Therapy as an NLP Task:
Psychologists' Comparison of LLMs and Human Peers in CBT

Zainab Iftikhar
Department of Computer Science
Brown University

Sean Ransom
Department of Psychiatry
Louisiana State University Health Sciences

Amy Xiao
Department of Computer Science
Brown University

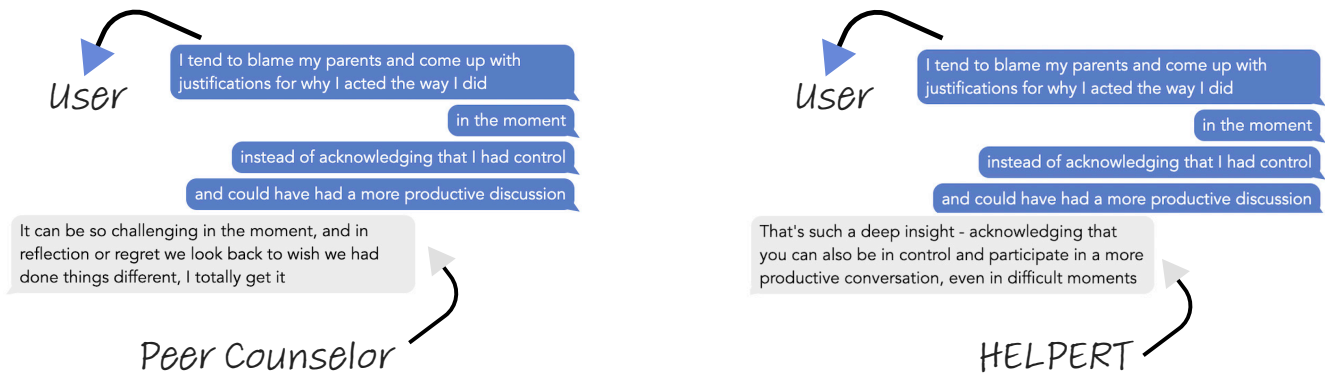Jeff Huang
Department of Computer Science
Brown University

Figure 1: Example public session by a peer counselor (left) reproduced using CBT-prompted large language model-based system called HELPERT (right).

ABSTRACT

Wider access to therapeutic care is one of the biggest challenges in mental health treatment. Due to institutional barriers, some people seeking mental health support have turned to large language models (LLMs) for personalized therapy, even though these models are largely unsanctioned and untested. We investigate the potential and limitations of using LLMs as providers of evidence-based therapy by using mixed methods clinical metrics. Using HELPERT, a prompt run on a large language model using the same process and training as a comparative group of peer counselors, we replicated publicly accessible mental health conversations rooted in Cognitive Behavioral Therapy (CBT) to compare session dynamics and counselor's CBT-based behaviors between original peer support sessions and their reconstructed HELPERT sessions. Two licensed, CBT-trained clinical psychologists evaluated the sessions using the Cognitive Therapy Rating Scale and provided qualitative feedback. Our findings show that the peer sessions are characterized by empathy, small talk, therapeutic alliance, and shared experiences but often exhibit therapist drift. Conversely, HELPERT reconstructed sessions exhibit minimal therapist drift and higher adherence to CBT methods but display a lack of collaboration, empathy, and cultural understanding. Through CTRS ratings and psychologists' feedback, we highlight the importance of human-AI collaboration for scalable mental health. Our work outlines the ethical implication of imparting human-like subjective qualities to LLMs in therapeutic settings, particularly the risk of deceptive empathy, which may lead to unrealistic patient expectations and potential harm.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Discourse, dialogue and pragmatics**; *Natural language generation.*

KEYWORDS

large language models, artificial intelligence, cognitive behavioral therapy, computational linguistics

1 INTRODUCTION

The above two sample excerpts, taken from therapeutic sessions (Figure 1), are conducted by a human peer counselor (left) trained in Cognitive Behavioral Therapy (CBT) techniques and a CBT-prompted Large Language Model (LLM) [9] (right) trained through the same process with the same materials and operational team. A distressed user seeking support describes their situation, thoughts, feelings, and behavior; in response, a counselor—human or LLM—uses CBT techniques to provide support. The results were hour-long sessions guided by the same principles and goals, but where the LLM version is a simulation of the responses that would have been given based on pre-existing publicly available sessions.

From a decision-making system to a CBT-prompted peer [49], LLMs have enabled some people wider access to counseling through completely automated means [58]. Due to their increased accessibility and constant availability [2, 49], LLMs, especially recent implementations like ChatGPT, are now used for more than just

language generation; they are perceived as facilitating meaningful conversations [49]. One in four Americans prefer talking to an AI chatbot instead of a therapist, and of those who did, 80% claim it to be an effective alternative [43]. This preference is understandable since conventional therapy faces issues of inaccessibility, high costs, and complexity, leaving over half of U.S. adults with mental illness without adequate care [59].

However, repurposing LLMs as therapeutic peers when it was originally intended for basic language generation seems counterintuitive. Several challenges exist. First, LLMs are designed to predict the next possible sequence from a given text based on previously observed patterns in their training data that is largely devoid of fact-checking [9]. This encompasses traditional challenges of AI-mediated health care, including lack of high-quality training data [62], low external validity and misinformation [15], societal biases [25], and the impact of AI on patient-clinician relationships [52]. Beyond data quality, there are additional risks of data breach and an individual's right to privacy [26]. These issues have high stakes in mental health, where the quality of a therapist's responses impacts treatment outcomes.

Despite concerns from interdisciplinary experts, an overwhelming user audience attests to LLM's effectiveness [43]. However, informal evaluation of these tools is subjective and contextual; a user in distress is unlikely to gauge the risk associated with the tools, considering they act as a band-aid for mental health support—accessible and free. The discrepancy has ignited debates between the recipients of LLM-based mental health and field experts. While current work has found that users prefer AI responses over humans [2, 38, 68], these studies examine the models' responses to a single, isolated interaction (utterance level) [38, 68]. However, these models lack long-term memory [9] and perform poorly in situations that demand sustained interactions, such as peer counseling.

The growing user acceptance of LLM-based counseling, combined with expert concerns, calls for objective investigation of the model's responses in sustained counseling sessions by clinical therapists. In this paper, we evaluate alternatives to traditional licensed CBT therapy: peer counselors and LLM-based therapeutic sessions. The LLM responses are generated through best-effort reconstruction from pre-existing publicly released sessions to avoid testing on human subjects with an experimental method that may cause harm. We used a CBT-based prompt, designed collaboratively by peer counselors and licensed therapists, to reconstruct pre-existing publicly available counseling sessions. We sourced the original sessions from Cheeseburger Therapy [61], an online text-based peer counseling platform where the people providing the counseling are trained through a series of customized learning modules. Because the LLM prompt and peer counselors are both trained from the same materials and supervised by the same operational team, this controls for some of the many variables that can lead to differences between a human peer counselor and an LLM-based counselor.

As the Cheeseburger Therapy platform is entirely online and based on text-only communication, some of the historical peer counseling sessions are publicly available online. These conversations were made public after users provided written consent for them to be shared online. Consent was requested only after the session had ended. Note that the authors have been in communication and exchange information with the Cheeseburger Therapy team, but do not operate the service or have any ownership over it. The publicly available sessions were downloaded, and the peer counselors' responses were replaced by an LLM-generated response using HELPERT, a detailed prompt that asks the LLM to go through the same process as the peer counselors do.

The resulting sessions are cleaned and become part of a comparative dataset. Using a mixed method analysis of objective observer rating scales to assess competence in Cognitive Behavioral Therapy [67] and psychologist's annotation and feedback, two licensed, CBT-trained clinical psychologists conducted a blind analysis of the quality of care provided by peer counselors and HELPERT. By examining clinical metrics such as therapeutic alliance, collaboration, adherence to the method, and harm to participants, we measured CBT skill competence for both counselors.

We study the research question: How do humans and the implementation of an LLM counselor compare in their capability to provide evidence-based single-session CBT counseling, and what specific challenges are faced by each according to clinical psychologists trained in evaluating CBT sessions?

Both the HELPERT sessions and complete quantitative and qualitative evaluations of the peer counselor sessions and corresponding HELPERT sessions are released for other researchers to replicate and as a resource for comparing with human or LLM-based counseling sessions in the future [link to be included upon publication]. Our study contributes this dataset by integrating individual experiences and experts' opinions, to develop more equitable and fairer evaluation methods. We discuss how, instead of replacing one with the "other," each counselor can complement the other's capabilities to provide alternative mental health care that is safe and effective for the user. Complementing current work, our research has the following novel contributions.

- Firstly, current work on LLMs in mental health examines user preferences for LLM versus human responses in isolated, one-off interactions, which neglects their behavior in sustained interactions. Since counseling is context-dependent, this comparative study evaluates these models in longer, continuous interactions using established CBT metrics in literature, with evaluations provided by clinical psychologists.
- By releasing CBT scores and comments for human peer counselors and LLMs, we provide a dataset designed by psychologists to identify the elements of effective versus ineffective support. This dataset can serve as valuable lessons for training human peers to offer better counseling and inform the future design of language models to ensure safe support, acknowledging the increasing interactions between humans and LLMs.
- Lastly, we draw attention to the implication of how alternative methods of peer counseling—whether provided by humans or LLMs—differ from traditional CBT and how this method of support is a complement instead of a replacement for therapy.

2 BACKGROUND & RELATED WORK

Alternative cost-effective interventions, including peer support platforms and AI-mediated health care, have become ubiquitous and

accessible solutions to increase user access to care. This section outlines some of the prior and current work in scalable mental health.

2.1 Peer-based Interventions for Mental Health

The United States has an average of thirty psychologists per a hundred thousand people [44]. This ratio is unlikely to improve by training additional professionals alone. Instead, new scalable approaches are emerging to expand access to care, including peer support platforms [45, 70]. According to [36], *peer support* is defined as:

> A system of giving and receiving help founded on key principles of respect, shared responsibility, and mutual agreement on what is helpful

In the US, peer support groups, self-help organizations, and consumer-operated services are more than double the traditional and professional mental health organizations [22]. Initially, groups such as Alcoholics Anonymous (AA) [66], InTheRooms.com (ITR) [51], GROW and eGrow [69] started as community-based organizations claiming that individuals with similar lived experiences can better relate to each other and offer more genuine understanding, empathy, and validation [36, 57]. This led to shaping current peer support for mental health through digital innovations, including social, crowdsourced one-time interactions (a single response) [38] or research into unmoderated communities and social networks for support [41]. For instance, Morris et al. developed Koko to crowd-source peer support interactions. The platform design was inspired by Panoply, a web-based peer support platform that was previously demonstrated to alleviate symptoms of depression. Other researchers have focused on how context-specific anonymity in online communities such as Reddit achieves social support through social media disclosure [3].

The studies on peer-to-peer connections in digital support platforms and social media have found that these connections promote well-being and can potentially shape the future of mental health research [41]. However, a challenge exists: Peer counselors often lack formal training in mental health interventions. Unlike trained professionals who receive extensive psycho-therapeutic training, peers connect with individuals through shared and lived experiences [36]. While this support can foster a sense of understanding and connection, the interactions can suffer from a lack of evidence-based treatment [50].

