AI-Augmented Psychotherapy Education: 16-Week Classroom Deployment Compared With Two Traditional Practice Methods

Ryan Louie PhD
Department of Computer Science, Stanford University
rylouie@stanford.edu
ORCID: 0000-0001-7266-3688

Ellen Converse MS
Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine; Palo Alto University
ellencon@stanford.edu
ORCID: 0009-0000-6583-6544

Debra L Safer MD
Professor, Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine
dlsafer@stanford.edu
ORCID: 0000-0002-9874-2277

Juan Pablo Pacheco
Department of Computer Science, Stanford University
School of Engineering, Stanford University
Department of Psychiatry and Behavioral Sciences, Stanford School of Medicine
pacheco7@stanford.edu
ORCID: 0009-0009-9949-9174

William Fang
Department of Computer Science, Stanford University
wfang03@stanford.edu
ORCID: 0009-0001-8150-0386

Jamie Kent PhD
Associate Professor
Co-Associate Director of Clinical Training, PGSP-Stanford Psy.D. Consortium
Practicum Co-Coordinator, PGSP-Stanford PsyD Consortium
jkent@paloaltou.edu

Bruce Arnow PhD
Professor and Associate Chair

Co-Chief Adult Division of Psychiatry and Clinical Psychology
Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine
arnow@stanford.edu
ORCID: 0000-0003-1645-857X

Diyi Yang PhD
Department of Computer Science, Stanford University
diyiy@stanford.edu
ORCID: 0000-0003-1220-3983

Abstract

Background: Efforts to enhance psychotherapy education may increasingly emphasize technology-assisted methods that extend practice beyond human supervision. Recent advances in artificial intelligence (AI) have enabled the automated assessment of therapeutic skills and the simulation of patient interactions. However, few studies have examined the feasibility, acceptability, and instructional integration of AI-augmented training tools in real-world training settings.

Objective: The primary objective of this study was to evaluate the feasibility and perceived value of deploying an AI psychotherapy training platform (CARE) within a doctoral-level psychotherapy course sequence. Secondarily, we explore implementation barriers, ethical and pedagogical considerations, and factors influencing students' acceptance and engagement with AI-assisted learning.

Methods: Participants were 29 first-year clinical psychology doctoral students enrolled in a two-quarter introductory psychotherapy sequence at an APA-accredited clinical training program. In the first course, we developed an AI-augmented peer-roleplay assignment that supplemented peer feedback with AI feedback after each session. In the second course, students practiced with traditional video vignettes and voice-based AI-simulated patient scenarios, with AI feedback provided afterwards. Two web-based surveys—containing Likert-type and free-response items—were completed by 9 students (31%) in the first course and 25 (86%) in the second. Analyses included descriptive statistics, matched-pairs hypothesis testing, and thematic analysis, with memos of classroom interactions used to contextualize the data.

Results: In the first course, many students were initially hesitant to use AI feedback due to privacy concerns; hence, only 9 students opted to use AI feedback. While this group valued AI feedback for offering alternative phrasing and immediate suggestions, peer feedback was rated more helpful and nuanced than AI; AI was limited by transcription errors and a narrow focus on empathy

statements. In the second course, students opting to participate (n = 25) were assigned weekly structured video vignette practice, alongside voice-based AI-simulated patients to compare modalities. Students preferred video-based practice for realism and emotional expressiveness, though many noted that AI-simulated patients enabled interactive, back-and-forth practice not possible in video-vignette practice. Instructor feedback was added in the second half of the course in response to concerns that AI feedback focused too narrowly on microskills, rather than promoting therapeutic alliance during the session. Students expressed mixed views on the perceived value of AI in clinical training—while some endorsed its utility, most also raised concerns related to human relational dynamics being supplanted, non-consensual use of their data for improving AI, and the environmental impacts of the generative AI industry writ large.

Conclusions: Results indicate that implementation was feasible; however, students consistently preferred human feedback over AI-generated feedback and favored video-based vignettes over voice-based AI interactions. Students also reported concerns about the quality of feedback, data transparency, and broader ethical issues.

Findings suggest that perceptions of AI's instructional utility are contingent on the degree of trust students place in the specific AI program and on more general views of AI impact. Accordingly, future deployments must consider the sample's attitudes toward AI, trust in the technology, and value relative to existing training resources before making general claims about AI's utility in psychotherapy education.

Keywords: artificial intelligence, psychotherapy, deliberate practice, classroom deployment, patient simulations

Introduction

Psychotherapy is a structured, clinical intervention wherein trained professionals employ established psychological methods, primarily through dialogue and relational engagement, to alleviate distress, reduce psychiatric symptoms, and promote psychological functioning and well-being [1, 2]. Psychotherapy is a collaborative process in which the therapeutic alliance, defined as the emotional bond between the therapist and client, as well as agreement on the goals of therapy and the tasks required to achieve the therapeutic goals [3], is consistently associated with clinical improvement [4].

Therapeutic communication skills cannot be fully developed through lectures or reading alone; instead, they are cultivated through experiential practice, structured feedback, and critical reflection. Accordingly, traditional

psychotherapy training emphasizes didactic instruction, structured, instructor-led teaching methods such as lectures, seminars, or presentations that convey foundational theories and practical guidelines, alongside supervised clinical practice, role-play, and reflective exercises. Collectively, these methods facilitate real-time interaction and iterative refinement of core therapeutic skills [5].

Training psychotherapists, therefore, presents a distinctive pedagogical challenge: students must master not only theoretical and technical knowledge but also deeply interpersonal competencies, including empathic attunement, alliance-building, and moment-to-moment responsiveness. A small yet steadily growing evidence base suggests that Deliberate Practice (DP) is an effective learning method that can outperform standard supervision and training [6]. Unlike didactic instruction, deliberate practice focuses on structured and intentional skill development through activities that include clearly defined goals, repeated practice, systematic opportunities for reflection, and real-time feedback from a supervisor or coach.

A common form of DP in graduate psychotherapy training contexts consists of a real-time session with a peer or clinical supervisor roleplaying a simulated scenario, and/or providing feedback following a rehearsal or therapy recording. Empirical evidence supports the effectiveness of DP in enhancing these skills. For example, learners who engaged in DP showed greater improvements in empathic expression compared to those who were trained using didactic methods alone [7]. Despite its potential benefits, the widespread implementation of DP in clinical psychology training programs is limited by the significant time and expertise required for 1-on-1, real-time supervision and personalized feedback [6].

Technological Innovations in Psychotherapy Training

Researchers have sought to enhance the accessibility and efficiency of psychotherapy training by developing technology-enabled platforms. Early efforts focused on structured tools for DP, such as video recording systems, which facilitate asynchronous supervision and reflective practice by enabling both students and instructors to review therapist–client interactions outside of real-time sessions [8]. Similarly, video-response platforms like Skillsetter [9] offer structured opportunities for practice by presenting short patient vignettes and requiring trainees to record therapeutic responses to those vignettes, which can later be reviewed for feedback by peers or supervisors. These innovations have expanded opportunities for repeated skill rehearsal, supplementing traditional classroom and supervisory contexts.

Recent advances in large language models (LLMs) have further broadened this landscape, particularly through the development of AI-simulated patients that allow counselors to engage in practice dialogues within safe,

controlled environments [10, 11]. Parallel work has advanced AI feedback systems capable of automatically assessing microskills such as empathy, active listening, and motivational interviewing [14, 15, 16, 17, 18], while also generating alternative responses [8, 19, 20] and explanatory rationales [21, 222].

Within the emerging literature, AI-augmented psychotherapy training systems show both potential and methodological limitations. Recent experimental lab studies suggest that such simulations, when paired with feedback, may promote the acquisition of client-centered counseling behaviors [12, 13]. Avatar-based simulated patients can effectively support motivational interviewing practice, garnering high levels of usability and satisfaction among both student and professional counselors [23]. More recently, AI-simulated patients have been assigned within clinical education settings such as psychotherapy [24, 25] and medical education [26]. However, even the most promising implementations fall short of assessing how trainees engage with AI over extended periods alongside traditional coursework, supervision, and peer practice. Consequently, questions remain about how trainees engage with AI tools when integrated into coursework as part of ongoing real-world classroom settings rather than deployed as standalone exercises.

The Current Investigation

Addressing the limitations of prior work, this study seeks to advance the understanding of how AI-augmented DP functions when directly embedded into psychotherapy training. As reviewed above, existing research on AI tools has primarily been restricted to controlled laboratory settings, short intervention windows, or isolated evaluations conducted outside of the classroom context. Such studies cannot fully address how AI interacts with real pedagogical structures, classroom dynamics, or established training methods such as peer roleplay, supervised feedback, and video vignette exercises. Moreover, it remains unclear whether AI practice tools operate as meaningful complements to these traditional approaches, whether they inadvertently displace more valuable training modalities, or whether their integration reshapes learning in unanticipated ways.

To address these gaps, the present investigation examined students' experiences with CARE, an AI-augmented deliberate practice platform, when deployed across two academic quarters within a doctoral-level psychotherapy curriculum. CARE was integrated alongside existing training methods—including peer roleplays and the Skillsetter video-response platform—to enable structured comparison of AI-based approaches with established educational tools. The primary objective of the study was to evaluate the feasibility and perceived value

of implementing CARE in this classroom context. Secondarily, we explore implementation barriers, ethical and pedagogical considerations, and factors influencing students' acceptance and engagement with AI-assisted learning. Guided by these objectives, the study addressed four research questions (RQs):
RQ1: Comparative Value — From the perspective of students, a) what features of AI-based practice and feedback tools are perceived as distinct, and b) how do these compare with traditional counseling education methods?
RQ2: Usability and Technical Limitations— What factors about the AI-augmented prototype might enhance the pedagogical effectiveness of CARE within psychotherapy training?
RQ3: Course Integration Influence on Acceptability — How is utility and acceptability influenced by how AI is embedded into the psychotherapy classroom, including issues of onboarding, communication, and strategies to preserve instructor-student and peer-to-peer interaction within the deployment?
RQ4: Student Beliefs about Acceptance — a) How do students perceive the legitimacy and acceptability of AI in psychotherapy education, and b) how has their exposure to AI in general influenced their views on its potential role in clinical training and practice?

