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

Ryan Louie RYLOUIE@CS.STANFORD.EDU
Stanford University

Ellen Converse ELLENCON@STANFORD.EDU
Stanford University and Palo Alto University

Diyi Yang DIYIY@STANFORD.EDU
Stanford University

Emma Brunskill EBRUN@CS.STANFORD.EDU
Stanford University

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-simulated practice tool for use in psychotherapy classrooms that provides speaking-practice with LLM-based patients, followed by a post-practice activity that includes feedback from a fine-tuned LLM and written reflection exercises. We deployed the tool 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.

Keywords: Micro-randomized trial, AI-simulated practice, psychotherapy education, LLM, deliberate practice

1. Introduction

LLM-simulated roleplay practice is an AI education approach to training professional social skills (e.g., interpersonal skills Lin et al. (2024), and conflict resolution Shaikh et al. (2024)), with recent lab studies Louie et al. (2024); Wang et al. (2024); Steenstra et al. (2025) demonstrating how AI systems can provide realistic practice for training novices for mental health counseling skills. An important challenge when moving these tools from the lab to real-world classrooms is evaluating and improving their pedagogical utility within authentic classroom use. Few studies to date have deployed LLM-simulated roleplay practice in classrooms Thesen et al. (2025), and it is uncommon for such studies to evaluate the impact of design choices in real-world educational settings.

Across a two-quarter course sequence, we deployed an AI-augmented practice tool for therapeutic skills which provides utterance-level AI feedback on transcripts of roleplay practice sessions. We found that students were not passive recipients of AI feedback—they critically evaluated suggestions and debated merits of their responses versus AI alternatives. These formative findings during

the initial half of our deployment provided valuable insights into the design space of post-practice activities, and motivated a continued experiment (N=25 students, 5-weeks) where we experimentally compared different post-practice activities. These activities guide psychotherapy 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).

Our experiment employs micro-randomized trials (MRTs) Klasnja et al. (2015)—a method commonly used to evaluate mobile health interventions—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. Students in the AI feedback with utterance-reflection condition reported the greatest learning utility (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 mentioned integrating AI suggestions rather than passively accepting them, providing evidence that our reflection design, inspired by cognitive forcing functions Buçinca et al. (2021) and Kolb (2014) experiential learning theory, can be used to scaffold critical reflection to support learning with AI feedback.

In the remainder of this paper, we present the CARE platform that was deployed to a graduate psychotherapy course; the micro-randomized trial design for evaluating the impact of post-practice review activities; and detailed results of our micro-randomized trial. Finally, we discuss the implications for how novel experimental methods can be used during formative and continual educational deployments to evaluate and improve the pedagogical utility of AI educational approaches like LLM-simulated training systems.

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). Our study deployed a roleplay practice tool in a graduate psychotherapy course, to understand what design factors impact the learning experience of students using reviewing their roleplay practice sessions, either with AI-simulated partners or human peers.

