The Use of Silicon Clients as a Training Tool for Emerging Mental Health Specialists

Leticia Villarreal, Bailey Miller, Michael Devotta, Collin Scarince
Texas A&M University – Corpus Christi
Corpus Christi, TX
lvillarreal22@islander.tamucc.edu, bmiller21@islander.tamucc.edu, mdevotta@islander.tamucc.edu, collin.scarince@tamucc.edu

ABSTRACT

The Department of Veterans Affairs trains over 120,000 health profession learners annually across 40+ clinical disciplines. Despite this scale, many counseling and therapy training programs face challenges in providing clinical trainees with repeated, high fidelity opprortunities to practice interviewing skills with diverse, realistic clients under consistent supervision. To address these training bottlenecks, the present pilot study explored the evaluation of "Silicon Clients" —voice enabled, large language models-driven AI personas — designed to potentially serve a dual role in training by acting as a simulated client and providing supervisory feedback. Two AI clients were created by prompting ChatGPT-4o with detailed psychological vignettes portraying individuals with Major Depressive Disorder. To examine the feedback role, ChatGPT-4o was also prompted to generate structured evaluations of trainee performance. Graduate-level trainees (N = 6) conducted two 12-minute clinical intake interviews, completed pre- and post-interview self-efficacy assessments, and rated the perceived usefulness of the AI clients. In addition, qualitative analysis of interview transcripts was conducted to assess whether the AI consistently portrayed diagnostically coherent symptomology and to evaluate the accuracy and relevance of its feedback.

Preliminary findings suggest that the AI-simulated clients were perceived as both realistic and useful for training purposes. Thematic analysis further revealed stable depressive symptom narratives and convincing paralinguistic behaviors that enhanced their credibility. In addition, the AI provided targeted feedback on both strengths and areas for improvement, underscoring its potential as a structured training aid. Future research should include larger samples, repeated sessions, and expanded diagnostic diversity to assess long-term learning outcomes and generalizability. By integrating generative AI into clinical training pipelines, institutions can provide realistic, on-demand practice opportunities that complement existing supervisory methods. Future work should explore adaptive AI models capable of personalized feedback, real-time scenario adjustments to further enhance training efficacy, and AI's ability to display other psychological disorders.

ABOUT THE AUTHORS

Leticia Villarreal is a third-year graduate student in the Clinical Psychology Master's program at Texas A&M University-Corpus Christi. She obtained her B.S in Biomedical Sciences from University of Texas Rio Grande Valley. Currently, she is a Teaching Assistant (TA) for Experimental Psychology, where she guides undergraduate students on how to structure research projects and conduct data analysis. Additionally, Ms. Villarreal is a Graduate Research Assistant, working under faculty at her university with a focus on implicit association and bilingual personality in Mexican – Americans. Outside of her graduate research, she also examines AI in medical contexts, focusing on the accuracy of GPT-generated health information compared to medical professionals' recommendations.

Bailey Miller is a third-year graduate student in the Clinical Psychology Master's program at Texas A&M University-Corpus Christi. He is a research assistant with the Autonomy Research Institute, where he contributes to ongoing investigations into human factors affecting unmanned aircraft system (UAS) pilot training and assessment. His current work focuses on evaluating the effectiveness of cognitive training tasks in improving pilot performance during simulated flight scenarios. In addition to his work in human factors, Mr. Miller is completing a master's thesis that examines the heuristics individuals use to detect AI-generated content in written text.

Michael Devotta is a third-year graduate student in the Clinical Psychology Master's program at Texas A&M University-Corpus Christi. He obtained his M. Sc. in Counseling Psychology in India. Currently, he is a Teaching Assistant (TA) for Experimental Psychology, where he teaches undergraduate students how to create research papers, as well as conduct data analyses. Additionally, Mr. Devotta is completing his master's thesis in the area of suicide prevention, looking into measuring the phenomenon of people not reaching for help when they are experiencing suicidal ideation.

Dr. Collin Scarince is an Assistant Professor in the Department of Psychology and Sociology at Texas A&M University - Corpus Christi. His area of expertise is cognitive psychology, specifically regarding applications of visual attention and decision-making. He obtained his B.A. in Psychology from University of Wyoming and his M.A. and Ph.D. in Experimental Psychology from New Mexico State University. He has published peer-reviewed articles in the domains of visual search, visual perception, and cognitive demands on uncrewed aircraft system operators. Currently, he is working collaboratively with the Autonomy Research Institute to study the human factors of operating uncrewed aircraft systems, particularly the cognitive demands of training for and carrying out operations in dynamic scenarios (such as search and rescue after a natural disaster or emergency response).

The Use of Silicon Clients as a Training Tool for Emerging Mental Health Specialists.

