frontiers in Robotics and AI

TYPE Original Research
PUBLISHED 09 September 2026
DOI 10.3389/frobt.2026.1802622

Check for updates

OPEN ACCESS

EDITED BY
Changiz Mohiyeddini,
Oakland University William Beaumont School of Medicine, United States

REVIEWED BY
Joe Hasei,
Okayama University, Japan
Kathleen Belhassein,
UPR3346 Institut P′ Recherche et Ingénierie en Matériaux, Mécanique et Energétique (Pprime), France

*CORRESPONDENCE
Mina Kian,
kian@usc.edu

†These authors have contributed equally to this work

RECEIVED 03 February 2026
REVISED 16 June 2026
ACCEPTED 22 June 2026
PUBLISHED 09 September 2026

CITATION
Kian M, Zong M, Fischer K, Singh A, Velentza A-M, O'Connell A, Ignatova L, Sang P, Upadhyay S, Gupta A, Faruki MA, Browning W, Arnold SMR, Chen M, Krishnamachari B and Matarić MJ (2026) LLM-powered socially assistive robot-delivered cognitive behavioral therapy exercises: an exploratory study with university students.
*Front. Robot. AI* 13:1802622.
doi: 10.3389/frobt.2026.1802622

COPYRIGHT
© 2026 Kian, Zong, Fischer, Singh, Velentza, O'Connell, Ignatova, Sang, Upadhyay, Gupta, Faruki, Browning, Arnold, Chen, Krishnamachari and Matarić. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# LLM-powered socially assistive robot-delivered cognitive behavioral therapy exercises: an exploratory study with university students

Mina Kian[1]*, Mingyu Zong[2†], Katrin Fischer[1†], Abhyuday Singh[1†], Anna-Maria Velentza[3†], Amy O'Connell[1], Lydia Ignatova[1], Pau Sang[1], Shriya Upadhyay[1], Anika Gupta[1], Misha A. Faruki[1], Wallace Browning[1], Sébastien M. R. Arnold[1], Meng Chen[4], Bhaskar Krishnamachari[5] and Maja J. Matarić[1]

[1]Interaction Lab, Computer Science Department, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, United States, [2]Autonomous Networks Research Group, Computer Science Department, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, United States, [3]Computer Science Department, Brest National School of Engineering, Plouzané, France, [4]Department of Psychology, University of Southern California, Los Angeles, CA, United States, [5]Autonomous Networks Research Group, Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, United States

Mental health is a significant healthcare challenge, and cognitive behavioral therapy (CBT) is a widely used therapeutic method for treating anxiety and depression. However, traditional CBT often requires access to trained clinicians and can be cost-prohibitive or logistically difficult for many individuals. To address these barriers, we developed a low-cost socially assistive robot (SAR) that uses a large language model (LLM) to guide the user through interactive at-home CBT exercises. In this exploratory study, 38 university students completed CBT exercises across a 15-day period using one of three modalities: with a robot (using an LLM for dialogue), a chatbot (using the same LLM for dialogue), or traditional CBT worksheets. We measured weekly therapeutic outcomes, changes in pre-/post-session anxiety measures, and adherence to completing CBT exercises. Our findings indicate that self-reported general psychological distress significantly decreased over the study period in the robot and worksheet conditions but not in the chatbot condition. Additionally, the SAR enabled significant single-session improvements on more days than the other two conditions combined. Mixed-effects modeling further An analysis with a mixed-effects model also suggested that the robot and chatbot conditions better reduced post-session anxiety for those with elevated levels of anxiety. Our findings suggest that SAR-guided, LLM-powered CBT may be an effective method for supporting therapeutic progress and decreasing user anxiety immediately after completing the CBT exercise. The findings underscore the

potential for combining AI-driven personalization with socially assistive robotics to create accessible, scalable, and engaging mental health interventions.

KEYWORDS

**human robot interaction (HRI), socially assistive robotics (SAR), cognitive behavioral therapy (CBT), mental health, large-language model (LLM), longitudinal study, university students**

# 1 Introduction

University students face increased periods of vulnerability in mental health and consequently experience anxiety, depression, and suicidality (Liu et al., 2019). At the same time, mental health services are often inaccessible to students on university campuses due to high demand. A survey conducted across 108 U.S. institutions indicated an urgent need for mental health support, with stress and suicidality especially pronounced among racial, ethnic, sexual, and gender minorities (Liu et al., 2019).

Cognitive Behavioral Therapy (CBT) is an effective approach for addressing mental health challenges (Beck, 2020) and is a widely practiced, clinically validated form of therapy (Hofmann et al., 2012). The premise of CBT is that "psychological problems are based, in part, on faulty or unhelpful ways of thinking […] and learned patterns of unhelpful behavior" and that people can learn better ways to cope, enabling them to be more effective in their lives (American Psychological Association, 2017). CBT treatment guides people to recognize and address their distorted thinking and alter their thought patterns (Kuru et al., 2018) as well as learn to adjust their behavior to influence their emotional state (Boswell et al., 2017).

For CBT to be effective, those receiving it must practice applying CBT strategies in their daily lives. CBT therapists often assign *homework* (Beck, 1993; Dobson, 2021), exercises to be completed at home that create opportunities for regular practice (Prasko et al., 2022) and encourage independence in addressing unhelpful patterns of thought and behavior (Kazantzis, 2021). Such homework typically involves completing written worksheets that guide the user through a set of steps that apply a particular CBT strategy to a recent experience (Beck, 1993). Worksheets help users work through thoughts via clear written steps, but they lack the interactivity of conversations with a therapist and instead provide static instructions; their effectiveness therefore hinges on the user's willingness to share their thoughts. In contrast, in interactive discussions, a therapist's tailored follow-up questions can encourage greater reflection and new perspectives. Adherence to CBT homework is generally low, negatively affecting therapeutic outcomes (LeBeau et al., 2013). Studies have shown that more frequent use of CBT techniques and practices in patients outside of therapy is associated with fewer symptoms of mental illnesses Kazantzis et al. (2010), including depression (Hundt et al., 2013). In particular, Kazantzis et al. (2010) established that CBT homework is a core mechanism of therapeutic change, with meta-analytic evidence demonstrating a reliable relationship between homework engagement and clinical outcomes. Thus, investigating methods to support CBT homework completion is of interest to the field and motivates our work.

Innovations in natural language processing, in particular large language models (LLMs), have greatly expanded the capabilities of dialogue-based therapeutic technologies. Applications of LLMs have grown rapidly, especially in conversational agents (Zaib et al., 2020), and will grow further with improved task-specific controls (Wei et al., 2022). For example, (Ziems et al., 2022), found that LLMs can successfully complete CBT-based positive reframing tasks. However, using LLMs for mental health has raised some concerns in the scientific and clinical communities because they may inadvertently generate potentially harmful or inappropriate messages (Weng, 2023); studies that employ LLMs must follow ethical and clinical guidelines to safeguard users from harm (Schueller and Morris, 2023).

While LLMs have greatly advanced human-machine dialogue, motivating and sustaining engagement in therapeutic exercises is a well-known challenge. Socially assistive robots (SARs), physically embodied agents designed to provide support through social interaction (Feil-Seifer and Mataric, 2005), have been shown to be effective in supporting motivation and practice for a wide range of health and therapeutic goals, including rehabilitation after stroke (Swift-Spong et al., 2015; Matarić et al., 2007), cerebral palsy (Feingold Polak and Tzedek, 2020; Malik et al., 2016)), cognitive and social skill training (Scassellati et al., 2012; Pennisi et al., 2016), encouraging physical activity (Fasola and Mataric, 2012), mental health outcomes (Kabacińska et al., 2021; Scoglio et al., 2019; Guemghar et al., 2022), and other areas of support and behavior change (Scassellati et al., 2012; Ferrante et al., 2021; Robinson et al., 2020). In a systematic review, Deng et al. (2019) concluded that physically embodied agents, i.e., robots, demonstrate performance improvements over virtual agents in a broad range of settings, particularly for relationship-oriented tasks where social presence and engagement are critical for maintaining rapport (Deng et al., 2019; Segura et al., 2012). A longitudinal HRI study found that, even as the novelty of interacting with a robot wore off, users reported increased interaction quality and perceived communication competence (Laban et al., 2022). Studies indicate that children and adults alike are willing to share confidential or embarrassing information with a robot, often to the same extent they would with each other (Bethel et al., 2011; Barendregt et al., 2014; Pitardi et al., 2021). This greater social presence allows robots to offer better encouragement and guidance to users in challenging tasks (Matarić and Scassellati, 2016).

Combining the strengths of SAR and LLMs is therefore a natural direction for exploration. LLMs have enabled SARs to engage in more natural, context-aware dialogue with users across domains (Brown et al., 2020; Zhao et al., 2023). Several recent SAR studies incorporate LLMs to support more personalized interactions (Shi et al., 2024a; Abdel Hafez, 2024; Lima et al., 2025). Our study explores how an LLM-enabled SAR may support users in

completing interactive CBT homework exercises in their daily lives. This work seeks to leverage the benefits of physical embodiment by developing an LLM-powered in-home SAR to encourage individuals to complete daily CBT homework exercises.

Prior to bringing SARs and LLMs together, studies have explored each technology separately in the CBT context. For example, Dino et al. (2019) showed that elderly users receiving CBT from a SAR enjoyed interacting with the robot, had elevated moods, and found the CBT information valuable, but this experiment was limited by its use of a fixed pre-programmed dialogue tree. Stronger conversational agents, including over 40 chatbots, have been developed for therapeutic purposes such as administering therapy, training therapists, screening for mental health symptoms, and encouraging self-management (Abd-Alrazaq et al., 2019). Chatbots have been shown to reduce some symptoms of depression and anxiety (Fulmer et al., 2018; Ahmed et al., 2023). Recent work has developed chatbots with greater personalization and interactivity, aiming to enhance accessibility, adherence, and therapeutic outcomes (Fitzpatrick et al., 2017; Burton et al., 2016; Denecke et al., 2020). For example, Kumar et al. (2022) developed a constrained language model-powered chatbot that guided participants through 5-min therapy sessions, modulating the chatbot's identity, behavior, and personality to test participant preferences. Similarly, Jang et al. (2021) showed that a chatbot delivering CBT and psychoeducation for adults with ADHD could be effective in relieving their symptoms. Technology-based interventions have been shown to be beneficial because of their fast response times compared to human therapists, and their ability to bridge the vocabulary gap between therapists and patients (Kumar et al., 2016).

In this work, we explored the combined promise of LLMs and SARs to assess their effectiveness in CBT homework exercise delivery. We conducted a 15-day exploratory user study with a sample of university students (N = 38) to understand how they respond to LLM-powered and SAR-guided CBT homework exercises. In this mixed design between- and within-subjects study, participants were assigned to one of three conditions and accordingly completed at-home CBT homework exercises using a SAR, chatbot, or traditional worksheets, respectively. We collected qualitative feedback from participants about their experience and measured the daily and weekly therapeutic outcomes in each condition. In the last 7 days of the study, completing the homework exercises was optional, allowing us to observe patterns of user adherence to each CBT delivery method. We found that, over the study period, psychological distress significantly decreased in the robot and worksheet conditions. We also found that pre-to post-session anxiety significantly decreased on more days in the robot condition than in the chatbot and worksheet conditions. Our findings provide novel insights about the effectiveness of LLM-powered SAR-guided CBT and its impact on therapeutic outcomes.

This work aimed to address the following research questions with university students as the intended target (not convenience) user population:

RQ1 How does the CBT in-home exercise delivery method affect immediate and short-term therapeutic outcomes?

RQ2 How does the CBT in-home exercise delivery method affect adherence to regular CBT practice?

RQ3 How does the CBT in-home exercise delivery method affect students' perception and experience with the system?

1. RQ3a How do students respond to the platform structure as a whole?
2. RQ3b How do students respond to LLM-driven CBT homework exercises?
3. RQ3c How do students respond to robot-delivered CBT homework exercises?

# 2 Methodology

This study examined the effectiveness of CBT homework delivered through three modalities: (1) traditional worksheets completed without an agent; (2) a text-based chatbot powered by an LLM; and (3) a socially assistive robot (SAR) powered by the same LLM. We collected quantitative measures of therapeutic outcomes, anxiety, and adherence, as well as qualitative feedback from post-study interviews, to provide a comprehensive view of participants' experiences. The following section describes the system design, study procedures, participant recruitment, data collection instruments, and analysis methods.

## 2.1 System design

Standard CBT exercises were adapted for delivery through three distinct modalities: traditional worksheets, a text-based chatbot, and a SAR. We used GPT-3.5, OpenAI's most advanced model at the time, to deliver CBT exercises in the chatbot and SAR modalities. Earlier pilot testing was conducted with GPT-3, as GPT-3.5 had not yet been released.

### 2.1.1 CBT exercise development

CBT homework worksheets offer a structured approach to CBT exercises, with a primary goal of improving patients' skill acquisition. We examined several online databases of CBT homework worksheets that were designed by CBT therapists using CBT principles and intended for use by CBT therapists[1]. From these worksheets, we selected a subset of exercises that readily translated into interactive exercises guided by an agent (robot or chatbot, depending on the study condition). CBT worksheets for the same exercise vary in structure and phrasing, and when delivered by a clinician in a session, they are often dynamically adapted to meet the client's needs. Our LLM-based approach aimed to achieve the goals of active CBT practice and homework completion, since more frequent CBT practice has been linked with fewer symptoms of mental illness (Kazantzis et al., 2010).