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

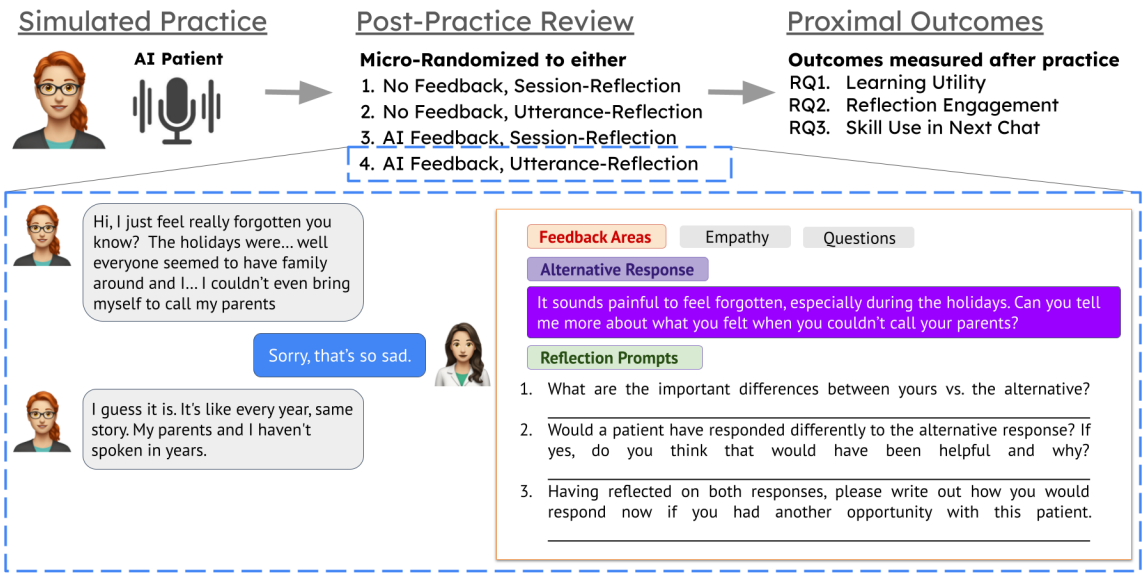

Figure 1: CARE is AI-based psychotherapy training system that provides speaking-practice with LLM-simulated patients, followed by post-practice review activities which vary in AI feedback (present/absent) and granularity of reflection (session-level/utterance-level). To test what type of review activity 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).

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. AI-Simulated Practice Tool for Psychotherapy Classrooms

Our classroom deployment of CARE, an AI-system for psychotherapy skills training, is a first step at translating previous lab-based findings on LLM-simulated practice and feedback into real-world educational setting Louie et al. (2025, 2024); Chaszczewicz et al. (2024). Our tool was offered as 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.

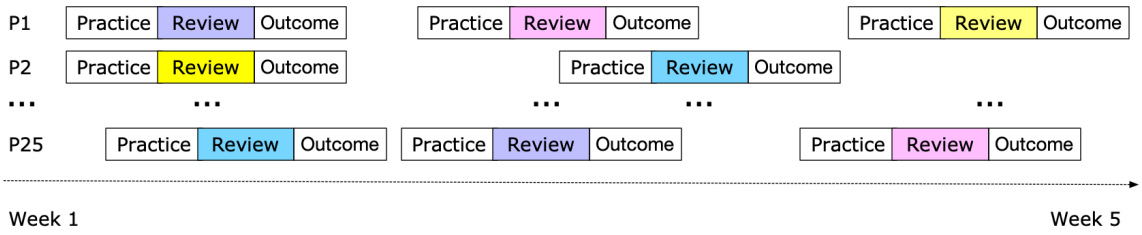

Figure 2: Students (n=25) practiced with AI-simulated patients; after each *student-initiated* session, the system micro-randomized one of four post-practice review activities: No Feedback, Session Reflection, No Feedback, Utterance Reflection, AI Feedback, Session Reflection, AI Feedback, Utterance Reflection. Because learners choose when to practice, decision points (and spacing between them) are irregular; the main text below summarizes measures and analysis.

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. In total, 30 voice-based LLM-simulated patients were deployed for the course's 10 week curriculum. The appendix provides more details on the creation of the voice-based LLM-simulated patients and the OpenAI Real-time API implementation of the patients.

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 ran a Micro-randomized Trial (MRT) Klasnja et al. (2015) in a psychotherapy class: treatment is randomized at many time points per participant, not once at enrollment. After each session, the system assigned one of four post-practice conditions with equal probability, independent of history (Figure 3).

Factorial, categorical MRT. The classroom required a 2×2 over components (AI feedback vs. no feedback $\times$ session- vs. utterance-level reflection), not a single intervention vs. control. We used uniform 25% randomization to each of four arms at every decision point and report contrasts on proximal outcomes with the *causal excursion effect* (CEE) for categorical treatments Lin and Qian (2025); Boruvka et al. (2018), summarizing *immediate, marginal* effects relative to a reference level. Full potential-outcomes definition, centering, and the weighted least-squares implementation we used are in Appendix A.8; immediate outcomes (vs. delayed $t+\Delta$) are standard in MRT practice Lin and Qian (2025).

Student-initiated (pull) sessions. In contrast to many mobile-health MRTs with clocked or push notifications, each decision point occurred when a student *started* a practice session, so the number and timing of randomizations differed by student. Inter-session gaps in our data were empirically irregular; we analyzed all person-decision points (Appendix A.5), so effects average over the engagement patterns that arose naturally, rather than a fixed weekly pull schedule. A software issue in session-reflection conditions meant some practices did not display the prompt; for those times we set availability $I_{i,t} = 0$ in estimation (Appendix A.6; see Table 5 for a sensitivity that sets $I_t{=}1$ everywhere).

RQs. MRTs compare intervention *components* rather than confirming a single package. We ask: *RQ1* perceived learning from review; *RQ2* reflection behavior; *RQ3* skill use in the next session; and *RQ4* course-end views of the review variants.