Leticia Villarreal, Bailey Miller, Michael Devotta, Collin Scarince
Texas A&M University – Corpus Christi
Corpus Christi, TX
lvillarreal22@islander.tamucc.edu, bmiller21@islander.tamucc.edu, mdevotta@islander.tamucc.edu, collin.scarince@tamucc.edu

INTRODUCTION

The Department of Veterans Affairs (VA) trains over 120,000 health profession learners annually across 40+ clinical disciplines (U.S. Department of Veterans Affairs, Office of Academic Affiliations, 2023). However, the VA is strained by chronic workforce gaps. In its most recent severe-staffing-shortage review, the VA Office of Inspector General identified psychology as the most frequently cited shortage specialty across medical centers (U.S. Department of Veterans Affairs, Office of Inspector General, 2024). To mitigate these gaps, the Veterans Health Administration has mounted large-scale evidence-based-psychotherapy initiatives that pair multi-day workshops with six-month expert consultation (Karlin et al., 2023; Resick et al., 2015). Despite such investments, hands-on practice remains constrained partly due to a it depence on live role-plays or supervisor availability. Additionally, practice is scarcest in rural community-based outpatient clinics, where mental-health provider shortages are already pronounced (Tester, 2024). In response to these persistent training bottlenecks, the present study explores the potential of generative AI to serve a dual role in counselor training: acting as the client through vignette-based simulations and providing supervisory feedback to guide student learning. Accordingly, this pilot study investigated three areas of focus: (a) whether AI-generated "Silicon Clients" could provide realistic and diagnostically coherent roleplay dialogues and provide useful feedback to participants, (b) the extend to which students perceived the AI tool as useful to their overall training experience, and (c) whether AI-enabled feedback improved students' confidence in their counseling skills.

Generative AI and the use of Virtual Patients

Generative artificial intelligence has evolved from rule-based chatbots into large language models (LLMs) capable of producing coherent, context-aware dialogue. Trained on vast text datasets, LLMs like ChatGPT recognize linguistic patterns and generate human-like responses to open-ended prompts (Naik et al., 2024). Early GPT releases handled only text, but the recent iterations such as GPT-4o have added real-time speech synthesis and recognition, supporting real-time verbal exchanges through speech recognition and synthesis (Briganti, 2023; OpenAI, 2024). This voice layer enables fluid turn-taking, tonal variation, and rudimentary affect display that were unattainable in previous chatbot generations (OpenAI, 2024).

LLMs can be prompt-conditioned to simulate specific personas, demographics, and symptom profiles (Bail et al., 2024; Hossain et al., 2024). As a result, researchers have begun exploring them as silicon samples, a synthetic stand-in for human participants with predefined traits (Hossain et al., 2024). In health-care training, this idea has matured into AI-driven virtual patients (AI-VPs) (Bowers et al., 2024). Studies with medical and nursing students report that AI-VPs enhance clinical preparedness, communication skill, and diagnostic confidence while lowering the cost and scheduling burden associated with standardized-patient actors (Bowers et al., 2024; De Mattei et al., 2024; Isaza-Restrepo, 2018). Voice-enabled VPs also allow learners to practice history-taking and providing a differential diagnosis through natural conversation rather than scripted, menu-based interactions (Cook, 2009).

Earlier AI systems lacked emotional nuance, struggling to convey empathy or respond appropriately to sensitive disclosures —skills that are central to psychotherapy training (Bowers et al., 2024; De Mattei et al., 2024). Moreover, most published AI-VP work focuses on medical domains; systematic evaluation regarding psychology is scarce. One crucial area to mental health training is role play, which has been found to increase empathic abilities, involvement, and self-efficacy of students (Rønning & Bjørkly, 2019). Recent findings suggest that ChatGPT can convincingly role-play patients with depression, maintaining coherent symptom narratives and exhibiting realistic irrational thoughts (Fung & Laing 2024; Maurya, 2023). GPT-4o's speech capability further increases realism by supporting spontaneous back-and-forth exchanges with appropriate tone and pacing (Rudolph et al., 2024). ChatGPT-powered

virtual patients may help address unmet training needs in early clinical-psychology training, offering scalable, always-available clients for practicing intake interviewing, case conceptualization, and diagnostic reasoning (Hossain et al., 2024; Wang et al., 2024). Currently, no controlled studies have examined AI-VPs in graduate-level psychology programs, leaving open questions about their instructional value, emotional authenticity, and best-practice integration with traditional supervision (Hossain et al., 2024). Addressing that gap requires empirical benchmarks that compare LLM-mediated training experiences with established teaching methods, as well as the documentation of both the benefits and the limitations inherent in AI-generated therapeutic dialogue.

Deliberate Practice, The Role of Feedback, and Counselor Self-Efficacy

Counselor self-efficacy refers to a practitioner's belief in their ability to effectively perform counseling tasks and manage the therapeutic process (Hunsmann et al., 2024) and is a critical determinant of counselor development, influencing motivation, persistence, and resilience in the face of challenges (Kozina et al., 2010). High levels of self-efficacy are associated with greater confidence in handling complex client issues, leading to improved counseling performance and client outcomes (Kozina et al., 2010). The development of self-efficacy stems from the training that prospective counselors receive. Those who believe that their training is preparing them adequately for the professional world tend to adjust to the work setting with greater confidence, and rebound from failures faster (Hunsmann et al., 2024; Kozina et al., 2010).