Through its design, the study provides the first longitudinal, classroom-based evaluation of an AI-augmented deliberate practice tool in psychotherapy education. By situating AI within authentic curricular structures rather than laboratory simulations, this work offers an ecologically valid assessment of both the pedagogical potential and the limitations of AI integration, thereby informing future directions for technology-mediated training in the mental health professions.

Ethical Considerations

The protocol for this psychotherapy classroom deployment study was reviewed by our Institution's Review Board and was assessed as Non-Medical Exempt (Protocol ID: 72546). This research was conducted as part of normal educational practice in an introductory psychotherapy course in a clinical psychology PsyD program at Palo Alto University. AI-augmented practice components were added to standard weekly assignments. The study included mid-class and end-class surveys to assess students' perceptions of AI feedback compared to peer feedback. Students received an information sheet explaining that while participation in course activities was expected as part of normal educational practice, they could opt out of having their data included in research analysis by emailing the protocol director. The instructor had no knowledge of which students opted out, ensuring no coercion or impact on grades. No

compensation was provided as the AI-augmented assignments and surveys were integrated into normal course requirements.

To protect privacy and confidentiality, all data transmission was encrypted via HTTPS and stored on password-protected AWS servers accessible only to investigators listed on the protocol. Survey responses were linked to de-identified participant IDs through Qualtrics. The research team committed that practice transcripts would not be publicly released or used to train autonomous AI therapy chatbots. Audio recordings and transcripts would be stored for up to 3 years before deletion. Only de-identified survey data, aggregated usage statistics, and feedback ratings could be presented at scientific meetings or published in journals. The AI feedback system underwent safety validations prior to deployment, including guard models to filter inappropriate content.

Methods

Participants

Participants were 29 first-year doctoral students enrolled in a highly selective PsyD program in clinical psychology at a university on the western coast of the United States. Of these participants, 4 students elected not to engage with CARE at any point during the study. The remaining 25 students participated in at least some component of the CARE activities and comprise the analytic sample reported here.

Students entered with heterogeneous training backgrounds: some held master's degrees and had accumulated significant supervised clinical hours, whereas others entered with bachelor's degrees supplemented by applied fieldwork or human services experience. This requirement for prior clinical experience differentiates PsyD training from many master's mental health programs and some PhD programs, which often emphasize research over early clinical immersion. As a result, these participants reflect a relatively advanced and motivated sample of trainees, with a strong commitment to developing clinical competencies at the outset of their doctoral training.

Demographic data (e.g., age, gender, ethnicity) were not collected, which limits analysis of subgroup variation. However, the advanced training context and selectivity of this PsyD program provide a valuable context for interpreting the findings and underscore the relevance of studying AI-augmented training in a population of highly prepared clinical psychology trainees.

Pedagogical Context and Course Goals

The study was embedded within a two-quarter sequence forming part of the core first-year PsyD curriculum. This sequence is designed to provide foundational training in psychotherapy competencies, with a structured

progression from introductory relational skills to more advanced therapeutic processes. The study was conducted over 16 weeks, spanning two consecutive courses (September 2024–March 2025)

The first course study emphasized pan-theoretical elements of psychotherapy, including basic listening skills, therapeutic alliance formation, therapist verbal response models and intentions, and consideration of cultural and ethical factors. The second course quarter built upon these foundations by targeting pan-theoretical competencies outlined by the APA Division 29 Task Force on Psychotherapy Competence, such as alliance maintenance (task, goal, and bond), rupture repair, transference and countertransference, and termination processes.

Historical Structure of the Learning Psychotherapy Course Series

In prior cohorts, the first course utilized weekly peer roleplays. Students alternated therapist and client roles in 10-minute sessions, followed by mutual feedback and written self-reflections. Confidentiality agreements were required, and each student submitted recorded sessions that were periodically reviewed in small group meetings with instructors.

In the second course, prior cohorts engaged with Skillsetter's video-based DP platform. Assignments required demonstration of specific therapeutic skills (e.g., empathy, cultural humility, alliance repair). Instructors or teaching assistants provided feedback on a weekly basis, and the quarter culminated in a live roleplay with an instructor or teaching assistant portraying a client, requiring integration of multiple therapeutic competencies. This established format of peer roleplay in the first course and Skillsetter video-based DP in the second course provided both relationally grounded and structured opportunities for iterative skill refinement.

Skillsetter Platform

Skillsetter is a web-based, DP video-based training platform designed to support deliberate practice in interpersonal and counseling skills [27]. It was developed to overcome logistical barriers commonly encountered in traditional psychotherapeutic training, such as the resource-intensive nature of instructor supervision and inconsistent feedback, by allowing learners to practice asynchronously, record responses to simulated prompts, and receive structured feedback. When using the platform, students submitted video-recorded responses to standardized patient vignettes performed by actors. Because feedback, rather than mere repetition, is central to DP, Skillsetter emphasizes scaffolded evaluation, either via instructor raters or automated scoring across domains.

The platform has achieved broad adoption across counseling, social work, and psychotherapy training programs, lending evidence of its acceptability and institutional integration. In the Learning Psychotherapy course series, feedback on Skillsetter practice was given by the course instructor and teaching assistants. In applying it to this study, Skillsetter served as an established, scalable platform for implementing component-level practice (e.g., empathy, reflections, open questions) in a remote environment. This enabled learners to engage in deliberate rehearsal outside class hours while receiving targeted feedback.

Integration of CARE

Building on this structure, the current study introduced CARE, an AI-powered simulation and feedback platform, to examine its potential role alongside traditional DP. Our design maintained the integrity of the existing pedagogy while creating opportunities for systematic comparison.

In the first course pilot study in 2024, the instructional team prioritized interpersonal skill-building and peer connection, continuing to rely on peer role-play as the primary experiential method. CARE was introduced as an optional feedback tool, allowing students to upload recorded peer sessions and receive automated feedback on the use of eight counseling skills (e.g., empathy, validation, reflections, questions, suggestions, self-disclosure, session structure, professionalism) derived from the Helping Skills and Motivational Interviewing model [29, 30]. This integration provided an opportunity to explore the additive value of AI-generated feedback while preserving the emphasis on human relational practice in the early stages of training.

In the second course (CLIN 702), CARE was deployed more comprehensively in parallel with Skillsetter to enable structured comparison of two distinct DP modalities. Students alternated between Skillsetter's single-turn vignette responses and CARE's multi-turn, dynamic dialogues with AI-simulated patients. Both platforms were aligned with weekly curricular objectives, matched for duration and credit, and graded using consistent rubrics. This dual-modality design created an ecologically valid setting for evaluating relative affordances and limitations of AI-augmented practice versus established methods.

CARE AI System

CARE was developed as a web-based training platform designed to support novice counselors in practicing counseling skills within simulated environments powered by large language models (LLMs). The platform enables trainees to engage in structured, text-based conversations with AI-simulated patients and to receive automated feedback on their therapeutic responses, generated by an AI "mentor." CARE builds on prior work in co-design with mental

health experts to enhance the realism of LLM-simulated patients [8, 10, 11, 26] and on research advancing the fine-tuning of domain-specific LLMs to incorporate therapeutic knowledge capable of producing both evaluative feedback and alternative counselor responses [6, 21, 31, 32, 33].

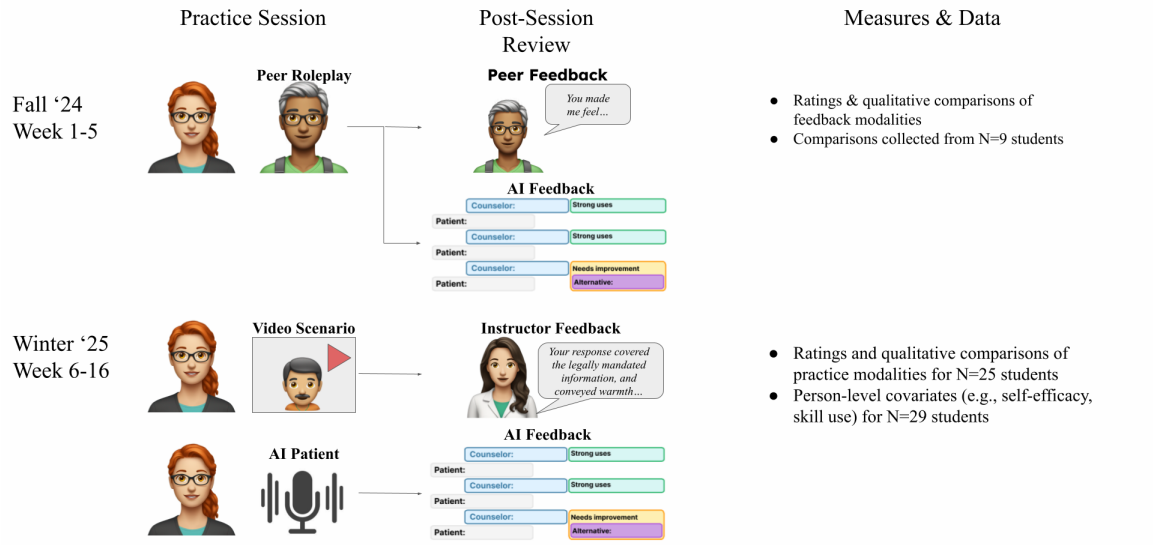


Figure 1. *Overview of the integration of CARE's practice, feedback, and data collection into the two courses. During the first course (Weeks 1–5), students engaged in peer role-play practice sessions followed by post-session review using either peer-provided or AI-generated feedback, with ratings and qualitative comparisons collected (N = 9). During the second course (Weeks 6–16), students completed practice using video-based scenarios and AI-simulated patients, followed by instructor or AI-generated feedback, with ratings, qualitative comparisons, and person-level covariates (e.g., self-efficacy, skill use) collected (n = 25). The figure illustrates the progression of practice modalities, feedback sources, and data collection across the two courses.*

From a technical standpoint, CARE was implemented as a web application with a Python Flask back end and a React JavaScript front end, accessible to students through a secure interface. AI-simulated patients were powered by the GPT-4o Realtime API, with scenario-specific prompts defining patient background, conversational style, and behavioral principles [10]. This infrastructure enabled integration into classroom deployment studies, allowing systematic investigation of CARE as a pedagogical supplement to established deliberate practice methods.