#### 2.1.1.1 Prompt generation for LLM delivery

To adapt the exercises for LLM delivery, we developed a series of prompts that allowed an LLM to guide users through each exercise. These prompts were refined and informally evaluated

1 https://www.therapistaid.com/; https://thinkcbt.com/think-cbt-worksheets; https://positivepsychology.com/cbt-cognitive-behavioral-therapy-techniques-worksheets/#6-evidence-based-cbt-worksheets

validated through in-lab pilot testing. In addition to prompt engineering, we compared the LLM's performance with zero-shot and few-shot learning examples. In the zero-shot learning setup, the LLM-powered agent received a description of the exercise and its role in the conversation to support the exercise, with no example conversations provided. In the few-shot learning setting, we additionally provided the LLM with 2-3 example conversations demonstrating how the CBT exercise should be conducted. After testing, we found that zero-shot learning with an appropriate prompt was enough for the model to properly and effectively deliver the exercises. With limited input tokens, it was challenging to select representative few-shot examples that enabled the agent to respond appropriately when the conversation topic differed from the example dialogue. We found that the model frequently repeated the language from the examples, regardless of the present context. We therefore adopted a zero-shot learning approach for the study. Additionally, omitting examples allowed the model to make use of the full session conversation history without reaching the token limit, enabling it to better adhere to the context. Research group members tested a variety of prompts corresponding to the list of candidate CBT exercises and identified the three best-performing exercises based on how closely the resulting interactions resembled a therapist-led session in terms of structure, flow, and clinical appropriateness.

#### 2.1.1.2 Prompt validation

Through an in-lab pilot study with 15 participants, we established that the LLM was able to effectively and safely manage two of the tested CBT exercises: Cognitive Restructuring[2] (Clark and Egan, 2015) and Coping Strategies[3] (Kuyken et al., 2008). Cognitive Restructuring supports identifying and altering cognitive distortions and negative/maladaptive thinking patterns (Clark and Egan, 2015). Coping Strategies is based on the CBT case conceptualization framework, helping individuals better understand the problems they face and identify healthy and productive ways to manage them (Kuyken et al., 2008).

The final prompts used in both the chatbot and SAR conditions are provided in Table 1 (model parameters are listed in Supplementary Table S1 in the Supplementary Materials).

#### 2.1.1.3 Moderator design

Due to the sensitive nature of therapy, we used GPT-3.5 to design a moderator to check the LLM-generated responses and flag and filter out problematic statements. Although the chatbot did not produce any problematic responses during in-lab pilot testing, we used the moderator as an additional safety measure to minimize potential harm to study participants. The GPT-3.5 moderator was designed using prompt engineering, as follows. We created over 200 examples of inappropriate therapeutic responses in various conversational contexts. We then tested various prompts (with both zero-shot and few-shot examples) to evaluate the resulting moderator's ability to recognize inappropriate responses.

TABLE 1 LLM Prompts. Model prompts used for the CBT exercises and the moderator.

| | |
|---|---|
| Cognitive restructuring | Cognitive restructuring is a strategy to help the patient identify cognitive distortion and find evidence to challenge the distortion. You are a therapist who uses cognitive restructuring to help your patient in this session |
| Coping strategies | Coping strategy is used to help patients identify problems they encountered and the triggers. When a problem is defined, a therapist will help the patient figure out ways to cope with it. You are a therapist who uses coping strategies to help your patient in this session |
| Moderator | "…Given the above context, is the following advice given by the therapist problematic? Reply with just a Yes or No." |

We selected the prompt with the best performance achieved through zero-shot learning. The selected moderator prompt is provided in Table 1 (for parameters, see Supplementary Table S1 in Supplementary Materials).

### 2.1.2 User interface design

For the user interface, we designed a web application (Figure 1) that was securely hosted on Amazon Web Services (AWS), an IRB-approved cloud service where participants consented to have their study data stored. Access to the web application was password-protected; each participant was assigned unique credentials for authentication. At the start of each CBT homework session, participants completed an anxiety scale survey. Participants were then prompted to click a "start" button to begin recording video and audio, giving their consent to be recorded as they completed the CBT homework exercise. They proceeded to complete the exercise through their assigned modality (robot, chatbot, or worksheet, based on the study condition). In the worksheet condition, participants typed responses to the original questions from the selected CBT worksheets (Cognitive Restructuring$_2$ or coping strategies$_3$). All questions were displayed on a single, non-interactive web page. In the robot and chatbot conditions, participants completed an interactive discussion with the LLM-powered agent by typing into a dialogue interface. The LLM-powered agent guided the participants through the exercise. In the robot condition, the robot verbalized the LLM response before it appeared on the screen. (The LLM-powered moderator described above reviewed all LLM-based responses by the robot or chatbot to determine their appropriateness before they appeared on screen.) After finishing the CBT exercise, participants completed a post-session anxiety scale survey. At the end of the session, the website displayed the university's Student Health helpline phone number, in case participants needed to access further mental health support. The participants in the robot condition were given additional instructions on how to power down the robot. The data from each session were securely stored on IRB-approved cloud services. The full session structure is shown in Figure 2.

2 Worksheet source: Cognitive Restructuring: Socratic Questions, https://www.therapistaid.com/therapy-worksheet/socratic-questioning/cbt/none

3 Worksheet source: Coping Styles Formulation, https://positive.b-cdn.net/wp-content/uploads/Coping-Styles-Formulation.pdf

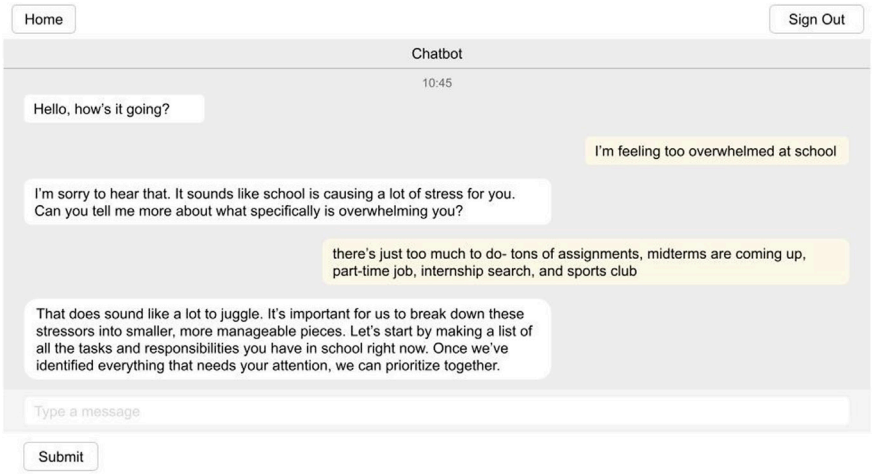



FIGURE 1
Web application interface. The user interface for the CBT homework interaction. The interface was identical in robot and chatbot conditions.

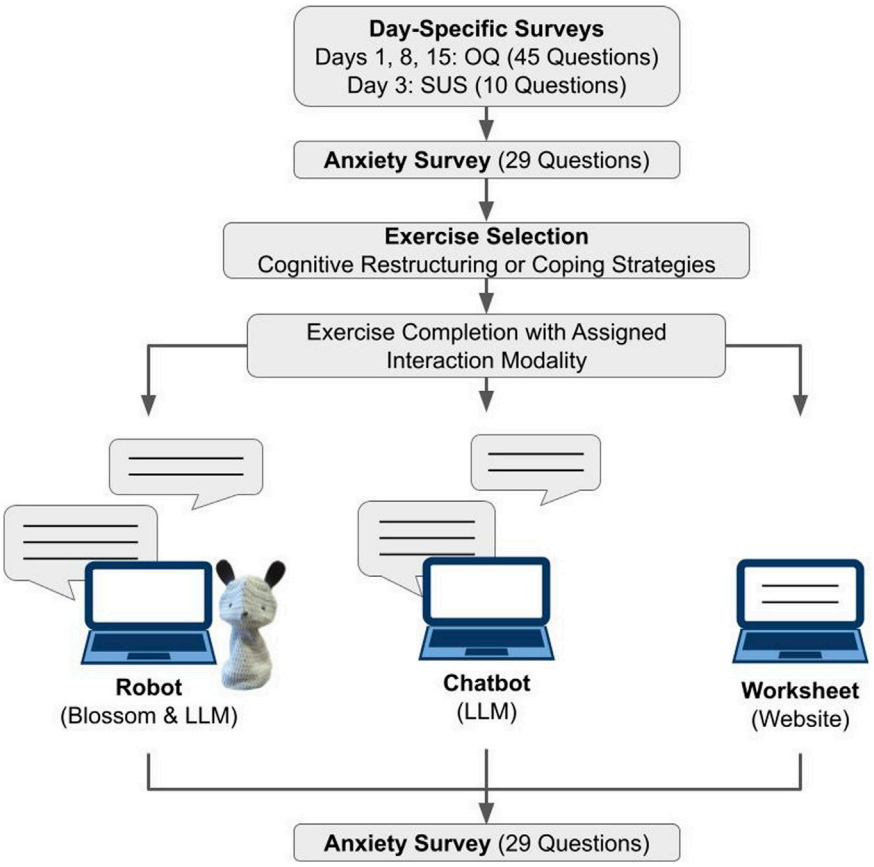



FIGURE 2
Interaction structure. Flow chart showing the sequence of surveys and CBT exercise for all three conditions.

### 2.1.3 Robot design

To promote replicability and accessibility, we used the low-cost open-source Blossom socially assistive robot (see Figure 3) (Suguitan and Hoffman, 2019; Shi et al., 2024b). Each robot cost approximately 350 US dollars to build and was connected to a Raspberry Pi that cost approximately 100 US dollars. The robot was covered with a neutral light gray crocheted exterior. The robot used the AWS Polly female text-to-speech voice to verbalize responses. We chose a female voice because Bhati (2014) observed that both

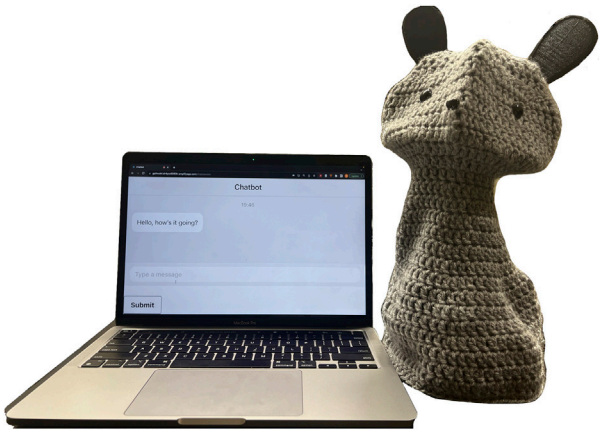



FIGURE 3
SAR Design. Blossom robot and laptop interface used in the robot condition in the study.

female and male clients reported higher levels of therapeutic alliance with female therapists than with male therapists.

The robot had two states: observing and speaking. In the observing state, the robot's head continuously raised and lowered to mimic breathing, indicating that the robot was awake and attentive to the user. We used the appearance of slow breathing to subtly encourage deep breathing relaxation by the user as it is often employed in therapeutic settings due to its efficacy in reducing stress (Andrayani et al., 2023). In the speaking state, the robot's head executed a gentle, oscillatory tilt from side to side, again indicating that it was awake and attentive (Fink, 2012).

In all three conditions, the participants provided their answers to the CBT exercise questions by typing them into the interface (or form, for the worksheet). We chose this for several reasons. We tested seven open-source speech-to-text (STT) packages and found that all had poor performance across various English accents representative of our university student population. In order to develop a system that was accessible to participants from a wide range of backgrounds, we chose to have participants enter their responses by typing instead of speaking. This choice also added a measure of privacy to the participants since their responses to CBT homework questions could not be overheard. Finally, it made all three experiment conditions more comparable by using the same response modality.

## 2.2 Study design

We designed a 15-day between- and within-subjects study between-subjects study with repeated measures to evaluate research outcomes. Participants were split into three study conditions for completing the CBT homework exercises: 1) with an LLM-powered SAR, 2) with an LLM-powered chatbot, and 3) using traditional worksheets. During the first half of the study (days 1–8), participants were required to complete daily CBT homework exercises through their assigned modality. The mandatory nature of the exercises intended to encourage the practice of completing the exercises daily as intended, and to collect data to compare outcomes among the three conditions. In the second half of the study (days 9–15), participants were not required to complete daily CBT homework exercises. The optional nature of the exercises intended to analyze how participants voluntarily interacted with their assigned delivery method. We used this approach because one objective of this study was to explore if the modality of delivering the CBT exercises impacts *intentional* non-adherence (Choi et al., 2015; Ingeson et al., 2019). To limit instances of unintentional non-adherence (Abdulrahman and Richards, 2021; Kelders et al., 2012), participants under all three conditions received daily emails throughout the study reminding them to complete the CBT homework exercise and questionnaires. Therapeutic outcomes were measured using the Anxiety Scale (Pilkonis et al., 2011), administered pre- and post-session as well as weekly measures administered on days 1, 8, and 15. The CBT exercise and surveys took approximately 30 min to complete. The usability of the interface was measured on Day 3.

