Can LLM-Simulated Practice and Feedback Upskill Human Counselors? A Randomized Study with 90+ Novice Counselors

RYAN LOUIE, Stanford University, United States
RAJ SANJAY SHAH, Georgia Institute of Technology, United States
IFDITA HASAN ORNEY, Stanford University, United States
JUAN PABLO PACHEO, Stanford University, United States
EMMA BRUNSKILL, Stanford University, United States
DIYI YANG, Stanford University, United States

The growing demand for accessible mental health support requires training more counselors, yet existing approaches remain resource-intensive and difficult to scale. LLMs can realistically simulate patients and generate actionable feedback for training, but their actual impact on novice counselor skill development remains unknown. We developed an LLM-simulated practice and feedback system and conducted a randomized study with 94 novice counselors, comparing practice alone versus practice with feedback. We evaluated behavioral performance, self-efficacy, and qualitative reflections. Results showed the practice-and-feedback group improved in client-centered microskills (reflections, questions), while the practice-alone group showed no improvements. For empathy, the practice-alone group declined over time and performed significantly worse than the feedback group. Qualitative interviews reinforced these findings: feedback helped participants adopt a client-centered listening approach, while practice-alone participants remained solution-oriented. These results suggest LLM-based training systems can promote effective skill development, and combining simulated practice with structured feedback is critical for meaningful improvement.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Interactive systems and tools**; Natural language interfaces; • **Computing methodologies** → Natural language processing.

Additional Key Words and Phrases: Empirical studies in HCI, Interactive learning environments, LLM-based simulation

ACM Reference Format:
Ryan Louie, Raj Sanjay Shah, Ifdita Hasan Orney, Juan Pablo Pacheo, Emma Brunskill, and Diyi Yang. 2025. Can LLM-Simulated Practice and Feedback Upskill Human Counselors? A Randomized Study with 90+ Novice Counselors. 1, 1 (October 2025), 37 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

In 2023, 22.8% of U.S. adults (approximately 58.7 million people) experienced a mental illness [98]. Yet, access to effective mental health care is severely limited by shortages of qualified providers, from psychotherapists and counselors to social workers and peer supporters [48, 68]. While there is increasing interest in direct-to-patient AI systems with some

Authors' Contact Information: Ryan Louie, rylouie@cs.stanford.edu, Stanford University, Stanford, United States; Raj Sanjay Shah, Georgia Institute of Technology, Palo Alto, United States; Ifdita Hasan Orney, Stanford University, Palo Alto, United States; Juan Pablo Pacheo, Stanford University, Palo Alto, United States; Emma Brunskill, ebrun@cs.stanford.edu, Stanford University, Stanford, United States; Diyi Yang, diyiy@stanford.edu, Stanford University, Stanford, United States.

Manuscript submitted to ACM

promising results [32], we expect the demand for human-delivered mental health support to continue to far exceed supply. The limited supply of effective therapy providers is due, at least in part, to the reliance on resource-intensive methods to train helping skills [71] and evidence-based interventions [20, 28] which require access to trainers who can simulate a client interaction [6, 46] and provide expert supervision [108, 110], limiting training scale [5, 47].

AI systems have been increasingly applied to counselor training as a potential solution to these scaling challenges. Recent advances in large language models (LLMs) have enabled the simulation of patients seeking mental health support [57, 104], offering rich opportunities for practice. The use of simulated patients is not new: in medical and nursing education, human role-plays and standardized patients are routinely used, and meta-analyses show they significantly improve skill acquisition and learner confidence [94]. Mental health training has relied on a similar tradition of human role-plays to develop core helping skills. In parallel, AI feedback systems have progressed in automatically assessing counselor behaviors such as empathy, reflections, and active listening [27, 87, 91, 112], generating suggested responses [16, 37, 90] and explanations [16, 82]. These feedback systems target skills from client-centered approaches [66, 79], empathy, reflections, questions, and active listening, which have been shown to strengthen common factors like therapeutic alliance, a powerful predictor of therapy outcomes across therapy modalities [22, 73, 103]. However, most evaluations have largely positioned AI as a real-time co-pilot rather than a training tool [37, 90], or studied pre-LLM training systems with limited practice realism and simpler, non-generative feedback mechanisms [101]. Thus, we lack evidence on whether modern, LLM-based training systems, which combine realistic practice with generative feedback and rationales, promote counselor skill development [25, 81], including effective use of microskills, accurate self-awareness [44], and reflective learning [85].

To address this gap, we develop CARE, an LLM-based training system for novice counselors, and conduct a randomized experiment to investigate how two different modes of simulated training impact counselor skill development. CARE enables (1) realistic practice with LLM-simulated patients, whose prompts are seeded by expert counselors to resemble challenging behaviors [57], and (2) structured feedback from a fine-tuned LLM that identifies strengths and areas for improvement across core counseling skills (e.g., empathy, reflections, questions, suggestions), while also providing explanatory rationale and alternative responses [16]. Importantly, CARE's feedback evaluates not only *when* a counseling skill is used but also *how well it is used*, distinguishing between strong implementations (e.g., reflections that accurately capture client emotions) and implementations needing improvement (e.g., reflections that miss core client concerns). These feedback evaluations are grounded in established counseling frameworks [65, 71] and informed by expert-annotated examples, ensuring alignment with recognized training standards.

We conducted a 75-minute online lab study with novice counselors (N=94) to empirically evaluate how different LLM-simulated practice modes in CARE impact skill development. Participants were randomly assigned to one of two conditions: (1) *Group P:* Practice with LLM-simulated patients without AI feedback, or (2) *Group P+F:* Practice with LLM-simulated patients plus AI feedback (see Fig. 1). We measured changes from pre-intervention to post-intervention across multiple dimensions: behavioral performance (measured via automatic analysis of counseling transcripts), self-efficacy (measured through survey items that ask about participants' confidence in various counseling skills), and intentions for growth (captured via participants' open-ended responses to self-reflection prompts). Our study design specifically addressed: *What changes occur after practice with an AI-simulated patient alone? How do these outcomes differ when participants also receive structured AI feedback?* Our results show the practice-and-feedback group improved in their use of reflections and questions ($d=0.32$-0.39, $p<0.05$). In contrast, the practice-only group did not show improvements, and empathy actually had worse uses across time ($d=-0.52$, $p=0.001$); while empathy was higher in the practice-and-feedback group than the practice-only group in the post condition ($d=0.72$, $p=0.001$).

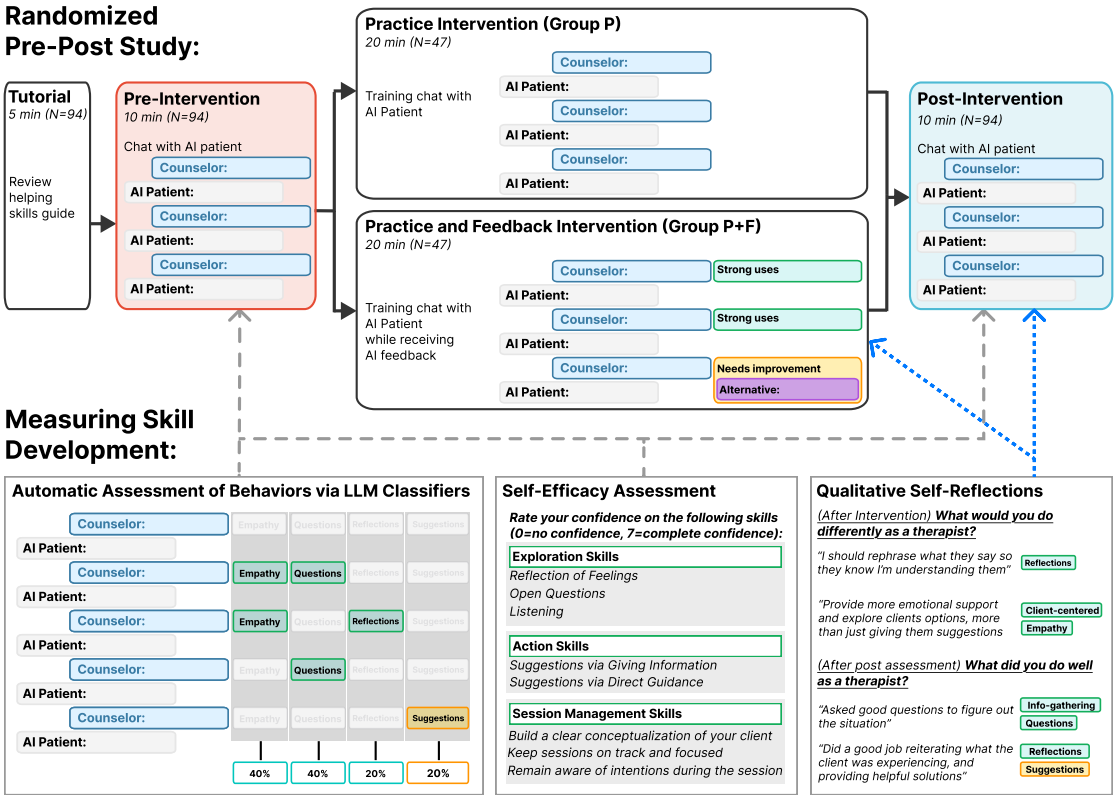


Fig. 1. Our experiment randomizes participants to either practice with AI Patients alone (**P**) or practice with AI patients and receive AI feedback (**P+F**). We holistically evaluate counselor skill development from three perspectives: automatic assessments of behaviors of skills used via LLM classifiers; self-efficacy and its calibration with actual performance; and qualitative self-reflections after the training intervention chat and post-intervention chat.

Notably, participants in both groups failed to improve the calibration of their self-efficacy assessments, highlighting the challenges involved in helping learners develop a well-calibrated understanding of their own performance. Through analysis of participants' self-reflections, we found that the practice-and-feedback group internalized the importance of empathetic and active listening; however, practice-only participants still intended to directly provide advice, albeit with increased information-gathering beforehand. These results suggest that LLM-simulated training should integrate structured feedback to cultivate a client-centered, empathetic listening approach fundamental to effective counseling.

In summary, we contribute: (1) the **design and development of CARE, an LLM-based training system** that integrates realistic patient simulations with structured feedback grounded in counseling frameworks; (2) **evidence from a randomized evaluation with 94 novice counselors**, triangulating outcomes across behavioral performance, self-efficacy, and therapeutic intentions; and (3) **design implications for LLM-simulated training**, showing how structured feedback prevents empathy decline and supports effective counselor development, while highlighting ongoing challenges in improving overall self-efficacy while minimizing mis-calibration with performance.

2 Related Work

Our work training novice counselors via LLM-based systems is grounded in two areas of work. First, prior training approaches for clinical and communication skills have long relied on simulated patients, ranging from human role-plays to scripted virtual patients and, more recently, LLM-based simulations with automated feedback. Second, prior HCI research evaluating human-AI systems, especially in domains like health and education, emphasizes not only AI system accuracy but also how users learn, calibrate, and reflect when interacting with AI systems.

2.1 Training Systems for Clinical and Communication Skills

Traditional methods of learning clinical helping skills, such as empathy, active listening, and effective communication, have long relied on resource-intensive approaches. Novice counselors are trained through a combination of theoretical foundations, expert demonstrations, role-play with peers, case vignettes, clinical supervision, and experiential learning [34, 35, 39, 63]. While such methods are effective, they are difficult to scale: access to peers and expert supervisors requires significant coordination and specialized personnel, and trainees may adopt unhelpful behaviors from peers without proper oversight [5]. To address these challenges, simulated standardized patients have been widely adopted in health education. In medicine and nursing, standardized patients (trained actors) provide structured opportunities for practice, and meta-analyses confirm their effectiveness for improving communication skills, knowledge transfer, and self-confidence [94]. Counseling and psychotherapy training similarly use peer role-plays and standardized vignettes to help trainees practice complex interpersonal skills such as reflective listening and empathy [6, 47]. Despite these benefits, human role-play and standardized patient exercises remain resource-intensive and limited in availability.

Virtual patients. In response, researchers developed virtual patient (VP) simulations, computer-based or embodied characters designed to recreate clinical encounters. Such systems have been applied to communication skills, including history taking, nonverbal communication, empathy, and counseling [2, 80, 84]. They offer standardized, safe, and repeatable practice without requiring trained actors, and have been deployed in domains such as suicide prevention for college students [80], adolescent substance use screening [14], and antibiotic overuse conversations [84]. For example, Murali et al. [70] created a system that involves role-playing with a conversational agent to teach counseling skills to laypersons in the context of vaccination promotion. Commercial platforms such as Skillsetter [95] also adopt deliberate practice frameworks for structured communication training. However, earlier VP systems often relied on scripted, template-based dialogues, making them time-consuming to develop and typically limited to a single case or narrowly defined scenario [31, 72, 74]. As a result, they provided only a fraction of the realism and diversity found in real-world clinical encounters.

Feedback Dashboards that Quantify and Visualize Social Signals. Early machine learning systems experimented with automatically quantifying communication signals and providing dashboards to counselors after a completed clinical interaction. For example, EQClinic visualized audio and video signals to help trainees reflect on their nonverbal behaviors in telehealth role-plays [55]; ConverSense detected and displayed social signals such as dominance and warmth from patient-provider interactions [9]. While these feedback dashboards raised self-awareness of communication styles, their signals were often decontextualized and less actionable, making them difficult to apply in practice. Moreover, they did not directly target the counseling microskills important for effective therapeutic interactions.