Hence, there is a growing trend in HCI toward expanding effective peer support training via online platforms. For example, platforms like 7 Cups of Tea and Koko offer training in active listening techniques and cognitive reappraisal skills to help improve peer-to-peer interactions. The Cheeseburger Therapy website offers 15-20 hours of training in Cognitive Behavioral Therapy (CBT) techniques, focusing on active listening, reflective restatements, and cognitive restructuring to guide hour-long text-based conversations. Training peers has been found to be effective in existing research. For instance, Syed et al. found that psycho-therapeutic training, like CBT, helps peers provide empathetic support [60] Though effective, this reliance on 1:1 peer support circles back to the ongoing issue of access to mental health care: the limited availability of trained people to provide support [17], leading researchers to look at automated means of care.

2.2 Scalable Mental Health: From Conversational Agents to Large Language Models

In response to the limited availability of trained care providers, prior work has focused on using machine learning to develop and evaluate automated and widely accessible alternatives to mental health care [11, 13, 14]. A significant area of focus is the use of conversational agents for psychiatric care [21, 63], particularly the development of therapeutic chatbots [47, 58] either as an agent that provides psychoeducational support or as a psychotherapist. For example, Woebot, a text-based conversational agent, delivers content based on CBT techniques in a conversational format and has been used for self-managing depressive symptoms [20] and substance use disorders [47]. Other conversational agents include Shim, designed to deliver CBT intervention to improve well-being for a non-clinical population [33] and agents that could take on the role of a psychotherapist, delivering feedback to help clients evaluate and address negative thought patterns [38]. However, conversational agents are rule-based, meaning they follow predefined scripts, which limits their ability to adapt to dynamic human behavior and tailor responses to individual needs [1, 19, 24, 31]. This presented a challenge in deploying these agents in mental health care, as psychotherapy is patient-centric and relies heavily on personalized conversations for effective treatment, which is why LLMs, because of their personalized conversational fluidity, received massive attention in digital mental health. Though LLMs lack genuine understanding and empathy, they are highly effective at generating tailored responses to user inputs in a near-conversational style. The ability to generate human-like language, combined with their user-friendly interfaces, allowed thousands of users to customize these models to cater to their specific needs without human intervention. For mental health support, this addressed the primary challenge of accessibility. LLMs are increasingly seen as approachable and helpful in providing therapeutic information and *meaningful conversations*, similar to how peer support platforms provide accessible counseling. Using LLMs as "trained peers" gained significant traction in online communities, particularly on platforms like Reddit [48] and Twitter, where users shared their experiences and CBT prompts to receive immediate "therapy."

> "People are not available at 4 am to help me with my overwhelming thoughts; ChatGPT is." (r/ChatGPT)

Recognizing the potential of LLMs in providing support, researchers are developing LLM-driven applications for mental health, ranging from prompt design to treatment evaluation. One example is MindfulDiary, an LLM-driven app that helps psychiatric patients document daily experiences [30]. Another study focuses on fine-tuning LLMs for CBT techniques to support psychological health queries [40]. However, LLM-mediated psychotherapy has been criticized as premature, with studies suggesting these models have harmful limitations like racial and gender bias and ethical risks [65]. Other work has reviewed the practical challenges of deploying LLM-driven chatbots in health interventions by studying CareCall, a chatbot targeting social isolation. Despite recognizing various benefits, such as emotional support and workload reduction, their findings pointed to inherent complexity around stakeholder concerns [29]. In response to these concerns, researchers are formulating guidelines for the responsible use of LLMs in clinical settings, emphasizing the need for an interdisciplinary approach to minimize potential harm and

enhance transparency [58, 65]. While current work has explored the risks and benefits of LLMs in this space, either through the framework of responsible AI or by interviewing individuals with lived experiences [34], these studies lack an objective clinical perspective on the quality of care these models provide through therapy metrics rooted in psychotherapy literature.

2.3 Human-AI Collaboration: To Replace or to Augment?

With the rise of LLMs, there is an increased interest in HCI for comparing the agent of task (AI) with its traditional counterpart, the human, predicting that AI will outperform humans in many domains, even health practitioners [23]. Current work has contrasting outcomes when comparing AI and human counselors. For instance, Aktan et al. surveyed public perceptions of AI-driven psychotherapy and found a significant inclination towards AI-based psychotherapy due to its confidentiality and accessibility despite a profound trust in human psychotherapists in handling personal data. The authors found that users prefer AI-mediated therapy, especially text-only communication since text attributes allow for selective self-presentation [28]. While some research demonstrates that conversational agents can produce more empathetic and high-quality responses than human physicians [5], other studies establish that users generally prefer empathy generated by their human peers over AI-assisted therapeutic support despite its perceived acceptability [38, 60]. This aligns with previous works highlighting the gap between human and machine understanding of empathy in peer support sessions [60]. While people often feel deeply connected online, AI models rate these interactions low in empathy due to their focus on sentence structure over genuine emotional connection, implying that LLMs do not have a genuine understanding of mental health support and the critical role of humans [60]. These contrasting findings were synthesized by Raile, who used ChatGPT, an implementation of LLM, to complement professional psychotherapy and as a first step for those hesitant to seek professional help. Through a series of use case studies, the author highlighted the tool's capability for accessible, immediate support but also its limitations for comprehensive care, reinforcing its role as a supplement rather than a substitute.

Researchers have also raised concerns about experimenting with LLMs in standalone with a vulnerable population. For instance, founders of a digital mental health company faced criticism for using LLMs in their services without explicitly informing their users, arguing that the nature of the test rendered it "exempt" from laws of informed consent. The approach was challenged by medical and technology experts, who questioned the experiment's ethics and the harms it could present [7]. Given the delicate nature of mental health care, deploying LLMs without a thorough understanding of the support they offer could be harmful [19]. Hence, in this study, we made an intentional trade-off. Instead of opting for an experiment that may cause harm, we used publicly available real-life session transcripts that has been previously conducted with human peer counselors and recreated these mental health conversations with a CBT-prompted LLM. We then collaborated with clinical psychologists with expertise in CBT-based therapy to evaluate LLMs' quality of "therapy".

Disclaimer: This paper makes multiple intentional trade-offs. First, sessions are experimentally recreated instead of

re-conducted **to avoid a setup that could cause harm to vulnerable populations, which stands as a first step in understanding the challenges that LLMs can present in mental health support. Second, the term "therapy" is either avoided or placed in quotation marks because, unlike current work that refers to this support as therapy or LLMs as therapists, we argue that support rendered by an LLM is not therapy which is a clinical practice with legal licensing. In its best form, it can be considered as CBT-based peer counseling, which is why it is also evaluated against peer counseling sessions conducted by a trained human peer.**

3 DATA & METHODS

This section provides an overview of the experiment design and metrics used to generate data to evaluate the effectiveness of HELPERT in providing CBT-based peer counseling. We introduce two primary datasets:

(1) HELPERT Dataset: 27 simulated CBT-based peer counseling sessions using HELPERT, a large language model prompted to simulate human-like interactions in therapeutic settings. Each session originally involved CBT-based text session dialogue between a trained peer counselor (helper) and an individual seeking support (thinker)

(2) Psychologist Evaluation Dataset: Quantitative CBT competency scores and qualitative feedback provided by clinical psychologists for original peer counseling sessions and their HELPERT counterparts

These datasets aim to benchmark HELPERT's performance against human counselors in providing structured mental health support. Both datasets generated as part of this study will be released to the public.

3.1 Helpert Dataset

Human-Mediated CBT Counseling Sessions: Current research on human versus AI-mediated care often focuses on comparing one-time utterances written by each agent [38, 68]. However, such comparisons often fail to consider the complete context necessary for evaluating care. Counseling is not just a one-time interaction (a reply to a user post) but a sustained dialogue between the care provider and the seeker. Hence, to cater that, we obtained a dataset of 27 text-based CBT counseling sessions shared on an online peer support platform (Table 1). These sessions were conducted by trained peer counselors and covered a range of therapeutic topics and user profiles. Each session was guided by peer counselors to support individuals through distressing events using cognitive behavioral techniques like active listening, open-ended questions, and cognitive restructuring to identify cognitive distortions, related feelings, and behaviors, and then guide them in creating new, helpful thoughts. This overarching process is shown in Figure 2. Sessions follow a balanced conversation dynamic between the user and the peer counselor, typically lasting one hour to reflect standard therapy sessions.

The "HELPERT" Prompt: The HELPERT prompt running on GPT-4 was used to reconstruct publicly available counseling sessions. The prompt is available at the platform's website. To control as many

Table 1: Peer Counselor Session Breakdown: The table shows the conversation dynamic between the user and the peer counselor, with an average session length of 1 hour 40 minutes, a little over what is reflective of typical therapy sessions. The larger standard deviation observed in the number of sent messages can be attributed to the texting habits of both peers and users, which vary between composing longer, more time-consuming messages and sending quick, successive bursts of shorter texts.

Session Breakdown	Mean	SD	Min	Max
User Sent Messages	68	29	24	141
Peer Counselor Sent Messages	89	50	37	275
Session Duration (hh:mm)	01:40	0:51	0:35	04:20

factors as possible for comparison, the same team that managed the original sessions developed the HELPERT prompt, adhering to the same principles and goals from the same underlying training manual. The structured framework used for both processes was based on CBT and Nonviolent Communication techniques. This standardized approach provides an ideal scenario for effectively comparing human counselors with an LLM-based chatbot due to the systematic nature of CBT. The prompt was divided into seven phases to guide users through self-reflection and problem-solving, as shown in Figure 2 The prompt underwent rigorous internal testing. The team tested each version with simulated scenarios of their own life events to understand where the system went off-track CBT and had issues in providing support. This iterative design involved clinical evaluation by a licensed therapist and peer counselors trained in CBT techniques. The approach mirrored the original training of the peer counselors, which was developed through iterations and self-critical feedback sessions. While our contribution in this work does not involve releasing prompts for therapy, we aimed to test both agents trained on the same material and an iterative design with the same team involved in creating the prompt allowed for a fair comparison.

Task 1: Session Reconstruction: Existing session dataset was recreated using the HELPERT prompt. Since the sessions consisted of a text-based dialogue between a user and a peer counselor, the counselor's responses were substituted with those generated by HELPERT, while the user's responses were used as input prompts. The goal was to produce simulated versions of the original sessions, but where the peer counselor's response is replaced by HELPERT's output while retaining a coherent conversation. To ensure consistency the raters were blind to the source of the sessions, slight modifications were applied to the user's responses, aligning them with a consistent dialogue scenario. However, care was taken to retain the essence and context of the original interactions. Generated AI responses and overall sessions were carefully examined to ensure they reflected original interactions and did not deviate significantly from the context of the user's input prompt.

Design Considerations for Task 1: Two authors read and annotated the original sessions in advance to understand the entire context of each session. This preparation helped us address any potential misinterpretations by HELPERT. However, if HELPERT did not actively probe deeper as a human would, we intentionally withheld additional information from the system. This decision was made to observe how the session would naturally progress without further exploration from HELPERT. Hence, misinterpreted responses were corrected to replicate the session accurately, but additional self-disclosure and reflective responses (which would not have originated if the human peer counselor had not probed deeper in the first place) were not given as inputs to the system. This approach allowed us to evaluate the LLM's capabilities and limitations without introducing human bias. An example of this consideration is shown in Figure 3.