## 2.3 Participants

We recruited 42 student participants from a major US university to participate in our study, a cohort size primarily determined by the availability of 15 robots and the logistical challenges of coordinating multi-week at-home deployments. This sample size is comparable to other in-home human-robot interaction SAR research studies, such as those by Clabaugh et al. (2019) (N = 17), Jeong et al. (2020) (N = 35), Matheus et al. (2022) (N = 43), Jeong et al. (2022) (N = 42), and O'Connell et al. (2024) (N = 11), and aims to provide useful exploratory insights.

### 2.3.1 Eligibility & ethical screening

This study was approved by the university's IRB. Participants were screened to be over 18 years of age, proficient in English, of normal or corrected-to-normal vision and hearing (due to the auditory and visual components of the study), and living near campus (to allow us to quickly address technical issues if any occurred).

After consulting the University's Institutional Review Board (IRB), we determined that, due to the sensitive nature of the CBT exercises, it was not appropriate to recruit severely depressed individuals in our sample of participants, and that recruiting severely depressed students to evaluate our previously untested system would pose an unnecessary risk to their safety and wellbeing. Participants were therefore screened using the PHQ-9 measure of depression (Kroenke et al., 2001); all responders who scored 15 or higher, indicating moderately severe to severe depression, were excluded from participating.

### 2.3.2 Recruitment and compensation

Students were recruited through emails to major academic departments as well as an e-bulletin post on a study promotion board. All students who filled out the study interest form were given information on available campus mental health resources. Participants were compensated with a US$150.00 Amazon gift card after completing the study. The amount was calculated based on the amount of time we anticipated participants would spend on the study and the local minimum wage. Participants were informed that

they would still be compensated in the full amount even if they chose not to complete any exercises in the optional week.

### 2.3.3 Cohort demographics

After the study began, two participants in the chatbot condition and two participants in the worksheet condition discontinued their participation. Complete data were therefore collected from 38 participants: 14 in the robot condition, 12 in the chatbot condition, and 12 in the worksheet condition.

The 38 participants ranged from ages 18 to 27 (M = 20.57895, SD = 2.1), and self-reported their gender as 16 female, 18 male, 1 non-binary, and 3 preferred not to answer. Ten of the participants were graduate students and 28 were undergraduates. Additional information about the participants is shown in Supplementary Figure S1 of the Supplementary Materials.

## 2.4 Procedure

Before the start of the data collection, each participant had an informed consent meeting with a researcher where they reviewed the study consent form and the requirements to maintain participation in the study. Participants were informed that for the first 8 days of the study, they would be required to complete at least one CBT homework exercise daily to continue participating. They were told that if they skipped the CBT homework exercises in the final 7 days of the study, it would not affect their participation in the study, and they would still receive full compensation. The researcher explained that the participants would receive a daily email reminding them to complete the CBT exercises and any required surveys every day of the study. The researcher set up the participants' login (and equipment for the robot condition) and guided them through a mock session to practice using the platform. At the conclusion of the study, all participants completed a semi-structured post-study interview with a researcher lasting approximately one hour. During the interview, they reflected on the study and gave feedback on the platform they used.

## 2.5 Data collection and instruments

To address our research questions, we followed a triangular approach (Östlund et al., 2011), collecting quantitative and qualitative data (Babbie, 2020) to build a holistic understanding of participants' experiences with, and responses to, their respective CBT homework delivery method.

The following survey instruments were used:

*PHQ-9.* The Patient Health Questionnaire-9 is a 9-item self-report diagnostic tool that measures an adult patient's severity of depression on a 4-point Likert scale (Kroenke et al., 2001). Study candidates were excluded from participating if they scored 15 or higher.

*OQ.* The Outcomes Questionnaire-45 (OQ) is a 45-item instrument where individuals rate their experiences with a list of symptoms shared across multiple adult mental disorders and syndromes, on a 5-point Likert scale (Lambert et al., 2004). The Full OQ score was administered on days 1, 8, and 15 to measure participants' therapeutic outcomes.

*SUS.* The System Usability Scale (SUS) is a 10-item instrument that evaluates the ease of use of the system (Brooke, 1996). Each question is on a 5-point Likert scale. The full SUS was administered on Day 3, after participants had gained familiarity with the study platform.

*Anxiety Scale.* Pilkonis et al. (Pilkonis et al., 2011) developed and calibrated an item bank for depression, anxiety, and anger as part of the Patient-Reported Outcomes Measurement Information System (PROMIS). The 29-item Calibrated Anxiety Items were administered immediately before and after each CBT homework session.

*NARS.* The Negative Attitude towards Robots Scale is a 14-item instrument that measures users' attitudes and emotions towards robots (Nomura et al., 2006). The NARS was administered on Day 1 of the study.

*Semi-Structured Interview.* We conducted a semi-structured interview with each participant after they completed the study to gather qualitative feedback about their experience using their assigned CBT homework delivery method. The interviews were conducted using a guide developed by the research team. Open-ended questions covered the following topics (platform refers to the robot/chatbot/worksheet): first impressions/beginning of the study, benefits of the platform for CBT homework, challenges with the platform, factors affecting their use of the platform for CBT homework, and overall experience during the study. The interview incorporated an introduction, a summary of the study's purpose, and a debriefing for participants to reflect on all aspects of the study and leave additional feedback. All interviews were conducted in English and recorded, with consent.

## 2.6 Data analysis

### 2.6.1 Statistical analysis of questionnaires

Questionnaire data were analyzed using statistical tests in Python 3.11.3 libraries and R 4.3.1 packages, as described in Supplementary Materials. The dependent variables were the instrument scores (Anxiety Scale, Outcome Questionnaire, System Usability) and adherence. The instrument data were internally consistent, with Cronbach $\alpha > 0.70$. The independent variables were the study condition, NARS scores, and time.

### 2.6.2 Therapeutic outcomes via weekly mental health assessments

The Outcomes Questionnaire (OQ) was sent to the participants to measure their psychotherapy progress throughout the duration of the study. Dependent t-tests were performed on the difference between the first and last day's OQ scores to check for significant improvement overall as well as within each condition.

### 2.6.3 Therapeutic outcomes via pre- and post-session anxiety scale scores

To check for single-session improvements under each condition, we compared the participants' pre- and post-session anxiety scores on each day of the required week (days 1–8). We performed a

dependent t-test (parametric) or Wilcoxon signed-rank test (non-parametric) between the pre-session and post-session anxiety scale scores on each day, depending on whether the data were normally distributed. To control the false discovery rate, we adjusted the alpha levels according to the Benjamini–Hochberg method (Benjamini and Hochberg, 1995), applying the correction separately within each condition across the eight daily tests.

To account for the nested structure of the data (repeated observations within participants), therapeutic outcomes were assessed using mixed-effects modeling. Linear mixed-effects models were estimated using the `lme4` package, with statistical significance evaluated using Satterthwaite's approximation for degrees of freedom as implemented in the `lmerTest` package. An unconditional model was first estimated to calculate the intraclass correlation coefficient (ICC). Fixed effects included day in the study, condition (using the worksheet condition as the reference group), and both between-participant (grand-mean centered pre-session anxiety: pre_gmc) and within-participant (cluster-mean centered pre-session anxiety: pre_cmc) components of pre-session anxiety. This centering approach allowed for the disaggregation of within- and between-participant effects. A bottom-up model building approach was used, where predictors were added sequentially. Nested models were compared using likelihood ratio tests to evaluate whether the inclusion of additional predictors significantly improved model fit. The final model included random intercepts and random slopes for pre_cmc across participants, allowing both baseline anxiety and within-participant associations to vary between individuals. Models were estimated using restricted maximum likelihood (REML). Missing data were handled using full information maximum likelihood (FIML), as implemented in `lme4`.

#### 2.6.4 Adherence to daily sessions

Adherence was calculated as the number of days in the second week of the study (days 9–15) that the participant completed a CBT homework session. To measure the difference in adherence to the CBT exercises between conditions, we performed a non-parametric version of one-way ANOVA (using Kruskal Wallis test) as the adherence data were not normally distributed but had homogeneous variance across conditions. To evaluate if participants' adherence was related to their attitudes towards robots (NARS), we fitted a simple linear regression model with adherence as the outcome. The assumptions of linear regression were met (linearity, independence of observations, normality of residuals (using the Shapiro-Wilk test), and homoscedasticity (using the Breusch-Pagan test)).

#### 2.6.5 Thematic analysis of exit interviews

Interview transcripts were analyzed in several iterations following thematic analysis methodology (Clarke and Braun, 2014). During an initial familiarization stage, seven team members reviewed all interview transcripts, then met together to identify patterns in the data and develop a set of codes, following an inductive approach (Clarke and Braun, 2014). After coding each transcript, the researchers reviewed the transcripts to develop and refine a thematic structure that reflected the participants' responses.

#### 2.6.6 LLM-based CBT exercise analysis

To determine how well the LLM-powered interactions aligned with the goals of the intended CBT exercise, two clinicians reviewed a randomly selected subsample (around 10%) of the collected transcripts. Since CBT worksheets are quite varied, we did not use an explicit rubric for this evaluation; each of the clinicians relied on their extensive CBT experience to gauge if the exercise was being followed.

## 3 Results

### 3.1 Baseline assessments

#### 3.1.1 Demographics

Baseline demographics were measured through comparisons of the participants' age and gender (when reported) among the three conditions. Levene's test was not significant but as the data were not normally distributed, a non-parametric test was used to compare adherence among conditions. A Kruskal–Wallis test was performed on the scores of the three conditions. The differences between the robot (mean = 20.07, rank mean = 16.82143), chatbot (mean = 20.75, rank mean = 20.83333), and worksheet (mean = 21, rank mean = 21.29167) conditions were not significant, $H(2, n = 38) = 1.3475$, $p = 0.5098$. To assess the distribution of gender across the three conditions, a Fisher's exact test was used due to violations of chi-square assumptions (i.e., low expected cell counts). The test indicated no significant association between gender and condition, $p = 0.351$. These results show that there were no significant difference among the three conditions for key demographic data.

#### 3.1.2 Psychological distress

The data were normally distributed in each condition and passed the homogeneity of variance assumption test. A one-way ANOVA was performed on the scores of the Outcomes Questionnaire-45 (OQ) and the results indicated no significant differences across the three experimental conditions in their positive affect at baseline, $F(2, 38) = 2.232$, $p = 0.1316$. The estimated marginal mean scores were 67.4 for the robot condition (95% CI [58.0, 76.7]), 53.5 for the chatbot condition (95% CI [43.4, 63.6]), and 64.1 for the worksheet condition (95% CI [53.5, 74.7]).

#### 3.1.3 Pre-session anxiety

We also compared pre-session participant anxiety scores for the three conditions. Assumptions of normality and homogeneity of variance were satisfied; therefore, a one-way ANOVA was conducted. Significant between-condition differences were observed in baseline participant anxiety scores: $F(2, 38) = 6.25$, $p = 0.007$. Post hoc comparisons using Tukey's HSD test indicated that anxiety scores were significantly lower in the chatbot condition (M = 46.8, 95% CI [36.0, 57.6]) compared to the robot condition (M = 66.4, 95% CI [56.7, 76.1]), $p = .025$. No other pairwise comparisons were significant ($p > .20$). The worksheet condition (M = 59.9, 95%

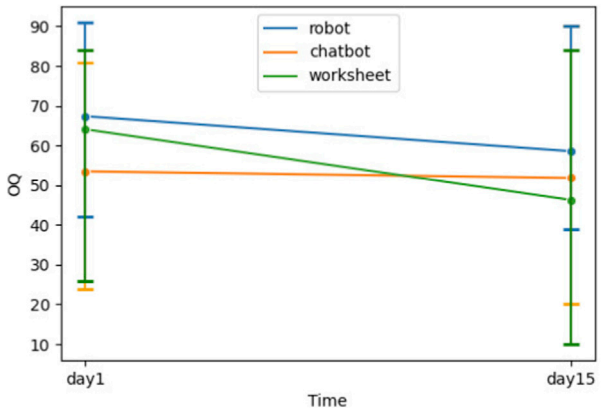



FIGURE 4
Marginal means of OQ scores from the first to last day of the study. There was a significant decrease in the OQ scores for the participants in the robot and worksheet conditions, but not in the chatbot condition.

CI [49.1, 70.7]) did not significantly differ from the other two conditions ($p > .2$).

## 3.2 Anxiety and distress measures

### 3.2.1 Statistical comparison of start to end of study general psychological distress

The Outcomes Questionnaire-45 (OQ) is a validated instrument designed to track weekly changes in mental health. It was selected for this experiment because it is commonly used to measure the effectiveness of psychotherapy. The OQ was used to assess self-reported measures of general psychological distress over the study period. The Cronbach's $\alpha$ showed high reliability for all days, $\alpha = 0.89$ for Day 1 and $\alpha = 0.93$ for Day 15. Dependent t-tests comparing Day 1 to Day 15 showed that there was a significant decrease in the OQ survey scores of participants across conditions at the end of the study ($M = 52.73$, $SD = 20.068$) compared to the beginning of the study ($M = 61.892$, $SD = 17.810$), $t(36) = 3.122$, $p = 0.004$, 95% CI [3.262,15.062]. The effect size, as measured by Cohen's d, was $d = 0.513$, a medium effect.

