Evaluating Therabot: A Randomized Control Trial Investigating the Feasibility and Effectiveness of a Generative AI Therapy Chatbot for Depression, Anxiety, and Eating Disorder Symptom Treatment

Michael V. Heinz[1,3], Daniel M. Mackin[1,3], Brianna M. Trudeau[1], Sukanya Bhattacharya[1], Yinzhou Wang[1], Haley A. Banta[1], Abi D. Jewett[1], Abigail J. Salzhauer[1], Tess Z Griffin[1], Nicholas C. Jacobson[1,2,3,4]

[1] Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States
[2] Quantitative Biomedical Sciences Program, Dartmouth College, Hanover, NH, United States
[3] Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States
[4] Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

Correspondence concerning this article should be addressed to:
Michael V. Heinz
michael.v.heinz@dartmouth.edu
Center for Technology and Behavioral Health
Dartmouth College
46 Centerra Pkwy, Lebanon, NH 03766

Acknowledgements

We are extraordinarily grateful for the efforts of the many dedicated people who made *Therabot* possible. We thank those who contributed in meaningful ways to the creation and curation of training data, including Victor A. Moreno, Chloe S. Park, Jimena Abejon Fuertes, Jonathan J. Cartwright, Anna C. St. Jean, Erica L. Simon, Isabel R. Hillman, Enoc A. Garza, Alexandra N. Limb, Dawson D. Haddox, Mingyue Zha, Camilla M. Lee, Rachita Batra, MK Song, Cameron M. Hasund, Avijit Singh, Daniel W. Shen, Rachel E. Quist, Kaitlyn I. Romanger, Chaehyun Lee, Anjali G. Dhar, Ivy N. Mayende, Eleanor M. Rodgers, Rachel Zhang, Jenny Song, Veronica E. Abreu, Russell T. Rapaport, Mary M. Basilious, Sofia M. Yawand-Wossen, Nathan J. Kung, Jenny Y. Oh, Ashna J. Kumar, Eda Naz Gokdemir, Janelle E. Annor, Ganza, Belise Aloysie Isingizwe, Chloe N. Malave, Ezinne E. Anozie, Tara L. Karim, Nhi D. Nguyen, Krista E. Schemitsch, Helen M. Young, Mia G. Russo, Rachel E. Quist, Tonya I. Tolino, Mckenzi B. Popper, Daniel G. Amoateng, and Dr. Seo Ho (Michael) Song.

We thank those who contributed in meaningful ways to the software development, including Jason Kim, John F. Keane, Dr. George D. Price, Dr. Matthew D. Nemesure, Ore E. James, Caroline C. Hall, Brendan W. Keane, Lisa Aeri Oh, Ly H. Nguyen, Dr. William R. Haslett, Vivian N. Tran, Alexander M. Ye, Atziri Enriquez, Sarah M. Chacko, Sofia Jayaswal, D.J. M. Matusz, Jose Hernandez Barbosa, Alyssia M. Salas, Ella J. Gates, and Tianwen Chen.

We thank the team from Amazon Web Services (AWS), especially Stefan Mationg and Dr. Jianjun Xu, who provided valuable technical support for *Therabot*.

Abstract

Background

Chatbots powered by generative AI (Gen-AI) hold promise for building highly personalized, effective mental health treatments at scale, while also addressing existing user engagement and retention issues common among digital therapeutics. We present the first RCT testing an expert-fine-tuned Gen-AI-powered chatbot, *Therabot*, for mental health treatment.

Methods

Participants (N=210) were randomized to a four-week Therabot intervention (N=106) or waitlist control (WLC; N=104). Subjects were clinically stratified into major depressive disorder (MDD), generalized anxiety disorder (GAD), or clinically high risk feeding and eating disorder (CHR-FED) groups using baseline symptom severity. Primary outcomes included disorder-specific symptom changes from baseline to four and eight weeks. Secondary outcomes included user engagement, acceptability, and therapeutic alliance. Cumulative link mixed models examined differential changes pre- to post-intervention and from pre-intervention to follow-up, between the Therabot and WLC groups.