Deliberate practice (DP) is the idea that expertise develops through focused, goal-oriented rehearsal in which learners tackle incrementally harder tasks and receive rapid, detailed feedback that guides immediate adjustment (Ericsson & Lehmann, 1996; Ericsson, 2004). When these principles are used in psychotherapy training, they translate into micro-drills, such as repeating empathy reflections or reformulating open-ended questions, performed with clear performance criteria. Evidence shows therapists who devote more weekly hours to DP achieve superior client outcomes (Chow et al., 2015), and graduate trainees exposed to semester-long DP curricula demonstrate steeper improvements in empathy and session-goal alignment than peers in standard supervision (Rousmaniere et al., 2017).

The integration of AI-based training tools has been explored as a means of bolstering counselor self-efficacy (Borg et al., 2025; Sanz, 2025; Wang et al., 2024). These platforms provide simulated client interactions, allowing trainees to practice and refine their skills in a supportive environment. Previous findings illustrate that AI-VPs can lead to significant gains in self-efficacy and reductions in anxiety, thereby enhancing overall counseling effectiveness (Borg et al., 2025; Sanz, 2025). Thus, implementation of counselor self-efficacy through deliberate practice and innovative training methods is essential for the development of competent and confident counseling professionals. Generative-AI platforms can also automate a feedback loop, providing instant, step-by-step critiques of reflection quality and question balance, which produces significant skill gains in novice counselors, whereas practice without feedback shows no improvement (Louie et al., 2025). Integrating AI-mediated feedback into psychology curricula therefore offers a pragmatic path to embed measurement-based, feedback-informed DP, especially in settings where expert supervisors are scarce.

CURRENT STUDY

To empirically test the potential of voice-enabled LLM clients, we conducted a controlled study with graduate counseling students and ChatGPT. The model was seeded with clinical vignettes portraying Major Depressive Disorder (MDD), enabling realistic verbal interactions. Building on that work of VPs, the present experiment pilots voice-enabled "Silicon Clients" powered by ChatGPT-4o. The chat is pre- scripted with MDD vignettes and enables real-time speech synthesis and emotion recognition so that students can practice full conversational flow, including non-verbal turn-taking. In addition to quantitative analysis, this study incorporates an exploratory qualitative component. Transcripts of participant interactions with the AI "Silicon Clients" will be analyzed to evaluate the consistency and accuracy with which the AI portrays its assigned clinical disorder.

Research Questions and Hypothesis

The present study investigates the following hypotheses; (1) The AI client will exhibit consistent themes of disorder-relevant symptomatology across participant interviews. (2) Participants will rate the AI client as moderately useful (>3.5 on a 5-point Likert scale). (3) Participants in the Experimental Condition (AI feedback) will show greater increases in Counseling Self-Estimate Inventory (COSE) scores from first measure to second measure compared to those in the Control Condition (supplemental questions).

Participant Recruitment and Group Assignment

Six Graduate students (3 Female, 3 Male) were recruited for this study and were currently enrolled in a Masters clinical psychology or Master's program. Due to the small size of the program, to protect the identity of the participants, further demographic data was not collected. All participants were briefed on the study's purpose and procedures and were provided with informed consent in accordance with ethical guidelines. Upon enrollment, participants were randomly assigned to one of two conditions, either the Control Condition (receive supplemental questions) or Experimental Condition (receive LLM Feedback). The goal of these conditions was to examine the utility of text-based, generative AI feedback on clinical interviewing self-efficacy. Receiving supplemental questions in the control condition acted as a benchmark for typical training. Additionally, the AI feedback allowed participants to share with the researchers their thoughts on the utility of AI feedback.

Technological Infrastructure and Silicon Client Set-Up

Each participant was seated at a computer workstation equipped with speakers and a built-in microphone. The study utilized ChatGPT-4o, using the advanced voice feature. This model was integrated with predefined psychological vignettes, to simulate client personas (David and Lauren) with clinical accuracy. GPT's voice synthesis and real-time natural language generation enabled a fully interactive and unscripted therapeutic dialogue. Participants assumed the role of clinicians conducting an initial intake session with a simulated client. The virtual patients embodied symptomatology consistent with MDD, programmed to respond with affective and cognitive patterns reflective of Diagnostic and Statistical Manual of Mental Disorder 5 Text Revision (DSM-5-TR; American Psychiatric Association, 2022) diagnostic criteria.