There was also a significant decrease in the OQ survey scores of participants in the robot condition at the end of the study ($M = 58.571$, $SD = 14.805$) compared to the beginning of the study ($M = 67.357$, $SD = 15.604$), $t(13) = 2.253$, $p = 0.042$, 95% CI [0.367, 17.468]. The effect size, as measured by Cohen's d, was $d = 0.602$, a medium effect. There was no significant decrease in the chatbot condition (95% CI [-11.307, 14.641]). There was a significant decrease in the OQ survey scores of participants in the worksheet condition at the end of the study ($M = 46.273$, $SD = 20.761$) compared to the beginning of the study ($M = 64.091$, $SD = 18.892$), $t(10) = 4.046$, $p = 0.002$, 95% CI [8.427, 29.095]. The effect size, as measured by Cohen's d, was $d = 1.22$, a large effect. Figure 4 shows Day 1 to Day 15 OQ score trends per condition.

TABLE 2 Pre-to post-session anxiety score improvements by modality. For each study condition, the days where the change from pre-to post-session anxiety scores significantly improved are shown in green.

| | Days | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Condition | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Robot | | ✓ | ✓ | | | ✓ | ✓ | ✓ |
| Chatbot | ✓ | | | | | | ✓ | |
| Worksheet | | | | | | | | |

✓ Significant
Non-significant

### 3.2.2 Statistical comparison of pre- to post-session anxiety assessments

The anxiety questionnaire (all $\alpha$'s $> 0.70$) was administered before and after each CBT session, twice daily during the first week and, depending on participants' adherence, an average of 4-6 more times during Week 2. To determine how often a single CBT session was associated with significant anxiety reduction within each condition, we conducted paired-samples t-tests comparing daily mean pre-post session anxiety scores. Given the repeated day-specific comparisons, p-values were adjusted using the Benjamini–Hochberg procedure to control the false discovery rate (FDR) at $q = 0.05$. The FDR correction was applied separately within each condition across the eight daily tests. This procedure was selected because it controls the expected proportion of false discoveries among significant findings while preserving greater statistical power than family-wise error corrections. Accordingly, only day-specific effects that survived the Benjamini–Hochberg correction were interpreted as statistically significant. Outcomes are summarized in Table 2, with detailed results in Table 3, which displays the days with significant pre-to post-session decreases and corresponding effect sizes.

Out of the eight mandatory days of CBT exercises, participants in the robot condition showed significant reductions in self-rated anxiety after Days 2, 3, 6, 7, and 8. Participants in the chatbot condition reported significant improvements on Days 1 and 7 (pre: $Mdn = 32.00$; post: $Mdn = 30.00$ at $z = -2.62$, $p = 0.009$, $r = 0.826$), and participants in the worksheet condition did not experience any significant anxiety reductions after any single session.

The data had a two-level structure, with repeated observations (Level 1) nested within participants (Level 2). The sample included 38 participants contributing a total of 297 observations, with an average of approximately 8 observations per participant. An unconditional (intercept-only) model indicated that 65.2% of the variance in post-session anxiety was attributable to between-participant differences (ICC = 0.65), suggesting substantial clustering of observations within participants and supporting the use of a mixed-effects modeling approach. A linear mixed-effects model was conducted to examine whether post-session anxiety was predicted by condition (worksheet, robot, and chatbot), day in the study, and both between-participant (grand-mean centered pre-session anxiety; pre_gmc) and within-participant (cluster-mean centered pre-session anxiety; pre_cmc) components of pre-session anxiety. Random intercepts and random slopes for pre_cmc were

TABLE 3 Pre-to post-session anxiety improvements. Only days with significant reductions in anxiety ($p < .05$) are shown. The corrected p values after the Benjamini–Hochberg procedure are given in column p-BH. The results for Chatbot Day 7 were computed using Wilcoxon signed-rank test because the assumption of homogeneity of variance was violated.

| Condition | Day | Pre-session M (SD) | Post-session M (SD) | Test statistic | p-value | p-BH | Cohen's d | 95% CI |
|---|---|---|---|---|---|---|---|---|
| Robot | 2 | 64.68 (18.93) | 58.75 (21.30) | t(13) = −2.981 | 0.011 | 0.024 | 0.797 | [-21.474, −3.428] |
| Robot | 3 | 61.93 (24.20) | 51.00 (11.99) | t(13) = −2.795 | 0.015 | 0.024 | 0.747 | [-17.597, −2.252] |
| Robot | 6 | 55.08 (17.74) | 47.62 (15.56) | t(12) = −3.757 | 0.003 | 0.022 | 1.04 | [-13.202, −3.510] |
| Robot | 7 | 55.79 (21.85) | 46.86 (16.13) | t(13) = −3.282 | 0.006 | 0.024 | 0.877 | [-14.806, −3.051] |
| Robot | 8 | 60.64 (29.40) | 51.00 (20.10) | t(13) = −2.838 | 0.014 | 0.024 | 0.758 | [-13.196, −1.788] |
| Chatbot | 1 | 46.83 (12.38) | 44.75 (18.21) | t(11) = −3.867 | 0.003 | 0.021 | 1.12 | [-9.344, −2.565] |
| Chatbot | 7 | 38.92 (14.16) | 35.00 (11.92) | z = −2.620 | 0.009 | 0.035 | 0.826 | [-11.000, −1.500] |

included for participants. There was a significant main effect of condition. Compared to the worksheet condition, participants in the robot condition reported significantly lower post-session anxiety (b = −6.26), (p = 0.001), as did participants in the chatbot condition, (b = −6.74), (p = 0.002). The effect of day was not significant, (b = −0.33), (p = 0.111). There was a significant positive effect of between-participant pre-session anxiety, such that individuals with higher average pre-session anxiety reported higher post-session anxiety, (b = 1.13)(p < 0.001). Similarly, within-participant fluctuations in pre-session anxiety were positively associated with post-session anxiety, (b = 0.63)(p < 0.001). Condition significantly moderated the between-participant effect of pre-session anxiety. The association between pre_gmc and post-session anxiety was lower in both the robot condition, (b = −0.44),(p < 0.001), and the chatbot condition, (b = −0.41), (p = 0.003), relative to the worksheet condition. However, the condition did not significantly moderate the within-participant effect of pre-session anxiety (p > 0.05). Random effects indicated substantial variability in baseline post-session anxiety across participants (($\sigma^2$ = 11.64), (SD = 3.41)), as well as variability in the within-participant effect of pre-session anxiety (($\sigma^2$ = 0.10), (SD = 0.32)). The residual variance was 55.23 ((SD = 7.43)). These results are further detailed in Table 4, and the model equations are outlined below.

The following equations detail the mathematical framework of the linear mixed-effects model, specifying how post-session anxiety is determined by the experimental condition, study duration, and the partitioned components of pre-session anxiety.

Level 1: Within-Participant

$$\begin{aligned}\text{post}_{ij} &= \beta_{0j} + \beta_1(\text{robot}_j) + \beta_2(\text{chatbot}_j) + \beta_3(\text{day}_{ij}) \\ &+ \beta_4(\text{pre_gmc}_j) + \beta_{\text{pre_cmc},j}(\text{pre_cmc}_{ij}) \\ &+ \beta_6(\text{robot}_j \times \text{pre_gmc}_j) + \beta_7(\text{chatbot}_j \times \text{pre_gmc}_j) \\ &+ \beta_8(\text{robot}_j \times \text{pre_cmc}_{ij}) + \beta_9(\text{chatbot}_j \times \text{pre_cmc}_{ij}) + r_{ij}\end{aligned}$$

Level 2: Between-Participant

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{\text{pre_cmc},j} &= \gamma_{50} + u_{1j}\end{aligned}$$

Combined model (with fixed effects coefficients):

$$\begin{aligned}\text{post}_{ij} &= 53.16 - 6.26(\text{robot}_j) - 6.74(\text{chatbot}_j) - 0.33(\text{day}_{ij}) \\ &+ 1.13(\text{pre_gmc}_j) + 0.63(\text{pre_cmc}_{ij}) \\ &- 0.44(\text{robot}_j \times \text{pre_gmc}_j) - 0.41(\text{chatbot}_j \times \text{pre_gmc}_j) \\ &- 0.11(\text{robot}_j \times \text{pre_cmc}_{ij}) - 0.21(\text{chatbot}_j \times \text{pre_cmc}_{ij}) \\ &+ u_{0j} + u_{1j}(\text{pre_cmc}_{ij}) + r_{ij}\end{aligned}$$

where $i$ is an observation from participant $j$, $u_{0j}$ is the random intercept for participant $j$, $u_{1j}$ is the random slope for pre_cmc for participant $j$, and $r_{ij}$ is the residual.

By combining these levels, the final equation accounts for the specific predictive power of the robot and chatbot interventions while simultaneously controlling for random error and participant-level variability.

To evaluate the sensitivity of the mixed-effects model, simulation-based power analyses were conducted using the simr package with 1,000 simulations. Power was estimated for the primary fixed effects of interest. The model demonstrated high power to detect the observed effect of condition (96.4%, 95% CI [95.1%, 97.5%]) and the observed condition × baseline anxiety interactions (97.0%, 95% CI [95.7%, 98.0%]). In contrast, power to detect the observed effect of study day was comparatively low (38.7%, 95% CI [35.7%, 41.8%]), suggesting that effects of this magnitude may have been difficult to detect with the current sample size n = 38. Because the observed power for the day-level effect was below conventional thresholds, we conducted a simulation-based minimum detectable effect size analysis using the simr package. Results indicated that, given the current sample size and model structure, an effect of approximately $\beta = 0.61$ would be required to achieve 80% power for detecting the day-level fixed effect.

## 3.3 Usability

### 3.3.1 Statistical comparison of SUS

A one-way ANOVA test was performed on the system usability scale (SUS) scores of the three conditions. The differences between the robot (mean = 76.250, rank mean = 20.786), chatbot (mean = 75.833, rank mean = 20.958) and worksheet (mean = 68.958, rank

TABLE 4 Linear Mixed-Effects Model Results Predicting Post-Session Anxiety. The worksheet condition was taken as the base model. 297 observations nested within 38 participants. pre_gmc = rand-mean centered pre-session anxiety; pre_cmc = cluster-mean centered pre-session anxiety. Significance codes: 0***0.001**0.01*0.05. 0.1" 1.

| Significance level | Estimate | Std. Error | df | t value | Pr( > \|t\|) | Term |
|---|---|---|---|---|---|---|
| (Intercept) | 53.16 | 1.45 | 56.89 | 36.73 | < 2e-16 | *** |
| Robot | −6.26 | 1.80 | 32.49 | −3.48 | 0.0014 | ** |
| chatbot | −6.74 | 2.04 | 33.01 | −3.31 | 0.0023 | ** |
| day_number | −0.33 | 0.21 | 255.50 | −1.60 | 0.1112 | |
| pre_gmc | 1.13 | 0.08 | 33.49 | 14.80 | 2.98e-16 | *** |
| pre_cmc | 0.63 | 0.14 | 40.54 | 4.63 | 3.74e-05 | *** |
| robot:pre_gmc | −0.44 | 0.10 | 32.45 | −4.35 | 0.0001 | *** |
| chatbot:pre_gmc | −0.41 | 0.13 | 33.35 | −3.17 | 0.0033 | ** |
| robot:pre_cmc | −0.11 | 0.17 | 34.97 | −0.65 | 0.5180 | |
| chatbot:pre_cmc | −0.21 | 0.20 | 40.02 | −1.09 | 0.2833 | |
| Random effects | | | | | | |
| participant_id (intercept) | 11.64 | 3.41 | | | | |
| participant_id (pre_cmc) | 0.10 | 0.32 | | | | |
| Correlation (intercept, pre_cmc) | 0.47 | | | | | |
| Residual | 55.23 | 7.43 | | | | |
| Observations: 297, Groups: 38 | | | | | | |

mean = 16.542) conditions were not significant, H(2, n = 38) = 0.992, $p = 0.381$, $\eta^2 = 0.054$, therefore, there was no effect.

### 3.3.2 Statistical comparison of session duration

A Kruskal–Wallis test was conducted to examine differences in session duration across conditions. There was no statistically significant difference between groups, H(2, N = 38) = 2.33, p = 0.312. Median session durations were 6.58 min (95% CI [3.89, 10.59]) for the robot condition, 4.79 min (95% CI [2.69, 5.88]) for the chatbot condition, and 5.57 min (95% CI [3.77, 7.50]) for the worksheet condition. The overlapping confidence intervals suggest comparable session durations across conditions.

## 3.4 Adherence

### 3.4.1 Statistical comparison of days of optional use

Adherence was measured as the number of days in Week 2 that participants continued to use their system for CBT exercises. Levene's test was not significant; as the data were not normally distributed, a non-parametric test was used to compare adherence between conditions. A Kruskal–Wallis test was performed on the scores of the three conditions. The differences between the robot (mean = 2.07, rank mean = 19.46), chatbot (mean = 3.67, rank mean = 23.96), and worksheet (mean = 1.08, rank mean = 15.08) conditions were not significant, H(2, n = 38) = 4.2, $p = 0.1$, $\eta^2 = 0.062$ hence, there was a medium effect (See Figure 5).