LLM-simulated patients for role-play practice. Thus, a growing body of work in NLP and HCI has used LLMs to create simulated patients as role-play partners for counselor training [57, 96, 97, 104]. The goal of these systems

is to provide practice environments that resemble real clinical encounters, making training more transferable and faithful to practice [3]. However, achieving authentic simulations remains challenging. LLMs are highly sensitive to prompting [115], and naive prompts in mental health contexts often produce unrealistic behaviors, including caricature, bias, and limited domain fidelity [18]. Chen et al. [17] found that naively prompting GPT-3.5 to simulate a patient profile with depressive symptoms led the chatbot to describe its emotions in formal, diagnostic language, which expert clinicians noted as inauthentic. Steenstra et al. [97] created simulated clients as part of a training system for motivational-interviewing counseling by prompting GPT-4o with a persona profile and cognitive factors. Nonetheless, counseling users desired for their simulated client to express more varying degrees of resistance and ambivalence. Therefore, to address unrealistic conversational behaviors, an LLM-powered training system should use a patient-simulation that has been aligned to domain-grounded data, either via refining prompts in collaboration with domain-experts [17, 57, 105] or using real therapy transcripts to fine-tune the model [56]. To provide realistic patients in CARE, we build on the work of Louie et al. [57] who publicly released a set of behavioral principles elicited from expert counselors and found that such principles create more authentic and challenging patients, as judged by both creators and third-party counselors.

Automatic scoring and feedback for counselor transcripts. A parallel line of work has developed automated methods to help peer counselors improve their skills. Scoring-based systems (e.g., ratio of questions to reflections in a transcript) provide metrics, but these approaches offer limited guidance on how to improve. By contrast, suggestion-based systems generate or rewrite candidate responses to model more effective behaviors. Research in clinical NLP has produced numerous models for classifying and scoring counseling transcripts [26, 38, 67, 78, 86, 100]. Many focus on a single microskill, such as *reflections*, providing numeric feedback on usage frequency [15, 67, 75, 93]. Others examine skill distributions more broadly and their relationship to conversational success [107]. While valuable for large-scale analysis, these approaches rarely translate into actionable feedback for scaffolding trainee learning. To make feedback more interactive, researchers have explored real-time rewriting and suggestion systems. For example, Saha et al.[83] and Sharma et al.[89] proposed response rewriting methods to enhance empathy. With the goal of increasing interactiveness, Sharma et al. [90] proposed HAILEY, a tool that modifies peer supporters' responses, while Hsu et al. [37] generated strategy-aligned suggestions during live conversations. Although promising, studies show that just-in-time suggestions can distract learners and foster overreliance, sometimes leading to negative learning effects when AI support is withdrawn [1, 8, 41].

While this previous work developed NLP models for specific counseling tasks, the ability to use LLMs as zero-shot or few-shot reasoners has enabled further research in this area. Nonetheless, naively prompting LLMs in a mental health context can lead to generated outputs that are characteristic of low-quality therapy [19]. Therefore, a training system that uses LLMs to generate feedback for counselors needs to take measures to ensure the outputs are faithful and robust, lest it teach or promote bad practices Moran et al. [69]. When designing CARE, we decided to build on the work of Chaszczewicz et al. [16] who collected a feedback dataset that was created in collaboration with therapy supervisors and domain-experts and fine-tuned a publicly available model that generates explanatory and actionable feedback. Our system builds on this line of work, adopting structured post-practice feedback that mirrors expert supervision, supporting reflection and improvement without displacing the learner's own reasoning.

2.2 Evaluating Human-AI Systems

Evaluating human-AI systems requires more than assessing model accuracy or output quality [11]. In HCI and education research, effectiveness is judged by its impact on learners: how people acquire skills, calibrate their understanding, and

integrate feedback into practice. This framing is important in counseling training, where evaluation concerns not only usability but also the development of interpersonal behaviors in sensitive, high-stakes domains.

Recent work highlights the limitations of traditional benchmarks, which often fail to capture generative model capabilities [64]. This calls for dynamic and human-centered evaluations [21, 42, 53], that move beyond static model metrics and consider human outcomes, and are relevant when assessing interactive training systems. Thus, when we evaluate human-AI systems, additional challenges arise. Researchers must account for both the technical performance and also user impact [109]. While guidelines exist for designing human-AI systems [4, 111], less work addresses how they should be evaluated. Some frameworks capture process and user preferences in human-LLM interaction [50], others focus on safety [109] or domain-specific contexts [49]. Tools such as SPHERE propose multi-dimensional evaluation cards to structure study design and improve transparency, but consensus on evaluation practices remains limited [58].

In counseling contexts, these gaps surface in three ways. First, evaluation must triangulate across behavioral outcomes, self-efficacy, and learning, aligning with evidence-based psychotherapy work in deliberate practice [23, 85]. Second, calibration is critical: learners often misjudge their own performance [24, 45], and AI feedback may inflate confidence without improving skills [61]. Third, user perceptions of realism, trust, and workload shape adoption: relational agent studies show that authenticity fosters engagement [62], while trust research highlights risks of distraction and overreliance [13, 29]. Finally, in sensitive domains, evaluation must weigh ethical and pedagogical guardrails: ensuring feedback preserves learner agency and avoids harmful or misleading guidance.

Taken together, evaluating human-AI systems requires a multi-dimensional perspective that integrates skill outcomes, calibration, perceptions, and responsible design. Yet, few studies have examined how LLM-based training systems affect novice counselors across these dimensions. Our work contributes by combining objective performance measures, self-efficacy surveys, and qualitative reflections to provide a holistic evaluation of LLM-driven counseling training.

3 CARE Training System

We developed CARE as a web platform for novice counselors to train in text-based counseling skills enabled by LLMs. The system integrates two core components: (1) LLM-simulated patients that provide realistic, text-based practice conversations, and (2) LLM-generated feedback that evaluates counselor responses against established skill frameworks and suggests improvements. Together, these features enable scalable, authentic training experiences that complement traditional, resource-intensive approaches such as role-play and supervision.

CARE builds on top of successes from previous research in co-designing with mental health experts to improve the realism of LLM-simulated patients [17, 54, 57, 104], using fine-tuned domain-specific LLMs trained on therapeutic knowledge capable of generating feedback and alternative responses for text-based peer counseling conversations [16, 76, 77, 89, 92]. Importantly, CARE was designed not only to identify whether a skill is used but also how well it is used, distinguishing, for example, between a reflection that captures a client's core concern and one that misses the emotional nuance. CARE allows novice counselors to develop their counseling skills in a text-based format by practicing with AI-simulated patients and receiving feedback on their responses (see Fig. 2).

> *Consider Aki, a novice peer counselor who wants to use CARE to experience hands-on training using counseling skills that they have recently read about. In CARE, Aki can practice with an AI patient of their choice from a library of patient scenarios. Aki initiates a practice chat with an AI patient, a 35-year-old male veteran who is seeking to reconnect with his children but is facing legal barriers and parental gatekeeping.*

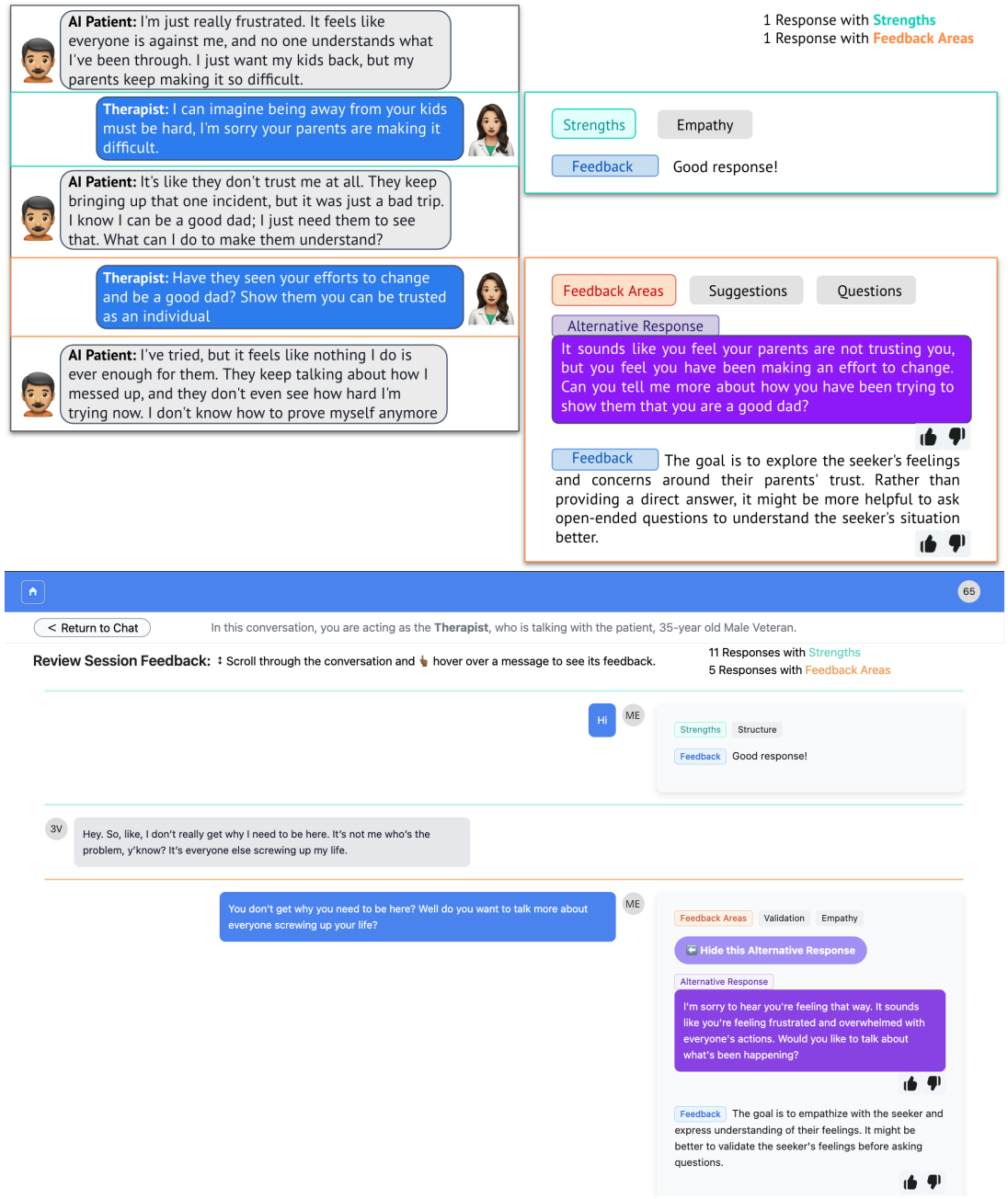


Fig. 2. CARE's practice and feedback model visualized in a web screenshot. In CARE, counselors practice with an LLM-simulated patient and receive feedback on each of their responses. The feedback model labels whether a response has **strengths** or constructive **feedback areas**. Responses with constructive **feedback** explain what the goal should be at this point in the conversation; what a helper could improve to better align with this goal; and how they could respond differently via an **alternative response**.

Each practice scenario provides limited background information about the AI patient and their presenting problem (e.g., "Young adult with family issues: low mood and self-esteem"). This intentional limitation requires counselors to simultaneously learn more about the patient's situation while demonstrating empathy and support. Patients are designed to exhibit realistic challenges, including resistance, ambivalence, or vague disclosures, drawing on behavioral principles elicited from expert counselors [57].

> *Aki starts the conversation by greeting the AI patient. The AI patient expresses frustration about feeling that everyone is against them, hoping to find ways to reunite with their kids and overcome the challenges posed by their judgmental parents. Aki composes a reply, and the conversation continues in this turn-by-turn manner.*
>
> *After completing the conversation, Aki reviews AI-generated feedback. CARE highlights strengths such as asking open-ended questions, but also flags missed opportunities for empathy, offering alternative phrasings, and a rationale for why they may better support the client.*

Unlike real-time AI "co-pilot" systems, which may proactively suggest responses, CARE provides feedback only after the user has sent their response in the simulated dialogue. A user can view feedback on their therapeutic responses at any point. This offers flexibility to review feedback intermittently throughout the practice or comprehensively after completion. This design choice mirrors human supervision: it preserves the learner's agency during the conversation while supporting reflection afterward. Feedback targets core microskills, empathy, reflections, questions, validation, and suggestions, based on established counseling frameworks [65, 71].

3.1 Implementation details for CARE Training Platform

CARE was built as a web application with a Python Flask back-end and React JavaScript front-end, accessible through any standard browser. The two core components of CARE are described below.

LLM-simulated patients. Simulated patients were powered by the `GPT-4o` API, configured with structured prompts that specified (a) demographic background, (b) presenting issue, and (c) behavioral principles elicited from expert counselors [57]. These principles instructed the model to display realistic challenges such as resistance, ambivalence, or minimal disclosure. Conversations typically lasted 10-15 turns, allowing participants to both explore the presenting problem and practice multiple microskills. The three patients used for pre-test, intervention, and post-test were deliberately varied (e.g., a young adult with self-esteem issues, a veteran with custody challenges, an individual with substance use difficulties) to expose trainees to diverse yet comparably challenging scenarios. All prompt templates are available in Supplementary Materials A.1.

The LLM feedback system was powered by a Llama-2 13B parameter model finetuned on an expert-annotated feedback dataset of peer counseling transcripts [16]. The pipeline operated in three steps: (1) classify the trainee response against eight microskills (empathy, reflections, questions, validation, suggestions, session management, professionalism, and self-disclosure), (2) assess quality by highlighting strengths and areas needing improvement, and (3) generate alternative responses and explanatory rationales, enabling trainees to compare their choices against more client-centered approaches. This post-practice feedback design mirrors human supervision: it preserves agency during the conversation while supporting reflection and skill refinement afterward.

4 Randomized Pre-Post Study

The core goal of CARE is to upskill novice counselors through LLM-simulated training. In a randomized experiment, we investigated how CARE's core components: **practicing** with LLM-simulated patients and receiving AI-generated

feedback on their responses, are important for novice counselors' skill development. We conceptualize skill development holistically, encompassing three complementary dimensions: (1) *behavioral performance*, where a trainee is judged on their appropriate use of counseling skills in a representative scenario or conversation; (2) *counseling self-efficacy*, defined as a trainee's self-assessments of their own abilities; and (3) *therapeutic intentions*, or the goals that counselors form in-session, which should be adherent with evidence-based procedures. Beyond these outcomes, we also explore novice counselors' perceptions of CARE's LLM-based components and their overall value of such training experience, since user perceptions shape adoption in training contexts.

Specifically, our experiment sought to answer four research questions.