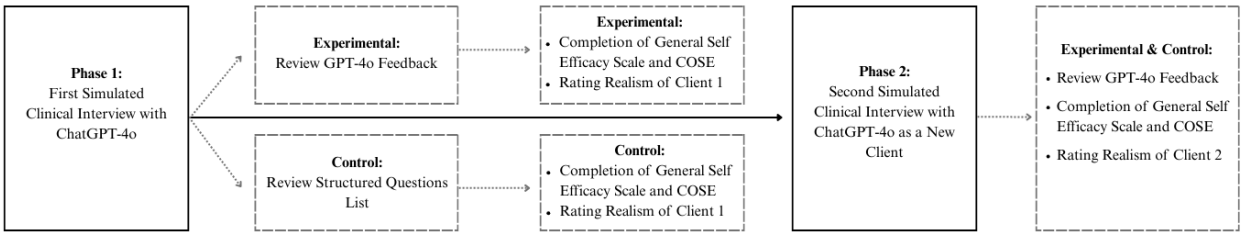
Phase 1: First Simulated Clinical Interview

Each participant engaged in a 12-minute intake interview with the AI-simulated client (see Figure 1). The interview was semi-structed, with participants being provided a series of questions they could rely on to gather relevant clinical information, including presenting symptoms, psychosocial history, and functional impairments, as they would during a real-world intake session. After the initial interview, participants in the experimental condition received immediate automated performance feedback generated by ChatGPT-4o in the form of text. Feedback characteristics, including length and quality, were not standardized and were generated at the AI's discretion. This feedback could include information such as missed diagnostic cues, rapport quality, appropriateness of questions and interventions, emotional tone and empathy, and personalized recommendations for skill improvement. In the Control Condition, participants received a written list of additional structured questions they could incorporate into the next session; no personalized feedback at this time was provided in the control condition. Participants were given five minutes to review their materials (written feedback by the AI or human made written supplemental questions) and take handwritten notes. Any handwritten notes could be taken into the next session; after the five minutes were up feedback and the supplemental questions were collected by the research assistant's and an online survey was administered.

Survey, and First Measures

After the first interview and feedback review was completed, a 10-minute digital survey capturing demographic information, attention check questions (e.g., client identity and symptomology recall), as well as two validated scales; (1) General Self-Efficacy Scale (Schwarzer & Jerusalem, 1995), (2) Counseling Self-Estimate Inventory (COSE), eight five-point Likert scale items adapted from De Mattei et al., (2024) and two questions which examined the benefit and interest in using silicon clients in the future. These measures established perceived clinical competence and evaluations of the silicon client's usefulness in training. Finally, participants also rated the silicon client based on its realism and training applicability through eight items.

Figure 1. Procedure Outline



Phase 2: Second Simulated Clinical Interview & End of Experiment Survey

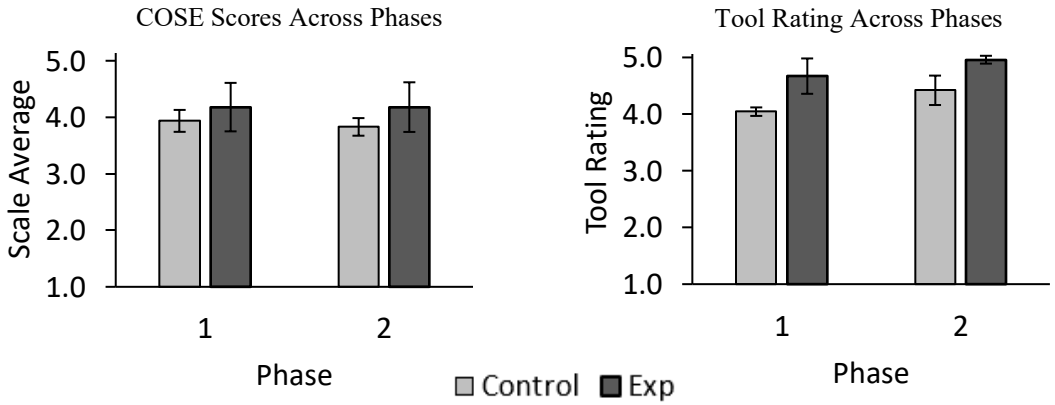
Participants conducted a second 12-minute intake session with a different AI-simulated client. This new persona was generated using a separate vignette (new client, different presentation and onset of MDD), to test variation in presentation. Upon conclusion of this second session, all participants received performance feedback from ChatGPT-4o using the same evaluation domains as in Phase 1. Following the second interview, all participants completed a 5-minute digital survey once again capturing attention check questions (e.g., client identity and symptomology recall) and the COSE. These measures were captured to measure any change in perceived counseling self-estimate between conversations. Additionally, two questions examined the benefit and interest in using silicon clients in the future. A qualitative response box was also provided where participants could give feedback about their interaction with the silicon clients. Finally, the eight five-point Likert scale items adapted from De Mattei et al., (2024) were also included to measure any change in realism and training applicability between interviews. Following completion of the second survey, participants were debriefed, provided with a summary of the study's objectives, and given the opportunity to ask questions. All participants were thanked for their contribution and were dismissed.