Within CARE, practice is organized around a library of patient scenarios, each defined by a brief background vignette (e.g., "Young adult with family issues: low mood and self-esteem") These intentionally limited descriptions are designed to approximate authentic counseling encounters, where the clinician

must simultaneously build rapport, elicit client context, and apply therapeutic skills (See Figure 3.) For example, a novice peer counselor might initiate a practice session with an AI patient modeled as a 35-year-old veteran struggling to reconnect with his children amid legal and interpersonal challenges. The dialogue proceeds turn by turn, requiring the trainee to apply core counseling skills such as empathy, reflection, and problem exploration, while managing relational complexity as the conversation unfolds.

The platform provided students with AI-generated feedback on each of their utterances at the end of the session. CARE's feedback implements an existing method for generating feedback that fine-tunes a Llama-2 13B parameter model on an expert-annotated dataset covering psychotherapy-relevant skill assessment [21]. Feedback focuses on strengths, areas for improvement, feedback explanation, and a suggested alternative response that addresses any issues.

Integration of CARE into the Learning Psychotherapy Course Series

The Learning Psychotherapy sequence spanned two quarters, sixteen weeks total, and was structured to provide developmentally sequenced training in core psychotherapy competencies. The first course implementation spanned six weeks, whereas the second course implementation extended across ten weeks. The study team began the first course in week four of the first course instead of week one to allow more time for finalizing implementation procedures.

The first course 2024 pilot, described in more detail below, served as an initial, formative phase of the study, designed to compare the educational utility of AI-generated feedback with traditional peer feedback following role-play exercises. This phase was crucial in identifying themes that shaped the subsequent second course study, including the relative value of different feedback modalities (RQ1), early design limitations (RQ2), challenges of classroom integration (RQ3), and students' broader attitudes about the role of AI in psychotherapy training and daily life (RQ4)

First Course: Foundational Skills and Peer Roleplay

The first course provides a foundation for training beginning psychotherapists in the crucial elements of psychotherapy. Across Weeks 1-7, instruction progressed via the gradual introduction of the core listening skills and verbal response modes, with particular attention in Week 4 to how sexual, gender, and racial diversity can impact the therapeutic alliance. Weeks 8–9 introduced principles and strategies from Motivational Interviewing, with the course concluding in Week 10 with a focus on ethics, professional boundaries, and therapist self-care.

Students engaged in 10-minute peer roleplays each week, alternating between therapist and client roles, with assigned therapeutic skills tailored to that week's topic. Recordings of these peer roleplays were uploaded to CARE, which generated automated feedback across eight core counseling skill domains derived from the Helping Skills and Motivational Interviewing models. This process enabled the instructional team to evaluate the educational value of AI-generated feedback as a supplement to traditional experiential exercises, while maintaining the central role of in-person peer practice in early clinical training.

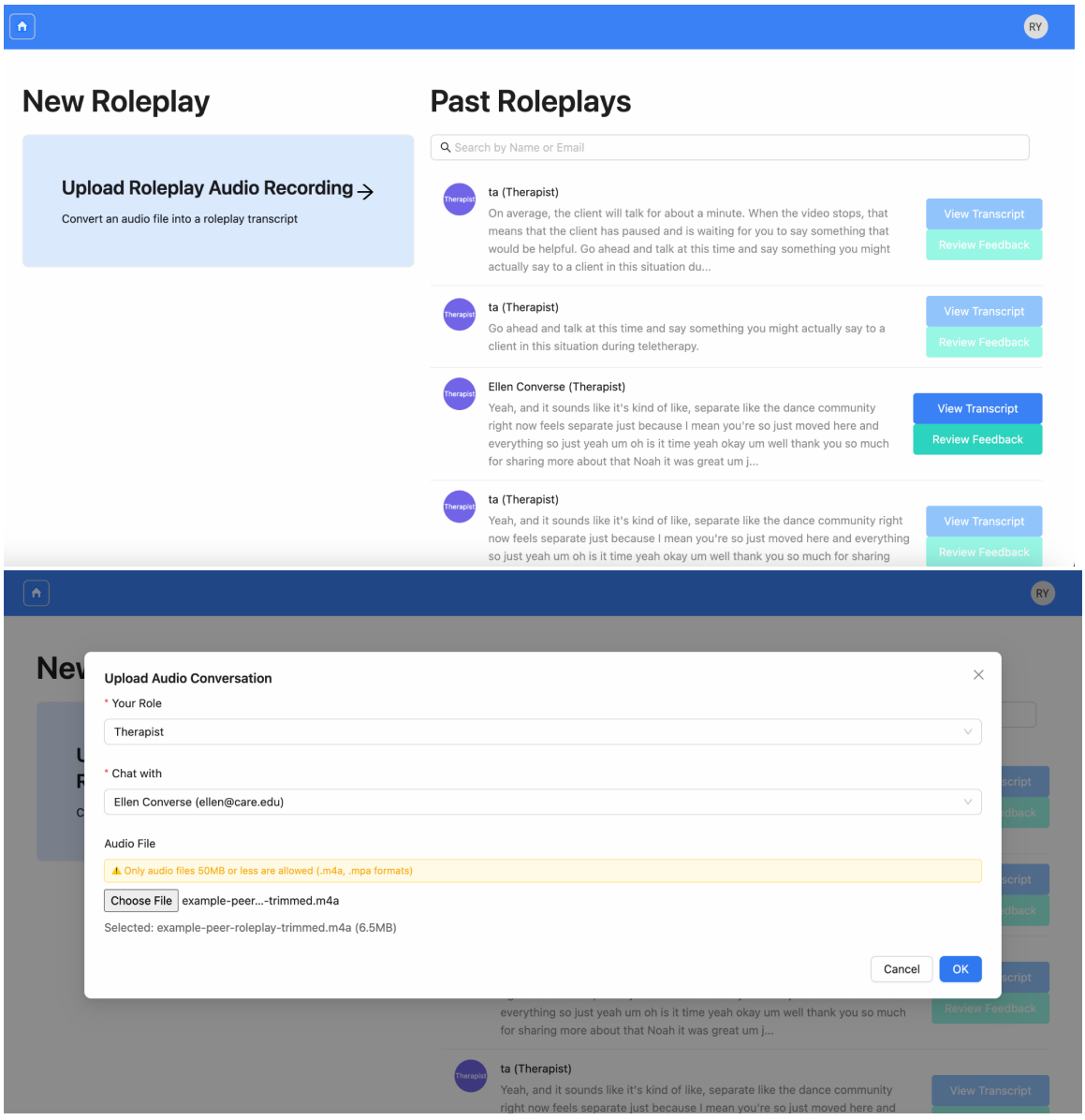


Figure 2. *In the first course, students recorded peer-to-peer roleplays where they acted as both therapists and patients for their peers. Students uploaded their roleplay audio files to CARE, which transcribes and diarizes the audio into speaker turns and generates feedback on their utterances.*