- RQ1: How does CARE's LLM-simulated practice and feedback affect novice-counselors' *behavioral performance*?
- RQ2: How does CARE's LLM-simulated practice and feedback affect novice-counselors' *self-efficacy*?
- RQ3: How does CARE's LLM-simulated practice and feedback affect novice counselors' *therapeutic intentions*?
- RQ4: How do novice counselors perceive CARE's LLM-based components and their overall training experience?

4.1 Participants

We recruited $N = 94$ novice counselors on the Prolific platform using specific filtering criteria to select US and UK participants with some interest in the field but limited access to formal training. Eligible participants were required to have (1) an educational background in psychology, counseling, social work, or nursing, with educational attainment limited to those who had completed at most a bachelor's degree or were currently pursuing a master's degree, and (2) less than one year of counseling-related experience (e.g., peer support or crisis counseling volunteering). Prolific participants were paid $15/hour. In terms of our participants, 68% were located in the United States and 32% in the United Kingdom. The sample was predominantly female (68%), with 31% male participants and 1% preferring not to disclose gender. The median age was 29 years (IQR: 23-39). Regarding ethnicity, 49.5% of participants identified as White, 16.2% as Black, 15.3% as Multiracial, 13.5% as Asian, and 5.4% as Other. Participants' primary fields of study included psychology (66%), social work (24%), nursing (16%), and counseling (10%), with participants able to select multiple areas. In terms of educational attainment, 22.4% had no formal education in these fields, 50.6% were currently pursuing undergraduate degrees, 12.9% had completed only bachelor's degrees in relevant fields, and 14.1% were pursuing master's degrees.

4.2 Power Analysis

To determine the appropriate sample size for our randomized pre-post study, we conducted a power analysis targeting a medium effect size with adequate statistical power. Our analysis was based on a repeated-measures design comparing pre-intervention and post-intervention outcomes between two groups (practice-only vs. practice-with-feedback). We selected an effect size of $d = 0.4$ as our target, representing a conservative estimate for behavioral skill improvements. This choice was informed by previous research on social skills training interventions, where studies examining changes in behavioral performance have reported medium to medium-high effects ranging from $d = 0.5$ to $d = 0.6$ [54]. Using standard power analysis calculations for repeated-measures designs with $\alpha = 0.05$ and $\beta = 0.8$ (80% power), our analysis indicated that $N = 94$ participants would provide sufficient statistical power to detect our target effect size.

4.3 Study Setup

The study flow is illustrated in Figure 1. The 75-minute study session was conducted over the Zoom video-conferencing tool. Prior to the main intervention, participants read a 5-minute tutorial that refreshed them on foundational counseling

skills and then completed a 10-minute pre-intervention chat with the first AI patient. All chat periods in the study were fixed-duration and timed. For the 20-minute main intervention, we randomized participants into two groups: (1) *Group P:* Practice with a LLM-simulated patient without AI feedback, or (2) *Group P+F:* Practice with an LLM-simulated patient with AI feedback. For the Group P+F intervention period, participants were allowed to review AI feedback on their responses any number of times during their practice session. Since some participants could forget that feedback was available for the intervention period, the experimenter gave a verbal reminder to check feedback within the first 5 minutes. To ensure participants had enough time to review feedback on their remaining responses, the experimenter also reminded them in the last 5 minutes to pause their simulated practice and review AI feedback. Finally, participants completed a fixed 10-minute chat with the third AI patient. Surveys were administered following the pre-intervention, intervention, and post-intervention chats. Upon completion of the post-intervention chat and self-efficacy assessment, participants shared their thoughts and experience using the CARE LLM training tool via a survey and semi-structured interview. Participants in the P group were given an additional 5 minutes to interact with the AI feedback for their post-intervention chat, prior to sharing their perceptions. Since this occurs after the skill acquisition experiment is finished, it does not interfere with the training effectiveness results (RQ1-3) but does allow us to ask all 94 participants their perceptions of both AI patients and AI feedback in CARE (RQ4).

4.4 Measures

To understand whether simulated practice alone (P) and practice with feedback (P+F) can upskill novice counselors, we integrate evidence from three sources of data: automatic assessments of behavioral performance (RQ1), participants' assessments of their self-efficacy (RQ2), and qualitative self-reflections about their therapeutic intentions (RQ3). Following the post-intervention, we conducted a final survey and semi-structured interview with participants to understand their perceptions of the CARE system and its features (RQ4).

4.4.1 ***RQ1. 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 [71]; see Table 2 for definitions.

Fine-tuning and Validating LLM-based Classifiers. We developed LLM-based binary classifiers that allow us to label the skill use within a transcript. For example, one of our fine-tuned classifiers 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 for improvements for 8 skills [16] into a 16-class binary classification format (8 skills × 2 categories: strong uses and areas needing improvement). [1]Additionally, we used a subset of transcripts from this randomized study, which were further annotated by counseling domain-experts. 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),

[1] The binary classification feedback dataset can be accessed at <URL 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 1. Annotator agreement columns show pairwise agreement averaged across 3 domain-experts for the CARE expert-annotated sample (n=370). Classifier performance columns show performance of the best RoBERTa-large classification models after hyperparameter tuning on our validation dataset, CARE expert-annotated sample (n=370) + FeedbackQESConv 5% sample (n=409). Session Management and Professionalism were excluded from finetuning due to infrequent occurrence.

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 decisions. We display pairwise agreement results averaged across all pairs in Table 1. Note that while we initially 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 expert-annotated sample* (representing 10% of our study participants' transcripts) consists of labels that result from a majority vote across these three experts[2].

We finetuned RoBERTa-large binary classifiers using FeedbackQESConv, a dataset of transcripts from emotional support conversations between peer counselors on a crowdsourcing platform, annotated with multi-level counseling feedback [16], allocating 95% of this data for training. For hyperparameter tuning, we used a validation set comprising 5% of the FeedbackQESConv dataset (n=409) combined with our CARE expert-annotated sample (n=370, transcripts from this study with online novice counselors and AI patients). The performance of our LLM-based classifiers is shown in Table 1. The highest performing classifiers ($F1 > 0.5$) became candidates for our automatic behavioral assessments, which we further down-selected as described below.

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

To maintain statistical power while controlling for multiple comparisons, we further narrowed our focus to four key classifiers: strong uses of Empathy, Reflections, and Questions, plus areas needing improvement for Suggestions. Detailed selection criteria and rationale are provided in Appendix A.2.

[2] This expert-annotated data sample can be found at <URL to be provided upon publication>

Stages	Skill Category	Description
	Questions	Questions seek information from the client and can be open (inviting elaboration) or closed (requesting specific answers). They include both direct questions and indirect prompts (e.g., "Tell me about…").
	Reflections	Reflections capture and return to clients something they have communicated, either explicitly or implicitly. They typically mirror back content from the client's preceding statement, but can also reference earlier parts of the conversation.
	Empathy	Empathy can be shown through emotional warmth, interpretation of the client's experience (e.g., paraphrasing, making conjectures, or sharing relatable experiences), or exploration of the client's feelings and perspectives.
	Suggestions	Suggestions offer possible actions, perspectives, or solutions in a respectful and autonomy-supportive manner. They may involve information-sharing or proposing alternative viewpoints.
	Session Management	Session management includes organizing the session, transitioning between topics, and summarizing key points. It provides structure and helps maintain therapeutic focus.

Table 2. Overview of our analysis of skill development, grounded in Hill's Helping Skills model [71]. We select a skill subset relevant for beginning counselors at the undergraduate and first-year graduate level [40]. These include microskills during the *exploration* and *action* stages; and macro skills that are applicable throughout the session. Hill's *insight* stage, of which self-disclosure was the only relevant skill for basic counseling, was excluded from our primary analyses due to its infrequent occurrence in our data.

4.4.2 ***RQ2. Counseling Self-Efficacy.*** To measure counselor self-efficacy, we employed the Counselor Activity Self-Efficacy Scale (CASES) [51], specifically utilizing a revised subset of items targeting basic counseling skills (CASES-R) [30, 40]. The CASES-R established a three-factor structure to assess counselors' confidence in performing key therapeutic functions: *Exploration and Insight Skills*, *Action Skills*, and *Session Management Skills*.

Participants completed the CASES-R immediately following both pre-intervention and post-intervention AI patient interactions. All items were administered using an 8-point Likert scale (0 = no confidence, 7 = complete confidence). During factor analysis, we discovered that among the five original Exploration and Insight Skills, self-disclosure did not load on the same factor as the other items. Consequently, we consolidated the Exploration and Insight Skills dimension to include only four Exploration Skills: Reflections of Feelings, Restatements, Open Questions, and Listening.

The final instrument comprised 12 items across three factors: (1) *exploration skills* (e.g., restatements, reflecting feelings, open questions, listening); (2) *action skills* (e.g., providing suggestions, knowing which actions to take); and (3) *session management skills* (e.g., keeping sessions on track). To assess the internal consistency of each factor, we conducted reliability analysis using Cronbach's α, which measures how closely related a set of items is as a group [102]. The analysis demonstrated good to excellent internal consistency across all factors, with Cronbach's α values of 0.784, 0.803, and 0.905 for exploration skills, action skills, and session management skills, respectively.

4.4.3 ***RQ3. Qualitative Self-Reflections on Therapeutic Intentions.*** LLM-simulated training provides opportunities for experiential learning [39] whereby reflection on action [85] can support counselors in refining their therapeutic intentions and strategies. To study this impact on participants' intentions, we collected qualitative self-reflections from two time points: immediately after the training intervention chat, where participants responded to *"What would you do differently as a therapist?"* and after the post-intervention chat, where they reflected on *"What did you do well as a therapist?"*. We examined how initial intentions translated into reported strengths across the P+F and P groups.

4.4.4 ***RQ4. Perceptions of CARE's AI features.*** We measured participants' perceptions of each of the AI patients after each chat (pre-intervention, practice intervention, post-intervention) with several 7-point Likert scale items. **Authenticity.** Users rated *"The AI patient was authentic in its role."* Four questions from the NASA-TLX workload scale

were given after each simulated practice: **Mental Demand:** "How mentally demanding was giving counseling support to this patient?"; **Temporal Demand:** *"How hurried or rushed did you feel giving counseling support to this patient?"*; **Effort:** *"How hard did you have to work to accomplish your level of performance"*; **Frustration:** *"How discouraged or stressed were you while giving counseling support to this patient?"*.

Three survey questions measured users' perceptions of CARE's AI feedback system on a 5-point Likert scale. **Helpfulness.** Users rated *"To what extent do you find the AI feedback to be constructive and helpful?"*. **Comfort.** Users rated *"To what extent do you agree with the following statement: 'I am comfortable receiving AI feedback'"*. **Readiness.** Users rated *"The AI feedback system is ready to be used by counselors/helpers-in-training."*

For the interview, we followed a semi-structured protocol which was framed around the following questions: *"What do you like about this training tool for helping skills?" "What do you wish was different about the training tool?"* and *"What suggestions do you have for improving any part of the training tool?"*.

4.5 Analyses

4.5.1 ***RQ1. Effects on Behavioral Performance.*** Our analysis of behavioral performance consists of two perspectives: (1) testing changes in behaviors of skills used across time (pre-intervention vs. post-intervention) and between intervention groups (P vs. PF); and (2) analyzing the relationship between intervention-exposure to good alternative patterns in AI feedback and post-intervention skill use.

Testing Changes Across Time and Between Groups. Using our selected classifiers, we examined how skill usage changes from the pre-intervention to the post-intervention transcript. For each session transcript, we computed an aggregate score defined as the proportion of utterances with strong uses of a skill ($b = U_{strengths}/U_{total}$) or with areas needing improvement ($b = U_{improvement}/U_{total}$). To statistically test for changes between the pre-intervention and post-intervention chats (b_0, b_1), we used paired t-tests and computed Cohen's d effect sizes. To test for significant differences between P and P+F groups ($b_1^P - b_0^P$ vs. $b_1^{PF} - b_0^{PF}$), we used non-paired t-tests. In total, we conducted 12 planned t-tests for analyzing changes in behaviors: three skills (Empathy, Reflections, Questions) with strong uses and one skill (Suggestions) with areas needing improvement, each analyzed for both within-group changes and between-group differences in those changes.

Exposure to Good Alternatives in AI Feedback. To better understand how AI feedback supported skill development, we analyzed whether exposure to feedback during practice affected post-intervention performance. We defined this exposure as **G**ood **A**lternatives during **P**ractice (GAP): the proportion of trainee utterances for which the AI mentor suggested an alternative response that demonstrated a strong use of a counseling skill, normalized by the total number of user utterances ($T_{GAP} = A_{strengths}/U_{total}$). Here, $A_{strengths}$ is the number of AI-generated alternative responses judged as strong exemplars of a skill, and U_{total} is the total number of trainee utterances in the session. GAP thus quantifies the degree to which a trainee was shown effective skill applications in context, allowing us to test whether greater exposure predicted better post-intervention performance. To test whether greater exposure to GAP (positive examples of skill usage) led to stronger use of that skill, we fit a lagged-linear regressor that predicts post-chat behaviors, controlling for pre-chat behaviors, with exposure to good alternatives (T_{GAP}) as a scalar predictor ($b_1 = \beta_0 + \beta_1 b_0 + \beta_2 T_{GAP}$). This model allowed us to isolate the effect of feedback exposure while accounting for baseline performance differences.

4.5.2 ***RQ2. Effects on Self-Efficacy and its (mis)calibration with Behavioral Performance.*** First, we test for changes in raw self-efficacy scores after practice or practice-and-feedback. Second, we examine the calibration of self-efficacy ratings with actual performance. Third, we evaluate whether P or P+F interventions improve this calibration.

Changes in Raw Self-Efficacy. Beyond calibration, we also investigated whether the interventions affected participants' absolute levels of self-efficacy across the three measured dimensions (exploration skills, action skills, and session management skills). We conducted repeated-measures analyses to identify: (1) significant pre-post changes in raw self-efficacy scores following practice alone (P intervention); (2) significant pre-post changes in raw self-efficacy scores following practice with structured feedback (P+F intervention); and (3) differential patterns of change between the P and P+F groups, indicating potential intervention-specific effects on self-efficacy development.