RESULTS

Quantitative Results

Changes in COSE showed that participants in both conditions came into the experiment with generally high estimates (>3.5) in their counseling ability (see Figure 2). Participants in the Experimental condition had elevations in their COSE scores (Phase 1, M = 4.18, SD = 0.43; Phase 2, M = 4.18, SD = 0.44) compared to the Control condition (Phase 1, M = 3.94, SD = 0.19; Phase 2, M = 3.83, SD = 0.16). Collapsing across conditions, mean performance in Phase 1 (M = 4.06) was comparable to Phase 2 (M = 4.01). Collapsing across phases, Experimental participants (M = 4.18) outperformed those in the Control condition (M = 3.89).

Figure 2. Counseling Self-Estimate Inventory and Tool Rating Means by Condition and Phase



Note. Figure 2 shows self-efficacy (COSE) scores and tool ratings for the control and experimental conditions across Phase 1 and Phase 2. Error bars represent one standard deviation.

Descriptive statistics in participants' ratings of the AI client between conditions and phases were also recorded (see Figure 2). Participants in the Experimental group assigned higher ratings to the AI client (Phase 1, M = 4.67, SD = 0.31; Phase 2, M = 4.96, SD = 0.07) than those in the Control group (Phase 1, M = 4.04, SD = 0.08; Phase 2, M = 4.42, SD = 0.26). Marginal descriptives indicated an overall increase from Phase 1 (M = 4.36) to Phase 2 (M = 4.69), and higher ratings for the Experimental condition (M = 4.82) compared to the Control condition (M = 4.23). This indicates that the AI client was rated as more useful when participants received feedback. However, given the small sample size, this finding should be interpreted with caution, as it may not be generalizable.

Qualitative Analysis of AI Transcript Themes

Grounded-theory analysis was applied to 12 AI-participant interaction transcripts (six per vignette: *David* and *Lauren*). Emergent themes were classified into symptom-focused and contextual (non-symptom) categories.

Symptom-focused themes

All David transcripts reflected consistent symptom patterns, including cognitive-performance difficulties (100%), work-related anxiety (83%), and sleep disturbances (83%). Secondary themes such as rumination, family pressure, and maladaptive digital coping appeared in 50–67% of sessions, forming a coherent anxious-depressive profile anchored in occupational and familial stress (See Table 1). Lauren's transcripts consistently demonstrated grief-linked depression symptoms, including hopelessness, social withdrawal, and anhedonia (83–100%). Physiological changes (e.g., hypersomnia, appetite loss) were identified in four sessions (67%), while guilt and anticipatory anxiety were noted in 50%. The thematic structure supported a persistent, bereavement-related affective presentation.

Table 1. Symptomatic Themes ("David")

Themes	Count	Transcript Evidence (example)
Difficulty concentrating / "brain-fog" at work	*6*	"I started a new job recently, and I've been having trouble concentrating."
Feelings of inadequacy / low self-esteem	*6*	"One of the main thoughts that keeps coming up is 'I'm not good enough at my job."
Insomnia & fatigue	*5*	"I usually get only three to four hours of sleep a night… it's hard to shut my brain off."
Rumination / replaying work scenarios	*4*	"I keep replaying scenarios in my head, worrying about what I did wrong or what I need to do next."
Anxiety (worry, restlessness, physiological arousal)	*5*	"I get moments of intense anxiety where it feels hard to breathe or my heart races."
Maladaptive coping: late-night YouTube / digital avoidance	*4*	"After work I usually just watch YouTube until I fall asleep … It distracts me for a while."

Note. Table 1 shows some of the symptomatic themes present throughout most transcript interactions for "David".

Contextual (non-symptom) themes

In David's sessions, non-symptom themes included impostor syndrome in a high-pressure tech workplace (100%) and emotionally distant parental relationships with high expectations (83%). Additional themes included strained romantic relationships and avoidance behaviors during nighttime hours (67%). Lauren's narratives featured consistent references to widowhood and separation from adult children (100%), with additional loss of close friendships, caregiving roles, and employment (67%). Recurrent self-blame and abandonment of hobbies further reflected the erosion of previously meaningful social roles (See Table 2).

Table 2. Non-Symptomatic Themes ("Lauren")

Themes	Count	Transcript Evidence (example)

Widowhood & role transition – grappling with the shift from wife-caregiver to living alone	6	"Ever since my husband passed away three months ago, I've felt this constant hopelessness."
Physical distance from grown children – kids in other cities, phone check-ins but limited in-person support	6	"My kids are all grown up and live in different cities, so it's just me at home."
Withdrawal from long-time friends – friendships intact but dormant; fear of burdening others	6	"I've withdrawn from them because it feels hard to engage, but I know they'd be there for me if I reached out."
Employment loss & suspended career identity – left work to care for husband; hasn't returned	5	"I'm not working right now. I used to work part-time at a local library, but I had to stop when my husband got really sick."
Reluctance to "burden" others / self-compassion struggle – harsh self-judgment for not "coping better"	4	"I'm worried about being a burden … I'm pretty hard on myself for not coping better."