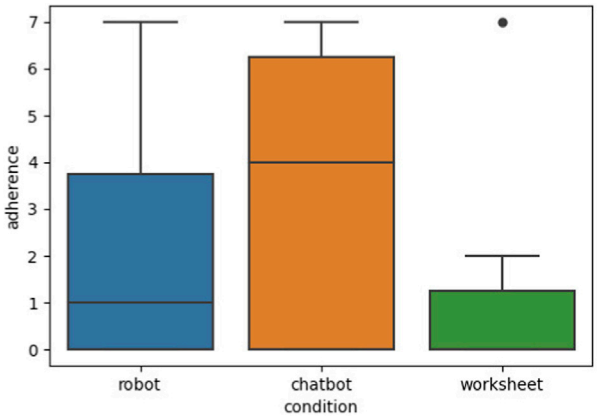



FIGURE 5
Adherence means. The distribution of adherence, measured by the number of days participants completed the exercises during the optional week (Days 9–15 of the study), for each condition.

### 3.4.2 Correlation between attitude towards robots and adherence

A simple linear regression indicated that there was a significant effect between the adherence and a participant's negative attitude towards robots (NARS) ($F(1, 9) = 9.139$, $p = 0.014$, $R^2 = 0.449$). The independent variable NARS was examined further and was a significant predictor of adherence ($t = -3.02$, $b = -0.339$, $p = 0.014$, 95% CI [-0.593, −0.085]).

## 3.5 LLM-based CBT exercise analysis

Two clinicians evaluated around 10% of the LLM-based condition transcripts. They found that the LLM initiated exercises aligned closely with the exercise structure (for both cognitive restructuring and coping strategies) at the beginning of interactions and was able to incorporate a variety of CBT skills and techniques beyond the worksheet's requirements, such as including psychoeducation to facilitate participant insight and understanding. After beginning an exercise, the LLM demonstrated active collaboration with the participant and could appropriately shift away from the worksheet structure to provide needed psychoeducation or discuss barriers to treatment that were preventing the participant from engaging in the designated, structured CBT technique. These findings are reflective of similar traditional in-person psychotherapy practices that focus on building rapport and encouraging treatment "buy-in" to facilitate stronger treatment adherence Leach (2005).

We also compared the conversation lengths in the chatbot condition and the robot condition. The results of a Welch's two-sample t-test indicated that there was no significant difference in the average number of turns between the chatbot condition (M = 11.28) and the robot condition (M = 12.01), t(275.78) = −0.93, p = 0.35, 95% CI [-2.27, 0.81]. A second Welch's two-sample t-test indicated that the average word count did not differ significantly between the chatbot condition (M = 168.56) and the robot condition (M = 225.23), t(23.01) = −1.03, p = 0.31, 95% CI [-170.13, 56.80].

## 3.6 Thematic analysis

### 3.6.1 CBT delivery method

#### 3.6.1.1 Worksheet

Participants completing CBT worksheets generally found the web-based worksheet interface easy to use, with little to no learning required. They reported that completing daily worksheets led them to reserve time each day for reflection. However, many participants found that static questions did not always align with their specific problems and wanted more guidance or dynamic follow-up questions. For example, P38 stated: *"It is a little one-sided: I am just writing what I am feeling I am not able to discuss with anyone. I am not able to get an alternate version of what I am feeling."*.

#### 3.6.1.2 Chatbot

In the chatbot condition, participants found the chatbot useful and simpler to use than they expected. Many found it engaging, and some participants even used it beyond the required 8 days during the remainder of the 2-week study period. Like those in the worksheet condition, some participants enjoyed writing out their thoughts. They appreciated the informal, conversational style, comparing it to texting, and valued the opportunity to write out their thoughts. However, many participants reported a lack of social connection with the chatbot, feeling that, as P16 stated, *"There is nobody there. It's just a bot."* Other participants commented that the chatbot could be helpful for others, but not for them.

#### 3.6.1.3 Robot

Participants in the robot condition reported feeling nervous on the first day. They were concerned about using it incorrectly and somewhat hesitant to interact with the robot, which was an unfamiliar process for them. P4 recounted, *"I was a little bit hesitant at first. It was a little interesting to be talking to a robot"*. After the first session, participants in the robot condition generally reflected positively about their interactions, reporting that completing exercises with a physical robot was fun and engaging. Blossom's zoomorphic, non-threatening design was often described as cute, approachable, and supportive. For example, P10 affectionately saying *"It is like a friend or a pet."* Many participants appreciated the robot's movements and deep breathing, which made it feel more lifelike, though a few initially found these movements uncomfortable. Some expressed a desire to customize the robot's appearance or voice. P3 was *"shocked to hear Blossom's deep voice, […] expect[ing] a cartoon character voice"* instead. Overall, participants emphasized that the robot's physical presence and verbal output enhanced the experience and made the CBT exercises feel more human. P5 said: *"It was essential to have the robot rather than just the computer system […] when it's just [chatbot messages], it doesn't feel as human as it would be with a robot."*

### 3.6.2 LLM performance

Both the chatbot and the robot were powered by OpenAI's GPT-3.5 LLM. We evaluated how participants perceived the quality, engagement, and usefulness of their conversations to account for the fact that LLMs can sometimes produce inconsistent or unexpected responses.

#### 3.6.2.1 Chatbot

Participants in the chatbot condition were generally impressed by its conversational abilities, appreciating the personalized responses and noting that it felt surprisingly human-like and flexible. P16 even said, *"At first questioning if I was actually talking to AI. It felt like a person."*

However, several challenges emerged. Some participants felt the chatbot offered generic advice or lacked empathy, and a few wanted it to take a leading role in suggesting coping strategies. Two participants reported feeling disheartened because they perceived that the chatbot did not care about what they were saying. Participants sometimes had to work to sustain conversations, and the chatbot occasionally ended sessions prematurely. P18 complained that *"Sometimes it does cut you off at the end."* While the chatbot did not retain any information between sessions, occasional hallucinations led some participants to believe it remembered prior conversations, and many expressed a desire for it to do so. Over time, participants learned that providing the chatbot with more details elicited richer, more relevant responses.

#### 3.6.2.2 Robot

Similarly, participants in the robot condition were impressed with the robot's conversational abilities, but encountered a couple of challenges. They sometimes had to take the lead in carrying the conversation and were frustrated when the robot tried to end the dialogue prematurely. Participants also wanted deeper discussions and more guidance beyond the robot's generic suggestions. As P1 expressed it, *"It was the same responses every time, it always suggested to calm down."*. As with the chatbot, the robot occasionally implied memory of prior sessions or planning for future ones (e.g., "Let's

talk about this next time"), leading some participants to believe it could remember past conversations. Many suggested improving the robot by enabling it to reliably recall previous sessions to support continuity and personalized guidance.

### 3.6.3 Therapeutic experience and outcomes

The study was intentionally conducted during the last 2 weeks of the academic semester, a period of naturally heightened stress for students. In this section, we discuss the participants' experiences with the CBT exercises across the three delivery methods, examining how they engaged with the activities, reflected on their stressors, and applied the strategies in their daily lives.

#### 3.6.3.1 Worksheets

Participants in the worksheet condition emphasized that completing the daily exercises encouraged self-reflection, a core component of CBT (Beck, 2020). Many noted that writing about their thoughts helped them to identify illogical patterns of thinking and to explore solutions to their challenges. P36 shared: *"It was beneficial to see exactly what was just written. If I saw something that was not a logical thought that I had just written, I could catch myself."* Several participants reported that the exercises reduced stress and helped them to apply coping strategies in real time. P32 said: *"One time, I was studying for a test I had to do and I was very frustrated because I didn't get much time. I got reminded about the survey and I did it right there and then. I got through a problem-solving [exercise] […] and realized I could use the strategies."*

However, some participants felt that reflecting on their problems did not relieve stress or address long-term challenges. P37 noted: *"I guess I felt a bit better talking about it, but it didn't relieve the stress because it didn't do my final or assignments for me […] I kind of felt stressed afterwards because I know it's still a problem."* A few participants who already considered themselves analytical or self-aware found the worksheets less helpful. P36 suggested: *"CBT is good especially for people who can't see through their own thoughts, but for people who are self-aware and able to see things from different points of view and take them to heart, CBT might not be the way to go. They might need more of a solutions-based approach."*

#### 3.6.3.2 Chatbot

Participants in the chatbot condition generally found the CBT exercises effective and appreciated talking through their stressors, with P23 reporting, *"After the end of the session I was completely relieved and less stressed."* Similarly, many participants stated that using the chatbot decreased their stress by helping them explore their feelings and find a more positive outlook. P18 reported: *"It's not the end of the world; it was helpful to remind myself of that sometimes."* Participants familiar with CBT reported that the exercises met their expectations, while newcomers felt the chatbot provided a helpful introduction. Some participants reported improved goal-setting, better reframing of negative experiences, and increased openness to seeking therapy.

Although most participants appreciated the interactions, some found the chatbot frustrating, reporting that it focused too heavily on negative emotions, offered overly generic solutions, or did not provide sufficient empathy. P21 recalled: *"The chatbot was too abrupt and very focused on just negative emotions, just trying to get the negative emotions out of me and nothing else."* One participant similarly felt that the chatbot could not provide much emotional gratification and lacked a sense of understanding. Others noted that exercises sometimes increased stress by prompting reflection on negative experiences. For example, P22 said: *"I would leave feeling more upset because I would be thinking about more negative thoughts […] issues I wasn't already thinking about."* A few participants also reported that the exercises had little effect because they did not apply the coping strategies or internalize the process.

#### 3.6.3.3 Robot

Participants in the robot condition overall reported a positive experience, often noting reduced stress and anxiety. For example, P5 reported: *"I had fewer panic attacks in general. If I was feeling really anxious I could sit down with Blossom, and it overall reduced my anxious behavior."* The robot facilitated reflection on daily stressors and helped participants restructure thoughts and identify solutions. Participants new to CBT found it to be an accessible introduction, with P6 noting: *"[CBT] feels a lot less intimidating to have as a resource and it has less stigma in [their] mind"*, and other participants echoing this sentiment. Many participants felt the robot validated their experiences, provided reassurance, and offered empathetic support. They reported applying coping strategies learned with the robot in everyday situations and becoming more socially engaged, such as P14, who stated: *Blossom made me feel less socially anxious"*. P4 noted: *One thing Blossom said was 'go and talk to people who really care about you,' so then I went to my best friend and it helped me address problems I have been putting off."*

However, some participants reported similar frustrations as those in the chatbot condition, including decreased mood after reflecting on stressors and negative thoughts, insufficient comfort, or perceived disregard for their problems. One participant also pointed out that a human therapist could provide more nuance.

### 3.6.4 Motivation and frequency of use

Participants' engagement with the exercises varied across the three delivery methods, influenced by both personal preferences and external factors such as time constraints. Next, we examine how often participants chose to complete the exercises, their motivations, and factors that encouraged or prevented use in each condition.

#### 3.6.4.1 Worksheet

Participants had mixed preferences for the frequency of worksheet completion. Many preferred to use the worksheets on an as-needed basis, stating that some days they had no pressing issues to reflect on. According to P30: *"Some days there was no need, I was doing them only when I needed to structure my thoughts."* Those who found the worksheets helpful in addressing daily challenges tended to complete them daily, and some even developed a habit of using them consistently, with P39 noting, *"It became a habit for me to sit down and do it […] I'm pretty sure I'll be using this for a long time."*

Many participants did not continue to complete worksheets into the second week. Motivation waned for several participants

due to external factors such as busy schedules, difficulty finding a private space, and the effort required to complete the exercises. Some participants also noted a lack of accountability, including the fact that there were no measures in place to keep them from speeding through the activity.

#### 3.6.4.2 Chatbot

Similar to the worksheet condition, many participants reported using the chatbot on an as-needed basis, particularly when they experienced stress. Some ran out of topics to discuss, while others continued because new issues arose. Time constraints were a common reason for reduced engagement in the second week. P28 stated: *"I did not continue with the exercises, mostly because as a college student, I don't have time to do what I want to do. I'm a biomedical engineering major. That takes up a lot of my time."* Novelty and curiosity about the chatbot's responses encouraged some participants to continue, while others suggested adding accountability mechanisms like follow-ups with a human therapist to increase motivation.

#### 3.6.4.3 Robot

Participants in the robot condition reflected more positively on completing daily sessions than participants in the chatbot and worksheet conditions, observing that they felt better with consistent use. Participants were generally excited and curious to interact with the robot, finding it novel and engaging. Many reported that daily interactions helped them establish a routine. For example, P6 stated: *"Continuing CBT with this system really encourages you to do it every day."* Still, some participants reported that they only used the robot when stressful events had occurred, such as P7, who said: *"When I feel crappy that day, I go to Blossom. It was a good resource for me"*. Only two participants reported wanting less frequent interactions with the robot.

The physical presence of the robot served as a reminder to complete exercises, though some participants noted that it lacked the authority of a human therapist. P8 said: *"A robot that cute doesn't give as much authority as a licensed therapist."* During the second week, many participants continued using the robot regularly, and some found that repeated interactions allowed for deeper reflection. A few participants adjusted the duration or frequency to suit their schedules, emphasizing the flexibility of the robot as an advantage. Participants who decreased or discontinued the use of the robot in the second week often cited external reasons, such as being busy with schoolwork.