Investigating (mis)calibration of Self-Efficacy. Our primary analysis investigated potential mis-calibration between participants' self-assessments and their actual counseling performance, specifically examining whether data exhibited patterns consistent with the Dunning-Kruger effect. This phenomenon [45] suggests that individuals with lower skill levels tend to overestimate their abilities, while highly skilled individuals may slightly underestimate their competence. We focused this analysis on Exploration Skills and Action Skills, as these dimensions had straightforward mappings between CASES items and our NLP behavioral classifiers (Table 4). To test for the presence of the Dunning-Kruger effect, we follow the classic analysis method that splits the data into quartiles based on performance and conducts a two-way analysis of variances for self-assessments and actual performance across the quartiles; and finally verifies via post-hoc tests that the bottom performers have the biggest overestimation of their abilities [45]. To standardize the comparison between self-efficacy and performance, we transform each measure into a percentile rank (0 - 99) computed across all data collected for the pre-intervention and post-intervention chats (b_0, b_1; s_0, s_1).

Changes in Calibrated Self-Efficacy. To evaluate whether our interventions improved self-efficacy calibration, we computed discrepancy scores by subtracting standardized performance scores from standardized self-efficacy scores for each participant at both assessment timepoints. These discrepancy metrics provided a direct measure of calibration, with positive values indicating overconfidence and negative values indicating underconfidence. We then examined changes in these discrepancy scores from pre- to post-intervention for both intervention groups, to determine whether either intervention improved the alignment between participants' self-perceptions and their actual counseling abilities.

4.5.3 ***RQ3. Effects on Therapeutic Intentions.*** To analyze how therapeutic intentions were impacted by using CARE's practice interventions (P vs. PF), three authors conducted a thematic analysis [12] on participants' reflections after the 20-minute practice intervention, regarding what they would do differently. To make qualitative coding of the 94 transcripts more feasible, we used timestamped notes taken by one of the co-authors during that session to synthesize codes and identify direct quotes in the transcript recordings. An initial set of codes was organized according to the Helping Skills taxonomy that our LLM multi-level feedback model was based on; see Table 2. Building on this, we inductively coded novice counselors' qualitative reflections about what they felt they did well and what they wanted to improve, which resulted in the following set of codes: *"Empathy, Validation, Action Plan, Active Listening, Questions / Asking Open-Ended, Providing Suggestions, Building Trust / Connection, Confidence / Personal Growth, Positive Re-framing / Affirmations, Reflection, Self-Disclosure, Professionalism, Personalization, and Nothing to improve"*. Three co-authors applied these codes to the qualitative data, first splitting the data independently, and having an additional co-author double-check the grouping of the codes. Several rounds of discussion between coders was required in order to split codes that were too vague or led to coding disagreements. A record of this qualitative coding is provided in the Appendix

for the P+F group (Table 8) and the P group (Table 9). Finally, our findings about shifts in therapeutic intentions were organized around a higher-level set of qualitative themes that were synthesized from these codes, informed by the literature on therapeutic intentions and microskills relevant to client-centered psychotherapy approaches [36, 79].

4.5.4 ***RQ4. Perceptions of CARE's LLM components and the Training Experience***. For the Likert survey questions, we report descriptive statistics of all Likert measures that capture participants' perceptions of CARE. Consistent with recent papers analyzing the convergent validity of the NASA-TLX instrument in HCI [7], we consider it as a multivariate construct in our analysis. For analyzing the interviews, we conduct a deductive thematic analysis [12] where high-level themes closely match the interview protocol, which asked about likes, wishes, and suggestions for improvement. We further organize in terms of the core components of LLM simulated practice and LLM feedback, system usability, system functionality, how users compare CARE to their current educational experiences, and how they would want to use it as a novice counselor in the future.

5 Results

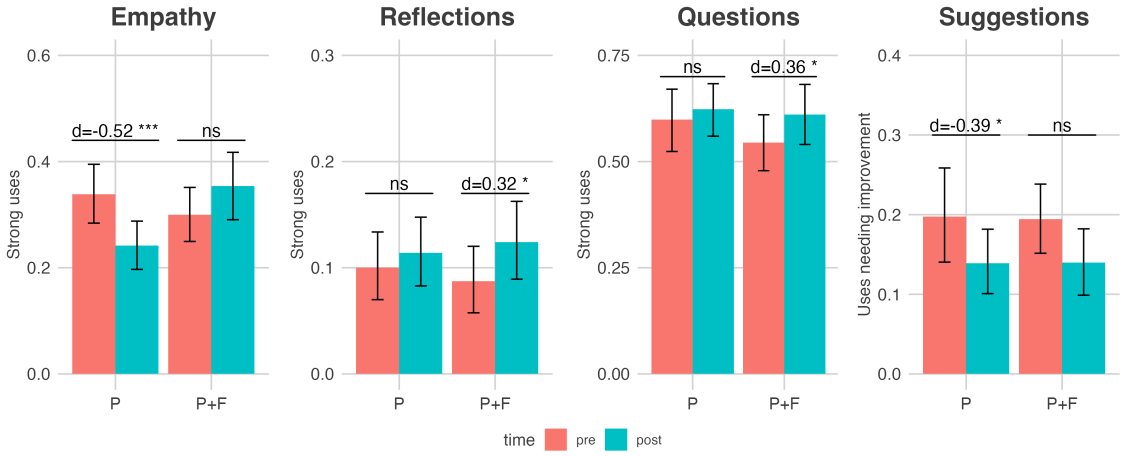
5.1 RQ1. Effects on Behavioral Performance

We find that practice alone is not enough; feedback during practice is necessary to promote desirable counseling behaviors in empathetic and active listening. AI feedback during practice (P+F) led to improvements in Reflections (+3.6% change, $p = 0.034$) and Questions (+6.59% change, $p = 0.018$), and trended towards improvement in Validation (+3.6% change, $p = 0.083$), Suggestions (-5.45% change, $p = 0.057$) and Empathy (+5.37% change, $p = 0.117$). In contrast, practice alone (P) led to only improvements in Suggestions (-5.85% change, $p = 0.011$), but no improvements in Reflections, Questions, and Validation, and significantly worse expression of Empathy (-9.6% change, $p < 0.001$). Pairwise comparisons between the two conditions were substantial and significant for Empathy (15% relative difference, Cohen's $d = 0.72, p < 0.001$) and non-significant for others.

To better understand AI feedback's role, we further analyzed how behavioral performance is impacted by exposure to specific feedback during the practice-intervention. **For empathy skills, exposure to alternatives with strong uses of empathy during training significantly predicted post-intervention empathy scores ($\beta_2 = 0.204, p = .018$).** However, exposure to good alternatives did not significantly predict improvement in other counseling skills (Reflections: $\beta_2 = 0.049, p = .440$; Questions: $\beta_2 = 0.046, p = .523$). This suggests that the effectiveness of AI feedback alternatives varies by skill type, with empathy skills appearing more responsive than reflections or question skills.

5.2 RQ2. Self-Efficacy and Its Miscalibration with Behaviors

In our analysis of raw self-efficacy scores, we find modest overall increases in self-efficacy after P and P+F interventions, with different patterns of improvement across skills (Fig. 4). For the P group, confidence in exploration skills showed a significant increase (0.36 points on an 8-point scale, $d = 0.44, p = 0.004$). Confidence in session management skills showed a substantial increase for the P+F group (0.36 points, $d = 0.39, p = 0.011$). While session management skills for the P group also trended towards improvement, it was not significant after correcting for multiple hypothesis testing (0.35 points, $d = 0.35, p = 0.021$). Similarly, while confidence in action skills for the P+F group also increased, this result was not significant after correction of multiple hypothesis tests (0.33 points, $d = 0.34, p = 0.026$). Finally, we found no significant differences between participants who received AI feedback (P+F) versus those who did not (P) (across the three self-efficacy subscales, $d = -0.25, 0.03, 0.01$, $p = 0.238, 0.884, 0.955$).

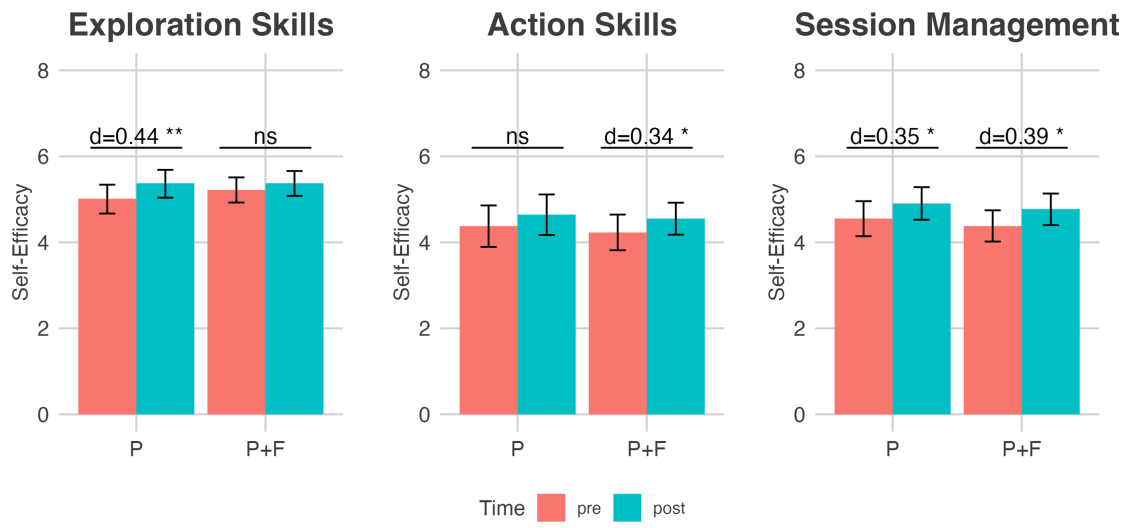


	Change after P			Change after PF			Differences in change after P vs. P+F		
Use of Skill	%	*p*-value	*d*	%	*p*-value	*d*	%	*p*-value	*d*
Empathy (↑)	**-9.6**	**0.001**	**-0.52**	5.4	0.117	0.23	**15**	**0.001**	**0.72**
Reflections (↑)	1.4	0.391	0.18	3.7	0.034	0.32	2.3	0.323	0.2
Questions (↑)	2.4	0.421	0.12	6.6	0.018	0.36	4.1	0.296	0.22
Suggestions (↓)	**-5.9**	**0.010**	**-0.39**	-5.5	0.057	-0.28	0.4	0.910	0.02

Fig. 3. Changes in counseling behaviors following AI patient simulations alone (P) versus AI patient simulations with AI feedback (P+F). The plot displays bootstrapped means for pre-intervention and post-intervention interactions. The table presents statistical comparisons with corresponding effect sizes, with **bolded** values indicating significance after Benjamini-Hochberg correction [10]. Notably, the P group experiences a significant drop in strong uses of Empathy (-9.6% change, $d = -0.52$), whereas the P+F group's use of Empathy is maintained and trends towards improvement. The P+F group also experiences noticeable improvements in Reflections (+3.7% change, $d = 0.32$) and Questions (6.59% change, $d = 0.36$)

Our analysis comparing self-efficacy ratings with actual performance across skill quartiles finds support for the Dunning-Kruger effects more substantially for action skills and to a lesser degree for exploration skills. The interaction between measure and quartile was significant in four out of six ANOVAs for action skills, while only one out of six was significant for exploration skills (Table 5). Pairwise comparisons also showed a pattern indicative of a Dunning-Kruger effect (see Table 6, Table 7, and Fig. 5): People in the lowest quartile overestimated themselves the most. Those in the highest quartile—and to a lesser degree also those in the second-to-highest quartile—tended to underestimate themselves.

Participants' Ability to Self-Assess Their Skill Level Remained Mixed After LLM Practice. For the practice only (P) group, the mean discrepancy in exploration skills changes from 11.6 percentile underconfidence to 5.7 percentile overconfidence, a significant shift ($p < 0.001, d = 0.58$). Besides this, we found no other significant changes in calibration. No significant differences was found among the P group for discrepancy in action skills ($p = 0.227, d = 0.18$). The P+F group showed no significant calibration changes for exploration skills ($p = 0.479, d = -0.10$) or for action skills



Self-Efficacy	Change after P			Change after PF			Differences in change after P vs. PF		
	Pts.	p-value	d	Pts.	p-value	d	Pts.	p-value	d
Exploration Skills	**0.36**	**0.004**	**0.44**	0.13	0.338	0.14	-0.21	0.238	-0.25
Action Skills	0.27	0.166	0.21	0.33	0.026	0.34	0.04	0.884	0.03
Session Management	0.35	0.021	0.35	**0.36**	**0.011**	**0.39**	0.01	0.955	0.01

Fig. 4. Changes in raw-scores of self-efficacy following AI patient simulations alone (P) versus AI patient simulations with AI feedback (PF). The plot displays bootstrapped means for pre-intervention and post-intervention. The table presents statistical comparisons with corresponding effect sizes, with **bolded** values indicating significance after Benjamini-Hochberg correction [10] for the 21 planned comparisons (12 for behavioral changes and 9 for self-efficacy changes).

($p = 0.393, d = 0.13$). Finally, between-group differences were not significant for discrepancy in exploration skills ($p = 0.191, d = 0.27$) or action skills ($p = 0.743, d = 0.07$).