Measures. **Perceived learning utility** (RQ1) was the mean of two 7-point items administered immediately after review (item wording, Appendix A.7). **Reflection engagement** (RQ2) was the log of total reflection text length, normalized by the number of prompts in that review type. **Skill use** (RQ3) used fine-tuned RoBERTa classifiers for four high-frequency course skills Louie et al. (2025) (Appendix A.11). **End-course perceptions** (RQ4) used multi-select and free text answers.

Analyses. We estimated separate CEEs for the three continuous proximal outcomes, using all decision points as above. We analyzed utterance-level reflections in depth (session-level reflection completion was low); we report revision patterns, visualized in Figure 3 (Appendix A.10), and a thematic read of RQ4.

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.

	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 $p < 0.05$.

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 14 (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 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 microskills 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 the first course teaching micro-skills. 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 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 classroom deployment evaluated through an MRT provided causal evidence for what design factors impacted learning utility—and revealed unexpected mechanisms. Utterance-level reflections with side-by-side comparison received nine-fold more engagement than session-level alternatives. Students used comparison to create integrative responses rather than accepting AI suggestions wholesale. Designs that structure analystical reflection about AI feedback will be important given growing AI over-reliance concerns Zhai et al. (2024). Survey data revealed 47% of participants wished they could choose which utterances to reflect on. This suggests that learners may want more control over the reflection process, and that the system should provide more flexibility in the reflection activities. Future deployments and randomized experiments could test learner-initited reflections compared to system or mixed-initited

reflections. Pilot observations revealed some students initially perceived AI feedback as authoritative. Our comparison-based design deliberately reframed AI suggestions as one perspective among many, and integrative revision patterns confirmed this reframing succeeded. 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. The constraints of small-classroom deployments—limited N, ongoing coursework, naturalistic engagement patterns—demand evaluation methods beyond traditional RCT designs. We borrowed micro-randomized trials from mobile health, where similar constraints had already been solved. With a median of 7 randomizations per participant over 5 weeks, we detected significant effects ($p<0.05$) despite only 25 students. This suggests a broader opportunity: AI education researchers should actively scout adjacent fields for methodological innovations suited to transitioning from lab-based to deployment-based research that maintains experimental rigor. The key insight is that ecological validity and rigorous evaluation need not be opposed—methods exist to achieve both, but they may not already used in education research or other mixed-method design-based research studies.

7. Limitations and Future Work

Our results are context-bound: our study with 25 doctoral psychotherapy trainees should be followed by new studies with non-doctoral students, that validate behavioral outcomes against independent human-expert raters, and that estimate the effect of feedback and utterance-reflection in other social skill training domains. Our study had a compromised reference arm where the software bug affected 48% of decision points. While we detect statistically-significant effects, our estimates of the utility still have wide confidence intervals. The deployed system used utterance-level feedback and is thus incapable of assessing session- and topic-level skills (e.g., confidentiality, alliance repair). Future work should validate feedback and rubrics with independent human-raters, separate chats used for informal practice from chats used as periodic standard assessments, and align models to specific curricular units 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

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

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.

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.

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.

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.

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. URL https://doi.org/10.1038/s41539-024-00289-9.

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.

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.

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.

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.

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. ISBN 9798400705052. doi: 10.1145/3640457.3691715. URL https://doi.org/10.1145/3640457.3691715.

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. URL https://link.springer.com/article/10.1007/s11251-013-9275-4.

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.

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.

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.

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.

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.

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

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.

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.

Inna Wanyin Lin, Ashish Sharma, Christopher Michael Rytting, Adam S Miner, Jina Suh, and Tim Althoff. Imbue: improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. *arXiv preprint arXiv:2402.12556*, 2024. doi: 10.48550/arXiv.2402.12556.

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

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.

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.

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.

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

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.

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. URL https://doi.org/10.1093/biomet/asaa070.

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.

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

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.

Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642159. URL https://doi.org/10.1145/3613904.3642159.

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.

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. ISBN 9798400716188. doi: 10.1145/3636555.3636853. URL https://doi.org/10.1145/3636555.3636853.

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.

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. ISSN 2156-8650. doi: 10.1007/s40670-025-02473-x. URL https://doi.org/10.1007/s40670-025-02473-x.