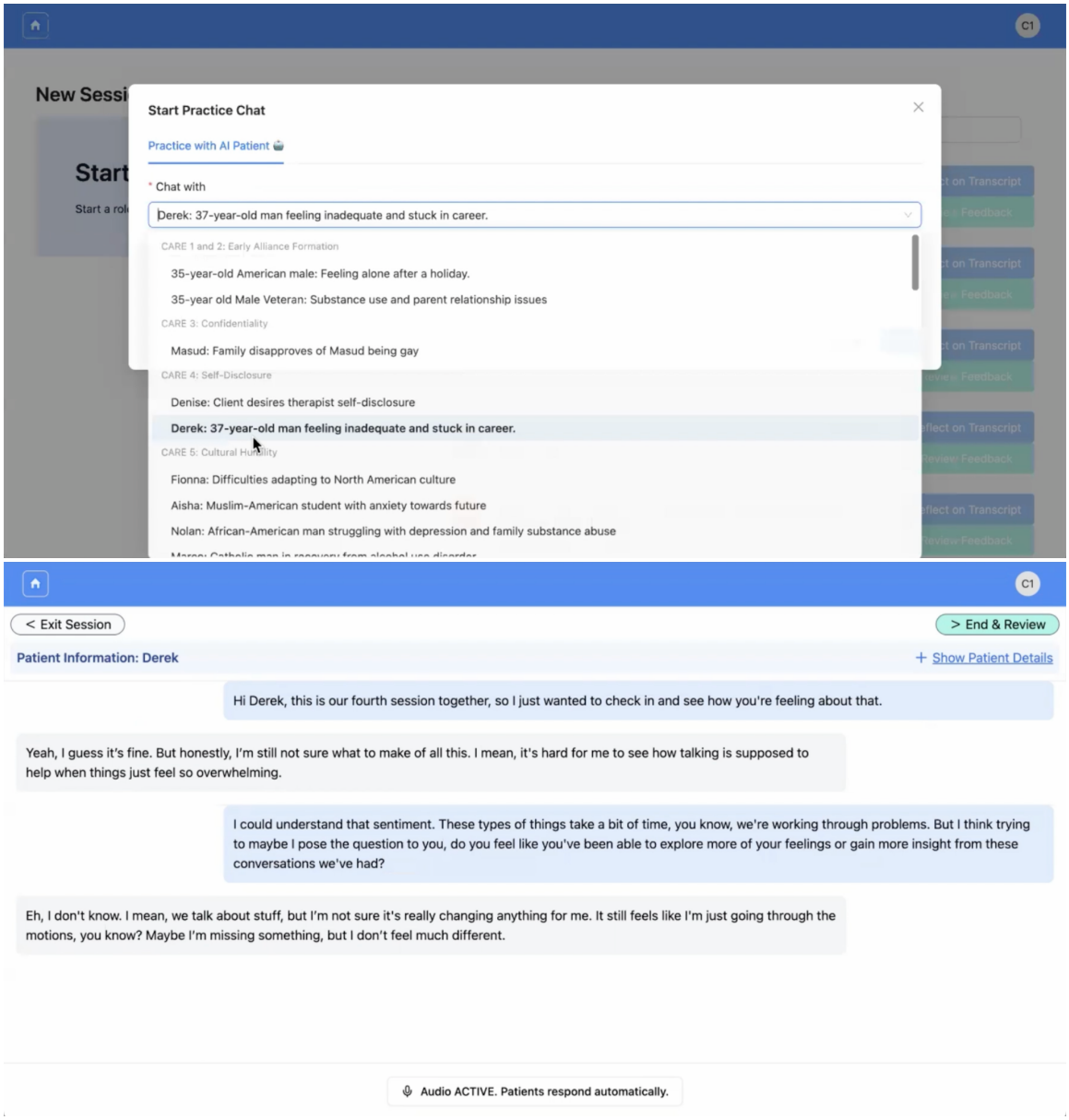
Second Course: Comparative Deployment of Skillsetter and CARE

The second course introduced students to the principles of DP, values-based care, and basic therapeutic skills. Instruction progressed from the concept of deliberate practice (Week 1) to the therapeutic alliance, drawing on Bordin's (1979) task–goal–bond framework (Week 2) [28], before moving to therapist self-disclosure, cultural humility, and the influence of power and identity within therapeutic relationships (Weeks 3–4.) The middle phase emphasized countertransference, resistance (Week 5), alliance ruptures and repair (Week 6), and clinical challenges such as mandated reporting and suicide risk (Week 7.) The sequence concluded with advanced relational concerns, including sexual attraction and boundary maintenance (Week 8.)

The course was designed to directly compare two distinct DP modalities: Skillsetter and CARE. This parallel design created a side-by-side comparison of vignette-based practice and conversational AI simulations, with both modalities aligned to weekly curricular goals and equivalent in duration, evaluation criteria, and grading weight.

To integrate CARE into the curriculum, the study team systematically mapped AI patient simulations onto weekly course topics at Week 4, enabling students to engage with simulated clinical encounters that corresponded directly with concepts under instruction. Students practiced with twelve AI-generated patient scenarios across the quarter, nine of which were adapted from Skillsetter transcripts (75%) and three adapted from documented examples of sexually inappropriate behaviors toward clinicians reported in prior empirical work (25%; see Table 3) [34]. CARE's scenarios encompassed alliance ruptures, mandated reporting, and crisis management related to suicidal ideation and self-injury, thereby anchoring AI practice in clinically relevant challenges.

Prior to classroom deployment, CARE's simulated patients underwent an iterative review process conducted by a TA to ensure clinical appropriateness, ethical safeguards, and pedagogical alignment. Each AI patient was programmed to follow structured behavioral principles, with particular emphasis placed on scenarios involving emotionally sensitive content areas (e.g., trauma, self-harm) to verify that AI-generated responses were trauma-informed, ethically sound, and supportive of therapeutic learning objectives. Additional realism was introduced through varied voice models designed to convey emotional depth and natural conversational flow. Some identified limitations, including emotional flatness in voice-based responses, were addressed by adding principles to the voice-based AI patient's prompt (e.g., "Mention past attempts to please family that failed, with resigned sighs, and even breaths where the nose is runny as in crying")


C1
New Sessi
Start Practice Chat
Practice with AI Patient
* Chat with
Derek: 37-year-old man feeling inadequate and stuck in career.
CARE 1 and 2: Early Alliance Formation
35-year-old American male: Feeling alone after a holiday.
35-year old Male Veteran: Substance use and parent relationship issues
CARE 3: Confidentiality
Masud: Family disapproves of Masud being gay
CARE 4: Self-Disclosure
Denise: Client desires therapist self-disclosure
Derek: 37-year-old man feeling inadequate and stuck in career.
CARE 5: Cultural Humility
Fionna: Difficulties adapting to North American culture
Aisha: Muslim-American student with anxiety towards future
Nolan: African-American man struggling with depression and family substance abuse
C1
< Exit Session
> End & Review
Patient Information: Derek
+ Show Patient Details
Hi Derek, this is our fourth session together, so I just wanted to check in and see how you're feeling about that.
Yeah, I guess it's fine. But honestly, I'm still not sure what to make of all this. I mean, it's hard for me to see how talking is supposed to help when things just feel so overwhelming.
I could understand that sentiment. These types of things take a bit of time, you know, we're working through problems. But I think trying to maybe I pose the question to you, do you feel like you've been able to explore more of your feelings or gain more insight from these conversations we've had?
Eh, I don't know. I mean, we talk about stuff, but I'm not sure it's really changing anything for me. It still feels like I'm just going through the motions, you know? Maybe I'm missing something, but I don't feel much different.
Audio ACTIVE. Patients respond automatically.

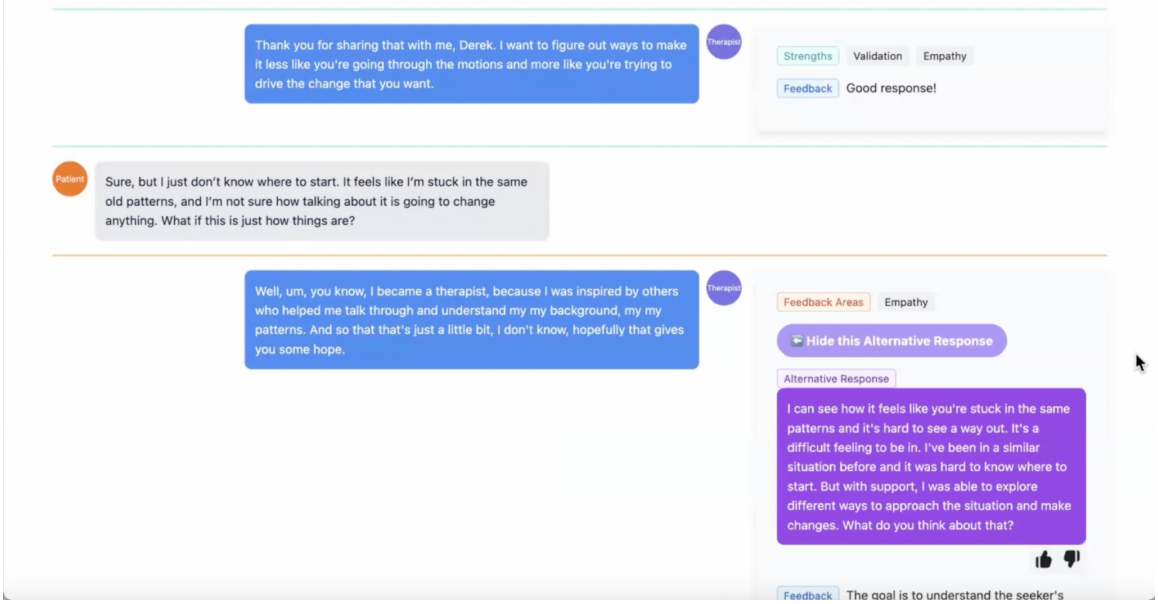


Figure 3: *In the second course, CARE provided students weekly practice options with AI patients that reinforced course learning goals (e.g., therapist self-disclosure; cultural humility) and generated feedback after their practice sessions on each of the utterances.*

Measurements

Table 1. Example survey items from our web-based survey administered to students in the second psychotherapy course during the second course.

Domain	Example Items
Comparative Value (RQ1)	"The way I could use [CARE/Skillsetter] allowed me to make mistakes I could learn from." "After conversing with the patients in [CARE/Skillsetter], I feel more equipped to build and maintain a therapeutic alliance."
AI Feedback (RQ1)	"Receiving AI feedback shortly after completing CARE practice enhanced my learning experience." "I know when to ignore AI feedback when it is incorrect or does not make sense."
Self-Reflection	"I found the self-reflections after a [CARE/Skillsetter] session valuable for understanding how to improve."

Limitations & Design (RQ2)	"My concerns about CARE AI's patients include: voices sound robotic; transcripts are inaccurate." "My concerns about CARE AI's feedback include: wish it tracked progress over time."
Course Integration (RQ3)	"Would you recommend CARE AI to next year's PsyD cohort as a deliberate practice option?"
Attitudes about AI (RQ4)	"Overall, would you say the increased use of AI in daily life makes you feel… (concerned vs. excited)?" "Do you think AI in the classroom would lead to better or worse helping skills for therapists?"