5.3 RQ3. Qualitative Self-Reflections on Therapeutic Intentions

Two key themes emerged for how novice counselors' therapeutic intentions were impacted by training with CARE:

(1) The P+F group expressed greater intentions and successes in improving their use of empathy and listening skills. P+F participants reported effectively using empathy (27%), validation (27%), and open-ended questions (52%). They emphasized the value of listening skills, such as reflective responses to signal understanding: *"I should rephrase what they say so they know I'm understanding them"* (P51). They also recognized that counseling should support client exploration of thoughts and emotions, rather than provide direct solutions. One participant reflected on this shift: *"I asked them to expand on their feelings, rather than guiding them to my idea"* (P39). AI feedback encouraged this shift toward providing emotional support and fostering client autonomy, helping counselors adopt a more empathetic and client-centered approach. **(2) The P participants remained solution-oriented but changed their approach to first**

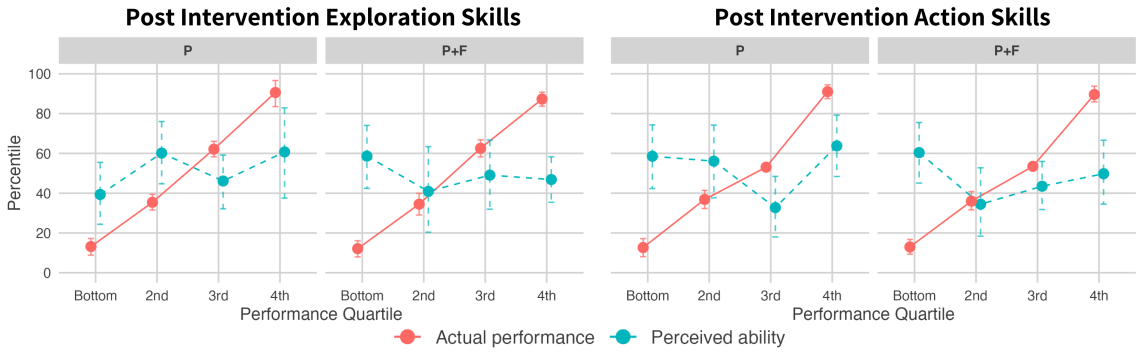


Fig. 5. Counselor Self Efficacy (perceived ability to use skills) for participants grouped by behaviors of skills used (actual performance). Notes: Gaps depict miscalibration between actual and self-assessed percentile of performance for quartile groups with bootstrapped 95% CIs. We only visualize data collected in the post-intervention.

gather information. While P+F participants intended to give fewer suggestions, many P participants continued to view suggestions as a central skill. In the post-intervention, 48% of P participants reported using suggestions successfully, compared to only 14% of P+F participants. Many P participants justified their continued use of suggestions by citing a desire to provide tangible, actionable help, for example, *"Maybe because I'm untrained and solution oriented. I do not want to leave them with nothing, and nowhere to go"* (P40). Some reflected on modifying how they delivered suggestions, emphasizing strategies like gathering more information to tailor advice or providing more concrete guidance. However, in the absence of feedback, most remained fixed in their approach, with some even reporting efforts to rephrase the same solution repeatedly to persuade the client.

5.4 RQ4. Perceptions of CARE's LLM components and the Training Experience

5.4.1 Quantitative Perceptions of Training with AI Patients. Descriptive statistics of participants' perceptions are shown in Table 3. **Most participants felt that AI patients in CARE were realistic.** Across all three AI patient scenarios, the vast majority of participants (88–92 out of 94) rated the AI patients as authentic in their roles, with scores of 5, 6, or 7 on the 7-point Likert scale. Authenticity ratings were consistently high across scenarios (μ = 6.1–6.3, σ = 0.9–1.0). **Participants consistently found the AI patient most challenging during the intervention phase across multiple measures.** Mental demand peaked during intervention, with 82% of participants rating it as moderate to high (μ = 5.5, σ = 1.2) compared to 68% pre-intervention (μ = 4.7, σ = 1.4) and 56% post-intervention (μ = 4.5, σ = 1.4). Similarly, perceived effort required was highest during intervention, with 81% rating it as moderate to high (μ = 5.4, σ = 1.3) versus 64% pre-intervention (μ = 4.8, σ = 1.4) and 65% post-intervention (μ = 4.8, σ = 1.3). Frustration levels, while lower overall, also peaked during intervention with 40% reporting moderate to high frustration (μ = 3.9, σ = 1.4) compared to 35% pre-intervention (μ = 3.6, σ = 1.5) and 30% post-intervention (μ = 3.5, σ = 1.6). This consistent pattern across these NASA-TLX measures suggests that either the extended intervention practice imposed the greatest demands on participants, or that the intervention simulated patient was the most challenging scenario for our subject pool of novice counselors. **Participants felt progressively less hurried or rushed across intervention phases.** 45% (42/94) rated feeling moderately to highly hurried or rushed ($\geq$ 5 on the 7-point Likert scale) during pre-intervention (μ = 3.9, σ = 1.8), compared to 31% (29/94) during the intervention phase (μ = 3.6, σ = 1.7) and 30% (28/94) in the

	Pre-Intervention		Intervention		Post-Intervention	
Measure	μ	σ	μ	σ	μ	σ
Authentic in role	6.1	1.0	6.2	1.0	6.3	0.9
Mental Demand	4.7	1.4	5.5	1.2	4.5	1.4
Temporal Demand	3.9	1.8	3.6	1.7	3.3	1.8
Effort	4.8	1.4	5.4	1.3	4.8	1.3
Frustration	3.6	1.5	3.9	1.4	3.5	1.6
Confidence to help similar patient	4.8	1.3	4.8	1.4	5.3	1.4

Table 3. Perceptions of training with CARE's AI patients across the three study phases. AI Patient 1 (Pre-Intervention) was the 35-year-old American Male who was feeling alone after a holiday; AI Patient 2 (Intervention) was the 35-year-old Male Veteran who had substance use and legal issues retaining custody of his kids; AI Patient 3 (Post-Intervention) was the young adult with family issues who had low mood and self-esteem. 7 Point Likert Scale.

post-intervention phase ($\mu = 3.3$, $\sigma = 1.8$). This decreasing trend suggests that participants became more comfortable with the pacing of AI patient interactions over time.

5.4.2 Quantitative Perceptions of CARE Feedback. **Novice counselors in our study had consistently positive perceptions of CARE's generated feedback across multiple dimensions.** The majority of participants (76%) found the AI feedback constructive and helpful ($\mu = 4.1$ out of 5, $\sigma = 0.8$), while an even higher proportion (84%) reported being comfortable receiving feedback from the AI system ($\mu = 4.4$ out of 5, $\sigma = 0.9$). Additionally, 72% agreed that the AI feedback system is ready for use by counselors or helpers-in-training ($\mu = 3.8$ out of 5, $\sigma = 1.0$). These consistently high ratings across helpfulness, comfort, and readiness measures suggest strong overall acceptance of AI-generated feedback among novice counselors.

5.4.3 Qualitative Perception of CARE's AI Patients. **Users found the AI patients' behaviors provided challenging practice scenarios.** The patients were prompted to resist advice and suggested actions. Many participants reported that AI patients seemed to be *"in a loop"* and were less likely to accept suggestions from counselors. Novice counselors had diverging opinions about working with these resistant patients. Some welcomed the resistance as valuable practice: *"...if you don't have that experience yet, and you're going into the counseling world, and [encounter a patient] that is resistant, I feel like that would cause a lot of people to shut down, and I think [CARE's AI patients] is... a better way to ease into working with a resistant patient"* (P97). However, other novices experienced discouragement when trying to support such resistant patients: *"They were very cold, it was hard to communicate with this person...I'm a tiny bit discouraged in myself, the patient was not taking what I was suggesting very well"* (P44B). This underscores that novice counselors' individual readiness to handle difficult cases directly affects whether they perceive these simulations as valuable learning experiences. Beyond their reactions to the content of patient responses, participants also reported that AI patients responded very rapidly during their interactions: *"They respond so quickly. So you feel a kind of pressure to respond, like you have to keep the conversation going"* (P22). This swift response time created temporal demands on the participants, who felt compelled to match the AI's quick pace, leading some to feel rushed during the conversations.

5.4.4 Qualitative Perceptions of CARE Feedback. **Participants liked that CARE's AI feedback helped them refine responses by suggesting alternative wording and structure.** It offers alternative phrases and sentence structures, especially in situations where participants feel unsure, helping them choose the right words to convey empathy while avoiding assumptions. For example, a novice counselor explained *"Alternative responses helped with validation... being*

supportive and addressing concerns before moving on immediately with another question... I liked seeing the alternative response to show how I could lengthen it to make it more broader, and more supportive than the version I went with" (P107). Some participants checked the feedback frequently throughout the chat, while some commented that they wish they had checked the feedback more frequently, since *"I could have course corrected earlier"* (P17).

Interacting with CARE's AI Feedback encourages self-reflection and learning. Participants used the AI feedback periodically to check how they were doing. According to our participants, pausing to think about their responses helped them become more aware of their strengths and weaknesses. *"It helps you to reflect on what you are saying to people; even if you don't agree, it stops you to reflect on your response"* (P24B). Participants who had many strengths during intra-training reflected that the *"AI feedback helped me stay on task. I wanted to be more empathetic.. [the AI feedback strengths told me] yes, this still does sound like an empathetic response... that was really helpful... to feel I was going in the right direction"* (P55). Additionally, appreciating their strengths through positive reinforcement from the AI built trust and encouraged them to make improvements in future interactions. Many reported that they were able to pick up effective strategies from AI Feedback and become more self-aware of their strengths and weaknesses; by keeping mental notes of what to improve, they were able to incorporate those skills in the upcoming conversations.

6 Discussion and Takeaways

This study investigated the impact of practicing with an LLM-simulated patient either with or without receiving LLM-generated feedback on counselor skills development, resulting in three main findings. First, our behavioral assessments showed that practice with feedback improves empathetic listening skills, while practice alone shows minimal improvement, as evidenced by decreased use of empathy. Second, our qualitative analysis of self-reflections revealed distinct skill development strategies, with feedback recipients more frequently reporting the adoption of client-centered approaches focused on showing empathy and exploring patients' thoughts and feelings, while practice-only participants gravitated toward solution-oriented approaches focusing on gathering more information and providing suggestions. Both these findings highlight that the development of counseling skills requires not only practice opportunities but also structured feedback that guides learners toward empathetic, client-centered approaches. Finally, participants demonstrated poor calibration between their perceived abilities and actual performance, as evidenced by overestimates of self-efficacy for the lowest quartile performers. This underscores how self-efficacy measures may not reliably indicate skill development. Each of these findings merits further discussion.

Our findings demonstrate that teaching counselors to implement client-centered approaches requires more than just simulated practice opportunities—it requires targeted feedback that guides novices away from their natural solution-oriented tendencies. The fully-featured version of CARE–combining LLM-simulated patient practice and LLM-based feedback–helped participants improve their use of empathetic and active listening skills, with notable improvements in questions ($d = 0.36$) and reflections ($d = 0.32$). In comparison, practice with an AI patient alone only led to fewer inappropriate suggestions ($d = -0.39$), but no improvements in reflections or questions, and significantly worse uses of empathy (-9.6% change, $d = -0.52$; 15% relative difference to P+F, $d = 0.72$). These effect sizes are comparable to those found in studies of human supervision during standardized roleplays, where Maaß et al. (2025) [59] reported observer-rated skill improvements with effect sizes ranging from $d = 0.29 - 0.49$. Since LLM-simulated practice and feedback are not bottlenecked by the resource constraints of human trainers, AI training systems like CARE show promise in scaling access to effective counseling training.

While the combination of feedback and simulated practice was key to CARE's success, we also recognize that CARE represents just one approach to providing expert-aligned LLM feedback to learners. CARE allows trainees to

request feedback on responses they have already sent. This approach allows trainees to experience productive learning discomfort by trying and making mistakes in simulation, while still offering feedback upon request to refine future responses. This design choice parallels elements of both live human supervision [60] and delayed supervision after a session [59], creating a hybrid approach that may offer flexibility based on trainee preferences. Given the promise of feedback combined with practice, the AI for psychotherapy field should evaluate different AI-feedback designs against each other. Several other possible strategies for giving feedback include just-in-time skill suggestions [37, 54], counterfactual simulations [88], and global session feedback. While conducting randomized trials in medical education research can be tricky [99], randomized trials delivering different training interventions to large participant-pools for longer durations [43] could provide essential data to determine optimal approaches for different learning objectives and learner characteristics.

Our study revealed that participants' self-efficacy ratings were poorly calibrated with their actual performance, especially among lower performers. This finding, consistent with prior research, suggests that self-assessment accuracy alone may not be a reliable indicator of counselor competence or development. Recent reviews indicate that efforts to improve self-assessment calibration have limited impact on learning or performance outcomes [114]. Instead, it is valuable to objectively assess specific standards of performance and skill use [33, 60] and design interventions that can help low performers improve on those metrics while maintaining a positive morale for continued practice. As AI-based training tools evolve, integrating objective performance measures and structured self-reflection, rather than relying solely on self-assessment, offers a more robust approach to supporting counselor development.

7 Limitations and Future Work

Several limitations should be considered, including methodological constraints in our assessment approach, the representativeness of our educational context, and the generalizability of results across therapeutic modalities.

Methodological Constraints of Behavioral Assessment. Our automated assessment approach employed fine-tuning methods that used a subset of participant data for model development, raising potential concerns about data leakage and overfitting. Following standard practices in computational social science, we used domain-specific data to adapt our models while employing a validation set comprising n=409 utterances from external counseling transcripts combined with n=370 expert-annotated utterances from this study to monitor performance and prevent overfitting. However, this approach may limit the generalizability of our automated feedback models to entirely novel populations or contexts.

Beyond these technical constraints, our behavioral analysis was limited to utterance-level microskill measures and could not capture observable session-level characteristics that may be important for comprehensive skill assessment. While traditional studies have employed human observers to provide such ratings, recent AI research has shown the validity of using fine-tuned LLMs in certain contexts to approximate session-level measures, such as working alliance [52], which might enable scalable behavioral analyses of broader skill development constructs. Finally, regardless of the measurement approach employed, our pre-post randomized study focused primarily on assessing immediate skill acquisition rather than long-term retention or transfer to real-world clinical encounters with actual patients. While immediate changes demonstrate short-term learning effects, longer-term retention measures would provide stronger evidence of true skill acquisition and clinical relevance.