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. doi: 10.1145/3579605. URL https://doi.org/10.1145/3579605.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, et al. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*, 2024. doi: https://doi.org/10.48550/arXiv.2405.19660.

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: nonrandomized controlled trial. *JMIR medical education*, 10(1):e58753, 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.

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. ISBN 9798400710582. doi: 10.1145/3675094.3678373. URL https://doi.org/10.1145/3675094.3678373.

Appendix A. Appendix

A.1. Creating Voice-based LLM Patient for Course Topics

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 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 Appendix Section 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).

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

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)

```
**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
```

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.

```
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 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 micro-randomization 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. MRT: Student-initiated practice and which decision points we analyze

Unlike many micro-randomized trials in mobile behavioral health, where decision points are often *push* or clocked, each decision in our class occurred when a student *started* a new practice session (*pull*). Learners chose when and how often to practice, so gaps between person-decision points were irregular. The distribution of time between points was bimodal, with mass near one day and near seven days—roughly matching same-day multiple sessions and roughly weekly work; the number of sessions in the MRT window varied (e.g., from one to fourteen).

One may instead restrict the analysis to well-spaced decision points (e.g., at least 5–7 days apart) to down-weight frequent practicers and align with a weekly course cadence. **Our primary analysis uses all person-decision points,** producing causal excursion effects that average over the schedules and session counts that arose naturally. One reading is: *"How do the four review activities compare in proximal outcomes when averaging over how often and when each student actually practiced?"*

A.6. MRT: Availability, session-level reflection prompts, and the $I_{i,t}$ indicator

A software issue affected 48% of decision points in *both* Session Reflection arms, where participants who were micro-randomized to those conditions did not always see the session-reflection interface. In CEE estimation we set availability to $I_{i,t} = 0$ for such decision points, reflecting that the assigned review was not *delivered*. Section A.9 reports a sensitivity analysis that instead sets $I_t = 1$ at those rows.

A.7. MRT: Wording and construction of proximal measures

Perceived learning utility (RQ1). Measured immediately after the review activity as the unweighted mean of two 7-point items ("Strongly disagree" to "Strongly agree"): *"This practice session was valuable for my development as a therapist"* and *"I gained useful insights from this practice experience"*.

Reflection engagement (RQ2). The metric was the natural logarithm of total written reflection length for that review, divided by the number of reflection prompts (Session Reflection: two prompts; Utterance Reflection: $3 \times N_{\text{utterances}}$). We examined engagement under the hypothesis that students may write more when they value an activity, while treating length as an imperfect proxy for depth.

End-course views of review variants (RQ4). The end-course survey asked for likes and dislikes for the utterance-level review conditions, with multi-select options and an "Other" free response; it did not emphasize the session-level reflection conditions because in-product completion of session-level reflection was low.

A.8. MRT: Causal Excursion Effect for Categorical Treatments