Surveys

First Course

Students completed a mid-course survey in the first course, halfway through the introduction of CARE, and again at the end of the course. The 37-item questionnaire was specifically designed by the investigators for this study to gather insights into the challenges and obstacles students faced while using CARE, as well as their experiences with peer feedback and AI feedback on their peer roleplays. The full set of 37 items for the first course study is provided in the supplementary materials.

Second Course

To evaluate the comparative value of CARE and Skillsetter, students completed a 39-item investigator-designed survey at the end of Week 16 in the second course. Items were adapted from national and professional surveys where possible (e.g., Pew Research Center, 2023; American Medical Association, 2023). Students rated their experiences across domains using 5-point Likert scales (1 = strongly disagree, 5 = strongly agree) and provided open-ended responses. The full set of 37 items for the second course is provided in the supplementary materials.

Data Analyses

We conducted non-parametric paired hypothesis tests to evaluate students' comparative perceptions of CARE and Skillsetter (RQ1). We report descriptive statistics for the remaining quantitative survey items to highlight the frequency of perceived limitations and design needs of CARE (RQ2), perceptions of CARE's course integration (RQ3), and attitudes towards AI (RQ4)

A thematic analysis was conducted to examine students' open-ended responses related to each research question. Two authors independently reviewed the qualitative data to familiarize themselves with the content. Following several rounds of coding, comparison, and discussion, a finalized coding

framework was established and applied to all responses. The qualitative findings were integrated into the presentation of results to complement the quantitative analyses.

Results

This section outlines the findings from the 16-week deployment, structured to provide a clear progression of insights. We begin with the first course 2024 pilot study, which served as a preliminary investigation into AI versus peer feedback. We then move to the more extensive second course study, which builds upon the first course to explore AI patient simulations in greater depth. This ordering reflects the evolving nature of the research and the increasing complexity of the questions addressed.

First Course 2024 Pilot: AI vs. Peer Feedback

The first course was a formative phase of the study in which only 9 of the 30 students in the course opted in to using CARE. Students in the first course compared the educational value of AI-generated feedback against traditional peer feedback following role-play exercises.

RQ1: AI Feedback vs. Peer Roleplay.

Quantitative comparisons from the first course pilot (Table 2) show that students rated peer feedback significantly higher than AI feedback in terms of being constructive and helpful (AI feedback from CARE rated to a moderate extent; peer feedback rated to a great extent)

Qualitatively, students valued peer feedback for its ability to capture non-verbal cues and provide a client-centered perspective. One student noted, “I think peer feedback considers the whole session, including tone, nonverbals, use of silence, etc., so it feels more relevant as a whole.”

Table 2. Comparing peer roleplay practice followed by AI feedback (CARE) vs. Peer Feedback (Peer Partner) rated by students in the first course (n=9)

Question	Modality	Mean	Median	SD	Effect Size (*d*)	IQR	p
Constructive and helpful	CARE	2	2	0.707	1.155	0	0.0244
	Peer Partner	3	3	0.707	1.155	0	

Comfort receiving feedback	CARE	5.444	6	0.726	0.612	1	0.1025
	Peer Partner	5.889	6	0.333	0.612	0	
Recognize areas to improve.	CARE	4.333	4	0.866	0.843	1	0.0461
	Peer Partner	5.222	5	0.833	0.843	1	

1 p-values come from Wilcoxon signed-rank tests comparing CARE and Peer Partner responses. Constructive and Helpful was rated on a 0 - 4 scale: 0=Not at all, 1=To a small extent, 2=To a moderate extent, 3=To a great extent, 4=To a very great extent. Comfort receiving feedback and recognizing areas to improve was rated on a 1 - 6 forced Likert scale: 1=Disagree Strongly, 2=Disagree Moderately, 3=Disagree Slightly, 4=Agree Slightly, 5=Agree Moderately, 6=Agree Strongly.

RQ2: Design Limitations.

Students identified both technical and substantive limitations in CARE's design. Transcription errors, often linked to audio quality, sometimes undermined the utility of feedback: "There are occasional issues with the transcript… CARE is unable to identify or differentiate between patient and client at times, resulting in occasionally inaccurate feedback." Beyond technical issues, participants described the scope of feedback as narrow, with repeated emphasis on empathy and validation while overlooking other therapeutic strategies. As one noted, "Most of the feedback was suggesting that I empathize with and validate the client, but there are many other techniques… which the AI doesn't seem to recognize." Others emphasized that responses were evaluated in isolation: "It feels like each response is being considered separate and apart from the rest, so it feels less relevant to the content overall."

Despite these concerns, students highlighted the value of alternative phrasing and immediate suggestions. One remarked, "I like the alternative responses… it gives me ideas for how to better and more concisely word my responses." Another added, "I appreciate how content-oriented AI is! I find the alternative responses extremely helpful as I can use certain phrases in future

sessions." While some described the system as "robotic" or "less personal," others appreciated its consistency, contrasting it with the subjectivity of human feedback: "Human feedback might only apply for that individual client… I think AI gives broader and more objective feedback that applies to the situation instead of the person."

RQ3: Course Integration.

The integration of CARE into the curriculum was shaped by instructional stance and student perceptions of the research team. In the first course, CARE was introduced under exploratory conditions, which encouraged discussion. Use of CARE's role-play transcript feedback was made optional in response to pronounced student concerns, and its role was described as exploratory and loosely integrated. Of the 29 students enrolled in the first course, 14 completed surveys about the integration of CARE into the course, and 9 provided comparative ratings about peer roleplay versus CARE.

When asked what prevented them from using CARE during the course, students most frequently cited concerns about privacy and data security, a preference for receiving feedback from a human rather than AI, uncertainty or lack of guidance on how to use the tool, and apprehension about the nature of the research being conducted by the research team (e.g., "making a chatbot therapist.")

RQ4: Acceptance.

A key concern raised by students in the first course was the security and intended use of their video transcripts when uploaded to the CARE platform. To address these concerns, the study team provided detailed explanations during class and in the Frequently Asked Questions (FAQ) document (see Supplementary Materials). Students were informed that all video transcripts would be stored on a secure, password-protected cloud server. In rare cases where transcription errors occurred, a member of the research team would access the server to identify and troubleshoot the affected audio file. Aside from this limited quality control function, students were assured their audio files would not be used for any other purpose and would be deleted upon the completion of research analyses. Similarly, text transcripts were to be stored on a secure, password-protected server within CARE's internal database. These transcripts were accessible only to the student and their selected peer partner. If students opted to use the "Review Feedback" tool, their transcripts were processed by CARE's AI feedback model. Importantly, CARE's model is entirely self-hosted and does not rely on commercial AI services (e.g., ChatGPT), ensuring that student data is not used to train external proprietary models.

In addition to these measures, students were provided with a formal study information sheet outlining all data protections and ethical commitments associated with the project.

Second Course: Deliberate Practice

Building on the first course, the second course study shifted focus to the use of CARE's AI-simulated patients for deliberate practice, comparing this to Skillsetter. CARE enabled dynamic, back-and-forth dialogue with simulated patients, whereas Skillsetter required a single, one-time response to a video prompt.

Students had already been introduced to the CARE platform in the prior term. The difference for the second course was that students could utilize CARE for DP, speaking with AI patient simulations. Students could also receive AI feedback on the basic counseling skills used in their interactions with the AI patients; however, the focus was on the utility of the deliberate practice itself rather than the feedback.

RQ1: Comparative Value of AI Patients vs. Video Vignettes

As shown in Table 3, students rated Skillsetter significantly higher across several dimensions, including learning from mistakes, immersion, and practicing communication beyond words. However, there was no significant difference in students' confidence in responding spontaneously, suggesting both tools were comparable in that regard.

Qualitatively, students appreciated that CARE's AI patients provided a "good opportunity to practice using the skills learned in class" and allowed for interactive dialogue, which is not possible with Skillsetter. One student noted they used the AI-augmented patient simulation because "you don't get to explore the back and forth with Skillsetter". CARE was also valued for simulating difficult scenarios that would be hard to practice with peers, such as dealing with an "angry client, silent client." Despite these benefits, a strong preference for practicing with humans remained. Students felt that AI practice could not replicate the nuances of human interaction, with one stating, "Nothing replaces role plays." There were also concerns that skills developed with the AI might not transfer to real therapy, with one student worrying they had "learned how to orient towards receiving positive feedback from CARE rather than towards providing thoughtful and nuanced support for a human person."

Table 3. Perceived value of challenging patient scenario practice with traditional video vignettes and instructor feedback (Skillsetter) compared to voice-based AI patients and AI feedback (CARE) rated by students in the second course (N=25)

Question	Modality	Mean	Median	SD	Cohen's d	IQR	P-Value
Learning from mistakes	CARE	3.2	3	2.041	0.488	4	0.0257
	Skillsetter	4.52	5	2.002		3	
Immersion in interaction	CARE	3.88	4	2.048	0.101	3	0.6578
	Skillsetter	4.16	4	1.972		4	
Communication beyond words	CARE	4.2	4	1.958	-0.126	3	0.5822
	Skillsetter	3.8	4	2.021		3	
Confidence responding spontaneously	CARE	4.32	4	1.819	0.075	3	0.8035
	Skillsetter	4.52	5	2.104		5	
Building a therapeutic alliance	CARE	4.52	5	1.851	-0.289	3	0.1298
	Skillsetter	3.8	4	2.255		4	

1 p-values come from Wilcoxon signed-rank tests comparing CARE and Skillsetter modalities. Items were rated on a 0-7 scale, 1 = Strongly Disagree, 2 = Disagree, 3 = Somewhat Disagree / Slightly Disagree, 4 = Slightly Disagree, 5 = Somewhat Agree / Slightly Agree, 6 = Agree, 7 = Strongly Agree.