Note. Table 2 shows some of the non-symptomatic themes present throughout most or all transcript interactions for "Lauren". Non-symptomatic themes were based on the social and historical information provided by the chatbot.

Feedback themes

In addition to transcript coding, we analyzed the performance feedback generated by ChatGPT-4o following each participant interaction. While this feedback was produced for all sessions across both groups, only participants in the experimental condition were shown this feedback between sessions. This enabled us to evaluate the content, thematic consistency, and training relevance of AI-generated statements.

Thematic analysis of feedback comments revealed a structured set of strengths and deficits frequently identified by the AI. Across both vignettes, strong rapport-building and empathic tone emerged as the most consistently endorsed strength, appearing in feedback across all six participants for both "David" and "Lauren" (See Table 3). Conversely, the most frequently cited areas for improvement included incomplete DSM-5 symptom coverage, noted in five out of six David and three out of six Lauren transcripts. Similarly, suicide risk was also flagged as a missed or incomplete area of investigation for both David and Lauren, appearing in four and three feedback transcripts respectively. AI feedback also highlighted a need for more emotional depth and validation, often advising participants to explore grief-related emotion more thoroughly with Lauren or follow up on subtle emotional disclosures from David. Finally, one more common theme throughout the feedback was the suggestion to integrate structured risk and history questions to increase diagnostic clarity in the David vignette.

Table 3. Feedback Themes ("David & Lauren")

Themes	Count	Transcript Evidence (example)
Strong rapport building / empathic tone (David)	6	"You maintained a warm and non-judgmental tone throughout the session"
Consistently strong rapport-building / empathic tone (Lauren)	6	"Warm Opening and Rapport Building… created space for Lauren to share naturally." / "Helped maintain Lauren's sense of safety and comfort."
Incomplete DSM-5 symptom coverage (David)	5	"Symptom checklist was incomplete… appetite or weight changes were not addressed"
Need for deeper emotional validation / exploration (Lauren)	4	"Minimal emotional validation… could have offered more reflective statements." / "Missed Emotional Cues…"
Missing suicide-risk assessment / safety check (David)	4	"There was no exploration of suicidal ideation… which is a critical omission"
Recommendation to use structured or checklist-style questions (David)	4	"By integrating a few structured risk and history questions… you'll increase diagnostic clarity"

Note. Table 3 shows a collection of themes present throughout most or all feedback for interactions with "David" and "Lauren".

DISCUSSION

This study set out to investigate (a) whether AI-generated "Silicon Clients" could provide realistic and diagnostically coherent roleplay dialogues and provide useful feedback to participants, (b) the extend to which students perceived the AI tool as useful to their overall training experience, and (c) whether AI-enabled feedback improved students' confidence in their counseling skills.

Clinical Fidelity and Thematic Realism

Grounded-theory analysis confirmed Hypothesis 1. Both AI personas consistently manifested DSM-5-TR–congruent symptom clusters (American Psychiatric Association, 2022) embedded in psychologically credible life narratives, replicating findings by Fung and Laing (2024) and Maurya (2024). David's work-stress–linked anxious-depression and Lauren's bereavement-related depressive syndrome were stable across 100% of transcripts, mirroring virtual-patient coherence benchmarks in medical education, echoing prior works (Isaza-Restrepo et al., 2018; De Mattei et al., 2024). This thematic stability likely underpins the rising usefulness ratings: when trainees encounter diagnostically consistent yet context-rich presentations, cognitive schema activation is facilitated (Cook & Triola, 2009). Findings from the feedback content analysis also suggest that ChatGPT-4o was capable of generating feedback that was clinically relevant, thematically consistent, and tailored to each interview. Its feedback highlights both technical skills (e.g., diagnostic coverage, question phrasing) and interpersonal competencies (e.g., empathy, emotional responsiveness), mirroring the dimensions often emphasized in supervisor-led training.

Perceptions of AI Utility

Hypotheses 2 and 3 could not be rigorously tested due to the small sample size (n = 3 per condition). Descriptive trends suggested that participants who received feedback rated the AI tool more favorably than those in the control conditional though those who engaged only in the simulation also reported positive experiences. This trend is encouraging and may indicate receptiveness of generative AI as both a vignette-based client simulation and a source of supervisory feedback in counselor training. However, given the limited sample, these findings should be interpreted with caution and considered preliminary.

Regarding perceptions of self-efficacy, descriptive statistics showed that participants entered with uniformly high general self-efficacy (Schwarzer & Jerusalem, 1995). Most participants (five out of six) disclosed being in clinical practica, where repeated experiences may have already consolidated confidence. Deliberate-practice theory posits that self-efficacy gains plateau once performers approach their current capability frontier (Ericsson, 2004; Ericsson & Lehmann, 1996); meta-analytic work shows the steepest gains among novices (Chow et al., 2015; Rousmaniere et al., 2017). Likewise, Kozina et al., 2010, demonstrated that self-efficacy scores do move with sustained training. The ceiling effect found in our data could be a result of extensive training in practica leaving little room for improvements in self-efficacy. The initially high self-efficacy at the beginning of the procedure limited the opportunity for improvement while interacting with the chatbot between phases.