This section provides an overview of the causal excursion effect methodology for micro-randomized trials with categorical treatments, following the framework developed by Lin and Qian (2025), including the notation, estimand, and weighted-centered least squares implementation. It builds on the potential outcomes and excursion definitions in Boruvka et al. (2018); Rubin (1974). The full CEE formulation can condition on context for heterogeneity, which we do not pursue here. We use the *immediate* (rather than $t+\Delta$ delayed) proximal-outcome case that matches our measures and the most common MRT reporting practice Lin and Qian (2025). We implement the estimators in the open-source R package of Lin and Qian (2025).[1] For delayed outcomes, see Lin and Qian (2025).

A.8.1. Notation and Setup

Consider a micro-randomized trial with n individuals, each observed for T decision points where treatments are randomized.

Symbol	Definition
A_t	Treatment assignment at decision point t (categorical: $A_t \in \{0, 1, \ldots, K\}$ where 0 is reference)
X_t	Contextual information collected between decision points $t-1$ and t
I_t	Availability indicator at time t (if $I_t = 0$, then $A_t = 0$ deterministically)
H_t	History information up to time t: $H_t = (\bar{X}_t, \bar{A}_{t-1})$
$p_t(k\|H_t)$	Randomization probability: $\Pr(A_t = k \mid H_t)$ for $k \in \{0, 1, \ldots, K\}$
$Y_{t,\Delta}$	Proximal outcome following treatment at t, measured over window of length Δ
S_t	Effect modifiers of interest at time t (subset of H_t)

Table 4: Basic notation for micro-randomized trials with categorical treatments

A.8.2. Causal Excursion Effect (CEE) Definition

The causal excursion effect compares potential outcomes under two different "excursions" from the trial's treatment protocol. For treatment level k at time t, Lin and Qian (2025) define the CEE as:

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

This definition captures the contrast between:

- **Excursion 1**: Receiving treatment k at time t, then no treatment for the next $\Delta - 1$ periods

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

- **Excursion 2**: Receiving no treatment at time t, then no treatment for the next $\Delta - 1$ periods

Both excursions condition on the observed treatment history up to time $t-1$ and on being available at time t.

A.8.3. GENERAL ESTIMATOR FOR CATEGORICAL TREATMENTS

Following Lin and Qian (2025), under the parametric model $\text{CEE}_{tk}(S_t) = f_t(S_t)^T \beta_k$, the estimator $\hat{\beta} = (\hat{\beta}_1^T, \ldots, \hat{\beta}_K^T)^T$ is obtained by solving $P_n m(\alpha, \beta) = 0$, where:

$$m(\alpha, \beta) := \sum_t I_t J_t \left\{ Y_{t,\Delta} - \sum_{k=1}^{K} C_k(A_t) f_t(S_t)^T \beta_k - g_t(H_t)^T \alpha \right\} \begin{pmatrix} g_t(H_t) \\ C_1(A_t) f_t(S_t) \\ \vdots \\ C_K(A_t) f_t(S_t) \end{pmatrix} \tag{2}$$

The key components are:

- **Weight**: $J_t := \frac{\tilde{p}_t(A_t|S_t)}{p_t(A_t|H_t)} \prod_{j=t+1}^{t+\Delta-1} \frac{\mathbf{1}(A_j=0)}{p_j(0|H_j)}$
- **Centering**: $C_k(A_t) := \mathbf{1}(A_t = k) - \tilde{p}_t(k|S_t)$
- **Control model**: $g_t(H_t)^T \alpha$ for $(Y_{t,\Delta}|H_t, I_t = 1)$

The centering term $C_k(A_t)$ ensures robustness: the estimator remains consistent even when the outcome regression model is misspecified, a crucial property given the high-dimensional nature of the history information H_t.

A.8.4. FOCUS ON IMMEDIATE, MARGINAL EFFECTS

In our analysis, we focus on the **immediate, marginal** causal excursion effect with:

1. **Immediate effect**: $\Delta = 1$ (outcome measured immediately after treatment). Focusing on the immediate effect is reasonable for RQ1 because the proximal outcome for learning utility is measured immediately after the post-practice review treatment. Similarly, the immediate proximal outcome for reflection engagement (RQ2) corresponds to the length of a students reflection for the current practice, rather than for future practices. Finally, the immediate proximal outcome for behavioral performance (RQ3) is for the next practice following the post-practice review.
2. **Marginal effect**: $S_t = \emptyset$ (averaging over all effect modifiers). In this MRT study, we average over all effect modifiers, and do not explore how this effect is moderated by other time-varying covariates.

This leads to the **Marginal Excursion Effect (MEE)**:

$$\text{MEE}_k(t) := \{Y_t(\bar{A}_{t-1}, k) \mid I_t(\bar{A}_{t-1}) = 1\} - \{Y_t(\bar{A}_{t-1}, 0) \mid I_t(\bar{A}_{t-1}) = 1\} \tag{3}$$

where $Y_t := Y_{t,1}$ represents the proximal outcome measured immediately following the treatment decision at time t.