RQ2: Design Limitations.

AI Patient Functionality: The most prevalent technical concern (20/25, 80% of respondents) was that the system doesn't detect vocal pauses, indicating a significant gap in the speech recognition capabilities. More than half of respondents (14/25, 56%) felt that the AI patients did not feel realistic, with a similar number (13/25, 52%) noting limited emotional expression in the voices. Transcript inaccuracy (10/25, 40%), robotic-sounding voices (9/25, 36%), lack of visual avatar representation (7/25, 28%), and voice recognition issues (6/25, 24%) were additional functionality concerns raised by participants.

Students expressed consistent concerns about the quality and personalization of AI feedback. The three top concerns received the same high level of response (17 out of 25, 68%): Students wished the AI patient could provide feedback from its perspective on how the student, as a therapist, made them feel. Students desired a system that tracked and referenced their

progression over time. Many reported that the AI was providing unhelpful feedback. Nearly two-thirds of respondents (16/25) wished the feedback addressed the session as a whole, suggesting a need for more holistic assessment. There's a desire for greater variety in evaluation criteria (15/25) and feedback that incorporates vocal elements rather than just written transcripts (13/25)

In response to students' concerns about the limited utility of CARE's feedback, the teaching team implemented a supplemental feedback process in which the TA provided individualized written feedback on students' CARE transcripts. Each week, three to five students were randomly selected to receive this feedback, which was designed to enhance learning by integrating human insight with AI-generated suggestions. The TA feedback was posted publicly, so all students could see and potentially benefit from it. The decision to supplement CARE in this way was informed by feedback indicating that the AI's responses alone were often perceived as insufficient for skill development.

One student reflected on the benefits of this hybrid model: "Additionally, I think everything involved with AI would have a better outcome when there is that human aspect. [TA] ended up providing additional feedback for us weekly for CARE. This combined the AI aspect with the human aspect, so it wasn't only AI that we were interacting with. I feel like this maximized our learning. It would also have been interesting if we actually looked at some of the AI feedback and analyzed those using a forum, in addition to human feedback. I think doing so would again emphasize the aspect of human and AI (or AI-assisted human feedback), instead of just AI doing all the work. AI's benefits wouldn't come alive if there weren't human input to modulate it." This perspective underscores students' desire for feedback that reflects both the analytical consistency of AI and the nuance of human clinical judgment.

RQ3: Factors about Course Integration that Impacted Acceptability

Table 4. Concerns about how CARE was integrated into the classroom experience.

Concern	**Choice Count (n=25)**
AI replacing peer support and feedback	**22**
AI replacing peer interaction	**20**
Am I just doing user testing for a product	**19**

Environmental impact	**16**
How my data is used by the research team	**14**
What if they are making an AI therapist to replace me	12
I feel that my data is going to be used to make an AI therapist	12
Other concerns	7

Bolded values indicate fields where over half of the 25 respondents had concerns.

The strongest concerns about CARE's integration into the classroom centered around how AI might replace human elements of education, with 22/25 (88%) respondents worried about AI replacing peer support and feedback, and 20/25 (80%) concerned about AI replacing peer interaction. Many students (19/25, (76%)) questioned whether they were simply user-testing a product rather than engaging in an educational experience. Environmental impact was a notable concern (16/25, (64%)), suggesting awareness of the computational resources required for AI systems, including potential environmental impacts.

Data usage concerns were prominent, with 14/25 (56%) worried about how their data is being used by the research team, and 12/25 (48%) concerned about their data potentially being used to train AI therapists.

RQ4: Acceptance.

Students articulated a range of perspectives regarding the presence and future role of AI in psychotherapy education and clinical practice. Their reflections coalesced around three major thematic domains: (1) openness to AI integration in training and administrative contexts, (2) boundary concerns regarding application of AI in clinical care, and (3) broader ethical apprehensions related to AI's societal impact.

Several students expressed cautious optimism about the integration of AI into psychotherapy, particularly in training contexts and administrative functions. For these individuals, AI was not viewed as a threat to the therapeutic profession, but rather as a supplemental tool that could enhance learning and reduce burdens associated with non-clinical tasks. As one student noted, "I do not personally hold the fear of AI replacing therapists, so the training and administrative help resonates with me." Another student highlighted the relevance of technological adaptability in a rapidly evolving professional landscape: "As trainees in therapy, I truly see the importance of adapting to these changes and evolving with our society at large. Eventually, there will be other

therapists who utilize AI, and we will be competing with those who know how to utilize it".

Administrative utility was a recurring subtheme, with several students noting the promise of AI in alleviating logistical tasks such as billing, insurance coordination, and website maintenance. One student proposed that "a better focus for AI in psychotherapy would be having an AI handle the billing, insurance, outreach, and website management that is non-billable time for therapists, rather than training." Others concurred, suggesting that AI could support clinical infrastructure without compromising the therapeutic relationship.

Despite general receptivity to AI in administrative and educational settings, many students expressed strong reservations about potential clinical applications affecting direct patient care. A recurrent concern was that incorporating AI into training contexts might normalize its use in psychotherapy delivery, thereby facilitating a negative chain of future events. One student remarked, "Honestly, [I am] a bit more concerned, as I think while the benefits are there, it is a slippery slope." Another reflected, "I still feel uncertain about... the possibility of AI eclipsing in-person therapy in the future."

This apprehension was often anchored in the perceived irreplaceability of human relational dynamics. As one student articulated, "Nothing can replace in-person practice and human interaction. Beyond certain admin tasks, I do not think AI should be used in clinical practice." Another student emphasized the importance of ethical containment: "It's still important to be cautious with it, and it should never replace an actual human in any context other than training". These sentiments underscore a shared appreciation for the therapeutic alliance and a desire to establish clear boundaries for emerging technologies.

A third thematic domain reflected broader ethical considerations about AI's proliferation across societal and global systems. For some students, skepticism about AI in psychotherapy was rooted not solely in clinical concerns but also in a larger critique of AI's sociopolitical implications. One student remarked, "It has increased my concerns about the increase of AI use throughout the world". Another described their initial discomfort with the study's AI component, stating, "As someone who has bigger-picture ethical concerns about AI, in relation to environment, jobs, and broader social/societal impact, I was initially pretty upset about having CARE be a required part of this course".

Concerns about commercialization, technological determinism, and the disconnect between developers and end-users were also prominent. One student stated, "The lack of awareness about therapy for those who are developing models frankly terrifies me," while another wrote, "I feel very hesitant about the use of AI in any space in psychotherapy. To me, it feels like too slippery of a slope, and from history, it seems like things like this inevitably end up in the

wrong hands, even if that is not the starting intention". Reflecting these reservations, some students reported electing not to use the AI platform for training, citing "broader ethical concerns around AI and its rapid expansion".

Together, these themes reflect a complex and multidimensional stance among emerging psychotherapists. While students acknowledged the potential utility of AI in specific educational and administrative contexts, they also articulated clear boundaries for its use in clinical care and raised critical ethical questions about its integration into psychotherapy and society more broadly.

Discussion

This study explored the integration of CARE, an AI-based platform for psychotherapy training, across two successive doctoral-level psychotherapy courses. The first course examined how AI feedback could augment traditional peer role-play, whereas the second examined how AI-simulated patients could augment traditional video-based deliberate practice. Together, these course deployments provided an opportunity to explore four questions: (1) Comparative value: how students perceive the distinctive features of AI-based practice and feedback tools, and how these compare with traditional counseling education methods; (2) Limitations and design: what technical and functional constraints emerged, and what design opportunities could enhance CARE's pedagogical effectiveness; (3) Course integration: what lessons were drawn from embedding CARE into the classroom, particularly regarding onboarding, communication, and preserving instructor-to-student and student-to-student interactions; and (4) Acceptance: how students perceive the acceptability of AI in psychotherapy education, and in what ways the deployment shaped broader views of AI's role in training and clinical practice.

Insights from the First Course Pilot

Students consistently rated peer feedback as more constructive and helpful than CARE's AI feedback, citing peers' capacity to integrate tone, nonverbal information, and a client-centered interpretive stance. Technical limitations in CARE, particularly transcription errors and the limited scope of feedback, undermined perceptions of utility, reflecting broader limitations identified in current LLM-based simulation and feedback systems that rely primarily on textual representations of therapy interactions [13,14,35].

Comparative Value of AI-Simulated Patients and Video Vignettes

In the second course, CARE's AI-simulated patients provided possibilities that video vignettes could not, including opportunities for interactive dialogue, consistent with prior work demonstrating the value of simulated patients and

LLM-based systems for interactive, conversational practice in mental health training [10–13, 26]. However, quantitative ratings favored Skillsetter across several domains, including immersion and learning from mistakes, while confidence in spontaneous responding (i.e., generating in-the-moment responses within the natural flow of conversation) did not differ between tools. Student reflections reinforced a strong preference for practicing with humans, with some expressing concern that practicing with AI might lead them to prioritize system-specific feedback over cultivating the relational skills required in therapy. These findings suggest that AI patients may augment deliberate practice opportunities but are not yet perceived as replacements for human-based training.

Design Limitations and Opportunities

Feedback on CARE's design pointed to recurring technical and functional limitations. Students highlighted that the system had difficulty detecting vocal pauses, limited emotional expression, and inaccuracies transcribing speech. Similar to the first course, students' concerns about AI-generated feedback centered on its narrow scope, lack of personalization, and absence of holistic evaluation across an entire interaction. At the same time, participants valued CARE's immediacy and clarity in suggesting alternative phrasing, which several described as transferable to future practice.

