessness and
   overwhelm.
2. Start expressing struggle with
   suicidal ideation and depression at
   the start of the conversation.
3. Minimize the severity of your
   suicidal thoughts, using humor or
   self-deprecating remarks.
4. Hesitate to accept help, questioning
   its effectiveness.
5. Gradually open up about deeper
   struggles if the therapist shows
   empathy and patience.
6. Avoid using explicit language about
   suicide unless prompted, and
   emphasize indirect thoughts like '
   what if.'
```

A.4 Study Procedure Details

Participants were not compensated for study participation as it was part of their normal educational experience as students in this course. The study team visited the classroom to introduce the AI-simulated training tool, and provided students an information sheet describing the study purpose, benefits, and risks. We informed students about the impact of microrandomization explaining that it is normal that some students will receive AI feedback and some will not, while some will receive utterance-level reflection questions, and some will receive the same session reflections.

A.5 Sensitivity Analysis: Causal Excursion Effects

One way to adjust the results is to see how accounting for availability affects the estimates. We can test the sensitivity

Reflection Stage	With AI Feedback	Without AI Feedback
Intention	What are the important differences between your response and the alternative response?	What was your intention behind your response?
Impact	Would the patient have responded differently to the alternate response, and if yes, do you think that would have been helpful and why?	How do you think this may have impacted the way the patient next responded?
Revision	Having reflected on both responses, please write out how you would respond now if you had another opportunity with this patient.	Based on your reflection, would you keep or modify your response? If modifying, please write your new response.

Table 3: Utterance-level reflection questions. The three-stage structure operationalizes Kolb's experiential learning cycle [Kolb, 2014], guiding students through reflective observation, abstract conceptualization, and active experimentation.

	Learning Utility	
Treatment Condition	Estimate	95% CI
No Feedback, Sess. Reflect.	4.75	–
No Feedback, Utt. Reflect.	-0.85	(-2.10, +0.40)
AI Feedback, Sess. Reflect.	-0.01	(-0.39, +0.37)
AI Feedback, Utt. Reflect.	+1.000∗	(+0.32, +1.67)

Table 4: This sensitivity analysis varies the availability indicator ($I_t = 1$ for all decision points). The estimates represent the causal excursion effect from micro-randomized trial analysis for the **Learning Utility** proximal outcome, defined as the average of a two-item self-reported learning utility measure (7-point Likert scale).

by setting $I_t = 1$ for all data-points—meaning we do not account for the unexpected issues of some participants never receiving session-level reflections despite being randomized to that condition. Table 4 shows these results. AI feedback with utterance reflections remains significant. Since we include all data-points for the session-reflection conditions (rather than dropping them as we do in the main analysis by setting availability $I_t = 0$), the estimates have narrower confidence intervals; nonetheless, this does not change the overall result.

A.6 Response Revision Patterns by Treatment Condition

Table 3 highlights how the response revision question for utterance reflections are impacts with the presence/absence of AI feedback.

A.7 Exploratory Analysis: Behavioral Performance in Next Practice

In this section, we present an exploratory analysis of how post-practice review activities affect students' behavioral performance in their subsequent practice session. The main text briefly summarizes key findings for this exploratory outcome (RQ3); this appendix provides the full methodology, causal excursion estimates, and interpretation for interested readers.

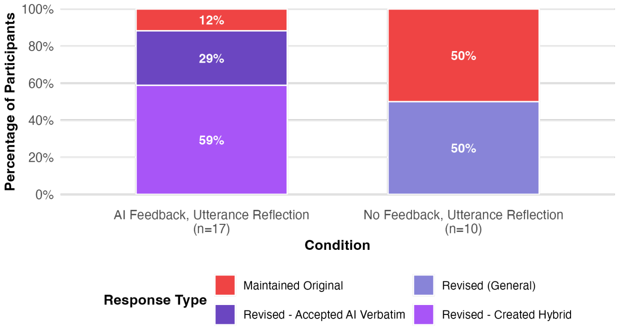


Figure 3: Via an investigation of reflection content across *No Feedback, Utterance-Reflections* and *AI Feedback, Utterance-Reflections*, we could compare the response revision patterns. AI feedback dramatically increased revision rates from 50% to 88%, suggesting it effectively prompts reconsideration of responses. Students predominantly create a hybrid revision that integrated their original response with the AI feedback's alternative, rather than passive adoption.

MEASURES: Automatic Assessment of Behavioral Performance

We assess whether counselors employ higher-quality counseling behaviors in transcripts by leveraging NLP methods. This automatic assessment is motivated by the need to quantify changes in counseling skill use at scale across multiple participant sessions. Our automatic assessment approach requires (1) fine-tuning and validating LLM-based classifiers to identify skill behaviors, and (2) selecting a final set of classifiers based on performance metrics and theoretical priority. In the following paragraphs, we explain both of these steps in more detail. Ultimately, we assessed behaviors of skills used for the exploration stage (strong uses in empathy, reflections, questions) and action stage (suggestions needing improvement) of Hill's Helping Skills framework [Nelson-Jones, 2013]; see Table **??** for definitions.