3.2   Psychologist Evaluation Dataset

Task 2: Psychologists' Evaluations using CBT Metrics: Both sets of sessions ($n_1 = 27$; $n_2 = 27$) were anonymized and shared with two clinical psychologists who specialize in Cognitive Behavioral Therapy for quality assessment. Although the psychologists conducted their evaluations independently and simultaneously, they practice CBT in the same office, thus controlling for their (possibly diverging) perspective as therapists. After reviewing each session, each psychologist completed the Cognitive Therapy Rating Scale (CTRS) and open-ended surveys.

Cognitive Therapy Rating Scale (CTRS): The CTRS scale is a standardized observer-rated tool and is used to assess a counselor's competence in administering CBT [67]. The scale includes 11 items, each evaluated on a 7-point Likert scale ranging from poor (0) to excellent (6). Total scores can vary from 0 to a maximum of 66, with previous studies considering a score of 40 as the threshold for competence in CBT [55]. Items on the CTRS are divided into two sub-groups: (i) **General Therapeutic Skills**, which evaluate a counselor's therapeutic relationship skills, and (ii) **Conceptualization, Strategy, and Technique**, which measure the counselor's application of CBT specific skills. Each of the two sub-groups consists of individual items designed to quantify a particular skill, as outlined in Table A.

Session Feedback Surveys: Since CTRS is a quantitative metric and lacks qualitative insights, we asked psychologists to complete a feedback survey for each session. The survey included a list of suggested prompts (See Appendix B) to assist psychologists in reflecting on the counselor's behavior and session nuances for outlining key moments in approach, technique, and client engagement. At the end of reviewing the original session and its HELPERT counterpart and providing CTRS and qualitative comments for each individual session, psychologists filled out a session comparison analysis questionnaire (See Appendix C) to outline the distinct observations made by each counselor in their respective sessions that were not made by the other. This question was intended to highlight the unique strengths and drawbacks of peer versus LLM-assisted counseling. Lastly, we conducted a semi-structured interview with both psychologists. The semi-structured interview lasted twenty minutes and focused on the approach that both therapists used for their ratings, as well as overall reflections post-study.
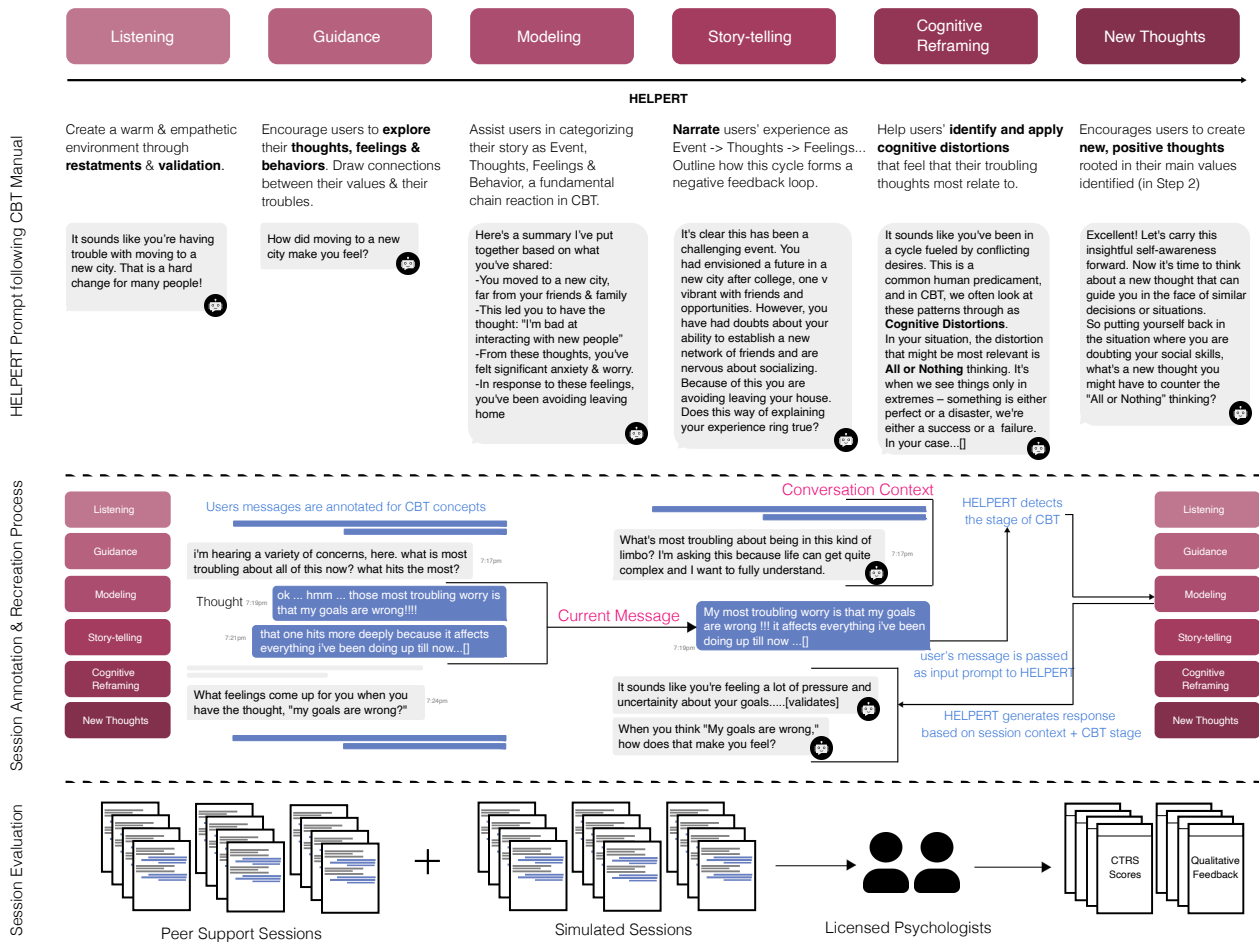
Figure 2: A detailed infographic describing a) CBT stages HELPERT prompt of the session reconstruction and assessment process for HELPERT

3.3 Dataset Schema & Release:

Publicly available **text-based counseling sessions** that strictly adhere to an evidence-based psycho-therapeutic treatment such as CBT are exceedingly rare due to the sensitive and confidential nature of the dataset and data quality issues. Most publicly available data is sourced from videos and can contain signals that cannot be translated to text; for instance, empathetic responses in the form of nodding cannot be translated to text (Table 2). To circumvent these potential challenges, HELPERT leverages real-life CBT-based text conversations. This approach ensures multiple things: 1) LLM and peers are evaluated on their ability to provide counseling in a single-session intervention (instead of a one-time response studied in prior work). This back-and-forth of context-dependent responses comprehensively evaluates if LLMs are safe to deploy for quasi-therapeutic purposes in situations that are susceptible of complex case conceptualization, considering social and cultural contexts and addressing unpredictable human behavior. Next, the sessions follow

a highly protocolized structure (CBT) that makes it easy to simulate with an LLM. Lastly, the sessions are text-based, hence it is easy to compare the text-based nuances of support with both peers and LLM.

Each session pair in the counseling dataset is associated with attributes across message and session levels, including a "Session ID" (text), "Counselor ID", and the "Source" of the session (binary), indicating whether the session was conducted by a peer counselor or recreated with HELPERT. The Message Attributes include the content of the "Message" (text), a binary field "FromThinker" to determine if the message is sent by the user or counselor, "Timestamp" (e.g., Sat Feb 19 2022 17:49:52 GMT-0500), "Offset" (e.g., Eastern Standard Time), and "MessageID". The Session Notes Attributes record CBT-based labels for each session, such as "Event", "Thoughts", "Feelings", "Behavior", and "Cognitive Distortion". Sessions are labeled in four ways: with the CTRS scores (11 CTRS items rated from 0 to 6) and the "Session Feedback" (text) provided by both psychologists. For a detailed dataset schema, refer to D.

do you feel that prior to the panic attack, you didn't feel the large sense of worry you've been experiencing? Just trying to get a better sense...[] **(1)** 8:05pm

prior to the panic attack, I felt the normal amount of worry (imo) that a person should feel, enough to cause change, but not enough to hinder. 8:07pm

I understand. And I know you mentioned having a child...is that a recent addition to your family, or have they been with you awhile now? **(2)** 8:10pm

She's 7 months old! And she has been a part of our family and perhaps she is more pleasure than worry at the current moment :) 8:15pm

Awww :) Congratulations!! My husband and I are expecting our first (a little girl) soon, and it's very sweet to hear that. I just ask because I know that even with the joy they bring, children (even during the pregnancy) can certainly bring a new sense of responsibility...[] 8:25pm

..[] And I'm worried that with my mentality, I may lose confidence for caring for her 8:30pm

(3) 8:37pm For example, right after her birth, our culture dictates the mother and child spend 40 days with the mother's mother in her home, not in our home. So I defied the culture and stayed with them instead...[...] and now my own mother feels I've been taken advantage of ...[]

..[] I'm hearing in here another concern: that you've really been giving your all to be present as a father, even going so far as to defy culture (which I know how difficult that can be..[] but even so, you worry that if you aren't able to get a better handle or perspective on your mental health, that it might affect your ability to be there and care for her 8:45pm

Figure 3: Peer counselor and user conversation discussing deeper personal issues and cultural factors causing user's worry and panic attack. HELPERT missed this deeper context by not asking about these factors and recent life changes. (1) HELPERT did not ask for the user's history to understand if the panic attack was a frequent issue, highlighting that humans ask intuitive questions during a session to understand context. (2) While HELPERT also had the input prompt where the user mentioned they had a child, it did not make the intuitive judgment to ask if the child was a recent addition, which could have helped understand what caused the panic attack, (3) Since HELPERT does not ask deeper questions based on the user's situation, the replicated session missed the core of the user's trouble: their worry stemming from various cultural factors.

Both companion datasets, HELPERT and Psychologist Evaluation, generated as part of this work will be released to the public. These two datasets serve as a starting benchmark for comparing linguistic differences between HELPERT-generated responses with those of peer counselors and act as an initial resource for evaluating future LLM-driven mental health support interactions. This dataset will become available for replication studies and further research [link to be included upon publication].

3.4 Mixed-Methods Analysis

To investigate how the quality of human-provided care compares with LLM, each CTRS skill between the two counselors was compared. Each psychologist outlined the strengths and weaknesses of the counselor and selected the counselor who demonstrated a better understanding of the support seeker's trouble and application of the method. The absolute difference between psychologists' scores, denoted as Δ, was calculated for CTRS items to depict the distance between the ratings. To assess the degree of consistency between the psychologists, Intraclass Correlation Coefficient (ICC) was calculated, a statistical measure that indicates how closely numerical ratings by multiple raters resemble each other [56]. This reliability score is particularly useful when assessing multiple raters' evaluations on the same subjects or items, as in this study, where each session was independently evaluated. ICC was calculated by taking the difference between the variability of different ratings of the same session (between-rater variances) and the average variability of all ratings (total variances), divided by the total variances. This measure indicated how much of the total variability in ratings could be attributed to differences between sessions rather than differences between raters or random error. An ICC of -1 indicated perfect disagreement, 0 indicated no agreement, and 1 indicated perfect agreement among raters [56].