Results

The Therabot group showed large and significantly greater reductions in MDD (d = 0.845-0.903), GAD (d = 0.794-0.840), and CHR-FED (d = 0.627-0.819) symptoms relative to controls at post-intervention and follow-up. Therabot was well received and well-utilized (average use $\geq$6 hours), and participants rated the therapeutic alliance comparable to human therapists.

Conclusions

This is the first RCT demonstrating the effectiveness of a fully Gen-AI therapy chatbot for treating mental health disorders. Results are promising for MDD, GAD, and CHR-FED symptom reduction. Participants were engaged with Therabot, reported exceptional therapeutic alliance, and rated the intervention highly. Fine-tuned Gen-AI chatbots are a feasible method for creating scalable, personalized interventions in mental health.

Keywords

Therapy chatbot; Generative Artificial Intelligence; Digital Therapeutics; Depression; Anxiety; Eating Disorders

Trial Registration Number

NCT06013137

Introduction

The prevalence and burden of mental health disorders, including depressive, anxiety, and eating disorders, has increased significantly over the past three decades.[1] Despite the adverse impact of these disorders,[2] the mental health infrastructure is inadequately resourced to meet the current and growing demand for care.[3–5] While empirically-validated psychosocial treatments exist,[6–8] they are resource intensive, limiting scalability and accessibility.[9] Indeed, fewer than half of persons with a mental health disorder receive necessary care.[5] Digital therapeutics (DTx), automated, evidence-based software for the treatment or diagnosis of medical conditions,[10] offer a solution to bridge the mental health treatment gap.

While the automated, on-demand nature of DTx may improve accessibility and scalability of evidence-based mental health interventions,[11] these approaches have been plagued by attrition and low engagement.[12] Within established psychotherapies, there is evidence for the benefit of nonspecific factors, such as alliance, empathy, and shared goals.[13] DTx's relative lack of personalization and alliance compared to human-delivered psychosocial interventions likely contributes to reduced engagement.[14] These nonspecific factors have been difficult to emulate via automated technologies and may be fundamentally different or remain unachievable in automated software.[15]

Even so, artificial intelligence (AI) – the branch of computer science concerned with creating systems capable of emulating human intelligence[16]– represents a promising direction for improving personalization and engagement in DTx. Within the space of AI, chatbots hold particular promise given their capacity to emulate human conversation and dialogue, long known to be integral parts to psychotherapeutic treatments – "the talking cure".[17] Indeed, chatbots applied to mental health and wellness is not a new phenomenon, with Eliza (1966),[18] an early rule-based chatbot, used to emulate a Rogerian therapist. Since then, structured, rule-based chatbots have found application in many narrow commercial (e.g., customer service) applications and only more recently have moved into wellness, companionship, and recreational spaces.[19] The study of chatbots for mental health, however, remains nascent, with exclusively rule-based conversational agents evaluated for mental health treatment to date. While such chatbots (e.g., Woebot) have shown benefits in clinical trials[20], and in some cases a capacity to promote a therapeutic alliance,[21] they are inherently limited by their reliance on an explicitly programmed decision tree and restricted inputs.

Recent advances in computing and machine learning have pushed the bounds of modern AI, allowing for sophisticated systems capable of learning, adapting, and understanding context in natural language, removing the necessity for explicit programming. Further pushing the bounds in the language domain, the advent of generative AI (Gen-AI), recently popularized by ChatGPT, [22] has enabled automated production of novel and highly personalized responses to human input. To date, conversational agents using Gen-AI have fallen under general purpose, wellness, or companion applications,[23] rather than software intended for diagnosis and treatment of mental health disorders. While some Gen-AI-powered chatbots have shown both wide appeal and the capacity to form human-like bonds,[24] they are not intended and have not been evaluated for mental health treatment. The nondeterministic and open-ended nature of Gen-AI, enabling the possibility of incorrect or harmful responses, has given rational pause to adoption in mental health. Such risks associated with Gen-AI and related chatbots underscore the need for a systematic approach to exploring the safety of chatbots for use in mental health.