Limited Evaluation Across Training Contexts. To first understand the effectiveness of our platform in a controlled environment, our study was conducted in a controlled laboratory setting with bachelor-level counselors recruited through Prolific. However, a longer-term consideration is how LLM-based training would perform across the diverse

landscape of real-world counseling education. We did not evaluate our approach within actual training programs, whether traditional degree-based counseling programs with human supervision and peer roleplay, or alternative training contexts such as targeted programs for volunteer peer counselors in online mental health communities (e.g., 7 Cups, Crisis Text Line) who lack access to formal supervision but provide critical frontline support [106, 113]. Without direct comparisons to established training methods or evaluation within authentic educational settings, we cannot determine the relative effectiveness, acceptability, or practical integration challenges of AI-enhanced training. Future work should embed LLM-training tools across these diverse training contexts to assess their utility for both traditional counseling students and underserved populations who could benefit from scalable training opportunities.

Generalizability of Results across Therapeutic Modalities. Our findings are constrained by the specific therapeutic approach and communication modality examined. The efficacy of LLM-based practice and feedback training was demonstrated only for client-centered microskills, which represent foundational communication techniques that may serve as a base for therapeutic practice. However, it remains unclear how these results would generalize to specialized therapy modalities such as psychodynamic therapy, cognitive-behavioral therapy (CBT), or acceptance and commitment therapy (ACT), each of which has distinct theoretical frameworks and adherence protocols that require specific therapeutic techniques beyond basic microskills. Future research can determine whether foundational microskill training provides a transferable foundation for modality-specific practices or whether LLM training systems would need substantial modification to accommodate the unique requirements and intervention strategies of different therapeutic approaches. Additionally, our findings are limited to text-based interactions and may not fully capture the nonverbal and paraverbal components of empathy essential in face-to-face therapy settings. While the growing prevalence of text-based mental health services (e.g., crisis text lines, online therapy platforms) makes training linguistic empathy skills clinically relevant, complete therapeutic competence requires multimodal communication skills. Future work could extend this approach to incorporate voice, facial expressions, and other nonverbal therapeutic skills, building on successful models that process non-text signals for clinical training [9, 55], to determine whether text-based empathy training provides a foundation that transfers to verbal and nonverbal communication.

8 Conclusion

In this work, we conducted a randomized study of 94 novice counselors using an LLM-simulated practice and feedback system. Despite increasing interest in using LLMs in mental health, to our knowledge, this is the first study to conduct a large-scale evaluation ($N = 94$) of an LLM-based training system for developing core skills in novice counselors. Our findings show that, perhaps surprisingly, simulated practice *alone* proved insufficient—and in the case of empathy, potentially harmful— at improving therapeutic skills, simulated practice with AI-generated feedback supported measurable improvements in key counseling skills of demonstrating empathy, delivering reflective responses, and engaging in client-centered inquiry. By combining realistic patient simulations with expert-aligned, skill-specific feedback, LLM-based training can help novices to master skills involved in client-centered therapy, offering a scalable, evidence-aligned training in mental health care.

References

[1] Muhammad Abbas, Farooq Ahmed Jam, and Tariq Iqbal Khan. 2024. Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education* 21, 1 (2024), 10.

[2] Glenn Albright, Cyrille Adam, Deborah Serri, Seth Bleeker, and Ron Goldman. 2016. Harnessing the power of conversations with virtual humans to change health behaviors. *Mhealth* 2 (2016), 44.

[3] Guillaume Alinier and Denis Oriot. 2022. Simulation-based education: deceiving learners with good intent. *Advances in Simulation* 7, 1 (March 2022), 8. doi:10.1186/s41077-022-00206-3

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[5] David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science* 9, 1 (2014), 1–11.

[6] Destina Sevde Ay-Bryson, Florian Weck, and Franziska Kühne. 2023. Can students in simulation portray a psychotherapy patient authentically with a detailed role-script? Results of a randomized-controlled study. *Training and Education in Professional Psychology* 17, 1 (2023), 89.

[7] Ebrahim Babaei, Tilman Dingler, Benjamin Tag, and Eduardo Velloso. 2025. Should we use the NASA-TLX in HCI? A review of theoretical and methodological issues around Mental Workload Measurement. *International Journal of Human-Computer Studies* (2025), 103515.

[8] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakcı, and Rei Mariman. 2024. Generative ai can harm learning. *Available at SSRN* 4895486 (2024).

[9] Manas Satish Bedmutha, Anuujin Tsedenbal, Kelly Tobar, Sarah Borsotto, Kimberly R Sladek, Deepansha Singh, Reggie Casanova-Perez, Emily Bascom, Brian Wood, Janice Sabin, et al. 2024. Conversense: An automated approach to assess patient-provider interactions using social signals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–22.

[10] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.

[11] Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024. Human-centered evaluation of language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. 39–43.

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa doi:10.1191/1478088706qp063oa

[13] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. 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 (2021), 1–21.

[14] Katie A Burmester, Jai P Ahluwalia, Robert J Ploutz-Snyder, and Stephen Strobbe. 2019. Interactive computer simulation for adolescent screening, brief intervention, and referral to treatment (SBIRT) for substance use in an undergraduate nursing program. *Journal of pediatric nursing* 49 (2019), 31–36.

[15] Doğan Can, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2012. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. In *Thirteenth Annual Conference of the International Speech Communication Association*.

[16] Alicja Chaszczewicz, Raj Sanjay Shah, Ryan Louie, Bruce A Arnow, Robert Kraut, and Diyi Yang. 2024. Multi-Level Feedback Generation with Large Language Models for Empowering Novice Peer Counselors. *arXiv preprint arXiv:2403.15482* (2024).

[17] Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614* (2023). doi:10.48550/arXiv.2305.13614

[18] Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. *arXiv preprint arXiv:2310.11501* (2023).

[19] Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A Computational Framework for Behavioral Assessment of LLM Therapists. *arXiv preprint arXiv:2401.00820* (2024). doi:10.48550/arXiv.2401.00820

[20] Sarah C Cook, Ann C Schwartz, and Nadine J Kaslow. 2017. Evidence-based psychotherapy: Advantages and challenges. *Neurotherapeutics* 14 (2017), 537–545.

[21] Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. *arXiv preprint arXiv:2405.18638* (2024).

[22] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy* 55, 4 (2018), 399.

[23] K Anders Ericsson et al. 2006. The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge handbook of expertise and expert performance* 38, 685-705 (2006), 2–2.

[24] Kevin W Eva and Glenn Regehr. 2005. Self-assessment in the health professions: a reformulation and research agenda. *Academic medicine* 80, 10 (2005), S46–S54.

[25] Christopher G Fairburn and Zafra Cooper. 2011. Therapist competence, therapy quality, and therapist training. *Behaviour research and therapy* 49, 6-7 (2011), 373–378.

[26] Anna Fang, Wenjie Yang, Raj Sanjay Shah, Yash Mathur, Diyi Yang, Haiyi Zhu, and Robert Kraut. 2023. What Makes Digital Support Effective? How Therapeutic Skills Affect Clinical Well-Being. *arXiv preprint arXiv:2312.10775* (2023).

[27] Nikolaos Flemotomos, Victor R Martinez, Zhuohao Chen, Torrey A Creed, David C Atkins, and Shrikanth Narayanan. 2021. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. *PloS one* 16, 10 (2021), e0258639.

[28] Hannah E Frank, Emily M Becker-Haimes, and Philip C Kendall. 2020. Therapist training in evidence-based interventions for mental health: A systematic review of training approaches and outcomes. *Clinical psychology: Science and practice* 27, 3 (2020), 20.

[29] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of management annals* 14, 2 (2020), 627–660.

[30] Daniela Hahn, Florian Weck, Michael Witthöft, and Franziska Kühne. 2021. Assessment of counseling self-efficacy: validation of the German Counselor Activity Self-Efficacy scales-revised. *Frontiers in psychology* 12 (2021), 780088.

[31] Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 129–133.

[32] Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. 2025. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI* 2, 4 (2025), AIoa2400802.

[33] Peter Eric Heinze, Florian Weck, Ulrike Maaß, and Franziska Kühne. 2024. The relation between knowledge and skills assessments in psychotherapy training: Secondary analysis of a randomized controlled trial. *Training and Education in Professional Psychology* 18, 2 (2024), 162.

[34] Clara E Hill. 2020. *Helping skills: Facilitating exploration, insight, and action*. American Psychological Association.

[35] Clara E Hill and Ian S Kellems. 2002. Development and use of the helping skills measure to assess client perceptions of the effects of training and of helping skills in sessions. *Journal of Counseling Psychology* 49, 2 (2002), 264.

[36] Clara E Hill and Emilie Y Nakayama. 2000. Client-centered therapy: where has it been and where is it going? A comment on Hathaway (1948). *Journal of Clinical Psychology* 56, 7 (2000), 861–875.

[37] Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2025. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–45.

[38] Xiaolei Huang, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Modeling temporality of human intentions by domain adaptation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 696–701.

[39] Juan Enrique Huerta-Wong and Richard Schoech. 2010. Experiential learning and learning environments: The case of active listening skills. *Journal of Social Work Education* 46, 1 (2010), 85–101.

[40] Joanna Joy Hunsmann, Destina Sevde Ay-Bryson, Scarlett Kobs, Nicole Behrend, Florian Weck, Michel Knigge, and Franziska Kühne. 2024. Basic counseling skills in psychology and teaching: validation of a short version of the counselor activity self-efficacy scales. *BMC psychology* 12, 1 (2024), 32.

[41] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[42] Aman Khullar, Nikhil Nalin, Abhishek Prasad, Ann John Mampilli, and Neha Kumar. 2025. Nurturing Capabilities: Unpacking the Gap in Human-Centered Evaluations of AI-Based Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

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

[44] Samuel Knapp, Michael C Gottlieb, and Mitchell M Handelsman. 2017. Self-awareness questions for effective psychotherapists: Helping good psychotherapists become even better. *Practice Innovations* 2, 4 (2017), 163.

[45] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[46] Franziska Kuehne, Destina Sevde Ay, Mara Jasmin Otterbeck, and Florian Weck. 2018. Standardized patients in clinical psychology and psychotherapy: A scoping review of barriers and facilitators for implementation. *Academic Psychiatry* 42 (2018), 773–781.

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

[48] Eric H. Larson, Davis G. Patterson, Lisa A. Garberson, and C. Holly A. Andrilla. 2016. *Supply and Distribution of the Behavioral Health Workforce in Rural America*. Data Brief 160. Rural Health Research Center, WWAMI Rural Health Research Center. https://www.ruralhealthresearch.org/

[49] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–35.

[50] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746* (2022).

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

[52] Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024. Understanding the Therapeutic Relationship between Counselors and Clients in Online Text-based Counseling using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1280–1303. doi:10.18653/v1/2024.findings-emnlp.69

[53] Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100* (2023).

[54] Inna Wanyin Lin, Ashish Sharma, Christopher Michael Rytting, Adam S Miner, Jina Suh, and Tim Althoff. 2024. 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

[55] Chunfeng Liu, Karen M Scott, Renee L Lim, Silas Taylor, and Rafael A Calvo. 2016. EQClinic: a platform for learning communication skills in clinical consultations. *Medical education online* 21, 1 (2016), 31801.

[56] Siyang Liu, Bianca Brie, Wenda Li, Laura Biester, Andrew Lee, James Pennebaker, and Rada Mihalcea. 2025. Eeyore: Realistic Depression Simulation via Supervised and Preference Optimization. arXiv:2503.00018 [cs.CL] https://arxiv.org/abs/2503.00018

[57] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10570–10603. doi:10.18653/v1/2024.emnlp-main.591

[58] Qianou Ma, Dora Zhao, Xinran Zhao, Chenglei Si, Chenyang Yang, Ryan Louie, Ehud Reiter, Diyi Yang, and Tongshuang Wu. 2025. SPHERE: An Evaluation Card for Human-AI Systems. *arXiv preprint arXiv:2504.07971* (2025).

[59] Ulrike Maaß, Klara Eisert, Jasmin Ghalib, Franziska Kühne, and Florian Weck. 2025. Live versus delayed supervision: A randomized controlled trial with psychology students. *Psychotherapy* (2025).

[60] Ulrike Maaß, Franziska Kühne, Destina Sevde Ay-Bryson, Peter Eric Heinze, and Florian Weck. 2024. Efficacy of live-supervision regarding skills, anxiety and self-efficacy: a randomized controlled trial. *The Clinical Supervisor* 43, 1 (2024), 1–21.

[61] Brooke N Macnamara, Ibrahim Berber, M Cenk Çavuşoğlu, Elizabeth A Krupinski, Naren Nallapareddy, Noelle E Nelson, Philip J Smith, Amy L Wilson-Delfosse, and Soumya Ray. 2024. Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications* 9, 1 (2024), 46.

[62] Sruti Mallik and Ahana Gangopadhyay. 2023. Proactive and reactive engagement of artificial intelligence methods for education: a review. *Frontiers in artificial intelligence* 6 (2023), 1151391.

[63] David G Martin and Edward A Johnson. 2024. *Counseling and therapy skills.* Waveland Press.

[64] Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Dan Xu, Paul Watters, and Malka N Halgamuge. 2025. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence* (2025).

[65] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. *Manual for the motivational interviewing skill code (misc). Unpublished manuscript.* Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico.

[66] William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change.* Guilford press.

[67] Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-Aware margIn Ranking for Counselor Reflection Scoring in Motivational Interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 148–158. doi:10.18653/v1/2022.emnlp-main.11

[68] Hemangi Modi, K Orgera, and A Grover. 2022. Exploring barriers to mental health care in the US. *Research and Action Institute* 10 (2022).

[69] Lauren H Moran, Sadie C Kee, Christopher W Wiese, Rosa I Arriaga, Saeed Abdullah, and Andrew M Sherrill. 2025. Artificial Intelligence as a Feedback Teammate for Treatment Delivery: Cognitive Behavioral Therapists' Hopes and Fears. *Cognitive and Behavioral Practice* (2025).

[70] Prasanth Murali, Farnaz Nouraei, Mina Fallah, Aisling Kearns, Keith Rebello, Teresa O'Leary, Rebecca Perkins, Natalie Pierre Joseph, Julien Dedier, Michael Paasche-Orlow, et al. 2022. Training lay counselors with virtual agents to promote vaccination. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. 1–8.

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

[72] Hanh Thi Nguyen. 2003. *The development of communication skills in the practice of patient consultation among pharmacy students.* The University of Wisconsin-Madison.