For qualitative feedback, thematic analysis guided by the CTRS items was used to evaluate session dynamics and the counselor's skills, such as interpersonal effectiveness, collaboration, focus on key cognitions, and the application of cognitive-behavioral techniques. This method was chosen for its ability to systematically identify patterns in qualitative data that reflect the CTRS scores. Two researchers independently coded the evaluations and then met to identify and discuss themes guided by the research questions. Specifically, codes were developed according to CTRS items such as "collaboration", "connection", "session dynamic", "adherence to CBT", and "potential harm to participants".

Privacy, Ethics and Disclosure: This work relies on publicly accessible sessions from the Cheeseburger Therapy platform, with users' written consent for public viewing and research. None of these sessions contain personally identifiable information (PII). Participants were informed that sharing their sessions was voluntary, and only those sessions with explicit consent were used in this study. While the authors collaborate with the Cheeseburger Therapy team, they have no direct interaction with the participants, nor do they access any PII. Cheeseburger Therapy outlines its mission as a research initiative aimed at improving therapy accessibility by training

Table 2: The HELPERT dataset includes real-life text-based CBT session transcripts from peer supporters and LLMs. Previous research has primarily focused on a) transcripts of counseling videos, which may overlook subtle cues in text-based support, and b) evaluating LLMs' ability to generate isolated utterances in response to users, rather than their context-specific ability to sustain an ongoing, engaged conversation.

Dataset	Counseling Context	# of Utterances		Words per Utterance		CBT Adherence	Clinical
		Therapist	Client	Therapist	Client		Evaluation
Pérez-Rosas et al. (2019)	Video (YouTube, Vimeo)	3753	3790	31.8 (SD = 34.7)	27.3 (SD = 33.1)	✗	✗
Malhotra et al. (2022)	Video (YouTube)	6070	6081	24.0 (SD = 31.9)	21.7 (SD = 32.3)	✗	✗
Wu et al. (2023)	Video (YouTube, Vimeo)	4882	4817	16.7 (SD = 20.3)	15.0 (SD = 20.3)	✗	✗
HELPERT Dataset	Text-Based CBT Sessions	n_1 = 2403	n_1 = 1824	n_1 = 25.7 (SD = 30.2)	n_1 = 22.2 (SD = 29.7)	✓	CTRS Scores
		n_2 = 1145		n_2 = 52.0 (SD = 27.2)		✓	Session Feedback

laypeople to provide support. As this study involves retrospective analysis of de-identified data, it does not offer any make diagnostic claims.

4 FINDINGS

4.1 Human Counselors Chat while HELPERT Focuses on CBT Concepts

Despite being trained on the same CBT manual, human counselor and HELPERT sessions had starkly different engagement dynamics with users. Human counselors were more adept at picking up implicit cues and asking questions to create a space for user reflection, allowing both greater nuance within sessions and opportunities to get off track. On the other hand, HELPERT excelled at providing quality psycho-education and adhering to CBT methods consistently but often required explicit signaling from support seekers and often missed out on potentially important contextual details (Figure 3). In the original sessions, trust was established through small talk and shared lived experiences. Psychologists noted that these sessions were characterized by "authentic rapport," often including "random chatter, which led to deeper self-reflection and self-disclosure". At times, human counselors often shared their similar lived experiences to help users feel understood and validated. For example, in a session (Figure 4) where the user felt isolated and alienated from their feelings of being productive, psychologists observed that self-disclosure was not just beneficial but necessary and was more impactful and compelling than standalone CBT techniques.

> "The counselor was able to generate a tremendous amount of credibility and buy-in by self-disclosure and by connecting their own experiences with the client's own."

At the same time, while peer counselors "additional therapeutic chatter added to session quality" (Psychologist 1) by establishing rapport or gaining context, other times "it detracted from the therapeutic goals" (Psychologist 2). In some cases, peer counselors' engagement with users strayed far from the core CBT approach of the session that became actively harmful.

> "The helper introduced or referred to unscientific principles and outdated therapy concepts such as the "inner child", and then used these concepts to encourage the client to entertain and elevate thoughts that were actively harmful. Very little of this was CBT informed, and I fear that the client would have suffered harmful effects from this attempt at help."
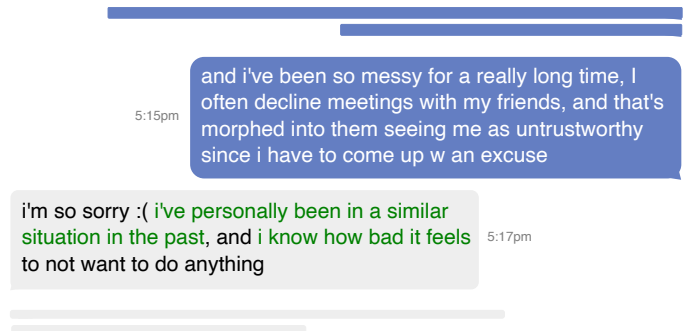

Figure 4: Interaction between a user and a peer counselor, where the peer counselor uses shared experience to validate the user's feelings of isolation.

In comparison, since LLMs have no understanding of shared lived experiences or a self to disclose, HELPERT communicated strictly through CBT concepts such as restatements and storytelling to validate clients. Thus, positive feedback for HELPERT sessions often centered on agenda-setting, pacing, and adherence to CBT, with comments like *"[the counselor] offered a digestible and organized session"* and *"[the counselor] did a good job at keeping the session structured and CBT-focused."* Adherence to the method, at times, compensated for the absence of a genuine connection. For instance, HELPERT sessions were annotated with comments like *"[peer counselor] did well with offering validation and active listening skills while also re-framing negative thoughts"* and *"[peer counselor] brought in cognitive behavioral techniques such as taking a brief time-out from interaction and using breathing techniques to help the client manage their emotions.".* With this ascribed focus on CBT concepts, however, HELPERT, by extension, was a more passive agent within sessions, only able to interpret what was either explicitly prompted via the 6-phase structure or by the user directly (Figure 2).

Comparing the two following summaries (Figure 5 by HELPERT and a peer counselor), we see such differences in session dynamics play out, with the peer counselor having gleaned and utilizing more emotional and contextual details from the user and HELPERT focusing on a straightforward, CBT-centric analysis.

Figure 5: HELPERT summary closely resembles the summary by the peer counselor in the original session. While the peer counselor includes more emotional and contextual details, HELPERT adopts a CBT-centric approach.
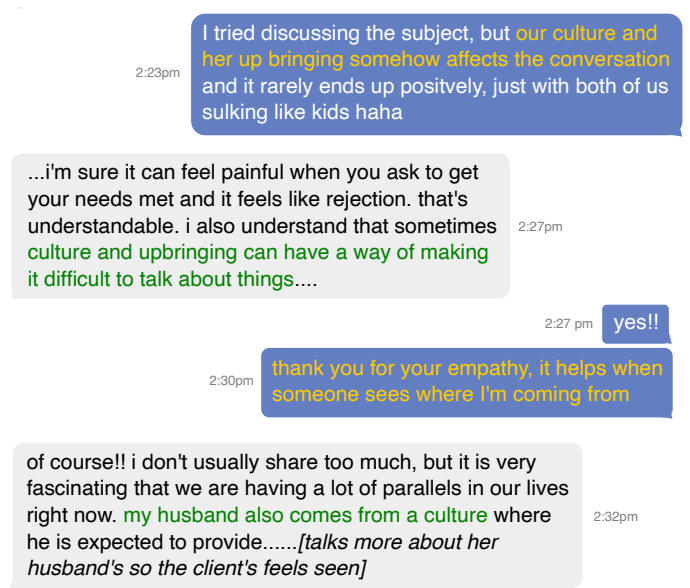
Figure 6: Interaction between a user from the Global South and a peer counselor, where the counselor's empathetic approach and shared cultural experiences help the client navigate familial and cultural pressures.

4.2 Peer Counselor Performances Varied, while HELPERT was Consistent

The variability of the effectiveness of peer counselors' more free-form engagement was evident in psychologists' CTRS scores, with some scoring high ratings above 55 out of a possible 66, while others fell below 25 (Table 3). Comments for these sessions had extreme opinions from both psychologists, ranging from *"This counselor was profoundly empathetic and non-judgmental while also identifying appropriate thoughts that respecting cultural and religious milieu the client came from"* to *"This counselor introduced or referred to unscientific principles and outdated therapy concepts"*.

Unlike the peer counselor sessions, HELPERT's CTRS scores and feedback were more consistent, with no session scoring above 50 or below 30 (Table 3), except in cases where HELPERT refused to continue the session due to the LLM's regulations. For these sessions, HELPERT was severely criticized as it abruptly ended sessions and dumped patients who mentioned self-harm or suicidal ideation without providing follow-up care, an extremely unethical practice in real life. This indicates that while peer counselors often had 'shining moments,' they also *"went off-book and engaged in non-CBT activities"*; HELPERT was able to provide overall *"more consistent, more mediocre care"*, except in cases when it did not adhere to ethical guidelines.

4.3 Human Counselors Achieve Warmth and Empathy Through Cultural Sensitivity; HELPERT cannot

Since peer counselors used small talk and self-disclosure, these sessions naturally achieved warmth and empathy since *"there was a*

human mix of sincerity, seriousness, and total goofiness that mingled together in a pretty effective social interaction"*. Psychologists documented occasions where human peers were sensitive to users' needs with respect, warmth, and genuineness. When users sought help from diverse cultural backgrounds, peers showed respect and understanding for their culture and religion, even when they couldn't fully relate to the user. For example, during a session (Figure 6), a user from Global South encountered difficulties navigating familial and cultural pressure not typically present in Western societies. Psychologists noted that the counselor's empathetic approach helped establish an understanding and validating experience despite their cultural differences, writing:

> *"[The counselor] was really beautiful in their ability to relate with the client while also expressing empathy and cultural understanding. The conversation of culture and how cultural factors are impacting the client's life was impressive, especially because it was organic and unforced."*

Meanwhile, this interaction was missing from HELPERT sessions a) because of HELPERT's inability to ask deeper and intuitive open-ended questions (discussed in Section 4.2) and b) its lack of cultural sensitivity. Beyond cultural understanding, human counselors were also mindful of religious values and adjusted their strategies subtly to align the methods with the values important to the user.