# 4 Discussion

This work leveraged SAR and LLM technologies to support CBT homework exercises. We compared three interfaces for facilitating regular at-home CBT homework exercises for college students. Most interestingly, we found a paradox where the least flexible interface, the worksheet condition, led to the greatest longitudinal improvement in therapeutic outcomes, while the interactive LLM-based conditions had the strongest pre-to post-session anxiety reductions. We ground our discussion of the results in both the quantitative and qualitative findings of this exploratory study.

## 4.1 Therapeutic outcomes

### 4.1.1 Weekly therapeutic outcomes

To address RQ1, which examined how the CBT in-home exercise delivery method affects short-term therapeutic outcomes, we compared students' Day 1 OQ scores with their Day 15 scores, because this validated instrument tracks weekly changes in mental health and measures the effectiveness of psychotherapy. A dependent t-test found a statistically significant decrease in OQ across all conditions, suggesting that sustained engagement with CBT exercises over a 2-week period is effective for decreasing psychological distress.

A condition-by-condition analysis of changes in OQ score by the end of the intervention provided interesting insights. The worksheet condition showed the largest effect on the students' differences in OQ scores, followed by the robot condition, and there was no statistically significant improvement found in the chatbot condition. Unsurprisingly, the worksheet condition produced large improvement in OQ scores, consistent with the extensive clinical use and empirical support for worksheet-based CBT interventions. The robot condition also produced statistically significant improvements, suggesting that SAR-delivered CBT is a viable modality, possibly because of an increased sense of participant accountability. In contrast, it was unexpected that the chatbot condition had no significant difference in OQ scores, as existing research has suggested that chatbot-guided therapy can lead to positive therapeutic outcomes (Jang et al., 2021). The chatbot appeared to suffer from the "worst of both worlds", lacking both the clinical validation of worksheets and the physical presence that made the robot engaging.

Since there was no statistically significant difference in Day 1 OQ scores or demographic makeup between the condition groups, it is unlikely that baseline group differences could account for the varied outcomes. However, there could be other baseline differences between groups that we failed to capture; for example, the time of day at which exercises were most frequently completed. Furthermore, because we did not have a content moderator to ensure that the robot and chatbot conditions stayed directly aligned with the worksheet content, we could have had differences in CBT content delivery across conditions, which would have influenced participant outcomes. Similarly, we had no formal measure of how deeply participants engaged with exercises or how motivated they felt, which could be confounding variables in achieving therapeutic outcomes. Overall, these results should be interpreted with caution, given the limited sample size across conditions. Nonetheless, these trends provide an intriguing look into the effectiveness of SAR-assisted CBT exercises.

### 4.1.2 Single-session therapeutic outcomes

While the previous analysis examined short-term therapeutic outcomes across the 2-week intervention, RQ1 also considers immediate therapeutic outcomes. To address this aspect of the research question, we analyzed anxiety ratings at the beginning and at the end of each session.

As a baseline assessment, we examined whether pre-session anxiety scores on Day 1 differed across the three conditions.

Results showed that participants in the chatbot condition reported significantly lower pre-session anxiety than those in the robot condition, while no significant differences were observed between the worksheet condition and either of the other two conditions. These findings are consistent with post-study interview feedback, in which more participants in the robot condition expressed feeling anticipatory nervousness about the upcoming interaction on Day 1 than participants in the other conditions. Given that this instrument was measured at the start of the first session (once the robot was set up and running on the participant's desk), we interpret this difference as part of settling into the study and getting used to using the robot. Participants in the other conditions reported greater familiarity and preparedness, drawing on prior experiences with worksheets or chatbots.

We calculated whether there was a decrease in participant anxiety from the start to the end of each mandatory session for each condition. Interestingly, we found that the worksheet condition showed no statistically significant immediate reductions in anxiety, the chatbot condition showed a statistically significant improvement in participants' anxiety on two of the eight mandatory days, and the robot condition showed improvement on five of the 8 days.

The lack of immediate anxiety reduction in the worksheet condition may have been because the worksheet exercises were more cognitively demanding and less exciting to complete than engaging with an LLM-based interactive system. However, participants still experienced long-term improvements in the worksheet condition, demonstrating that though individual sessions did not provide an improvement pre-/post-session, they enabled participants to internally process their emotions and learn CBT skills for sustained change, consistent with the literature on the value of CBT homework exercises (Kazantzis et al., 2010).

The chatbot modality had some immediate effects, but lacked significant long-term distress reduction. This highlights an important distinction between exercises that may feel satisfying or effective in the moment, but may not translate into long-term therapeutic benefit.

The robot condition demonstrated the most frequent statistically significant improvements, with over half of the mandatory sessions yielding a reduction in anxiety. This suggests two possibilities: 1) the robot-guided interaction was inherently more effective for immediate anxiety reduction; or 2) the robot-guided interaction caused the participants to start each session in a more elevated, anxious state that then was reduced through the CBT exercise. These insights merit further in-depth exploration, given that the robot condition was the only modality to have both session-level and long-term improvements.

Due to the hierarchical nature of our data, in addition to the single session anxiety assessment, we developed a mixed-effects model to assess the pre-to post-session anxiety scores. In our mixed-effects model, the significant main effect of study condition suggests that both the robot and chatbot conditions were associated with lower post-session anxiety compared to the worksheet condition. On average, participants in these conditions ended sessions with anxiety scores approximately 6–7 points lower than those in the worksheet condition, when controlling for other variables. There was no strong evidence of a systematic change in anxiety over time.

Pre-session anxiety was a strong predictor of post-session anxiety at both the between-participant and within-participant levels. Individuals with higher-than-average pre-session anxiety (relative to others) had post-session anxiety scores that were, on average, 1.13 points higher for each unit increase in pre-session anxiety. Additionally, on days when a participant's pre-session anxiety was higher than their own personal average, their post-session anxiety was also higher by approximately 0.63 points. This indicates that between-participant differences in anxiety are more strongly associated with outcomes than within-participant fluctuations. Furthermore, individuals with higher pre-session anxiety not only reported higher post-session anxiety overall, but also exhibited greater reactivity to fluctuations in their anxiety. In other words, deviations from their typical anxiety levels had a stronger impact on their post-session outcomes. This pattern is consistent with the idea that individuals with elevated anxiety may have more difficulty regulating momentary increases in distress, making therapeutic engagement more challenging.

Interestingly, there was a significant interaction between condition and average pre-session anxiety (between-participant effect). Although higher baseline anxiety was associated with higher post-session anxiety in general, this relationship was curbed in both the robot and chatbot conditions. Specifically, the increase in post-session anxiety associated with higher pre-session anxiety was reduced by approximately 0.4 points per unit increase in pre-session anxiety in these conditions. This suggests that the robot and chatbot interventions were particularly effective for individuals with higher-than-average anxiety. In contrast, for individuals with lower-than-average pre-session anxiety, the worksheet condition showed relatively better outcomes. Taken together, these findings suggest that the robot and chatbot interventions may be especially beneficial for individuals with elevated anxiety, while traditional approaches may be sufficient or even preferable for those with lower baseline anxiety. Given that CBT is often used to support individuals with various anxiety disorders, observed effectiveness of the robot and chatbot interaction among individuals with elevated anxiety is promising. However, further research is needed to better understand these effects.

As with the weekly therapeutic outcomes, there may be additional factors that contributed to changes in single-session therapeutic outcomes by condition. For example, participant motivation and session timing throughout the day were not captured in the current study, leaving opportunities for future research.

## 4.2 Adherence outcomes

To address RQ2, we examined participants' adherence to daily CBT exercises across the three delivery modalities. There was no statistically significant difference in adherence to daily CBT exercises across the three conditions. However, participants' post-study interviews provided insight into their experiences with the different modalities and the factors that influenced their willingness to continue engaging with CBT exercises. While these qualitative findings do not indicate differences in actual adherence behavior, they reveal that participants perceived the three modalities differently and cited different motivations and barriers related to continued use of each one after the required period.

Across all three conditions, participants were instructed to complete CBT exercises daily. In each modality, participants

suggested that the exercises might be more helpful at a reduced frequency or on an as-needed basis, an approach several participants enacted in the second week of the study, when practice was optional. In all conditions, some participants reported running out of topics to discuss, which may have led them to complete fewer sessions in the second week. Some participants also reported wanting to take time to practice applying the skills they learned from the exercises in their daily lives before completing additional sessions, decreasing daily participation. These themes suggest that daily CBT exercises may exceed a reasonable level of interaction for most participants.

Logistical constraints were another commonly reported reason for decreased use of the CBT exercise systems. For instance, the timing of the study partially overlapped with some participants' final exams, limiting their capacity to engage in the daily CBT exercises. Some participants also fell ill, negatively impacting their adherence. Additionally, the sessions had to be completed in a private setting due to IRB restrictions, and some participants found it challenging to find a time to be alone to complete sessions. These common logistical barriers may partly explain why adherence did not differ significantly between modalities, since participants faced similar practical constraints regardless of which system they used.

Interestingly, participants only cited the time-consuming nature of the pre- and post-surveys and the lack of accountability as deterring them from completing additional sessions in the worksheet and chatbot conditions. This may suggest that, in the robot condition, participants felt a greater sense of social obligation toward the physical robot than they did toward the other modalities, even if they did not act on it.

In the two LLM conditions (robot and chatbot), participants said that the ability to complete flexible-length sessions made the system easier to fit into their busy schedules, an advantage of these modalities suggesting that even though adherence rates were comparable across conditions, participants in the LLM conditions may have engaged more willingly and with greater enjoyment.

In the robot conditions, participants also reported that the robot's physical presence on their desk served as a visual reminder to complete the exercises. Even though this did not produce measurable increases in adherence, it points out the potential of embodiment to serve as a passive reminder toward promoting adherence.

Overall, although adherence trends for optional engagement with CBT exercises did not vary significantly by modality, the participants' perspectives on their interactions did, and users chose to speak about varying reasons that either encouraged or discouraged them from using the platforms. Another important insight from participants was that lower daily adherence did not inherently indicate a lack of interest, but may in some cases have reflected deeper engagement with the material.

## 4.3 User preferences

### 4.3.1 System usability

To address RQ3a, we examined participants' perceptions of the platform structure through usability ratings. High usability ratings across all conditions ($M > 68$) suggest that the participants found each CBT system to be simple to use. The lack of significant differences in usability across the three conditions indicates that the robot's physical complexity did not make it more difficult for participants to use the platform than the worksheet.

### 4.3.2 LLM-based exercises

To address RQ3b, we examined participants' perceptions and experiences with the LLM-driven CBT exercises through post-study interviews. Our post-study interviews highlighted that participants in the two LLM-based conditions enjoyed the flexibility of the dialog, while the participants in the worksheet condition sometimes felt limited by the static format. Participants in the worksheet condition mentioned more guidance for the worksheets would be helpful, complained that the questions felt too generic and sometimes did not align with the specific topic they wanted to discuss, and compared the system to journaling instead of to talk therapy. Participants in the LLM-based conditions, on the other hand, expressed that the interaction was fun and engaging, they were impressed with the flexible dialogue, and compared the systems to human therapists.

The human-like nature of the conversational LLM raised participants' expectations of the system, which, at times, led to subsequent disappointment and frustration when the LLM could not match the quality of a human therapist. Despite these occasional letdowns, participants appreciated the judgment-free nature of the LLM system, emphasizing that they felt comfortable being honest with a chatbot or robot without fear of being judged.

Overall, participants in the robot and chatbot exercise conditions reported positive experiences, indicating that LLMs can enable engaging CBT homework exercises. They present a viable and novel alternative to traditional CBT worksheets and can expand the set of tools available to therapists in their practice.

### 4.3.3 Robot delivery

To address RQ3c, we examined participants' perceptions and experiences with robot-delivered CBT exercises through post-study interviews. Despite the LLM system being identical across the chatbot and robot conditions, during our post-study interviews, we observed differences in participants' experiences.

Participant in both conditions reflected positively on the platform convenience, but the students in the robot conditions emphasized that the robot's physical presence was essential to their feeling of connectedness during the exercises. Participants in the chatbot condition expressed feeling that the interaction lacked social connection, with many participants sharing that they felt the chatbot did not care about them, while participants in the robot conditions were very positive about their daily interactions and relationship with the robot. Furthermore, participants in the robot condition were much less critical of LLM-related issues. Embodied agents can access diverse channels of communication, including non-verbal cues that are not available to text-based systems (Deng et al., 2019). These cues can elicit social, cognitive, and task outcomes (Mutlu, 2011), and allow a robot to be perceived as a social actor in providing a human-like conversation (Jung

and Lee, 2004; Bainbridge et al., 2008). This was reflected in our interviews, where participants talked about perceiving the robot as a friend/human and specifically linked its non-verbal cues, like breathing and head tilts, to making interactions with it feel like human conversations. It is evident that while text-based chatbots are convenient and can effectively guide users through CBT homework exercises, incorporating a physically embodied agent like an appealing robot can improve the user experience.

Overall, this study found strong promise for SAR-based systems as a viable, effective, and well-received alternative to traditional CBT homework exercises.

## 4.4 Limitations and suggestions for future work

### 4.4.1 CBT exercise variety

We implemented two CBT homework exercises and asked participants to choose one of those two to complete each day. We limited the number of exercises to two to simplify the familiarization process with the system. However, we found that the limited variety of exercises, combined with the expectation of daily use, led participants to become fatigued with the two exercises, negatively impacting their adherence to the daily sessions. Future studies should balance the number of activities with the frequency of use and duration of the study, taking care to avoid making participants repeat the same exercises many times. Shorter, flexible sessions can also positively impact adherence, as we observed in our study.