While prior work finds that pairing simulated practice with performance feedback is often desired by students [23,26] and an effective combination for achieving counseling skill improvement [12,13], students in our longitudinal classroom studies highlighted additional requirements for AI feedback. First, use of AI feedback over multiple weeks revealed that feedback needs to expand beyond client-centered listening skills, instead matching specific-criteria relevant for the pedagogical goals of a simulated practice (e.g., how to handle cultural humility issues). Second, students consistently highlighted that the AI feedback system in its current form is not a substitute for human feedback: in peer-to-peer roleplays, students valued peer feedback more than AI feedback; in simulated patient interactions, TA-provided supplemental feedback on select transcripts was also positively received. Students noted the promise of this hybrid feedback model combining the standardization of AI with the nuance of human judgment. Efforts that combine human and AI feedback have been explored more broadly in traditional and online education contexts (e.g., learner-sourcing [37, 38]) suggesting that adapting such approaches for the psychotherapy and medical education area may offer the ability to scale knowledge and human interactions which are especially valued in this learning environment.

Course Integration

Instructional stance played a critical role in shaping students' engagement. The first course revealed challenges with course integration; following strong initial negative reactions, the use of CARE's role-play transcript feedback was made optional, and participation was uneven. Concerns about data security and transcript use were salient, leading the team to develop a detailed FAQ and formal information sheet to clarify protections. These lessons guided adjustments in the second course, where CARE was more fully embedded as a core pedagogical component. Incorporating CARE as a weekly course requirement corresponded to higher participation and normalization of its use. However, after the second week of the second course, CARE became optional due to students' apprehension. They could choose to use CARE for deliberate practice or complete additional Skillsetter videos.

This phenomenon aligns with other recent course-based integrations of AI-simulated patients that vary in instructional requirements. Prior work has documented both optional and required approaches to AI-supported practice, with optional implementations associated with more limited uptake and required implementations designed to ensure broad exposure across learners [25, 26]. Taken together, these approaches suggest that optional integration may limit reach, whereas required engagement may increase participation and normalize use. At the same time, the present findings indicate that when AI-based practice is required, sustained engagement may depend on parallel instructional investment, including responsiveness to student concerns and the provision of human feedback to contextualize and support AI-augmented learning.

Acceptance and Broader Perspectives

Students articulated a range of views regarding the legitimacy and acceptability of AI in psychotherapy education. Some viewed AI as a potentially useful training supplement, particularly for administrative or preparatory tasks such as documentation and structured practice, reflecting broader trends in human–AI collaboration within mental health and educational contexts [19,20,21]. Others expressed reservations about the role of AI in clinical practice, emphasizing the centrality of human relational processes that are widely regarded as foundational to psychotherapy effectiveness and training [1,4,28]. Ethical concerns extended beyond psychotherapy to encompass the broader societal implications of AI adoption, including commercialization pressures, technological determinism, and the risk of devaluing human judgment—concerns echoed in recent critiques of AI-mediated counseling and simulated practice systems [24,25]. Students also raised environmental considerations related to the computational demands of AI systems, a theme that has received limited attention in the psychotherapy training literature to date. Notably, discussions of

environmental sustainability coincided with the January 2025 Southern California wildfires, during which some students had personal connections to affected communities. This context may have amplified the salience of environmental concerns in their reflections.

Limitations

Several limitations should be noted. First, this study was conducted within a single graduate training program, and the experiences of this cohort may reflect the particular pedagogical environment, instructional style, or program culture in which CARE was introduced. In addition, the sample consisted of first-year graduate students enrolled in a multi-year program characterized by intensive supervision, including regular review of recorded therapy sessions with experienced supervisors. This context likely shaped student perceptions of AI-assisted training and may limit generalizability to settings where trainees have fewer opportunities for expert feedback (e.g., peer counselors or paraprofessionals with limited supervision access).

Second, the CARE platform integrated existing technologies for voice-based patient simulation and feedback generation that were best performing in December 2024. Technical issues may have reflected the specific version deployed rather than inherent limitations of AI-assisted training tools: voice-based practice interactions had issues with transcription accuracy, speech recognition, and limited emotional expression, while feedback was limited to skill-based content on client-centered listening skills rather than scenario-specific guidance regarding therapeutic-alliance challenges.

Finally, the study relied primarily on self-reported perceptions rather than objective measures of skill acquisition. While qualitative data provide valuable insight into students' perspectives, future work should triangulate these findings with behavioral outcomes or supervisor ratings to better assess pedagogical effectiveness.

Future Work

The findings and limitations of this study suggest several priorities for future research on AI-assisted psychotherapy training. Short-term work should focus on improving the technical functionality of AI platforms. Students in this study consistently identified transcription errors as well as AI-simulated patients with limited vocal expressiveness as barriers to engagement. Research is needed to evaluate whether improvements in speech recognition, naturalistic prosody, and affective responsiveness increase the realism and pedagogical value of AI-simulated patients.

Parallel efforts should advance feedback systems toward greater context-awareness. Current utterance-level approaches offer timely guidance on specific

communication skills, but models that process turns independently can produce feedback that is repetitive, contradictory, or disconnected from the broader therapeutic process. Prior work on automated feedback and performance assessment in psychotherapy and counseling has similarly noted that utterance-level or narrowly scoped feedback may be insufficient for supporting integration of relational and process-oriented skills across the flow of a session [14,21,22]. Approaches that consider the entire session history and enforce global consistency during feedback generation could better support students in developing coherent, session-oriented improvements. Moreover, consistent with emerging evidence that human–AI collaborative feedback may enhance contextual understanding and learner trust [19,20,21], short-term research should further evaluate hybrid feedback models that combine AI-generated suggestions with human commentary—an approach that students in the present study viewed as particularly beneficial.

Beyond technical functionality, future studies should evaluate whether AI-assisted training translates into measurable improvements in clinical skill acquisition. Recent research on AI-supported counseling and psychotherapy training has highlighted the need to move beyond learner self-report toward objective indicators of competence, including supervisor ratings, behavioral coding of microskills, and structured performance assessments [6,12,13]. Critically, such assessments should extend beyond generic skill-based evaluations (e.g., client-centered listening) to include scenario-specific rubrics aligned with the particular competencies being taught—whether therapeutic-alliance navigation, crisis intervention, or other course-specific learning objectives. Experimental designs that directly compare AI-only, human-only, and hybrid feedback approaches would clarify the specific pedagogical contributions of each modality and help determine whether AI-supported practice confers additive benefits beyond established training methods. Multi-site studies are also essential to assess whether findings generalize across programs with differing curricula, instructional philosophies, and trainee populations, a limitation frequently noted in early evaluations of AI-based training tools [6,12].

At a broader level, long-term research should address how AI integration reshapes the pedagogy and culture of psychotherapy training. Questions include whether repeated practice with AI-simulated patients influences professional identity formation and sensitivity to relational nuance. Ethical considerations will remain central, particularly regarding transparency in data use, the potential displacement of human elements in training, and environmental sustainability. The salience of ecological concerns in this study, possibly amplified by the January 2025 Southern California wildfires, highlights the need for future work to consider how external sociocultural contexts intersect with classroom-based

experiences of AI. Ultimately, research must examine whether AI can be a useful adjunct to psychotherapy training, enhancing current methods, or whether its widespread adoption risks altering the relational foundations of psychotherapy education.

Recommendations for Classroom Implementation

These recommendations were informed by student feedback across both terms, as well as by the adjustments made between the first course pilot and the second course deployment.

Transparency

Clear, proactive communication about data use, research protocols, and pedagogical aims is essential to building student trust. The second course implementation benefited from refinements to communication strategies developed in response to concerns raised during the first course pilot.

Hybrid Models

Combining AI feedback with human oversight may mitigate concerns about impersonality and enhance learning outcomes. The addition of supplemental TA feedback during the second course exemplified this adjustment and was positively received by students. Future research may benefit from systematically examining the impact of required AI integration on students' perceptions of the teaching team, including measures of instructional alliance, trust, and perceived support, to assess whether and how AI-mediated feedback influences the relational dynamics of the learning environment.

Instructor Stance

Student reflections suggested that instructional framing influenced their engagement with CARE. In the first course, the tool was positioned as optional following student concerns and apprehension, and several students reported feeling more willing to try it under these conditions than when it was framed as required. In the second course, CARE was initially presented as a central, required component of the course, which corresponded with higher participation early in the term. However, after concerns were raised, the instructional team shifted to making it optional, where students could engage in additional Skillsetter practice instead of CARE. These observations suggest that how the AI tool is introduced and framed by instructors plays a role in shaping student engagement, and that providing choice in an assigned activity can foster greater willingness to engage with AI-assisted training.

Conclusion

The integration of CARE across two academic terms illustrates both the promise and challenges of AI-assisted psychotherapy training. The first course pilot surfaced key concerns regarding feedback quality, course integration, and data transparency, while the second course study demonstrated how iterative adjustments, such as clearer communication, hybrid feedback models, and shifts in instructional framing, enabled more effective implementation. At the same time, students expressed persistent preferences for human-based training and raised concerns about realism, data use, and broader societal implications, including environmental impact. Taken together, these findings suggest that at this stage of development, AI has the potential to serve as a valuable adjunct to psychotherapy education when framed as a supplement to human interaction, implemented transparently, and paired with human oversight.

Looking ahead, priorities for advancing this field can be organized across three levels. Short-term priorities involve addressing technical limitations and refining feedback mechanisms. Medium-term priorities focus on evaluating whether AI-assisted practice translates into measurable improvements in clinical performance, ideally through controlled, multi-site studies that include objective learning outcomes. Long-term priorities include examining how AI integration reshapes the pedagogy and culture of psychotherapy training, professional identity formation, and ethical considerations such as data use, role displacement, and environmental sustainability.