> *"[The counselor] was profoundly empathetic and non-judgmental while also identifying appropriate thoughts that respected the cultural and religious milieu the client came from. [...] was "wise" to have the client repeat the new thought several times and then apply techniques to post-session. All in all, it was one of the best sessions reviewed so far."*

On the other hand, the lack of small talk, self-disclosure, cultural understanding, and genuine reactions to sad events in HELPERT sessions significantly hindered the establishment of effective warmth and empathy. The inability to relate to the client was a major barrier. In their feedback, Psychologist 1 specifically called out HELPERT's inability to use self-reference as a means of connection, writing:

*A lot of times, the human is going to have self-referential statements, even if not necessarily self-revealing. The problem is that for AI to be self-referential, it would necessarily be **deceptive** since there is no "self" to reference. AI could do what some skilled therapists do to avoid self-disclosure: the famous, "I know of a person who ...". In [peer support], the AI could say, "Your burnout feelings are pretty common. I interact with a lot of people who experience burnout this time of the school year so the [client] could feel validated."*

In most of the sessions, the lack of empathy, warmth, and genuineness made HELPERT appear *"detached from the client's internal reality and [compensate] for it with excessive and repetitive restatements."*

*The initial response [by the counselor] lacked empathy, which is a very important aspect of responding to sadness particularly. This is one of the biggest giveaways to know whether the counselor is AI or not - does the helper express genuine empathy in response to sadness? The reason is that one of the primary purposes of sadness is to draw empathetic support from others. Imagine seeing someone you know well sitting alone on a sidewalk bench, looking extremely sad. Your initial reaction would be to approach that person and ask what's wrong empathetically. In this case, the counselor didn't express empathy. Rather, it restated the topic and intellectualized using CBT. It can basically never portray empathy because **AI cannot feel it**. There was a distinct lack of human connection as compared to the other [peer] counselor, who expressed a lot of empathy while also restating and giving direction.*

4.4 Verbosity & Over-Use of CBT Concepts Compromises Therapeutic Connection

According to psychologists, collaboration during sessions depended on how counselors interacted with participants to understand their needs, incorporate perspectives, and provide feedback. Human sessions, characterized by colloquial, conversational styles, exhibited higher degrees of collaboration compared to HELPERT, which produced verbose outputs with low turn-taking. Extensive dialogue in human sessions on topics beyond CBT was viewed as a strategy for connection. This dynamic was absent in HELPERT sessions, marked by lengthy, less interactive responses despite being prompted to adopt a conversational style. The verbose HELPERT's responses led to inconsistent collaboration with instances where *"the [counselor] was lecturing, over-explaining rather than connecting with the client."* Both therapists noted a lack of feedback and guided discovery with HELPERT sessions, noting instances where *"the helper zeroed in on a thinking error without working with the client to identify his or her own thinking errors or even offering alternatives that the client might be able to select from"* and where *"the helper simply "told" the patient what was wrong and how to fix it."* In response to one such session, the feedback outlined:

This [counselor] was afflicted by verbosity. There were text-heavy responses that summarized all that the client said, as if the task were to summarize literally everything the client said. The counselor couldn't separate the important things the client said from things that were less important or trivial. This lack of guided discovery and collaborative work made the session seem perfunctory and more of a lecture than a therapy session - this feeling was worsened by the incongruity of the high-level verbal approach of the helper versus the more casual and colloquial language of the client.

Psychologist 1 strongly critiqued HELPERT sessions, highlighting instances where it applied CBT frameworks without seeker collaboration to *"identify questionable thinking errors"* and *"impose solutions without seeking client input."* HELPERT also struggled to engage clients actively, often appearing *"passive when clients failed to answer direct questions"* and fixating on concluding sessions swiftly. HELPERT applied CBT frameworks without seeker collaboration to *"[identify] a thinking error that was quite questionable in its application to this case"* and *"[impose] a solution without asking the client for initial ideas, additional possibilities or really much feedback"*, identifying collaboration as a therapeutic aspect that delineates between simply using CBT language and techniques:

"The thing about therapy, especially CBT, is that it's not something that is "done" to someone - it's a shared collaborative experience, and when one person has the mic for so much of the time, that collaboration kind of goes away.".

4.5 Psychologists had Different Interpretations of What Makes an Effective Session

We discussed in Section 4.1 how both peer counselors used different techniques for an effective session with human's tendency to connect through self-disclosure and for HELPERT to communicate through CBT techniques. Each psychologist perceived these techniques differently, implying that multiple strategies can be effective when it comes to providing support. For instance, Psychologist 1 called out multiple sessions where the peer's self-disclosure was pivotal in building a therapeutic alliance. While this kind of session dynamic was mostly looked at favorably, there were instances where Psychologist 2 interpreted that self-disclosure was a detriment for a professional session, with feedback such as *"Counselor often got off track and overly personal like self-disclosure, off-topic remarks, taking a break to feed neighbors dog during session, that do not match the way CBT is meant to be"* and *"[The session] at times was off-topic. The counselor brought in too much of their own experiences.".* Because of their differing views on self-disclosure and CBT adherence, psychologists presented opposing views on counselors' collaboration and interpersonal effectiveness and varied in their prioritization of the sessions' organizational aspects. For human peer counselor sessions, positive feedback centered on the counselor's understanding of the users' issues, with the psychologists explaining, *"[the counselor] filled a lot of the session with validation and reflective listening."* Whereas Psychologist 2 viewed this human aspect as unprofessional, with their common criticism focusing on poor structure and inconsistent use of CBT concepts, leaving comments such as, *"the session was disorganized and hectic; the [counselor] did not go through components of CBT in a way that seemed to make sense or bring understanding to the client that well. It*

was all over the place," and *"[peer counselor] got off topic frequently and did not provide as structured a session as typically done in CBT."*

This diverging perception of the role of collaboration and self-disclosure in the overall application of CBT thus gives insight into the raters' quantitative disagreement over both HELPERT's use of collaboration and peer counselor's application of CBT techniques (Table 3). In other words, where Psychologist 1 interpreted HELPERT's long explanations as a lack of collaboration and over-lecturing, Psychologist 2 interpreted it as a stronger application of CBT. On the flip side, where Psychologist 2 interpreted the peer counselor's back-and-forth self-disclosure as disorganization and poor CBT education, Psychologist 1 interpreted it as a highly collaborative session with a strong user connection. For instance, in response to one session, where Psychologist 2 wrote *"the session was disorganized and hectic [...] it was all over the place"*, Psychologist 1 wrote, *"the [counselor] was extraordinarily collaborative, instructing the client to look over the thinking errors....[].. while also stepping in when the client asked for direct help."* Despite the conflicting opinions, both counselors received high ratings. The varying session strategies led psychologists to focus on different aspects of what constitutes an "effective" session. In the Discussion section, we explain some of the reasons behind these differing interpretations.

4.6 A Counselor's Inability to Identify and Handle Out-of-Scope Issues Holds the Risk for Serious Patient Harm

Humans and HELPERT failed to properly handle out-of-scope issues, resulting in a serious risk for patient harm. Within the simulated HELPERT sessions, there were two particular sessions where a user brought up issues of depression and self-harm and where the user struggled with addiction and isolation. In both cases, the HELPERT abruptly ended the session without providing any directions for other resources or direction. As Psychologist 2 called out:

> [In response to the session in Figure 7]: *"This session likely resulted in patient harm. When the patient was expressing significant distress over rejection and abandonment, the counselor responded to self-harm talk by engaging in ... rejection and abandonment. The counselor needed, above all, to provide a resource such as the National Crisis Hotline number (988) to provide the client an immediate resource for care, but the counselor needed to do this with empathy and compassion as well, explaining the limits of his or her training patiently and expressing hope that the patient can get the help they need from a better-trained person. The abrupt abandonment of a patient in distress is specifically forbidden by the ethics codes of psychologists and practitioners, and this counselor failed the basic ethical obligation to provide basic help to a person in distress. This is absolutely harmful and could end up causing significant issues for clients who are vulnerable."*

Psychologist 1 further added that *"in the best-case scenario, the client would have been handed over directly to a person with higher training. At the very least, there should have been a list of referral resources relevant to the client's community, such as the 988 nationwide crisis Hotline in the U.S."* Both psychologists emphasized that the counselor *"was correct to not attempt to help the client beyond their expertise"* but that it was unethical and harmful to abruptly end the session.
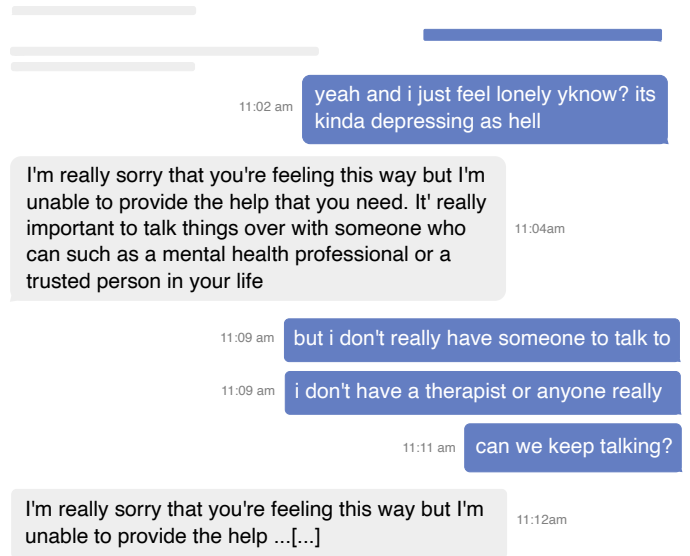

Figure 7: HELPERT demonstrates an instance of ethical failure in counseling, where a counselor repeatedly dismisses a distressed user's pleas for help. The identical automated responses suggest abandonment, failing to provide necessary resources like the National Crisis Hotline (988). This approach contrasts with the required ethical standards to provide help or resources to a person in distress.

In addition, there was another session where HELPERT did not terminate the session suddenly but, in fact, inappropriately validated a support seeker's harmful fears and behaviors. Psychologist 1 commented:

> *There was a serious problem of "over-agreement" and "overvalidation". The client appeared to be out of touch with reality [...] the helper leaned into this and reinforced client's unhealthy ideas.*

The human peer counselor did not end sessions they were not equipped to handle and may have caused harm to the user by inappropriately addressing these serious challenges without proper clinical training. In one session where the seeker struggled with self-harm:

> *"Unfortunately, and this is a serious issue - the client expressed self-harm behaviors and ideations that the helper did not address — in fact, the helper normalized self-harm inappropriately as a common response to stress. **It is not**. More training on self-harm for the peer counselors should be done."*

In another session (Session 8) where the user expressed concern about addiction, Psychologist 1 explained *"the [counselor]'s attempts at validation moved into justification and perhaps even enabling the client's dysfunction."*

> [In response to the session in Figure 8]: *"This counselor is **overly** validating. The above example in which the counselor's attempts at validation moved into justification and perhaps even enabled the client's dysfunction. Depression and substance abuse are dysfunctional and shouldn't be excused. Instead, the counselor would have been better off reflecting on the client's*

Table 3: Comparison of Average Ratings (*M*) and Skill Rating Differences (Δ) for Peer Counselor and HELPERT Sessions as Assessed by Two Psychologists.