Despite significant risks, there is also significant potential benefit from the use of therapeutic Gen-AI-powered chatbots. Paired with existing frameworks for DTx, Gen-AI powered chatbots have unprecedented potential to address existing problems with engagement while powering the development of new, personalized interventions. While literature supports the effectiveness of CBT-based DTx and rule-based AI chatbots for depression and anxiety, and Gen-AI chatbots exhibit promise for addressing issues of accessibility, scalability, engagement, and personalization in mental healthcare, no prior RCTs have investigated the effectiveness and safety of a Gen-AI-powered chatbot for the treatment of mental health symptoms. Therefore, starting in 2019, we developed *Therabot*, a Gen-AI chatbot trained using expertly written therapist-patient dialogues based on third-wave CBT approaches. Developed with over 100,000 human hours, *Therabot* is designed to augment and enhance conventional mental health treatment services by delivering personalized, evidenced-based mental health interventions to clinical populations at scale. In this RCT, we examined the effectiveness of *Therabot* for the treatment of major depressive disorder (MDD), generalized anxiety disorder (GAD), and chronic high-risk for eating disorders (CHR-FED) in a large, nationally representative sample of participants. We hypothesized that participants assigned to a four-week intervention with *Therabot* would measurably improve mental health

symptomatology across all symptom domains relative to patients assigned to the waitlist control (WLC) condition, at both post-intervention and an eight-week follow-up. Furthermore, we hypothesized that participants would demonstrate a high level of engagement with *Therabot*, rate *Therabot* positively, and develop a "therapeutic alliance" with *Therabot*.

Methods

Trial Design

The study was designed as a randomized, WLC trial, with a 1:1:1 allocation ratio across the MDD, GAD, and CHR-FED groups, and a Meta Ads campaign was used to recruit adults across the U.S. Based on responses to a baseline questionnaire, participants screening positive for MDD, GAD, or CHR-FED symptoms were stratified accordingly into one of the three pathology groups and then randomized into either the control or intervention group. Comorbidity was allowed, and outcomes were analyzed based on participants scoring within the respective groups at baseline, regardless of their primary presenting problem. Participants randomized to the intervention group were given access to *Therabot* for 8 weeks, and both groups received assessments at baseline, 4 weeks, and 8 weeks. After completing their final assessment, the control group was also provided access to *Therabot*. Study data were collected and managed using REDCap electronic data capture tools hosted at Dartmouth College.[25,26]