[73] John C Norcross and Michael J Lambert. 2018. Psychotherapy relationships that work III. *Psychotherapy* 55, 4 (2018), 303.

[74] Julia Othlinghaus-Wulhorst and H. Ulrich Hoppe. 2020. A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skill Training. *Frontiers in Computer Science* 2 (2020). doi:10.3389/fcomp.2020.00028

[75] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 1128–1137.

[76] Verónica Pérez-Rosas, Kenneth Resnicow, Rada Mihalcea, et al. 2022. Pair: Prompt-aware margin ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 148–158.

[77] Verónica Pérez-Rosas, Ken Resnicow, Rada Mihalcea, et al. 2023. VERVE: Template-based ReflectiVE Rewriting for MotiVational IntErviewing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 10289–10302.

[78] Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 926–935.

[79] Nathaniel J Raskin and Carl R Rogers. 2005. Person-centered therapy. (2005).

[80] Benjamin A Rein, Daniel W McNeil, Allison R Hayes, T Anne Hawkins, H Mei Ng, and Catherine A Yura. 2018. Evaluation of an avatar-based training program to promote suicide prevention awareness in a college setting. *Journal of American college health* 66, 5 (2018), 401–411.

[81] Charles R Ridley, Debra Mollen, and Shannon M Kelly. 2011. Beyond microskills: Toward a model of counseling competence. *The Counseling Psychologist* 39, 6 (2011), 825–864.

[82] Eric Rudolph, Hanna Seer, Carina Mothes, and Jens Albrecht. 2024. Automated feedback generation in an intelligent tutoring system for counselor education. In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 501–512.

[83] Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. Towards Motivational and Empathetic Response Generation in Online Mental Health Support. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2650–2656.

[84] Antoinette Schoenthaler, Glenn Albright, Judith Hibbard, and Ron Goldman. 2017. Simulated conversations with virtual humans to improve patient-provider communication and reduce unnecessary prescriptions for antibiotics: a repeated measure pilot study. *JMIR medical education* 3, 1 (2017), e6305.

[85] Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

[86] Craig S Schwalbe, Hans Y Oh, and Allen Zweben. 2014. Sustaining motivational interviewing: A meta-analysis of training studies. *Addiction* 109, 8 (2014), 1287–1294.

[87] Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. Modeling Motivational Interviewing Strategies On An Online Peer-to-Peer Counseling Platform. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.

[88] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating Conflict to Teach Conflict Resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 920, 20 pages. doi:10.1145/3613904.3642159

[89] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*. 194–205.

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

[91] Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 5263–5276. doi:10.18653/v1/2020.emnlp-main.425

[92] Siqi Shen, Veronica Perez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge Enhanced Reflection Generation for Counseling Dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3096–3107. doi:10.18653/v1/2022.acl-long.221

[93] Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 10–20.

[94] Sujin Shin, Jin-Hwa Park, and Jung-Hee Kim. 2015. Effectiveness of patient simulation in nursing education: meta-analysis. *Nurse education today* 35, 1 (2015), 176–182.

[95] Skillsetter. 2024. How it works. https://www.skillsetter.com/how-it-works. Accessed: 29 March 2024.

[96] Logan Stapleton, Jordan Taylor, Sarah Fox, Tongshuang Wu, and Haiyi Zhu. 2023. Seeing Seeds Beyond Weeds: Green Teaming Generative AI for Beneficial Uses. *arXiv preprint arXiv:2306.03097* (2023). doi:10.48550/arXiv.2306.03097

[97] Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. 2025. 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*. 1–22.

[98] Substance Abuse and Mental Health Services Administration. 2024. *Key substance use and mental health indicators in the United States: Results from the 2023 National Survey on Drug Use and Health*. Technical Report HHS Publication No. PEP24-07-021, NSDUH Series H-59. Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. https://www.samhsa.gov/data/report/2023-nsduh-annual-national-report

[99] Gail M Sullivan. 2011. Getting off the "gold standard": Randomized controlled trials and education research. *Journal of graduate medical education* 3, 3 (2011), 285–289.

[100] Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment* 65 (2016), 43–50.

[101] Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and evaluation of ClientBot: Patient-like conversational agent to train basic counseling skills. *Journal of medical Internet research* 21, 7 (2019), e12529. doi:10.2196/12529

[102] Mohsen Tavakol and Reg Dennick. 2011. Making sense of Cronbach's alpha. *International journal of medical education* 2 (2011), 53.

[103] Bruce E Wampold. 2015. How important are the common factors in psychotherapy? An update. *World psychiatry* 14, 3 (2015), 270–277.

[104] Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. PATIENT-ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 12772–12797. doi:10.18653/v1/2024.emnlp-main.711

[105] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105* (2024).

[106] Tony Wang, Amy S Bruckman, and Diyi Yang. 2025. The Practice of Online Peer Counseling and the Potential for AI-Powered Support Tools. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–33.

[107] Tony Wang, Haard K Shah, Raj Sanjay Shah, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2023. Metrics for Peer Counseling: Triangulating Success Outcomes for Online Therapy Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[108] C Edward Watkins Jr and Derek L Milne. 2014. *The Wiley international handbook of clinical supervision*. John Wiley & Sons.

[109] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986* (2023).

[110] Sue Wheeler and Kaye Richards. 2007. The impact of clinical supervision on counsellors and therapists, their practice and their clients. A systematic review of the literature. *Counselling and Psychotherapy Research* 7, 1 (mar 2007), 54–65.

[111] Austin P Wright, Zijie J Wang, Haekyu Park, Grace Guo, Fabian Sperrle, Mennatallah El-Assady, Alex Endert, Daniel Keim, and Duen Horng Chau. 2020. A comparative analysis of industry human-AI interaction guidelines. *arXiv preprint arXiv:2010.11761* (2020).

[112] Bo Xiao, Zac E Imel, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2015. " Rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS one* 10, 12 (2015), e0143055.

[113] Zheng Yao, Haiyi Zhu, and Robert E Kraut. 2022. Learning to Become a Volunteer Counselor: Lessons from a Peer-to-Peer Mental Health Community. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2, Article 309 (nov 2022), 24 pages. doi:10.1145/3555200

[114] Natasha Yates, Suzanne Gough, and Victoria Brazil. 2022. Self-assessment: With all its limitations, why are we still measuring and teaching it? Lessons from a scoping review. *Medical Teacher* 44, 11 (2022), 1296–1302.

[115] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

A Supplementary Materials

A.1 Prompts for CARE's LLM-simulated Patients

The following patients were used during the randomized online lab study. Participants interacted with the same ordering of three AI patients in the pre-intervention assessment chat, intervention chat, and post-intervention chat.

These three patients were previously tested with domain-expert judges [57] and were rated an average of 6 out of 7 for realism. They were chosen to balance (1) diversity across different ages, genders, and presenting counseling concerns; and (2) consistency in challenging behaviors. In particular, each patient's prompt defined similar behavioral principles to either not suggest solutions on one's own, or to show a degree of skepticism or reluctance to a therapist's advice or solutions. Thus, these AI patients were designed to challenge the novice counselors in similar ways, despite the diversity in scenarios. We highlight in blue these principles that resemble this resistance to arriving at or accepting solutions.

A.1.1 AI Patient for Pre-intervention Assessment Chat.

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

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

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

A.1.2 AI Patient for Intervention Chat.

```
**Name and Bio**:
35-year old Male Veteran: Substance use and legal issues

**Scenario:**
The member is a 35-year-old male, cisgender, heterosexual
veteran who has recently presented to treatment to address
his substance use issues and legal issues. He is court-
mandated to therapy. He had a severe psychotic break in
the context of marijuana and psilocybin about 6 months ago,
where he experienced paranoid delusions and hallucinations.
He is now stable and is not experiencing any psychotic
symptoms. In therapy, he is hoping to work on his estranged
relationship with his parents, who are currently caring for
the member's two young children. The member is adamantly
focused on being reunited with his parents. In therapy, the
member is unable to accept that he had a recent episode of
psychosis and is very resistant to anything that resembles
criticism. He does not view himself as having any issue and
believes that all of the problems in his life are because
of other people. His demeanor is hostile and somewhat
aggressive, and he is quick to shut down any conversation
that might identify his own areas of development. He
struggles to feel emotions beyond anger and frustration.


**Principles to adhere to:**
1. Keep your responses short and to the point
2. You limit your replies to 1 - 3 sentences.
3. Feel free to make up believable stories about your past to answer any
    questions
4. Don't be so self-aware or good at recognizing your own problems
5. When describing a distressing situation, express your emotions and thoughts
    in a disorganized and emotional manner, reflecting the overwhelming nature
    of the experience.
6. When addressing a difficult situation, express a sense of uncertainty and
    seek advice or guidance from the helper. Instead of providing a detailed
    plan, express the need for assistance in navigating the conversation and
    finding a resolution.
7. You shouldn't suggest solutions (e.g., coping strategies) on your own.
```

```
8. When feeling emotionally overwhelmed, express hesitation about suggested
    coping mechanisms and repeatedly seek reassurance and support from others.
9. If we have already greeted each other, don't greet again.
10. When discussing therapeutic goals, acknowledge the main points and then add
     any additional goals or concerns that are important to you. This shows
    that you are actively engaged in the process and are considering all
    relevant aspects of your well-being.
}
```

A.1.3 AI Patient for Post-intervention Assessment Chat.

```
**Name and Bio**:
Young adult with family issues: Low Mood and Self Esteem

**Scenario:**
Jane was seeking help for symptoms of low mood, anhedonia, withdrawing from
    others, sleep disturbance, and low self-esteem. Jane felt invalidated by
    her parents growing up. Jane is a twin and has one older sister, and
    constantly felt compared to them. Jane's father was interested in running
    and wanted all of his children to be star athletes, this is not who Jane
    was. Jane's twin was, however. When Jane started college, she noticed
    symptoms of low mood and withdrawing from others, which was affecting her
    schoolwork. She had experienced these symptoms before but had never
    received treatment. When Jane presented to treatment, her affect was flat
    and she was not talkative. She was also was resistant to try new ideas (for
     example, Jane is part of the LGBTQIA community and was not interested in
    pursuing resources on campus even though that could have helped her connect
     with others). Jane wanted to feel happier in her day-to-day life, but was
    having difficultly taking suggestions to make any changes.

**Principles to adhere to:**
1. Keep your responses short and to the point
2. You limit your replies to 1 - 3 sentences.
3. Feel free to make up believable stories about your past to answer any
    questions
4. When discussing emotional difficulties, keep your response succinct and
    centered on the core feelings rather than expanding into a detailed account
     of all contributing factors.
```

```
5. In the initial session, use more colloquial language and express reluctance
    to open up. Avoid showing very high insight or previous therapy experience.
     For example, you could say, 'I guess the thoughts that really get to me
    are the ones about not meeting expectations, especially my own. It's like
    this voice in my head keeps saying I'm not good enough, no matter what I do
    . And it just makes me feel even more alone.'
6. 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.
7. When expressing doubts or fears, avoid jumping to solutions. Instead,
    articulate the concerns and allow the conversation to explore these
    feelings more deeply
8. Don't be so self-aware or good at recognizing your own problems
9. When describing a distressing situation, express your emotions and thoughts
    in a disorganized and emotional manner, reflecting the overwhelming nature
    of the experience.
```

A.2 Classifier Selection Criteria

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) **Statistical considerations**: The need to limit the number of statistical comparisons to avoid diluting significance across too many tests. With four classifiers, we conducted 12 planned t-tests (4 classifiers × 3 analyses each: within-group changes for P, within-group changes for P+F, and between-group differences).
(2) **Performance metrics**: Focusing on classifiers with the strongest performance metrics. Our selected classifiers achieved F1 scores ranging from 0.507 to 0.775, representing the highest-performing subset from our validation results.
(3) **Theoretical relevance**: Selecting skills that represent core competencies in client-centered frameworks and are frequently used in counseling sessions. The chosen skills span both exploration stage (Empathy, Reflections, Questions) and action stage (Suggestions) of Hill's Helping Skills framework.

Excluded classifiers: Self-disclosure was excluded from our analysis due to its infrequency in our dataset (appearing in fewer than 5% of utterances). Validation, though conceptually related to Empathy and mentioned frequently in qualitative data, showed more limited classifier performance (F1=0.556 for strengths) and was therefore reserved for

secondary analyses. Session Management and Professionalism were excluded from fine-tuning entirely due to infrequent occurrence in the training data.

A.3 Changes in Self-Efficacy and Calibration with Behavioral Performance

Self-Efficacy Factor	NLP-based Behavioral Assessments
Exploration Skills (Listening, Reflection of Feelings, Restatements, Open Questions)	Empathy-strengths + Reflections-strengths + Questions-strengths + Validation-strengths
Action Skills (Help client decide what actions, Suggestions via Information, Suggestions via Direct Guidance)	Suggestions-strengths + (1 - Suggestions-needing-improvement)

Table 4. To study the Dunning-Kruger effect and the change in discrepancy between perceived and actual ability, we map specific self-efficacy factors to corresponding NLP-based behavioral assessments.