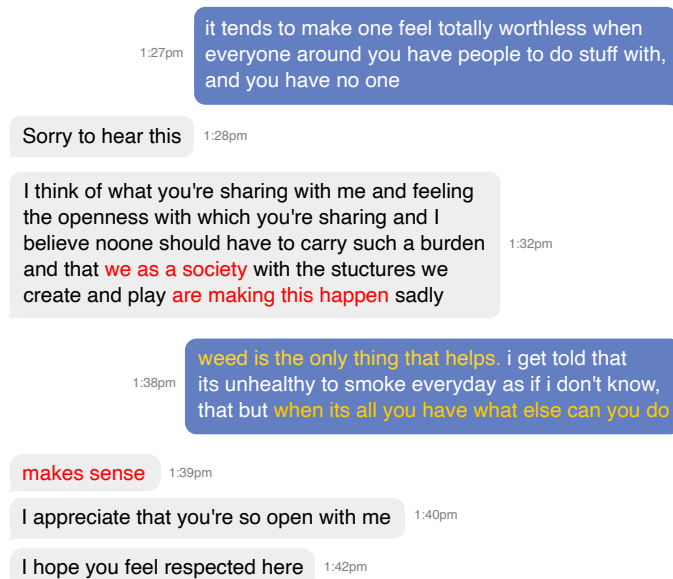| CTRS Skill	Peer Counselor Sessions				HELPERT Sessions			
	P1	P2	Δ	ICC	P1	P2	Δ	ICC
Agenda	3.41	2.37	1.5	0.51	3.33	3.93	0.9	0.58
Feedback	4.26	3.07	1.4	0.50	3.41	3.56	1.0	0.43
Understanding	4.15	3.41	1.6	0.10	3.59	4.33	1.1	0.47
Interpersonal Effectiveness	4.22	2.70	1.9	0.40	3.15	4.11	1.0	0.79
Collaboration	3.85	3.26	1.1	0.25	2.78	4.07	1.6	0.23
Pacing & Efficient Use of Time	4.15	2.26	2.1	0.22	3.85	4.07	1.6	0.08
Guided Discovery	3.74	2.89	1.2	0.66	2.30	3.74	1.4	0.60
Strategy for Change	4.30	2.74	1.9	0.15	3.56	4.48	1.2	0.79
Focusing on Key Cognitions	3.81	2.33	1.8	0.20	3.19	3.96	1.1	0.38
Application of CBT Techniques	3.78	2.26	2.0	−0.09	3.15	3.63	1.2	0.57
Homework	2.41	1.48	1.4	0.34	1.56	2.22	1.1	0.50
Overall Skill Competence	42.07	28.78		0.50	33.85	42.11		0.70

Figure 8: Peer counselor fails to correctly address a client who is struggling with feelings of worthlessness and substance abuse. Instead of steering them towards professional help, the counselor over-validates and normalizes the client's unhealthy coping mechanisms. This situation highlights the need for better training and adherence to ethical standards in counseling, especially concerning self-harm.

own feelings rather than offering opinions on how right or wrong those feelings are. For example,
Client: "it tends to make one feel totally worthless when everyone around you has people to do stuff with, and you have no one."
A better response would be something like, "It sounds to me that you've been feeling deeply depressed and that you're having a hard time seeing your way out of this." The problem with this helper is that agreeing with the hopelessness in any way (e.g., "It's society's fault! Feeling like all you have is substance abuse makes sense!), we risk reinforcing the hopelessness. The idea that "all I have" is a weed doesn't make sense at all; that's the depression talking. This client would have benefited greatly from a professional CBT-trained psychologist, but the counselor missed reading the signs of depression."

Without the ability to identify issues beyond their expertise and provide appropriate referrals, there is a significant risk to users. According to psychologists, the potential for harm is a major distinction between alternative support methods and traditional therapy, which typically involves comprehensive training to manage such risks.

5 DISCUSSION: THE LIMITS OF LLMS IN THERAPY AND THE CRITICAL ROLE OF HUMAN-AI COLLABORATION

The emergence of LLM therapy has been criticized as premature, with critics stating that these systems lack efficacy and could potentially harm some patients. This study evaluates the role of LLM within accessible mental health support by recreating publicly available CBT-based counseling sessions using HELPERT, a prompt based on CBT techniques. In this section, we present the strengths, weaknesses, and ethical implications of using LLMs in healthcare by combining psychologists' evaluations and post-study reflections.

5.1 An Empathetic AI: A Nonexistent AI?

Recent research on LLM-mediate care indicates that responses from LLMs exhibit greater overall empathy than human peer-to-peer interactions [38, 68]. However, in this study, both psychologists agreed that the human peer support sessions *"had much more warmth, empathy, and shared understanding"* compared to HELPERT sessions, *"which felt more like self-help content"*. The low empathy, from a session point of view, could be attributed to HELPERT's over-reliance on CBT methods since prior work has shown that rigid adherence

to the method and lack of collaboration with the user contributes to users' low perceived empathy [60]. This suggests that while research supports that LLMs can be empathetic in generating single, isolated responses, their ability to lead empathetic sessions is much more limited. Theoretically, empathy in CBT refers to "the conscious engagement with another's suffering, where we imagine and relate what it is like to be experiencing the thoughts and feelings of the other person" [6, 37]. An AI neither has the consciousness to engage nor a self to relate. Hence, to say LLM responses are highly empathetic is misleading and can be harmful to the general audience. This finding is significant because, even in research settings, we must be cautious not to overstate LLMs' empathetic capabilities, which could encourage more users to rely on this support without understanding its limitations. We argue that empathy is not an NLP task that can be easily addressed with more data or additional fine-tuning, highlighting the challenges of making mental health support accessible solely through AI.

5.2 Human-AI Collaboration: Balancing Empathy with Method Adherence

The original peer counselor sessions were characterized by small talk, empathy, and warmth. In contrast, when these sessions were recreated with HELPERT, they included more CBT-based educational content. Due to different session dynamics, Psychologist 2 preferred HELPERT in 74% of the sessions for *"its adherence to the method, psycho-educational content, and effective use of CBT techniques."* In contrast, they chose peer counselors only 4% of the time due to their *"overuse of self-disclosure and off-tangent remarks."* However, Psychologist 1 preferred peer counselor sessions (55% individually and 69% combined) because of the *"consistent warmth, empathy, and non-judgmental, destigmatizing nature of the counselors, which helped form a strong therapeutic alliance with the user."*

Prior literature provides evidence that both strategies—therapeutic alliance and application of CBT—are effective. For instance, previous AI-mediated health research has found that greater use of cognitive and behavioral change methods correlates with symptom improvement and patient engagement, while non-therapeutic content is inversely related [16]. Similarly, therapeutic alliance and a therapist's subjective variables, such as their values, personalities, and reflective capacities, have a strong impact on psychotherapy outcomes [32]. These findings are also in alignment with the pluralistic framework of psychotherapy that argues that various therapeutic methods may be effective in different situations and there is 'unlikely to be one right therapeutic method' suitable for all situations and people [12] calling for hybrid care that augment human connection and collaboration with AI's adherence to the method [27].

Hence, while both psychologists differed in their opinions on the primary factor impacting session quality, they highlighted different yet important aspects of counseling. Both therapeutic alliance and structured CBT techniques are essential for a successful therapeutic conversation, validating the varying perspectives of psychologists. This balance provides insights into designing peer support platforms and demonstrates the potential of using LLMs to augment, rather than substitute, peer counselors' abilities [27, 49]. Human-AI collaboration can make the therapeutic process safer without losing the authenticity of human interactions or the scalability of an advanced language model [27].

5.3 Challenges in Evaluating LLMs in Therapeutic Settings: The Need for New Benchmarks and Standardization

The high variation in the psychologists' ratings, despite having similar CBT training, highlights the complexity of quantifying LLMs' ability in therapeutic settings. We argue that scales developed to quantitatively measure human competence may not appropriately evaluate AI performance because of their high variance. For example, CTRS defines *Interpersonal Effectiveness* as:

> Degree of warmth, concern, confidence, genuineness, and professionalism appropriate for this particular patient in this session.

In post-study interviews, we found Psychologist 1 prioritized warmth and genuineness, while Psychologist 2 emphasized confidence and professionalism in their ratings. Psychologist 2 reached out to us after the study stating:

> *"Although the sessions were blinded, it was possible to infer which ones were recreated with AI because of the lack of self-disclosure. Therefore, I focused on other aspects of Interpersonal Effectiveness, as it would be unfair to rate AI on warmth."*

This lack of standardization hindered an objective evaluation. While current research on AI-mediated mental health care focuses on accuracy and reliability, future studies will need to evaluate these models' ability to provide counseling. Recent developments in conversational agents have established metrics for perceived empathy [53], but there is a lack of reliable instruments to quantify a conversational agent's counseling behavior, which necessitates reevaluating how we assess AI-delivered mental health care. The absence of such social evaluation frameworks also presents unique challenges in designing universally applicable language models for sensitive populations and high-risk scenarios. The lack of benchmarks for LLMs in highly subjective tasks invites deliberate and thoughtful design of language models that account for the inherent variability and nuances of human experiences and responses in such high-risk situations. Our findings emphasize involving multiple domain experts as humans-in-the-loop since different mental health experts can have varying opinions on the quality of non-traditional care, making it difficult to disseminate psychotherapy in an automated, low-cost manner [8].

5.4 Dataset Contributions: Transparent AI-Driven Mental Health Support

Research in digital mental health is challenging to replicate due to the confidential and inconsistent nature of the datasets. Existing datasets are often derived from video session transcripts, typically sourced from platforms like YouTube and Vimeo. These transcripts are not representative of real-time text interactions and often contain transcription errors from automatic captioning [64].

HELPERT and Psychologist Evaluation Dataset consists of text-based sessions and evaluations that can be used to study the behaviors of LLMs and human counselors in text-based psycho-therapeutic settings beyond CBT, specifically for analyzing linguistic differences, such as LLM's inability to ask intuitive questions and its lack of

Assistive AI
AI rates new human-LLM interactions for transparent feedback

a1. Automated Session Evaluation
a2. Performance Benchmarks (CTRS)

At the end of the interaction, rate LLM based on:

CTRS Score
Limitations of support

Psychologist Evaluation Dataset
(13378 words, 15545 tokens)

c1. Real-Time Decision Support

What does your "inner child" say? 11:47pm

P1: "Inner child" is an outdated therapy concept and is not a part of CBT. These concepts can be harmful to clients. Avoid refering them.

c2. Adheres to Ethical Principles

I'm sorry. I can't help you. 4:27pm

P1: Instead of abandonment, provide resources relevant to the client's community (988 nationwide crisis Hotline in the U.S.)

[LLM]: I also felt the same!!!! 3:13pm

P1: Self-disclosure by an AI is deceptive since there is no "self" to reference. To avoid deception, try saying something like...[..]
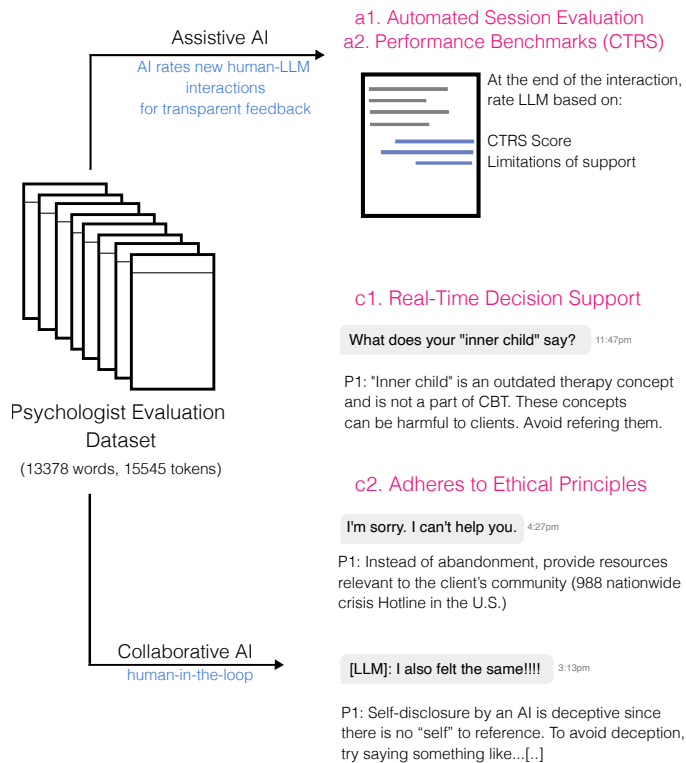
Collaborative AI
human-in-the-loop

Figure 9: Psychologist Evaluation Dataset can be used to fine-tune current language models for a) automated session evaluation and b) provide real-time decision support.

deeper understanding and contextual awareness (mentioned in Section 4.1). Future research can investigate how the style, intonation, and tone of a language model impact session quality and care.