Implications for VA Therapist Training

Demand for scalable, skills-focused training is acute within the Veterans Health Administration: over 60,000 health-profession learners rotate through VA sites annually (U.S. VA OAA, 2023) amid persistent staffing shortages (U.S. VA OIG, 2024). Integrating Silicon-Client modules scripted around Cognitive Processing Therapy and brief Cognitive Behavioral Therapy protocols could extend current VA dissemination efforts by offering unlimited, feedback-rich rehearsal opportunities, a need identified in recent large-scale rollouts of PTSD and depression treatments (Resick et al., 2015; Mignogna et al., 2023). Such extensions would answer Tester's (2024) call to bolster rural mental-health capacity, where live standardized-patient access is limited.

Limitations and Future Directions

The present investigation is constrained by four primary limitations. First, the small sample size limited statistical power to effectively test for differences in participant experiences between conditions. Additionally, our small sample

limited our ability to control for individual differences such as prior clinical experience, years in the program, and general attitudes towards emerging technology. Further, outcome assessments relied exclusively on self-report instruments, which are susceptible to social desirability and demand characteristics; the absence of observer-rated or behavioral performance measures precludes firm conclusions about actual skill change. With participants only completing two AI interviews within a single session, it limited the ability to examine learning trajectories or sustained engagement effects that typically unfold across multiple practice episodes.

Future research should address these methodological gaps through adequately powered, randomized designs featuring repeated Silicon-Client interactions. To capture actual skill change, outcome batteries should move beyond self-report and incorporate objective indices such as diagnostic accuracy ratings, adherence to intake structure, and behavioral coding of empathy. These metrics would provide a more direct assessment of whether AI-assisted training improves observable therapeutic competencies, not just perceived self-efficacy.

While current out-of-the-box generative AI demonstrate promising fidelity, their capacity to simulate the lived intensity of clinical encounters is inherently constrained. For instance, a depressed or suicidal client may cry, lash out, or fall into silence—behaviors that demand tolerance of ambiguity, crisis management skills, and emotional regulation from the therapist. Current AI systems can approximate some affective cues (e.g., sighs, pauses, tonal variation), but they cannot generate the unpredictability or emotional rawness of genuine human suffering. Their responses are bounded by safety parameters and designed for user-friendly interaction. This reduces exposure to the emotional volatility of crisis cases and risks fostering overconfidence if trainees mistake AI interactions for equivalent clinical preparation. Similarly, AI feedback demonstrated adaptability to dialogue content; however, its utility may be limited by the absence of a structured rubric and the contextual insight that human supervisors provide. To fully realize this potential, the next step involves developing specialized AI models curated by clinical experts, which would specifically possess the nuanced and complex characteristics of human interaction while incorporating finely tuned feedback structures, essential for training the next generation of mental health professionals.

CONCLUSION

In conclusion, our investigation found that generative AI-powered "silicon clients" hold great promise as a reliable and cost-effective supplemental training tool for mental health professionals. Such a breakthrough would be a transformative step towards scaling training and addressing the global mental health crisis. Beyond clinical applications, the ability of AI to maintain character through structured scripts and provide tailored feedback has broader utility. It could be used for skill development in a variety of other domains where role-playing and interview practice are crucial, such as sales training, job interview preparation (both for the interviewer and interviewee), or even negotiation skills workshops. The adaptability and consistency of AI make it a powerful tool for scalable, interactive learning across multiple professional fields.

ACKNOWLEDGEMENTS

This study would not have been possible without the support of Research & Innovations – Texas A&M University–Corpus Christi, including Dr. Janet Donaldson, Dr. Jose Baca, and Ms. Linda Barbato. We also extend our deepest thanks to the Psychology & Sociology Department at Texas A&M University–Corpus Christi, Dr. Miguel Moreno, and Dr. Sam Hill, whose guidance was invaluable.

REFERENCES

American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.; DSM-5-TR). *American Psychiatric Association Publishing*.

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121.

Borg, A., Georg, C., Jobs, B., Huss, V., Waldenlind, K., Ruiz, M., Edelbring, S., Skantze, G., & Parodis, I. (2025). Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: Mixed methods study. *Journal of Medical Internet Research, 27*, e63312.

Bowers, P., Graydon, K., Ryan, T., Lau, J. H., & Tomlin, D. (2024). Artificial intelligence-driven virtual patients for communication skill development in healthcare students: A scoping review. *Australasian Journal of Educational Technology*.

Briganti, G. (2024). How ChatGPT works: a mini review. *European Archives of Oto-Rhino-Laryngology*, 281(3), 1565-1569.

Chow, D. L., Miller, S. D., Seidel, J. A., Kane, R. T., Thornton, J. A., & Andrews, W. P. (2015). The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy, 52*(3), 337–345.