### 4.4.2 Treatment fidelity

The LLM-enabled exercises were guided by a single system prompt, with no programmed mechanism to enforce the completion of every specific CBT step in every session. Two clinicians' review of a subset of participant transcripts confirmed that while the LLM stayed aligned with the spirit of the CBT exercises, it frequently pivoted to address immediate participant needs, such as providing psychoeducation or rapport-building, rather than strictly following the worksheet's structure. Future work would benefit from implementing state-tracking moderators to ensure CBT exercise alignment in real-time.

### 4.4.3 Clinical context of CBT homework

The CBT exercises used in this work did not fully reflect the context in which CBT homework is typically administered. While there exist self-guided CBT homework exercises developed for those without access to clinician support (for example, therapy apps on smartphones, self-help books, etc.), in clinical practice, CBT homework exercises are commonly accompanied by therapist guidance and feedback. By presenting the exercises as a standalone intervention, the study excluded components that may contribute to the effectiveness of CBT-based practices. Consequently, this study should not be interpreted as representing a complete clinical CBT intervention. Rather, the goal of this study was to compare different modalities for delivering CBT homework activities within an exploratory research context. Future work would benefit from investigating therapist-in-the-loop approaches that better reflect current clinical practices.

### 4.4.4 Sample size

A primary limitation of the described study is its small sample size ($N = 38$), with 12–14 participants per condition. The sample size of this study is consistent with a large body of work in human-robot interaction that is constrained by the challenges of scaling up experiments with real robot hardware (as discussed in Section 2.3). Consequently, this work should be viewed as an exploratory study rather than a definitive hypothesis-testing experiment. We implemented a Mixed-Effects Model to account for the nested nature of the data and maximize statistical power, but future large-scale longitudinal studies are required to confirm these preliminary trends.

### 4.4.5 Study duration

Although it has been shown that even short-term CBT interventions can yield positive outcomes (Bertuzzi et al., 2021), and immediate gains following single-session interventions may predict longer-term symptom improvement (Schleider et al., 2019), the 15-day duration of our study is substantially shorter than typical CBT programs, which often span 6–10 weeks or longer and are commonly delivered through weekly sessions rather than near-daily exercises (Ruwaard et al., 2012). While our study participants were exposed to a substantial amount of CBT content, the compressed schedule may have influenced how they engaged with the exercises and limited opportunities to reflect on and integrate newly learned skills into their daily lives. As such, it is difficult to determine whether the observed effects reflect lasting therapeutic benefits, novelty effects associated with the technology, or factors related to participants' day-to-day circumstances. Future work should use longer study periods.

### 4.4.6 Interaction design

This study was one of the first (O'Connell et al., 2024) to use the Blossom research robot in a real-world in-home setting, deploying it in university dorms. As anticipated, there were some logistical difficulties associated with the non-commercial research SAR system. In future work, we aim to improve the system design to make it more robust for in-home use. Following participants' suggestions, we hope to explore how to create a more immersive interaction between the robot and the user, for example, by enabling the robot to not only suggest coping strategies, but to guide the user through an example strategy, such as meditative breathing.

Participants wanted the SAR or chatbot to recall information that they had previously disclosed so they could pick up topics from previous sessions and avoid repeating context information in each session. Future systems should incorporate basic strategies for between-session memory recall (Hatalis et al., 2023). Future work should also study the effect of between-session

memory recall on users' perceptions of mental health agents and therapeutic outcomes.

### 4.4.7 Personalization

Participants in our study expressed a desire to personalize the robot's appearance, consistent with prior findings showing that robot personalization along several dimensions can improve user perceptions (Shi et al., 2023; Sung et al., 2009; Lee et al., 2012). We observed a diverse set of preferences and expectations among the participants around the robot's voice and movement. We suggest that, when possible, future SAR deployments consider personalizing the robot to fit the preferences of each user, allowing them to control several aspects of the robot, including the color and texture of the exterior, the perceived age and gender of the voice, and the type of movements the robot performs during the interaction.

### 4.4.8 Isolating embodiment effects

In comparing the robot condition to the chatbot condition, two variables were introduced: physical embodiment and auditory output. The robot verbalized responses using a text-to-speech system, while the chatbot was entirely text-based. This design choice was made because disembodied voices in standard chat interfaces can be perceived by users as unnatural, and this creates a confound preventing us from definitively isolating the effect of physical embodiment. Future explorations could include a voice-enabled chatbot or a screen-based virtual agent as an additional control condition.

### 4.4.9 Repeated measurement effects

Participants completed the same anxiety instrument before and after each session across all study days. Repeated administration of the measure may have introduced a test–retest effect, whereby participants' responses could have been influenced by familiarity with the instrument or response habituation over time, potentially contributing to reductions in reported anxiety across the study period.

## Data availability statement

The datasets presented in this article are not readily available due to the sensitive nature of the collected data, and our IRB approved consent forms. Requests to access the datasets should be directed to MK, kian@usc.edu.

## Ethics statement

The study involving humans were approved by University of Southern California Social Behavioral Institutional Review Board. The study were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

MK: Writing – review and editing, Software, Conceptualization, Writing – original draft, Resources, Investigation, Project administration, Supervision, Formal Analysis, Methodology, Data curation. MZ: Writing – review and editing, Conceptualization, Formal Analysis, Software, Writing – original draft, Data curation, Visualization. KF: Formal Analysis, Writing – original draft, Investigation, Data curation, Writing – review and editing, Methodology. AS: Visualization, Investigation, Software, Data curation, Writing – review and editing, Formal Analysis, Writing – original draft. A-MV: Methodology, Investigation, Writing – review and editing, Conceptualization, Writing – original draft, Formal Analysis. AO': Writing – review and editing. LI: Writing – review and editing. PS: Investigation, Software, Writing – original draft. SU: Writing – original draft, Investigation, Writing – review and editing, Software. AG: Software, Writing – original draft, Investigation. MF: Investigation, Writing – original draft. WB: Writing – original draft, Software, Investigation. SA: Conceptualization, Writing – review and editing. MC: Formal analysis, Writing – review and editing. BK: Conceptualization, Writing – review and editing, Funding acquisition. MM: Conceptualization, Funding acquisition, Resources, Writing – review and editing.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. Research reported in this publication was supported in part by the University of Southern California (USC) Viterbi School of Engineering, which supported faculty salaries and infrastructure, and the Center for Undergraduate Research in Viterbi Engineering (CURVE) program, which supported undergraduate researcher salaries; the National Science Foundation (NSF) under grant number: NSF IIS-1925083, which supported the development and construction of the Blossom robots deployed in the user study; and the National Institute of Mental Health of the National Institutes of Health (NIH) under Award Number R01MH139134, which supported partial faculty summer salary, partial PhD student salary, and publication costs. Of the approximate $93,000 spent on the project, an estimated 90% of the costs were financed by non-governmental sources, an estimated 7.8% of the costs were financed by NIH, and 2.2% were financed by NSF.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2026.1802622/full#supplementary-material

## References


Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., and Househ, M. (2019). An overview of the features of chatbots in mental health: a scoping review. *Int. J. Med. Inf.* 132, 103978. doi:10.1016/j.ijmedinf.2019.103978

Abdel Hafez, R. (2024). Enhancing human-robot interaction: integrating large language models and advanced speech recognition into the pepper robot

Abdulrahman, A., and Richards, D. (2021). In search of embodied conversational and explainable agents for health behaviour change and adherence. *Multimodal Technol. Interact.* 5, 56. doi:10.3390/mti5090056

Ahmed, A., Hassan, A., Aziz, S., Abd-Alrazaq, A. A., Ali, N., Alzubaidi, M., et al. (2023). Chatbot features for anxiety and depression: a scoping review. *Health Informatics Journal* 29, 146045822211467. doi:10.1177/14604582221146719

American Psychological Association (2017). *What is Cognitive Behavioral Therapy?*

Andrayani, L. W., Wijayanti, G. S. P. W., Emilyani, D., and Malik, Q. M. (2023). Effectiveness of reminiscence therapy and deep breath relaxation therapy on stress levels of elderly during the covid-19 pandemic. *J. Keperawatan Terpadu Integr. Nurs. J.* 5, 38–47. doi:10.32807/jkt.v5i2.366

Babbie, E. R. (2020). *The Practice of Social Research (Cengage Learning)*.

Bainbridge, W. A., Hart, J., Kim, E. S., and Scassellati, B. (2008). "The effect of presence on human-robot interaction," in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication (IEEE)*, 701–706.

Barendregt, W., Paiva, A., Kappas, A., Vasalou, A., Heath, C., Serholt, S., et al. (2014). "Child–robot interaction: social bonding, learning and ethics," in Workshop proceedings of Interaction Design and Children Conference IDC14.

Beck, J. S. (1993). *Cbt Worksheet Packet*.

Beck, J. S. (2020). *Cognitive Behavior Therapy: Basics and beyond*. New York: The Guildford Publications Inc.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statistical Society Series B Methodol.* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bertuzzi, V., Fratini, G., Tarquinio, C., Cannistra, F., Granese, V., Giusti, E. M., et al. (2021). Single-session therapy by appointment for the treatment of anxiety disorders in youth and adults: a systematic review of the literature. *Front. Psychol.* 12, 721382. doi:10.3389/fpsyg.2021.721382

Bethel, C. L., Stevenson, M. R., and Scassellati, B. (2011). "Secret-sharing: interactions between a child, robot, and adult," in 2011 IEEE International Conference on systems, man, and cybernetics (IEEE), 2489–2494.

Bhati, K. S. (2014). Effect of client-therapist gender match on the therapeutic relationship: an exploratory analysis. *Psychol. Reports* 115, 565–583. doi:10.2466/21.02.PR0.115c23z1

Boswell, J. F., Iles, B. R., Gallagher, M. W., and Farchione, T. J. (2017). Behavioral activation strategies in cognitive-behavioral therapy for anxiety disorders. *Psychotherapy* 54, 231–236. doi:10.1037/pst0000119

Brooke, J. (1996). Sus: a quick and dirty'usability. *Usability Evaluation Industry* 189, 189–194.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Information Processing Systems* 33, 1877–1901. doi:10.48550/arXiv.2005.14165

Burton, C., Szentagotai Tatar, A., McKinstry, B., Matheson, C., Matu, S., Moldovan, R., et al. (2016). Pilot randomised controlled trial of help4mood, an embodied virtual agent-based system to support treatment of depression. *J. Telemedicine Telecare* 22, 348–355. doi:10.1177/1357633X15609793

Choi, A., Lovett, A. W., Kang, J., Lee, K., and Choi, L. (2015). Mobile applications to improve medication adherence: existing apps, quality of life and future directions. *Adv. Pharmacol. Pharm.* 3, 64–74. doi:10.13189/app.2015.030302

Clabaugh, C., Mahajan, K., Jain, S., Pakkar, R., Becerra, D., Shi, Z., et al. (2019). Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders. *Front. Robotics AI* 6, 6–2019. doi:10.3389/frobt.2019.00110

Clark, G. I., and Egan, S. J. (2015). The socratic method in cognitive behavioural therapy: a narrative review. *Cognitive Ther. Res.* 39, 863–879. doi:10.1007/s10608-015-9707-3

Clarke, V., and Braun, V. (2014). "Thematic analysis," in *Encyclopedia of Critical Psychology* (Springer), 1947–1952.

Denecke, K., Vaaheesan, S., and Arulnathan, A. (2020). A mental health chatbot for regulating emotions (sermo)-concept and usability test. *IEEE Trans. Emerg. Top. Comput.* 9, 1170–1182. doi:10.1109/tetc.2020.2974478

Deng, E., Mutlu, B., and Mataric, M. J. (2019). Embodiment in socially interactive robots. *Found. Trends® Robotics* 7, 251–356. doi:10.1561/2300000056

Dino, F., Zandie, R., Abdollahi, H., Schoeder, S., and Mahoor, M. H. (2019). "Delivering cognitive behavioral therapy using a conversational social robot," in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE), 2089–2095.

Dobson, K. S. (2021). A commentary on the science and practice of homework in cognitive behavioral therapy. *Cognitive Ther. Res.* 45, 303–309. doi:10.1007/s10608-021-10217-5

Fasola, J., and Mataric, M. J. (2012). Using socially assistive human–robot interaction to motivate physical exercise for older adults. *Proc. IEEE* 100, 2512–2526. doi:10.1109/jproc.2012.2200539

Feil-Seifer, D., and Mataric, M. J. (2005). "Defining socially assistive robotics," in 9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005 (IEEE), 465–468.

Feingold Polak, R., and Tzedek, S. L. (2020). "Social robot for rehabilitation: expert clinicians and post-stroke patients' evaluation following a long-term intervention," in *Proceedings of the 2020 ACM/IEEE International Conference on human-robot Interaction*, 151–160.

Ferrante, G., Vitale, G., Licari, A., Montalbano, L., Pilato, G., Infantino, I., et al. (2021). Social robots and therapeutic adherence: a new challenge in pediatric asthma? *Paediatr. Respir. Rev.* 40, 46–51. doi:10.1016/j.prrv.2020.11.001

Fink, J. (2012). "Anthropomorphism and human likeness in the design of robots and human-robot interaction," in Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings 4 (Springer), 199–208.