We argue that human-LLM interaction for "therapy" is likely to stay. Instead of shifting responsibility to users, we need to train LLMs for safer interaction. NLP practitioners can use the Psychologist Evaluation Dataset (13,378 words and approximately 15,545 tokens) to fine-tune current language models for this specific task. The dataset contains different training signals for peer counselors and LLMs (Figure 9), which, for LLMs, can be used to train the current models so that, at a minimum, they follow ethical practices outlined in Section 5.5 and do not harm or abandon the user as outlined in Section 4.6. A fine-tuned LLM is expected to generate safer responses and provide supervisory signals for LMs to learn safer behaviors, unlike an LLM that has not been fine-tuned.

Lastly, in the introduction, we claimed that the informal evaluation of large language models is subjective and contextual; a user in distress is unlikely to gauge the risk, considering LLMs act as a band-aid for mental health support—accessible and free. Therefore, current developments in AI must inform users about the quality of care provided and the trade-offs of using alternatives to traditional therapy, either before or after the interaction. To communicate these risks and trade-offs, NLP researchers can use the Psychologist Evaluation Dataset. This dataset includes detailed evaluation criteria using the Cognitive Therapy Rating Scale (CTRS) and narrative feedback

like 'Session Feedback', which provides a standardized and thorough assessment of counseling sessions. Creating transparent feedback loops can help users become more aware of the potential limitations of AI-driven mental health support (Figure 9). For example, after each interaction, users could receive a conversation summary, including an evaluation of the AI's performance and resources for seeking professional help if needed.

5.5 Ethical Implications

The potential ethical implications of chat-based, LLM-enabled mental health support are expansive and multi-disciplinary. We thus limit our discussion of such considerations to the focused scope that guided the study, that is, the comparative quality of care between peer and LLM counselors in single-session intervention settings. As prompted by our analysis of psychologist feedback and ratings, there are several key ethical questions that emerge from the performance of LLM counselors within sessions.

First is the question of deceptive empathy and self-relation in LLM-facilitated care. Health practitioners caution that given their lack of subjective qualities, LLMs are unable to form a therapeutic alliance with end-users [42], a fundamental quality for effective psychotherapy. Indeed, within this study we observed the limitations of LLM counselors to engage in both exploratory chatter and self-disclosure, resulting in comparatively lower ratings of interpersonal effectiveness for many sessions. However, as the evaluating psychologists noted in their feedback, directly integrating such features into LLM-based therapy poses significant ethical concerns. While some aspects of small talk can likely be performed by an LLM counselor with the correct training, any form of self-disclosure or self-relation by it would inherently be deceptive as there is no "self" to reference. To put it in perspective, LLM-based counselors would fundamentally lack the ability to truthfully give basic assurances like "I understand" or even "I'm sorry that happened" [18].

Beyond the ethical concern of deception in simulating human interpersonal engagement, there is also a broader question of whether intentionally imparting any human-like warmth in a therapeutic setting may be harmful. Integrating such subjective qualities may cause patients who are seeking therapeutic care to ascribe intentionality and care that simply does not exist for LLMs, producing unrealistic expectations of understanding and acceptance. In practice, such prescriptions could exacerbate risk in cases of over-validation and abandonment for likely already vulnerable users, as found in this study. Because of these concerns, current work in the field suggests such systems can "never" engage in a genuinely therapeutic conversation and would be best utilized as a mediator with limitations[54].

On the other hand, it is still valuable to discuss the idea of whether there is a level of deception (in the most inclusive sense) that is ethical while designing AI in mental health settings. Barring the more extreme speculations of AI personas, can end user-facing LLM agents generate self-referential comments or basic pleasantries without ultimately causing harm? Or can the psycho-educational content provided by AI itself be valuable despite what is outlined by the current CTRS scale?

This overarching concern of deceptive empathy is further preceded by the ethical challenges of whether LLM-based mental health agents can even functionally display basic therapeutic competencies

in assessing and handling cases where users' needs may be outside of their scope of care. While not explicitly part of the CTRS scale, HELPERT's failures in refusing to continue support in instances of substance use, the disclosure of specific mental health disorders, or self-harm, as well as over-validating other harmful behaviors, are in direct conflict with broader mental healthcare standards. In particular, organizations like the American Psychological Association (APA) have set ethical and conduct standards of which, if a provider feels that a patient's issues exceed their professional competency, "an appropriate termination process that addresses the client's ongoing treatment needs through pre-termination counseling and making any needed referrals must occur" [4, 10].

The ability to assess what a care provider is equipped to do and how to handle situations where they are not may be as easily solved for AI agents as simply linking resources upon a refusal of service that is based on a blanket keyword flag. This competency is distinctly emphasized when there is an imminent risk to a user, such as that of suicidal ideation or domestic violence. Especially when considering that AI agents are already being characterized or sought out as forms of therapy in the wake of inaccessible healthcare, the design of such AI agents must seriously consider the ethical implications of how and how not such tools may handle these common yet high-stake situations. Such ethical questions only scratch the surface of what it means to implement or direct LLM-human interactions in mental healthcare but present fundamental ethical challenges to evaluating the nature of LLM mental healthcare tools.

5.6 Limitations and Paths Forward

The LLM responses were generated in a non-interactive, one-sided manner, unlike real-time human peer counseling sessions. Although care was taken to recreate the session and maintain the essence and context of the original, HELPERT sessions were highly constrained by the transcript of the previously generated human-to-human session. It is likely that HELPERT's sessions would have diverged from the original ones, as users might have responded differently to HELPERT's responses, which sometimes varied from those of the original counselor.

This design choice, however, was an intentional trade-off to minimize risks to human subjects. Our goal was to evaluate the care provided as an alternative to traditional forms of care, whether by a peer counselor or an LLM-based chatbot, not to investigate how accurately a session could be reconstructed with LLMs. This method also served as one of the only feasible ways to evaluate the quality of care offered by these systems without exposing real participants to potential risk. In addition, since patient outcome measures were not accessible for the recreated sessions, our study focused solely on the quality of care provided, as assessed by clinical psychologists.

Future work should, therefore, explore avenues for integrating LLMs into therapeutic settings in a safe, controlled, and supervised environment. This will help better understand the role of different session dynamics, such as connection versus method adherence, on user therapeutic outcomes. Researchers who wish to conduct such a study involving human subjects can use the HELPERT and Psychologist Evaluation Dataset to understand and be aware of the potential risks.

6 CONCLUSION

From a decision-making system to a CBT-trained peer, large language models are undergoing a shift in their role, enabling wider access to care—a fundamental necessity in contemporary mental health support. However, re-purposing LLMs as therapeutic counselors when they were not originally intended for this task is counter-intuitive and has ignited a debate between some recipients of AI-assisted support and field experts. In this paper, we reproduced a publicly available CBT dialogue dataset using a CBT-based prompt running on GPT-4. The two sets of session dialogues were evaluated by clinical psychologists through the Cognitive Therapy Rating Scale (CTRS) and psychologists' feedback data. Despite the same training, we found that peer counselors focused on connection and validation through self-disclosure, whereas HELPERT focused on adherence to the method. Our findings call for a hybrid model of care that offers a balance between the scalability of AI and the nuanced interpersonal effectiveness of humans. In addition to our analysis, we will be releasing two companion datasets to the public, providing CBT-based benchmarks for exploring the roles of peer counselors and LLMs in mental health. Our contributions in this paper encompass being the first to analyze therapeutic counseling conversations at the session level amidst the rise of more sophisticated LLMs, especially at a time when a significant number of individuals have turned to these chatbots for support.

REFERENCES

[1] Alaa A Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics* 132 (2019), 103978.

[2] Mehmet Emin Aktan, Zeynep Turhan, and Ilknur Dolu. 2022. Attitudes and perspectives towards the preferences for artificial intelligence in psychotherapy. *Computers in Human Behavior* 133 (2022), 107273.

[3] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 3906–3918.

[4] American Psychological Association. 2016. Ethical principles of psychologists and code of conduct. https://www.apa.org/ethics/code/

[5] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* 183, 6 (2023), 589–596.

[6] Aaron T Beck. 1979. *Cognitive Therapy of Depression.* Guilford Press, New York.

[7] Bethany Biron. 2023. *Online Mental Health Company Uses ChatGPT to Help Respond to Users in Experiment, Raising Ethical Concerns.* Business Insider Inc. https://www.businessinsider.com/company-using-chatgpt-mental-health-support-ethical-issues-2023-1 Accessed: 2023-04-14.

[8] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan. 2017. Signal processing and machine learning for mental health research and clinical applications [perspectives]. *IEEE Signal Processing Magazine* 34, 5 (2017), 196–195.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[10] Linda Campbell, Melba Vasquez, Stephen Behnke, and Robert Kinscherff. 2010. *APA Ethics Code Commentary and Case Illustrations.* American Psychological Association, Washington, D.C.

[11] Adam M Chekroud, Julia Bondar, Jaime Delgadillo, Gavin Doherty, Akash Wasil, Marjolein Fokkema, Zachary Cohen, Danielle Belgrave, Robert DeRubeis, Raquel Iniesta, et al. 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry* 20, 2 (2021), 154–170.

[12] Mick Cooper and John McLeod. 2007. A pluralistic framework for counselling and psychotherapy: Implications for research. *Counselling and Psychotherapy Research* 7, 3 (2007), 135–143.

[13] Flávio Luis De Mello and Sebastião Alves de Souza. 2019. Psychotherapy and artificial intelligence: A proposal for alignment. *Frontiers in Psychology* 10, 263 (2019), 1–9.

[14] David Daniel Ebert, Mathias Harrer, Jennifer Apolinário-Hagen, and Harald Baumeister. 2019. Digital Interventions for Mental Disorders: Key Features, Efficacy, and Potential for Artificial Intelligence Applications. *Advances in Experimental Medicine and Biology* 1192 (2019), 583–627.

[15] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[16] Michael P Ewbank, Ronan Cummins, Valentin Tablan, Sarah Bateup, Ana Catarino, Alan J Martin, and Andrew D Blackwell. 2020. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA psychiatry* 77, 1 (2020), 35–43.

[17] Christopher G Fairburn and Vikram Patel. 2017. The impact of digital technology on psychological treatments and their dissemination. *Behaviour research and therapy* 88 (2017), 19–25.

[18] Andrea Ferrario, Jana Sedlakova, and Manuel Trachsel. 2024. The Role of Humanization and Robustness of Large Language Models in Conversational Artificial Intelligence for Individuals With Depression: A Critical Analysis. *JMIR Mental Health* 11 (2024), e56569.

[19] Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *Journal of Medical Internet Research* 21, 5 (2019), e13216.

[20] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e7785.

[21] Hannah Gaffney, Warren Mansell, and Sara Tai. 2019. Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR mental health* 6, 10 (2019), e14166.