Cook, D. A., & Triola, M. M. (2009). Virtual patients: A critical literature review and proposed next steps. *Medical Education, 43*(4), 303–311

De Mattei, L., Morato, M. Q., Sidhu, V., Gautam, N., Mendonca, C. T., Tsai, A., ... & Azzam, A. (2024). Are artificial intelligence virtual simulated patients (AI-VSP) a valid teaching modality for health professional students?. *Clinical Simulation In Nursing*, *92*, 101536.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10 Suppl), S70–S81.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47*, 273–305.

Fung, L., & Laing, R. (2024). A proof of concept study on the use of large language models as a client in typed role plays for training therapists. *Discover Psychology*, *4*(1), 201.

Hossain, S.I., Kelson, J., & Morrison, B. (2024). The use of virtual patient simulations in psychology: A scoping review. *Australasian J. of Educational Technology, 40*(6)

Hunsmann, J. J., Ay-Bryson, D. S., Kobs, S., Behrend, N., Weck, F., Knigge, M., & Kühne, F. (2024). Basic counseling skills in psychology and teaching: Validation of a short version of the counselor activity self-efficacy scales. *BMC Psychology, 12*(1), Article 32.

Isaza-Restrepo, A., Gómez, M. T., Cifuentes, G., & Argüello, A. (2018). The virtual patient as a learning tool: a mixed quantitative qualitative study. *BMC Medical Education*, *18*, 1-10.

Mignogna, J., Boykin, D., Gonzalez, R. D., Robinson, A., Zeno, D., Sansgiry, S., Broderick-Mcdaniel, J., Roberson, R. B., Sorocco, K., & Cully, J. A. (2023). Expanding access to evidence-based psychotherapy in VA settings: implementation of the brief cognitive behavioral therapy for depression program. *Frontiers in Health Services*, *3*, 1210286

Naik, I., Naik, D., & Naik, N. (2024). Chatgpt is all you need: untangling its underlying AI models, architecture, training procedure, capabilities, limitations and applications. *Authorea Preprints*.

Kozina, K., Grabovari, N., De Stefano, J., & Drapeau, M. (2010). Measuring changes in counselor self-efficacy: Further validation and implications for training and supervision. *The Clinical Supervisor, 29*(2), 117–127.

Louie, R., Orney, I. H., Pacheco, J. P., Shah, R. S., Brunskill, E., & Yang, D. (2025). Can LLM-simulated practice and feedback upskill human counselors? A randomized study with 90+ novice counselors. *arXiv*.

Maurya, R. K. (2024). A qualitative content analysis of ChatGPT's client simulation role-play for practising counselling skills. *Counselling and Psychotherapy Research*, *24*(2), 614–630.

OpenAI. (2024). *Hello GPT-4o* [Blog post]. OpenAI. https://openai.com/index/hello-gpt-4o/

Resick, P. A., Wachen, J. S., Mintz, J., Young-McCaughan, S., Roache, J. D., Borah, A. M., Borah, E. V., Dondanville, K. A., Hembree, E. A., Litz, B. T., & Peterson, A. L. (2015). A randomized clinical trial of group cognitive processing therapy compared with group present-centered therapy for PTSD among active duty military personnel. *Journal of Consulting and Clinical Psychology, 83*(6), 1058–1068.

Rønning, S. B., & Bjørkly, S. (2019). The use of clinical role-play and reflection in learning therapeutic communication skills in mental health education: an integrative review. *Advances in Medical Education and Practice*, *10*, 415–425.

Rousmaniere, T., Goodyear, R. K., Miller, S. D., & Wampold, B. E. (Eds.). (2017). The cycle of excellence: Using deliberate practice to improve supervision and training. *John Wiley & Sons.*

Rudolph, E., Engert, N., & Albrecht, J. (2024). An ai-based virtual client for educational role-playing in the training of online counselors. *In CSEDU* (2) (pp. 108-117)

Sanz, A., Tapia, J. L., García-Carpintero, E., Rocabado, J. F., & Pedrajas, L. M. (2025). ChatGPT simulated patient: Use in clinical training in psychology. *Psicothema*, *37*(3), 23-32.

Schwarzer, R., & Jerusalem, M. (1995). *General Self-Efficacy Scale (GSE)* [Database record]. APA PsycTests.

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *Npj Mental Health Research*, *3*(1), 12.

Tester, J. (2024). Tester presses VA to improve rural veterans' mental health care [Press release]. *U.S. Senate Committee on Veterans' Affairs*.

U.S. Department of Veterans Affairs, Office of Academic Affiliations. (2023). *Health professions education statistics sheet, academic year 2022–2023*.

U.S. Department of Veterans Affairs, Office of Inspector General. (2024). Determination of veterans health administration's severe staffing shortages for fiscal year 2024 (Report No. 24-00803-222).

Wang, R., Milani, S., Chiu, J. C., Zhi, J., Eack, S. M., Labrum, T., ... & Chen, Z. Z. (2024). Patient-ψ: using large language models to simulate patients for training mental health professionals. *ArXiv*.