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (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, e7785. doi:10.2196/mental.7785

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., and Rauws, M. (2018). Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR Mental Health* 5, e9782. doi:10.2196/mental.9782

Guemghar, I., Pires de Oliveira Padilha, P., Abdel-Baki, A., Jutras-Aswad, D., Paquette, J., and Pomey, M.-P. (2022). Social robot interventions in mental health care and their outcomes, barriers, and facilitators: scoping review. *JMIR Ment. Health* 9, e36094. doi:10.2196/36094

Hatalis, K., Christou, D., Myers, J., Jones, S., Lambert, K., Amos-Binks, A., et al. (2023). Memory matters: the need to improve long-term memory in llm-agents. *Proc. AAAI Symposium Ser.* 2, 277–280. doi:10.1609/aaaiss.v2i1.27688

Hofmann, S. G., Asnaani, A., Vonk, I. J., Sawyer, A. T., and Fang, A. (2012). The efficacy of cognitive behavioral therapy: a review of meta-analyses. *Cognitive Therapy Research* 36, 427–440. doi:10.1007/s10608-012-9476-1

Hundt, N. E., Mignogna, J., Underhill, C., and Cully, J. A. (2013). The relationship between use of cbt skills and depression treatment outcome: a theoretical and methodological review of the literature. *Behav. Therapy* 44, 12–26. doi:10.1016/j.beth.2012.10.001

Ingeson, M., Blusi, M., and Nieves, J. C. (2019). "Microsoft hololens - a mHealth solution for medication adherence," in Artificial Intelligence in Health. AIH 2018. Lecture Notes in Computer Science. Editors Koch, F., Koster, A., Riaño, D., Montagna, S., Schumacher, S., ten Teije, A., et al. (Cham: Springer), 11326. doi:10.1007/978-3-030-12738-1_8

Jang, S., Kim, J.-J., Kim, S.-J., Hong, J., Kim, S., and Kim, E. (2021). Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: a development and feasibility/usability study. *Int. Journal Medical Informatics* 150, 104440. doi:10.1016/j.ijmedinf.2021.104440

Jeong, S., Alghowinem, S., Aymerich-Franch, L., Arias, K., Lapedriza, A., Picard, R., et al. (2020). A robotic positive psychology coach to improve college students' wellbeing

Jeong, S., Aymerich-Franch, L., Arias, K., Alghowinem, S., Lapedriza, À., Picard, R. W., et al. (2022). Deploying a robotic positive psychology coach to improve college students' psychological well-being. *User Modeling User-Adapted Interaction* 33, 571–615. doi:10.1007/s11257-022-09337-8

Jung, Y., and Lee, K. M. (2004). Effects of physical embodiment on social presence of social robots. *Proc. PRESENCE* 2004, 80–87.

Kabacińska, K., Prescott, T. J., and Robillard, J. M. (2021). Socially assistive robots as mental health interventions for children: a scoping review. *Int. J. Soc. Robotics* 13, 919–935. doi:10.1007/s12369-020-00679-0

Kazantzis, N. (2021). Introduction to the special issue on homework in cognitive behavioral therapy: new clinical psychological science. *Cognitive Ther. Res.* 45, 205–208. doi:10.1007/s10608-021-10213-9

Kazantzis, N., Whittington, C., and Dattilio, F. (2010). Meta-analysis of homework effects in cognitive and behavioral therapy: a replication and extension. *Clin. Psychol. Sci. Pract.* 17, 144–156. doi:10.1111/j.1468-2850.2010.01204.x

Kelders, S. M., Kok, R. N., Ossebaard, H. C., and Van Gemert-Pijnen, J. E. (2012). Persuasive system design does matter: a systematic review of adherence to web-based interventions. *J. Medical Internet Research* 14, e152. doi:10.2196/jmir.2104

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16 (9), 606–613. doi:10.1046/j.1525-1497.2001.016009606.x

Kumar, V. M., Keerthana, A., Madhumitha, M., Valliammai, S., and Vinithasri, V. (2016). Sanative chatbot for health seekers. *Int. J. Of Eng. And Comput. Sci.* 5, 16022–16025. doi:10.18535/ijecs/v5i3.28

Kumar, H., Musabirov, I., Shi, J., Lauzon, A., Choy, K. K., Gross, O., et al. (2022). Exploring the design of prompts for applying gpt-3 based chatbots: a mental wellbeing case study on mechanical turk. *arXiv Preprint arXiv:2209.11344*. doi:10.48550/arXiv.2209.11344

Kuru, E., Safak, Y., Özdemir, İ., Tulacı, R., Özdel, K., Özkula, N., et al. (2018). Cognitive distortions in patients with social anxiety disorder: Comparison of a clinical group and healthy controls. *Eur. J. Psychiatry* 32, 97–104. doi:10.1016/j.ejpsy.2017.08.004

Kuyken, W., Padesky, C. A., and Dudley, R. (2008). The science and practice of case conceptualization. *Behav. Cognitive Psychotherapy* 36, 757–768. doi:10.1017/s1352465808004815

Laban, G., Kappas, A., Morrison, V., and Cross, E. S. (2022). "User experience of human-robot long-term interactions," in Proceedings of the 10th International Conference on Human-Agent Interaction (New York, NY, USA: Association for Computing Machinery), 287–289. doi:10.1145/3527188.3563927

Lambert, M. J., Gregersen, A. T., and Burlingame, G. M. (2004). *The Outcome questionnaire-45*.

Leach, M. J. (2005). Rapport: a key to treatment success. *Complementary Therapies Clinical Practice* 11, 262–265. doi:10.1016/j.ctcp.2005.05.005

LeBeau, R. T., Davies, C. D., Culver, N. C., and Craske, M. G. (2013). Homework compliance counts in cognitive-behavioral therapy. *Cogn. Behaviour Therapy* 42, 171–179. doi:10.1080/16506073.2013.763286

Lee, M. K., Forlizzi, J., Kiesler, S., Rybski, P., Antanitis, J., and Savetsila, S. (2012). Personalization in hri: a longitudinal field experiment. *Proc. Seventh Annual ACM/IEEE International Conference Human-Robot Interact.* 319–326. doi:10.1145/2157689.2157804

Lima, M. R., O'Connell, A., Zhou, F., Nagahara, A., Hulyalkar, A., Deshpande, A., et al. (2025). "Promoting cognitive health in elder care with large language model-powered socially assistive robots," in Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (New York, NY, USA: Association for Computing Machinery).

Liu, C. H., Stevens, C., Wong, S. H., Yasui, M., and Chen, J. A. (2019). The prevalence and predictors of mental health diagnoses and suicide among us college students: implications for addressing disparities in service use. *Depress. Anxiety* 36, 8–17. doi:10.1002/da.22830

Malik, N. A., Hanapiah, F. A., Rahman, R. A. A., and Yussof, H. (2016). Emergence of socially assistive robotics in rehabilitation for children with cerebral palsy: a review. *Int. J. Adv. Robotic Syst.* 13, 135. doi:10.5772/64163

Matarić, M. J., and Scassellati, B. (2016). *Socially Assistive Robotics*. Springer handbook of robotics, 1973–1994.

Matarić, M. J., Eriksson, J., Feil-Seifer, D. J., and Winstein, C. J. (2007). Socially assistive robotics for post-stroke rehabilitation. *J. Neuroengineering Rehabilitation* 4, 1–9. doi:10.1186/1743-0003-4-5

Matheus, K., Vázquez, M., and Scassellati, B. (2022). "A social robot for anxiety reduction *via* deep breathing," in 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (IEEE Press), 89–94. doi:10.1109/RO-MAN53752.2022.9900638

Mutlu, B. (2011). Designing embodied cues for dialog with robots. *AI Mag.* 32, 17–30. doi:10.1609/aimag.v32i4.2376

Nomura, T., Kanda, T., and Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *Ai and Soc.* 20, 138–150. doi:10.1007/s00146-005-0012-7

Östlund, U., Kidd, L., Wengström, Y., and Rowa-Dewar, N. (2011). Combining qualitative and quantitative research within mixed method research designs: a methodological review. *Int. Journal Nursing Studies* 48, 369–383. doi:10.1016/j.ijnurstu.2010.10.005

O'Connell, A., Banga, A., Ayissi, J., Yaminrafie, N., Ko, E., Le, A., et al. (2024). Design and evaluation of a socially assistive robot schoolwork companion for college students with adhd. *Proc. 2024 ACM/IEEE Int. Conf. Human-Robot Interact.* doi:10.1145/3610977.3634929

Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., et al. (2016). Autism and social robotics: a systematic review. *Autism Res.* 9, 165–183. doi:10.1002/aur.1527

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., et al. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (promis®): depression, anxiety, and anger. *Assessment* 18, 263–283. doi:10.1177/1073191111411667

Pitardi, V., Wirtz, J., Paluch, S., and Kunz, W. H. (2021). Service robots, agency and embarrassing service encounters. *J. Service Management* 33, 389–414. doi:10.1108/josm-12-2020-0435

Prasko, J., Krone, I., Burkauskas, J., Vanek, J., Abeltina, M., Juskiene, A., et al. (2022). Homework in cognitive behavioral supervision: theoretical background and clinical application. *Psychol. Res. Behav. Manag.* 15, 3809–3824. doi:10.2147/prbm.s382246

Robinson, N. L., Connolly, J., Hides, L., and Kavanagh, D. J. (2020). Social robots as treatment agents: pilot randomized controlled trial to deliver a behavior change intervention. *Internet Interv.* 21, 100320. doi:10.1016/j.invent.2020.100320

Ruwaard, J., Lange, A., Schrieken, B., Dolan, C. V., and Emmelkamp, P. (2012). The effectiveness of online cognitive behavioral treatment in routine clinical practice. *PloS One* 7, e40089. doi:10.1371/journal.pone.0040089

Scassellati, B., Admoni, H., and Matarić, M. (2012). Robots for use in autism research. *Annu. Review Biomedical Engineering* 14, 275–294. doi:10.1146/annurev-bioeng-071811-150036

Schleider, J. L., Abel, M. R., and Weisz, J. R. (2019). Do immediate gains predict long-term symptom change? Findings from a randomized trial of a single-session intervention for youth anxiety and depression. *Child Psychiatry Hum. Dev.* 50, 868–881. doi:10.1007/s10578-019-00889-2

Schueller, S. M., and Morris, R. R. (2023). Clinical science and practice in the age of large language models and generative artificial intelligence. *J. Consult. Clin. Psychol.* 91, 559–561. doi:10.1037/ccp0000848

Scoglio, A. A., Reilly, E. D., Gorman, J. A., and Drebing, C. E. (2019). Use of social robots in mental health and well-being research: systematic review. *J. Medical Internet Research* 21, e13322. doi:10.2196/13322

Segura, E. M., Kriegel, M., Aylett, R., Deshmukh, A., and Cramer, H. (2012). "How do you like me in this: user embodiment preferences for companion agents," in Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings 12 (Springer), 112–125.

Shi, Z., Chen, H., Velentza, A.-M., Liu, S., Dennler, N., O'Connell, A., et al. (2023). "Evaluating and personalizing user-perceived quality of text-to-speech voices for delivering mindfulness meditation with different physical embodiments," in Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, 516–524.

Shi, Z., Landrum, E., O'Connell, A., Kian, M., Pinto-Alva, L., Shrestha, K., et al. (2024a). How can large language models enable better socially assistive human-robot interaction: a brief survey. *Proc. AAAI Symposium Ser.* 3, 401–404. doi:10.1609/aaaiss.v3i1.31245

Shi, Z., O'Connell, A., Li, Z., Liu, S., Ayissi, J., Hoffman, G., et al. (2024b). "Build your own robot friend: an open-source learning module for accessible and engaging ai education," in Proceedings of the AAAI conference on artificial intelligence, 23137–23145. doi:10.1609/aaai.v38i21.30359

Suguitan, M., and Hoffman, G. (2019). Blossom: a handcrafted open-source robot. *ACM Trans. Human-Robot Interact. (THRI)* 8, 1–27. doi:10.1145/3310356

Sung, J., Grinter, R. E., and Christensen, H. I. (2009). "Pimp my roomba designing for personalization," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 193–196.

Swift-Spong, K., Short, E., Wade, E., and Matarić, M. J. (2015). "Effects of comparative feedback from a socially assistive robot on self-efficacy in post-stroke rehabilitation," in 2015 IEEE International Conference on Rehabilitation Robotics (ICORR) (Los Angeles: IEEE), 764–769.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* 35, 24824–24837. doi:10.48550/arXiv.2201.11903

Weng, L. (2023). Just had a quite emotional, personal conversation w/chatgpt in voice mode, talking about stress, work-life balance. *Interes. I Felt X(formerly Twitter)*.

Zaib, M., Sheng, Q. Z., and Emma Zhang, W. (2020). "A short survey of pre-trained language models for conversational ai-a new age in nlp," in *Proceedings of the Australasian Computer Science Week Multiconference*, 1–4.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. *arXiv Preprint arXiv:2303.18223 1*. doi:10.48550/arXiv.2303.18223

Ziems, C., Li, M., Zhang, A., and Yang, D. (2022). Inducing positive perspectives with text reframing. *arXiv Preprint arXiv:2204.02952*. doi:10.48550/arXiv.2204.02952