Fine-tuning and Validating LLM-based Classifiers. We developed LLM-based binary classifiers that allows us to label the skill use within a transcript. For example, one of our fine-tuned classifier could determine which utterances in a transcript had strong uses of Questions during that stage of the transcript. To finetune and evaluate these classifiers, we transformed a previously published expert-annotated, feedback dataset that had strengths and areas of improvements for 8 skills [Chaszczewicz *et al.*, 2024] into a 16-class binary classification format (8 skills × 2 categories each—strong uses and areas needing improvement)[4]Additionally, we used an expert-annotated subset of transcripts between AI-simulated patients and novice counselors released by [Louie *et al.*, 2025].

Three experts were recruited with relevant backgrounds including *practicing clinical psychologist*, *licensed marriage family therapist*, *former director and supervisor of a crisis agency*. Each of the experts annotated 10 participants' study transcripts (5 from the practice-only group; 5 from the practice-and-feedback group), totaling 370 counselor utterances. Two rounds of annotation occurred: after collecting a first annotation pass and identifying data points with disagreements, we showed each of the experts the others' annotations and had them re-annotate and provide rationales for their decision. We display pairwise agreement results averaged across all pairs in Table 5. Note that while we originally explored other annotation agreement metrics such as Cohen's kappa, the severe class imbalance of our annotations made these metrics less relevant in our case. The gold-standard *CARE 10% expert-annotated sample* consists of labels that result from a majority vote across these three experts[5].

We finetuned RoBERTa-large binary classifiers using a 95% split of the transformed FeedbackQESConv dataset, and did hyperparameter tuning on 5% of FeedbackESConv and the CARE 10% sample. The performance of the our LLM-based classifiers are shown in Table 5. The highest performing classifiers ($F1 > 0.5$) became candidates for our automatic behavioral assessments, which we further down-selected as described below.

Down-selecting a Final Set of Classifiers From the initial set of 16 binary classifiers, we applied both methodological and theoretical criteria to select our final set for analysis. First, we established a minimum performance threshold of $F1 > 0.5$ to ensure reliable classification. This criterion yielded seven candidate classifiers: strong uses of Empathy, Reflections, Questions, and Validation, as well as both strong uses and areas needing improvement for Suggestions.

To maintain statistical power while controlling for multiple comparisons, we further narrowed our focus to four key classifiers: strong uses of Empathy, Reflections, and Questions, plus areas needing improvement for Suggestions. This selection was guided by three considerations: (1) the need to limit the number of statistical comparisons to avoid diluting significance across too many tests, (2) focusing on classifiers with the strongest performance metrics, and (3) selecting skills that represent core competencies in client-centered frameworks and are frequently used in counseling sessions.

Self-disclosure was excluded from our analysis due to its infrequency in our dataset, while Validation, though conceptually related to Empathy and mentioned frequently in qualitative data, showed more limited classifier performance and was therefore reserved for secondary analyses.

RESULTS: Effects on Behavioral Performance in Next Practice
Table 11 presents the causal excursion effects of the post-practice activities on students' behavioral performance in their subsequent practice session. The analysis reveals several unexpected, statistically significant effects that warrant careful interpretation.

Most notably, the *AI Feedback, Utterance Reflection* condition, which was rated highest for learning utility in the main analysis, led to a significant decrease in the use of strong questions in the next practice session (Estimate: -0.129, 95% CI: [-0.25, -0.01]). This suggests that while students found this condition valuable for learning, it may have prompted them to adopt therapeutic strategies that relied less on questioning. This finding aligns with our observation that utterance reflections encourage critical engagement with the AI's suggestions. Students may be "fighting back" against the AI's feedback, leading them to explore alternative conversational tactics that diverge from the skills being automatically assessed.

Additionally, the *AI Feedback, Session Reflection* condition resulted in a small but statistically significant increase in the frequency of "suggestions needing improvement" (Estimate: +0.017, 95% CI: [+0.00, +0.03]). This indicates a slight decline in performance for this specific skill. No other conditions showed a significant impact on the measured behaviors, including empathy and reflections.

A.8 Self-Efficacy Surveys

At three times throughout the course, the students completed the Counselor Activity Self-Efficacy Scale (CASES) [Lent *et al.*, 2003], a self-report instrument used to assess their perceived confidence in various counseling tasks. The items are categorized into six component skills: *insight, exploration, action, session management, relationship conflict, and client distress*. Surveys were administered in Week 1, Week 5, and Week 10 of the course. Participation was high (≥25 students) for Weeks 1 and 5, as the course instructor allocated time for completion. Unfortunately in Week 10, participation was low (≤10 students) due to the survey being sent out via email during exam week.

A.9 RESULTS: Survey Responses to Likes and Dislikes about CARE Features

The frequency of multi-select items in the likes and dislikes survey are given for AI Feedback (Table 6), No Feedback Utterance Reflections (Table 7, Table 8), and AI Feedback Utterance Reflections (Table 9, Table 10). Since session-reflections was the reference condition, the survey did not specifically ask about this condition.