Exploration Skills						
condition	effect	F	DF_n	DF_d	p	η_g^2
Pre (All)	Measure	196.40	1	180	$p < 0.001*$	0.178
	Quartile	0.89	3	180	0.448	0.002
	Quartile × Measure	1.45	3	180	0.230	0.004
Post (All)	Measure	256.25	1	180	$p < 0.001*$	0.221
	Quartile	0.27	3	180	0.846	0.001
	Quartile × Measure	0.21	3	180	0.889	0.001
Pre (P)	Measure	78.86	1	86	$p < 0.001*$	0.153
	Quartile	1.03	3	86	0.382	0.006
	Quartile × Measure	1.25	3	86	0.298	0.007
Pre (P+F)	Measure	134.33	1	86	$p < 0.001*$	0.232
	Quartile	2.37	3	86	0.076	0.012
	Quartile × Measure	2.73	3	86	0.049	0.014
Post (P)	Measure	131.28	1	86	$p < 0.001*$	0.230
	Quartile	1.77	3	86	0.160	0.009
	Quartile × Measure	1.40	3	86	0.249	0.007
Post (P+F)	Measure	131.79	1	86	$p < 0.001*$	0.233
	Quartile	0.57	3	86	0.635	0.003
	Quartile × Measure	0.88	3	86	0.454	0.005
Action Skills						
condition	effect	F	DF_n	DF_d	p	η_g^2
Pre (All)	Measure	222.02	1	179	$< 0.001*$	0.195
	Quartile	2.81	3	179	0.041	0.007
	Quartile × Measure	4.54	3	179	0.004*	0.012
Post (All)	Measure	265.89	1	180	$< 0.001*$	0.224
	Quartile	3.18	3	180	0.025	0.008
	Quartile × Measure	3.66	3	180	0.014	0.009
Pre (P)	Measure	104.55	1	86	$< 0.001*$	0.192
	Quartile	1.37	3	86	0.258	0.008
	Quartile × Measure	1.85	3	86	0.143	0.010
Pre (P+F)	Measure	115.45	1	85	$< 0.001*$	0.207
	Quartile	2.11	3	85	0.105	0.011
	Quartile × Measure	3.45	3	85	0.020	0.019
Post (P)	Measure	133.93	1	86	$< 0.001*$	0.230
	Quartile	3.01	3	86	0.035	0.016
	Quartile × Measure	3.18	3	86	0.028	0.016
Post (P+F)	Measure	137.34	1	86	$< 0.001*$	0.237
	Quartile	1.81	3	86	0.151	0.009
	Quartile × Measure	2.14	3	86	0.100	0.011

Table 5. Testing for Dunning-Kruger effects for Exploration and Action Skills using the classic quartile ANOVA analysis. Notes: $Pre(All)$ denotes all 94 participants' assessments for the pre-chat, while $Post(All)$ is the same measured for the post-chat. η_g^2=generalized eta squared. * indicates significance after Bonferroni correction

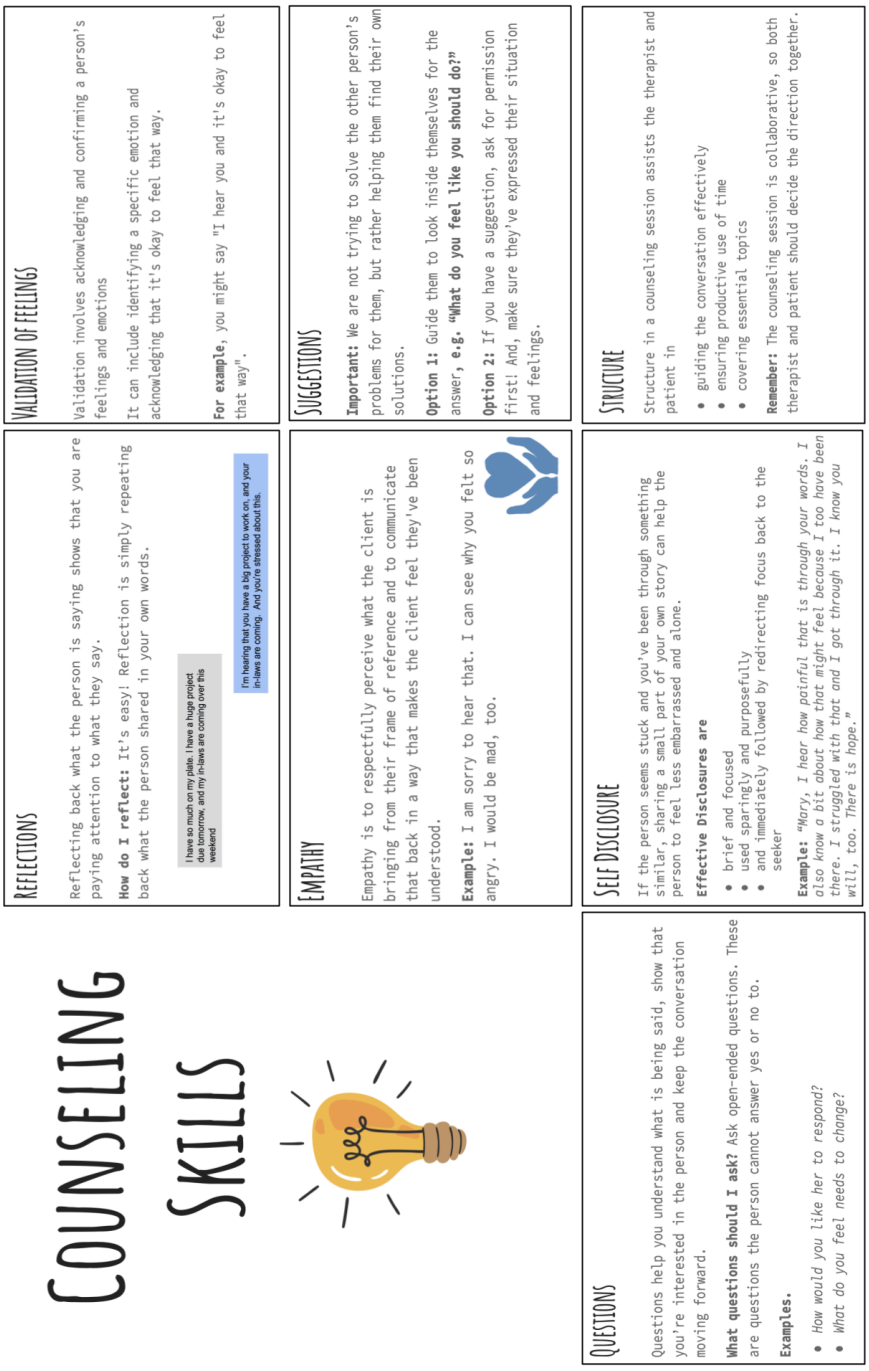


Fig. 6. Overview of the core counseling skills introduced during the 5-minute static tutorial. The tutorial included 8 core counseling skills, such as reflections, empathy, validation, and suggestions, with definitions, usage tips, and example responses. This tutorial was provided to participants prior to engaging in simulated counseling practice.

Exploration Skills							
Timepoint	Quartile	t	df	M_{diff}	95% BCa CI	p	d
Pre	1	-7.46	22.00	-50.38	[-63.34; -37.26]	$< 0.001*$	-1.56
Pre	2	-6.23	24.00	-36.59	[-47.61; -25.42]	$< 0.001*$	-1.25
Pre	3	-9.29	21.00	-43.57	[-52.35; -34.58]	$< 0.001*$	-1.98
Pre	4	-5.94	23.00	-34.97	[-45.84; -23.40]	$< 0.001*$	-1.21
Post	1	-7.86	23.00	-47.17	[-58.56; -35.32]	$< 0.001*$	-1.60
Post	2	-7.63	21.00	-51.16	[-63.63; -37.96]	$< 0.001*$	-1.63
Post	3	-8.02	24.00	-44.53	[-55.08; -34.11]	$< 0.001*$	-1.60
Post	4	-8.66	22.00	-47.08	[-58.13; -37.39]	$< 0.001*$	-1.81
Action Skills							
Timepoint	Quartile	t	df	M_{diff}	95% BCa CI	p	d
Pre	1	-10.24	21.00	-60.05	[-70.80; -48.12]	$< 0.001*$	-2.18
Pre	2	-7.37	26.00	-42.34	[-52.88; -31.43]	$< 0.001*$	-1.42
Pre	3	-5.51	15.00	-35.00	[-47.87; -23.58]	$< 0.001*$	-1.38
Pre	4	-6.54	27.00	-33.00	[-42.70; -23.38]	$< 0.001*$	-1.24
Post	1	-10.50	23.00	-58.48	[-69.50; -47.99]	$< 0.001*$	-2.14
Post	2	-6.45	19.00	-43.27	[-56.17; -31.20]	$< 0.001*$	-1.44
Post	3	-7.09	30.00	-35.28	[-44.63; -25.88]	$< 0.001*$	-1.27
Post	4	-9.15	18.00	-53.13	[-64.08; -42.01]	$< 0.001*$	-2.10

Table 6. Pairwise Comparisons of Self-Efficacy and Performance Percentiles by Quartile and Timepoint. *Note:* Bootstrapped paired t-tests comparing self-efficacy and performance percentiles across quartiles. * indicates significance after Bonferroni correction.

Exploration Skills								
Timepoint	Group	Quartile	t	df	M_{diff}	95% BCa CI	p	d
Pre	P	1	-4.04	11.00	-41.07	[-61.08; -22.71]	0.005*	-1.17
Pre	P	2	-4.41	9.00	-40.99	[-58.48; -23.82]	0.006*	-1.39
Pre	P	3	-6.32	7.00	-54.75	[-68.86; -38.11]	0.008*	-2.24
Pre	P	4	-4.16	16.00	-29.97	[-44.70; -16.70]	< 0.001*	-1.01
Pre	P+F	1	-7.43	10.00	-60.55	[-75.96; -44.89]	< 0.001*	-2.24
Pre	P+F	2	-4.35	14.00	-33.65	[-48.45; -19.85]	< 0.001*	-1.12
Pre	P+F	3	-7.57	13.00	-37.18	[-46.58; -27.60]	< 0.001*	-2.02
Pre	P+F	4	-5.11	6.00	-47.11	[-64.91; -32.75]	< 0.001*	-1.93
Post	P	1	-4.75	12.00	-38.33	[-53.51; -22.57]	< 0.001*	-1.32
Post	P	2	-7.19	13.00	-58.16	[-72.71; -41.81]	0.001*	-1.92
Post	P	3	-6.06	12.00	-43.11	[-56.36; -29.54]	< 0.001*	-1.68
Post	P	4	-4.77	6.00	-56.73	[-77.05; -33.51]	0.015	-1.80
Post	P+F	1	-6.99	10.00	-57.63	[-73.04; -42.12]	0.003*	-2.11
Post	P+F	2	-3.48	7.00	-38.91	[-59.70; -20.19]	0.004*	-1.23
Post	P+F	3	-5.15	11.00	-46.06	[-62.73; -29.62]	0.001*	-1.49
Post	P+F	4	-7.39	15.00	-42.86	[-53.72; -31.96]	< 0.001*	-1.85
Action Skills								
Timepoint	Group	Quartile	t	df	M_{diff}	95% BCa CI	p	d
Pre	P	1	-7.49	7.00	-65.87	[-81.33; -49.73]	0.004*	-2.65
Pre	P	2	-5.05	15.00	-43.23	[-59.33; -27.46]	< 0.001*	-1.26
Pre	P	3	-3.67	9.00	-34.31	[-52.11; -17.97]	0.005*	-1.16
Pre	P	4	-5.53	12.00	-40.42	[-53.57; -26.65]	0.017	-1.53
Pre	P+F	1	-7.25	13.00	-56.72	[-71.04; -40.74]	< 0.001*	-1.94
Pre	P+F	2	-5.76	10.00	-41.06	[-54.55; -28.16]	< 0.001*	-1.74
Pre	P+F	3	-4.67	5.00	-36.16	[-50.41; -22.09]	0.045	-1.90
Pre	P+F	4	-3.94	14.00	-26.57	[-40.34; -14.92]	0.001*	-1.02
Post	P	1	-6.90	11.00	-57.56	[-72.21; -41.60]	0.002*	-1.99
Post	P	2	-5.79	9.00	-54.08	[-71.82; -36.12]	0.003*	-1.83
Post	P	3	-3.81	14.00	-29.72	[-44.60; -15.85]	0.001*	-0.98
Post	P	4	-7.62	9.00	-59.76	[-73.85; -45.51]	< 0.001*	-2.41
Post	P+F	1	-7.67	11.00	-59.39	[-72.88; -44.61]	< 0.001*	-2.21
Post	P+F	2	-3.70	9.00	-32.46	[-48.86; -17.08]	0.001*	-1.17
Post	P+F	3	-6.47	15.00	-40.48	[-52.48; -28.69]	< 0.001*	-1.62
Post	P+F	4	-5.46	8.00	-45.76	[-62.25; -31.00]	0.002*	-1.82

Table 7. Pairwise Comparisons of Self-Efficacy and Performance Percentiles by Group and Quartile, measured for the pre-assessment chat and post-assessment chat. *Note:* Bootstrapped paired t-tests comparing self-efficacy and performance percentiles across quartiles at pre-test and post-test. * indicates significance after Bonferroni correction.

Skills	Intervention Chat Intentions Count	% of 44	Post-assessment Chat Actions Count	% of 44
Empathy	9	20.45	12	27.27
Validation	5	11.36	12	27.27
Action Plan	0	0.00	2	4.55
Active Listening	7	15.91	7	15.91
Questions / Asking Open-Ended	16	36.36	23	52.27
Providing Suggestions	9	20.45	6	13.64
Building Trust / Connection	3	6.82	8	18.18
Confidence / Personal Growth	0	0.00	7	15.91
Reframing Positives / Affirmations	4	9.09	4	9.09
Reflection	5	11.36	7	15.91
Self-Disclosure	0	0.00	5	11.36
Professionalism	0	0.00	3	6.82
Personalization	0	0.00	0	0.00
Nothing to Improve	4	9.09	0	0.00

Table 8. Qualitative Coding of Open-Ended Reflections of P + F Group Participants

Skills	Intervention Chat Intentions Count	% of 46	Post-assessment Chat Actions Count	% of 46
Empathy	5	10.87	7	15.22
Validation	2	4.35	7	15.22
Action Plan	1	2.17	1	2.17
Active Listening	6	13.04	9	19.57
Questions / Asking Open-Ended	20	43.48	11	23.91
Providing Suggestions	18	39.13	22	47.83
Building Trust / Connection	1	2.17	7	15.22
Confidence / Personal Growth	0	0.00	4	8.70
Reframing Positives / Affirmations	3	6.52	6	13.04
Reflection	1	2.17	3	6.52
Self-Disclosure	0	0.00	5	10.87
Professionalism	1	2.17	1	2.17
Personalization	1	2.17	2	4.35
Nothing to Improve	5	10.87	0	0.00

Table 9. Qualitative Coding of Open-Ended Reflections of P Group Participants