A.8.5. SIMPLIFIED ESTIMATOR FOR OUR SETTING

Under constant randomization probabilities $p_t(k) = P(A_t = k|I_t = 1)$ and the immediate, marginal effect assumptions, the estimating equation simplifies to:

$$m(\alpha, \beta) := \sum_t I_t \left\{ Y_t - \sum_{k=1}^{K} C_k(A_t) f_t^T \beta_k - g_t^T \alpha \right\} \begin{pmatrix} g_t \\ C_1(A_t) f_t \\ \vdots \\ C_K(A_t) f_t \end{pmatrix} \tag{4}$$

where:

- $C_k(A_t) = \mathbf{1}(A_t = k) - p_t(k)$ (simplified centering)
- f_t and g_t are functions of time t only
- The complex weighting reduces to unity due to $\Delta = 1$ and constant randomization

A.8.6. PARAMETRIC WORKING MODEL

For practical implementation, we assume:

$$\text{MEE}_k(t) = f_t^T \beta_k \quad \text{for } t \in [T], k \in [K] \tag{5}$$

Common choices include:

- **Constant effect**: $f_t = 1 \Rightarrow \text{MEE}_k(t) = \beta_k$ (effect constant over time)
- **Linear trend**: $f_t = (1, t)^T \Rightarrow \text{MEE}_k(t) = \beta_{k0} + \beta_{k1} \cdot t$ (linear change over time)
- **Quadratic trend**: $f_t = (1, t, t^2)^T$ (allowing for non-linear patterns)

The choice of f_t should reflect scientific understanding of how treatment effects might evolve throughout the study period, with simpler models (constant effects) often preferred for interpretability and statistical power.

This framework extends micro-randomized treatment comparisons to intensively collected longitudinal data from *push* mHealth or *pull* learning tools such as our classroom deployment.

A.9. 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 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 5 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.

	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 5: 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).

A.10. 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.

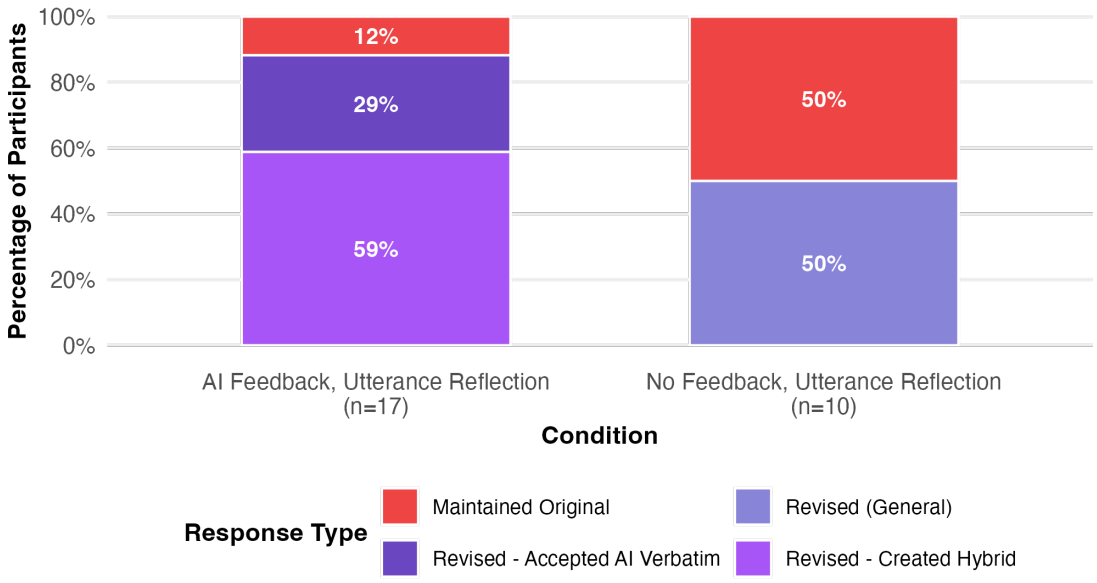

Figure 3: Via an investigation of reflection content across *No Feedback, Utterance-Reflections* and *AI Feedback, Utterance-Reflections*, we could compare the response revision patterns. AI feedback dramatically increased revision rates from 50% to 88%, suggesting it effectively prompts reconsideration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

A.11. 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.

A.11.1. 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)[2]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 6. 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[3].

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 6. The highest performing classifiers ($F1 > 0.5$) became candidates for our automatic behavioral assessments, which we further down-selected as described below.

[2] The binary classification feedback dataset can be accessed at ¡URL provided upon publication¿ [3] 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 6: 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.

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.

A.11.2. RESULTS: EFFECTS ON BEHAVIORAL PERFORMANCE IN NEXT PRACTICE

Table 14 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.12. 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 ($\geq$25 students) for Weeks 1 and 5, as the course instructor allocated time for completion. Unfortunately in Week 10, participation was low ($\leq$10 students) due to the survey being sent out via email during exam week.

A.13. 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 7), No Feedback Utterance Reflections (Table 8, Table 9), and AI Feedback Utterance Reflections (Table 10, Table 11). Since session-reflections was the reference condition, the survey did not specifically ask about this condition.

Table 7: 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 8: 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 9: 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 10: 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 11: 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

(*a*)

	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)

Table 12: Treatment Effects on Empathy and Reflections

(*a*)

	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)

Table 13: Treatment Effects on Questions and Suggestions

Table 14: 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.