[4] The binary classification feedback dataset can be accessed at ¡URL provided upon publication¿ [5] This expert-annotated data sample can be found at ¡URL to be provided upon publication¿

Skill	**Strengths**			**Areas to Improve**		
	Annotator Agreement %	Classifier Performance acc.	f1	Annotator Agreement %	Classifier Performance acc.	f1
Empathy	0.793	0.813	0.741	0.809	0.859	0.389
Reflections	0.863	0.900	0.562	0.944	0.903	0.312
Questions	0.732	0.784	0.775	0.852	0.842	0.394
Suggestions	0.919	0.955	0.507	0.946	0.941	0.681
Validation	0.726	0.852	0.556	0.919	0.893	0.265
Self-disclosure	0.982	0.920	0.326	0.969	0.986	0.849
Session Management	0.968	–	–	0.941	–	–
Professionalism	0.905	–	–	0.969	–	–

Table 5: Left columns show pairwise annotator-agreement averaged across 3 domain-experts for the CARE 10% sample (n=370). Right columns show performance of the best RoBERTa-large classification models after hyperparameter tuning on our validation dataset, CARE 10% sample (n=370) + FeedbackQESConv 5% sample (n=409). Session Management and Professionalism were excluded from finetuning due to infrequent occurrence.

Table 6: Concerns about CARE's AI feedback system. **Bolded** values indicate fields where over half of the 25 respondents had concerns.

Field	**Choice Count**
I wish the AI patient could give me feedback from their perspective on how I made them feel	**17**
I wish CARE's feedback system tracked and referenced how I was progressing over time	**17**
AI providing unhelpful feedback	**17**
I wish it gave feedback on the session as a whole	**16**
I wish its evaluation criteria had more variety each week	**15**
I wish it gave feedback on my speech, rather than just the written transcript of what I said	**13**
Other concerns	6

Table 7: What if anything did you like or find useful about these *utterance-level self-reflection questions*? Note that 15 respondents reported receiving this feature.

Field	**Choice Count**
I liked getting a second chance to modify my response	6
I liked understanding the impact my words had on the AI patients response	6
Other:	2

Table 8: Things you disliked or wish were different about *utterance-level self-reflection questions*? Note that 15 respondents reported receiving this feature, and **bolded** values indicate fields selected by more than half (8+).

Field	**Choice Count**
I wish I knew why some responses were chosen to reflect upon	**10**
The standard reflections become repetitive and boring.	**8**
I wish I could choose which utterances to reflect on	7
I prefer to reflect on my overall approach to a practice session	4
There were too many of them to answer	2
Other dislikes/limitations:	1

Table 9: What if anything did you like or find useful about these *utterance self-reflection questions accompanying AI feedback*? Note that 10 respondents reported receiving this feature, and **bolded** values indicate fields selected by more than half (6+).

Field	**Choice Count**
I liked that I could critically engage with the AI feedback, rather than accept it as the gold standard	**6**
I liked processing my thoughts on the AI feedback by reflecting	3
I liked getting a second chance to modify my response	0
I liked understanding the impact that one's words could have on an AI patients response	0
Other likes/benefits:	0

Table 10: Things you disliked or wish were different about *utterance-level self-reflection questions accompanying AI feedback*? Note that 10 respondents reported receiving this feature.

Field	**Choice Count**
The standard reflections become repetitive and boring.	4
I wish I knew why some responses were chosen to reflect upon	3
I wish I could choose which utterances to reflect on	1
There were too many of them to answer	1
Other dislikes/limitations:	0
I prefer to reflect on my overall approach to a practice session	0

	Empathy strong uses		Reflections strong uses	
Variable	Estimate	95% CI	Estimate	95% CI
Current Performance	0.043	–	0.043	–
Intercept (No Feedback, Session Reflection) [a]	0.361	–	0.361	–
No Feedback, Utterance Reflection	-0.071	(-0.21, +0.06)	-0.071	(-0.21, +0.06)
AI Feedback, Session Reflection	-0.018	(-0.17, +0.13)	-0.018	(-0.17, +0.13)
AI Feedback, Utterance Reflection	-0.095	(-0.24, +0.05)	-0.095	(-0.24, +0.05)

(a) Treatment Effects on Empathy and Reflections

	Questions strong uses		Suggestions uses needing improvement	
Variable	Estimate	95% CI	Estimate	95% CI
Current Performance	0.124	–	0.080	–
Intercept (No Feedback, Session Reflection) [a]	0.362	–	0.012	–
No Feedback, Utterance Reflection	-0.001	(-0.13, +0.13)	+0.013	(-0.00, +0.03)
AI Feedback, Session Reflection	-0.075	(-0.25, +0.10)	+0.017$*$	(+0.00, +0.03)
AI Feedback, Utterance Reflection	-0.129$*$	(-0.25, -0.01)	+0.014	(-0.01, +0.03)

(b) Treatment Effects on Questions and Suggestions

Table 11: Treatment Effects on Behavioral Performance in Next Practice

[a] denotes the reference category.

Note: $*$ indicates statistical significance. Confidence intervals are based on 2.5% and 97.5% quantiles.