The utility of AI-assisted psychotherapy training may differ across educational contexts. In this graduate program, students receive intensive supervision, which may include live or tape-review of clinical encounters, which may shape how they perceive the added value of AI. In other settings, such as programs with more limited supervisory resources, AI tools may assume a more prominent role in supporting deliberate practice and providing structured feedback. Thus, the pedagogical value of AI may not be uniform but instead contingent on the broader training environment, with its potential relevance varying according to the availability of human supervision and feedback.

By situating AI within authentic classroom contexts and iteratively adapting implementation based on student feedback, this study provides a strong foundation for future inquiry in psychotherapy pedagogy. Its contribution lies not only in documenting students' perceptions of AI-assisted training but also in demonstrating how responsive piloting can inform best practices for responsibly integrating emerging technologies into psychotherapy education.

Acknowledgments

We thank the course instructors, teaching assistants, and students for their collaboration. This research was funded by the Stanford Psychiatry and Behavioral Sciences Department Innovator Grant, Stanford HAI Seed Grant, and the Stanford Impact Labs.

Conflicts of Interest

None declared.

Abbreviations

AI — artificial intelligence
DP — deliberate practice
LLM — large language model
PsyD — Doctor of Psychology

References

[1] American Psychological Association. (2013) Recognition of psychotherapy effectiveness: The APA resolution. Psychotherapy, 50(1), 98–101. https://doi.org/10.1037/a0030276

[2] Shedler, J. (2018. Where is the evidence for “evidence-based” therapy? Psychiatric Clinics of North America, 41(2), 319–329. https://doi.org/10.1011016/j.psc.2018.02.001

[3] Knox, S., & Hill, C. E. (2021. Training and supervision in psychotherapy: What we know and where we need to go. American Psychological Association. https://www.apa.org/pubs/books/training-supervision-psychotherapy

[4] Wampold, B. E. (2015. How important are the common factors in psychotherapy? An update. World Psychiatry, 14(3), 270–277. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4592634/

[5] Hill, C. E., & Knox, S. (2023. Psychotherapy training and supervision with undergraduate and graduate students. https://www.apa.org/pubs/books

[6] Nurse, K., O’Shea, M., Ling, M., Castle, N., & Sheen, J. (2024.)The influence of deliberate practice on skill performance in therapeutic practice: A systematic review of early studies. Psychotherapy Research, 1–15. https://doi.org/10.1080/10503307.2024.2308159

[7] Larsson, J., Werthén, D., Carlsson, J., Salim, O., Davidsson, E., Vaz, A., Sousa, D., & Norberg, J. (2025). Does deliberate practice surpass didactic training in learning empathy skills? A randomized controlled study. Nordic Psychology, 77(1), 39–52. https://doi.org/10.1080/19012276.2023.2259170

[8] Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., & Cui, L. (2023.) LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation (arXiv:2305.13614). https://arxiv.org/abs/2305.13614
[9] Skillsetter. (2024.) How it works. https://www.skillsetter.com/how-it-works
[10] Louie, R., Nandi, A., Fang, W., Chang, C., Brunskill, E., & Yang, D. (2024.) Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles (arXiv:2407.00870). https://arxiv.org/abs/2407.00870
[11] Wang, R., Milani, S., Chiu, J. C., Zhi, J., Eack, S. M., Labrum, T., Murphy, S. M., Jones, N., Hardy, K., Shen, H., et al. (2024) Patient-Ψ: Using large language models to simulate patients for training mental health professionals (arXiv:2405.19660). https://arxiv.org/abs/2405.19660
[12] Louie, R., Hasan Orney, I., Pacheco, J. P., Shah, R. S., Brunskill, E., & Yang, D. (2025). Can LLM-Simulated Practice and Feedback Upskill Human Counselors? A Randomized Study with 90+ Novice Counselors (arXiv:2505.02428.) https://arxiv.org/abs/2505.02428
[13] Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2019). Development and evaluation of ClientBot: Patient-like conversational agent to train basic counseling skills. Journal of Medical Internet Research, 21(7), e12529. https://doi.org/10.2196/12529
[14] Flemotomos, N., Martinez, V. R., Chen, Z., Creed, T. A., Atkins, D. C., & Narayanan, S. (2021). Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. PLOS ONE, 16(10), e0258639. https://doi.org/10.1371/journal.pone.0258639
[15] Shah, R. S., Holt, F., Hayati, S. A., Agarwal, A., Wang, Y.-C., Kraut, R. E., & Yang, D. (2022). Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. Proceedings of the ACM on Human–Computer Interaction, 6(CSCW2), 1–24. https://dl.acm.org/doi/10.1145/3579615
[16] Sharma, A., Miner, A. S., Atkins, D. C., & Althoff, T. (2020). A computational approach to understanding empathy expressed in text-Based mental health support. EMNLP 2020. https://doi.org/10.18653/v1/2020.emnlp-main.425
[17] Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). “Rate my therapist”: Automated detection of empathy in drug and alcohol counseling via speech and language processing. PLOS ONE, 10(12), e0143055. https://doi.org/10.1371/journal.pone.0143055

[18] Lin, I. W., Sharma, A., Rytting, C. M., Miner, A. S., Suh, J., & Althoff, T. (2024) IMBUE: Improving interpersonal effectiveness (arXiv:2402.12556). https://arxiv.org/abs/2402.12556
[19] Hsu, S.-L., Shah, R. S., Senthil, P., Ashktorab, Z., Dugan, C., Geyer, W., & Yang, D. (2023.) Helping the helper: Supporting peer counselors via AI-empowered practice and feedback (arXiv:2305.08982). https://arxiv.org/abs/2305.08982
[20] Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence, 5(1), 46–57. https://doi.org/10.1038/s42256-022-00557-8
[21] Chaszczewicz, A., Shah, R. S., Louie, R., Arnow, B. A., Kraut, R., & Yang, D. (2024.) Multi-Level Feedback Generation with Large Language Models for Empowering Novice Peer Counselors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). http://doi.org/10.18653/v1/2024.acl-long.227
[22] Rudolph, E., Seer, H., Mothes, C., & Albrecht, J. (2024). Automated feedback generation in an intelligent tutoring system for counselor education. FedCSIS. https://ieeexplore.ieee.org/document/10697327
[23] Steenstra, I., Nouraei, F., & Bickmore, T. (2025, April.) 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 (pp. 1-22)
[24] Sim, K. Y. H., Lee, R. K. W., & Choo, K. T. W. (2025). " Is This Really a Human Peer Supporter?": Misalignments Between Peer Supporters and Experts in LLM-Supported Interactions. arXiv preprint arXiv:2506.09354.
[25] Beeson, E. T., Zhai, Y., Fulmer, R., Burck, A. M., & Maurya, R. (2025). A pilot study evaluating the fidelity of ChatGPT in client simulations. Journal of Counselor Preparation and Supervision, 19(3), 5.
[26] Thesen, T., O'Brien, W. N., Stone, S., & Pinto-Powell, R. (2025). Generative AI as the first patient: practice, feedback, and confidence. Medical Science Educator, 1-6.
[27] Skillsetter. (n.d.) Skillsetter: The deliberate practice system for interpersonal skills. https://www.skillsetter.com/
[28] Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. Psychotherapy, 16(3), 252–260. https://psycnet.apa.org/record/1979-25203-001
[29] Hill, C. E. (2020). Helping skills: Facilitating exploration, insight, and action (5th ed). APA. https://www.apa.org/pubs/books/9781433835056

[30] Miller, W. R., & Rollnick, S. (2013). Motivational interviewing: Helping people change (3rd ed.). Guilford Press. https://www.guilford.com/books/Motivational-Interviewing/Miller-Rollnick/9781609182274

[31] Pérez-Rosas, V., Resnicow, K., Mihalcea, R., et al. (2022.) PAIR: Prompt-aware margin ranking for counselor reflection scoring. EMNLP 2022. https://aclanthology.org/2022.emnlp-main.148/

[32] Pérez-Rosas, V., Resnicow, K., Mihalcea, R., et al. (2023.) VERVE: Template-based Reflective Rewriting. EMNLP Findings. https://aclanthology.org/2023.findings-emnlp.689/

[33] Shen, S., Perez-Rosas, V., Welch, C., Poria, S., & Mihalcea, R. (2022). Knowledge-enhanced reflection generation. ACL 2022. https://aclanthology.org/2022.acl-long.221/

[34] Hartl, T. L., Zeiss, R. A., Marino, C. M., Zeiss, A. M., Regev, L. G., & Leontis, C. (2007). Clients' sexually inappropriate behaviors directed toward clinicians. Professional Psychology, 38(6), 674. https://psycnet.apa.org/record/2007-17996-006

[35] García-Torres, D., Fernández, C., Mira, J. J., Morales, A., & Vicente, M. A. (2025). Using AI-Based Virtual Simulated Patients for Training in Psychopathological Interviewing: Cross-Sectional Observational Study. JMIR Medical Education, 11, e78857. http://doi.org/10.2196/78857

[36] Elhilali, A., Ngo, A. S. H., Reichenpfader, D., & Denecke, K. (2025). Large Language Model–Based Patient Simulation to Foster Communication Skills in Health Care Professionals: User-Centered Development and Usability Study. JMIR medical education, 11, e81271. http://doi.org/10.2196/81271

[37] Kim, J. (2015). Learnersourcing: improving learning with collective learner activity (Doctoral dissertation, Massachusetts Institute of Technology). http://hdl.handle.net/1721.1/101464

[38] Singh, A., Brooks, C., Wang, X., Li, W., Kim, J., & Wilson, D. (2024, March). Bridging learnersourcing and AI: Exploring the dynamics of student-AI collaborative feedback generation. In Proceedings of the 14th Learning Analytics and Knowledge Conference (pp. 742-748). https://doi.org/10.1145/3636555.3636853