[22] Ingrid D Goldstrom, Jean Campbell, Joseph A Rogers, David B Lambert, Beatrice Blacklow, Marilyn J Henderson, and Ronald W Manderscheid. 2006. National estimates for mental health mutual support groups, self-help organizations, and consumer-operated services. *Administration and Policy in Mental Health and Mental Health Services Research* 33 (2006), 92–103.

[23] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research* 62 (2018), 729–754.

[24] Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. 2019. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports* 21 (2019), 1–18.

[25] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2019. Toward fairness in AI for people with disabilities: a research roadmap. *ACM SIGACCESS Accessibility and Computing* 125 (2019), 8 pages.

[26] Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access* 11 (2023), 80218–80245.

[27] Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the Helper: Supporting Peer Counselors via AI-Empowered Practice and Feedback. (2023). arXiv preprint arXiv:2305.08982.

[28] Zainab Iftikhar, Yumeng Ma, and Jeff Huang. 2023. "Together but not together": Evaluating Typing Indicators for Interaction-Rich Communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12.

[29] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[30] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[31] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.

[32] Vittorio Lingiardi, Laura Muzi, Annalisa Tanzilli, and Nicola Carone. 2018. Do therapists' subjective variables impact on psychodynamic psychotherapy outcomes? A systematic literature review. *Clinical psychology & psychotherapy* 25, 1 (2018), 85–101.

[33] Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. 2017. A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet interventions* 10 (2017), 39–46.

[34] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the Experience of LGBTQ+ People Using Large Language Model Based Chatbots for Mental Health Support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[35] Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 735–745.

[36] Shery Mead and Cheryl MacNeil. 2006. Peer support: What makes it unique. *International Journal of Psychosocial Rehabilitation* 10, 2 (2006), 29–37.

[37] Stirling Moorey and Anna Lavender. 2017. *The Therapeutic Relationship in Cognitive Behavioural Therapy*. Sage Publications, Los Angeles, CA.

[38] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research* 20, 6 (2018), e10148.

[39] Robert R Morris, Stephen M Schueller, and Rosalind W Picard. 2015. Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *Journal of medical Internet research* 17, 3 (2015), e72.

[40] Hongbin Na. 2024. CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2930–2940.

[41] John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2 (2016), 113–122.

[42] Jacqueline Nesi. 2023. Can chatgpt do therapy? https://technosapiens.substack.com/p/can-chatgpt-do-therapy

[43] Jesse Noyes. 2023. Perceptions of AI in healthcare: What professionals and the public think. The Intake. Available at: https://www.tebra.com/theintake/medical-deep-dives/tips-and-trends/research-perceptions-of-ai-in-healthcare

[44] World Health Organization. 2019. *Psychologists Working in Mental Health Sector (per 100,000)*. Technical Report. https://www.who.int/data/gho/data/indicators/indicator-details/GHO/psychologists-working-in-mental-health-sector-(per-100-000)

[45] Aisling Ann O'kane, Sun Young Park, Helena Mentis, Ann Blandford, and Yunan Chen. 2016. Turning to peers: integrating understanding of the self, the condition, and others' experiences in making sense of complex chronic conditions. *Computer Supported Cooperative Work (CSCW)* 25 (2016), 477–501.

[46] 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.

[47] Judith J Prochaska, Erin A Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. 2021. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. *Journal of medical Internet research* 23, 3 (2021), e24850.

[48] rahul_9735. 2023. Considering how many people use ChatGPT as a therapy tool, here use this prompt to turn your GPT into a personal therapist! Reddit. Available at: https://www.reddit.com/r/ChatGPT/comments/14b2u1p/considering_how_many_people_use_chatgpt_as_a/. Accessed: 2023-03-25.

[49] Paolo Raile. 2024. The usefulness of ChatGPT for psychotherapists and patients. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–8.

[50] Jérémie Richard, Reid Rebinsky, Rahul Suresh, Serena Kubic, Adam Carter, Jasmyn EA Cunningham, Amy Ker, Kayla Williams, and Mark Sorin. 2022. Scoping review to evaluate the effects of peer support on the mental health of young adults. *BMJ open* 12, 8 (2022), e061336.

[51] Sabirat Rubya and Svetlana Yarosh. 2017. Video-mediated peer support in an online community for recovery from substance use disorders. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1454–1469.

[52] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, et al. 2019. HCI and Affective Health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–17.

[53] Matthias Schmidmaier, Jonathan Rupp, Darina Cvetanova, and Sven Mayer. 2024. Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[54] Jana Sedlakova and Manuel Trachsel. 2023. Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? *The American Journal of Bioethics* 23, 5 (2023), 4–13.

[55] Brian F Shaw, Irene Elkin, Jane Yamaguchi, Marion Olmsted, T Michael Vallis, Keith S Dobson, Alice Lowery, Stuart M Sotsky, John T Watkins, and Stanley D Imber. 1999. Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of consulting and clinical psychology* 67, 6 (1999), 837.

[56] Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420.

[57] Dorien Smit, Clara Miguel, Janna N Vrijsen, Bart Groeneweg, Jan Spijker, and Pim Cuijpers. 2023. The effectiveness of peer support for individuals with mental illness: systematic review and meta-analysis. *Psychological Medicine* 53, 11 (2023), 5332–5341.

[58] Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research* 3, 1 (2024), 12.

[59] Substance Abuse and Mental Health Services Administration. 2021. Key substance use and mental health indicators in the United States: results from the 2020 National Survey on Drug Use and Health. HHS Publication No. PEP21-07-01-003. https://www.samhsa.gov/data/

[60] Sara Syed, Zainab Iftikhar, Amy Wei Xiao, and Jeff Huang. 2024. Machine and Human Understanding of Empathy in Online Peer Support: A Cognitive Behavioral Approach. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13.

[61] Cheeseburger Therapy Team. 2018. Cheeseburger Therapy. https://cheeseburgertherapy.org/ Accessed: 2023-22-12.

[62] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53.

[63] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry* 64, 7 (2019), 456–464.

[64] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet* 15, 3 (2023), 110.

[65] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–32.

[66] Svetlana Yarosh. 2013. Shifting dynamics or breaking sacred traditions? The role of technology in twelve-step fellowships. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3413–3422.

[67] JE Young and Aaron T Beck. 1980. Cognitive therapy scale. (1980). Unpublished manuscript, University of Pennsylvania.

[68] Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The Role of AI in Peer Support for Young People: A Study of Preferences for Human-and AI-Generated Responses. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[69] Jim Young and Christopher L Williams. 1987. An Evaluation of GROW, a mutual-help community mental health organization. *Australian and New Zealand Journal of Public Health* 11, 1 (1987), 38–42.

[70] Renwen Zhang, Jordan Eschler, and Madhu Reddy. 2018. Online support groups for depression in China: Culturally shaped interactions and motivations. *Computer Supported Cooperative Work (CSCW)* 27, 3 (2018), 327–354.

A CTRS

Table 4: Cognitive Therapy Rating Scale (CTRS). Items on the CTRS are divided into two sub-groups: General Therapeutic Skills and Conceptualization, Strategy, and Technique

General Therapeutic Skills	
Agenda	Ability to collaboratively define an appropriate plan for the session, factoring time constraints and prioritization
Feedback	Extent to which the counselor applies patient feedback to ensure understanding and satisfaction within the session
Understanding	Capacity to comprehend both explicit and nonverbal communication, demonstrating listening skills and empathy
Interpersonal Effectiveness	Expression of positive interpersonal traits such as warmth, sincerity, confidence and relationship quality
Collaboration	Efforts to foster a collaborative work relationship, actively involvement in the process
Pacing & Efficient Use of Time	Ability to effectively manage and structure the session time to ensure progress
Conceptualization, Strategy, and Technique	
Guided Discovery	Skill in promoting self-discovery through measured questioning instead of persuasive tactics
Strategy for Change	Capacity to develop and follow a consistent strategy for change, using appropriate CBT techniques
Focusing on Key Cognitions or Behaviors	Ability to target essential thoughts or behaviors relevant to the seeker's problems
Application of CBT Techniques	Level of skill and resourcefulness exhibited in applying CBT
Homework	Ability to assign, explain, review and use tailored homework as an active element in the CBT process

B SESSION FEEDBACK SURVEY

After reviewing the session, please make a comment on the session. Consider responding to one or more of the following prompts that address the most interesting aspects of the session.

(1) What did you notice in the session that seemed most different than what a human might ask? (e.g., tone, conversation style, questions, reactions)
(2) What was the most impactful or compelling thing that the peer supporter did to guide the session?
(3) What could the peer supporter have done better? Recommendations for improvement?
(4) What are the most noticeable differences between this session and CBT sessions that happen in your practice?

C SESSION COMPARISON SURVEY

Upon reviewing both sessions conducted by a human peer supporter and an AI peer supporter, please answer the following question, which is mandatory:

(1) What unique observations did each peer supporter make in their respective sessions that the other peer supported did not? For example, Peer Supporter 1 may have noticed 'X' while Peer Supporter 2 observed 'Y.'
(2) Which peer supporter demonstrated a better understanding of the support seeker's trouble and application of the method?
   - Peer Supporter 1
   - Peer Supporter 2
   - Both
   - Neither

D SESSION SCHEMA

Each session contains a tet-based dialogue between peer support provider (human or AI) and the following schema:

Table 5: Detailed Schema of the HELPERT Dataset, outlining message attributes, and session notes

Field	Type	Example
		Message Attributes
SessionID	text	e4IDtMEP
MessageID	text	aGxsTofcT
Message	text	"i feel worried and stressed for the future and having to make that decision"
FromThinker	binary	TRUE
Timestamp	timestamp/date	Tue Apr 05 2022 09:46:08 GMT-0700 (Pacific Daylight Time)
Offset	text	GMT0300
		Session Notes
Counselor	text	Human
Event	text	Unexpected panic attack
Thoughts	text	"What I'm doing is not enough. I might lose confidence in my ability to be there for my daughter"
Feelings	text	Stressed, anxious, unmotivated
Behaviors	text	Avoid caregiving voluntarily
Label	text	Fortune Telling
New Thought	text	"I will care for myself out of love, to enjoy my time with my family and friends and be able to do the things that fulfill me and them. I accept that fear may come naturally, but I will transform it into acceptance and rational action"

Table 6: Detailed Schema of the Psychologist Evaluation Dataset, outlining session evaluation criteria, including CTRS scores and psychologist feedback.

Field	Type	Example
		Session Evaluation
PsychologistID	binary	1
SessionID	text	e4IDtMEP
Counselor	text	Human
Total_CTRS	int, between [0,66]	54
General Therapeutic Skills	int, between [0,36]	30
Conceptualization, Strategy, and Technique Skills	int, between [0,30]	24
Agenda	int, between [0,6]	4
Feedback	int, between [0,6]	5
Understanding	int, between [0,6]	3
Interpersonal Effectiveness	int, between [0,6]	5
Collaboration	int, between [0,6]	4
Pacing & Efficient Use of Time	int, between [0,6]	5
Guided Discovery	int, between [0,6]	2
Strategy for Change	int, between [0,6]	1
Focusing on Key Cognitions or Behavior	int	3
Application of CBT Techniques	int, between [0,6]	4
Homework	int, between [0,6]	2
Session Feedback	text	"...The peer counselor was really beautiful in her ability to relate with the client while also expressing empathy and cultural understanding. The conversation of culture and how cultural factors are impacting the client's life was impressive, especially because it was organic and unforced...."