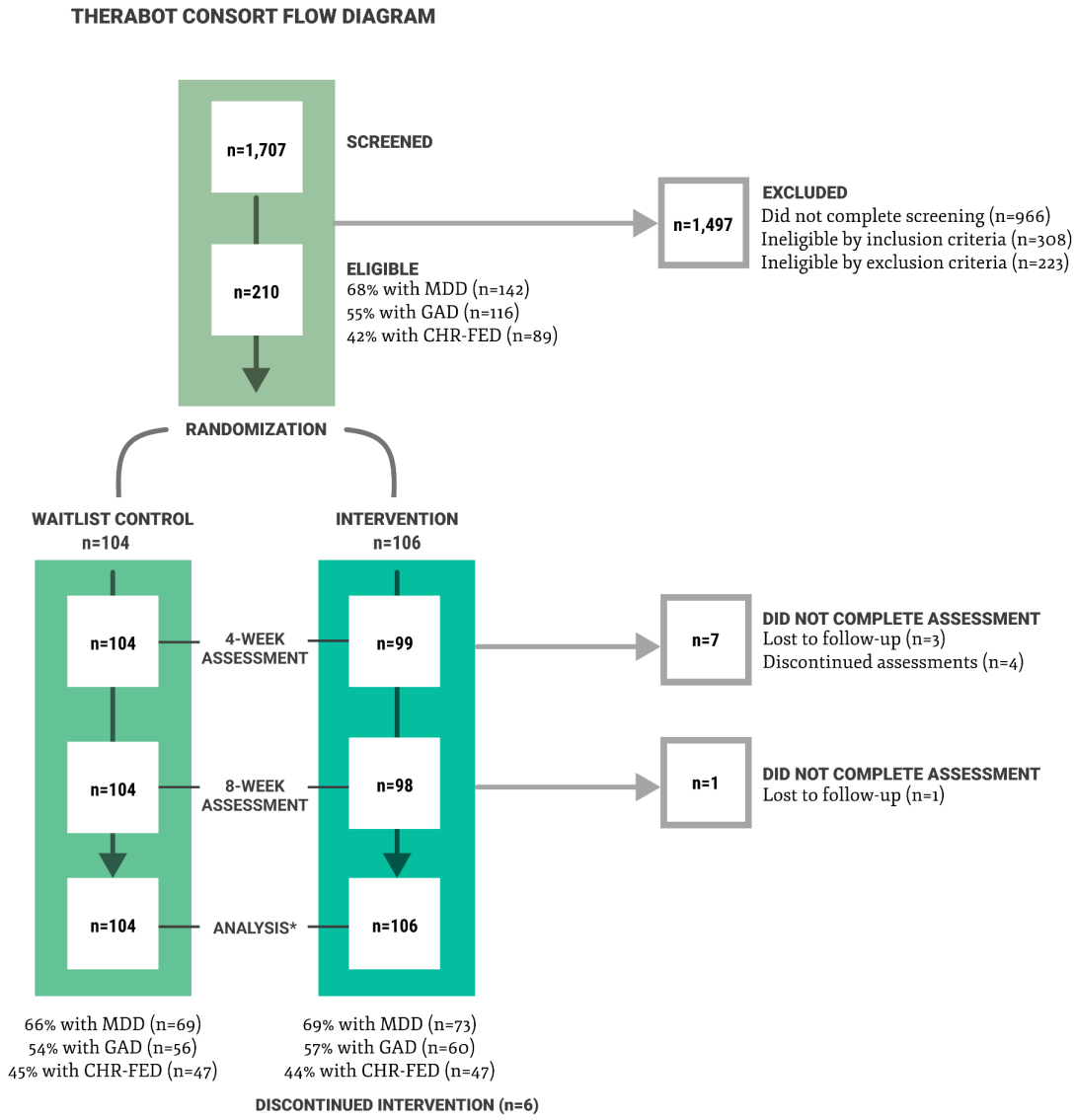

Figure 1. *CONSORT Flow Diagram showing movement of participants through the study, with associated counts. Because comorbidity is the norm rather than the exception, data were analyzed based on each participant's pathology group membership at baseline.*

Participants

Participants were required to be at least 18 years of age and to screen positive for at least one of the following disorders: MDD, GAD, or CHR-FED (see Supplemental Materials 1). Participants who screened positive for CHR-FED were given priority assignment until recruitment goals were met for that group. Otherwise, participants were assigned to the pathology group coinciding with their most severe screening measure. Exclusion criteria included active suicidality, mania, and psychosis (see, Supplemental Materials 1).

Intervention

Therabot comprised a text-based multi-thread chat application for iOS and Android capable of interacting with participants regarding their mental health problems in natural language. The intervention utilized a generative large language model (LLM), fine-tuned on expert-curated mental health dialogues. Dialogues were developed by our research team, including a board-certified psychiatrist and clinical psychologist, and peer-reviewed using evidence-based (primarily cognitive behavior therapy) modalities. Given the potential risks associated with Gen-AI, we added multiple guard rails, including a crisis classification model.

For the first four weeks, participants in the intervention group were prompted daily to engage with *Therabot*. Subsequently, participants engaged with *Therabot* as frequently as desired (weeks 4-8). The application allowed users to either initiate a session directly in the chat interface or to respond to scheduled notifications. In response to a user prompt, conversation history and most recent user message were combined and sent to the LLM. Common output included empathetic responses, validation, targeted interventions, or questions prompting elaboration (see Figure 2).

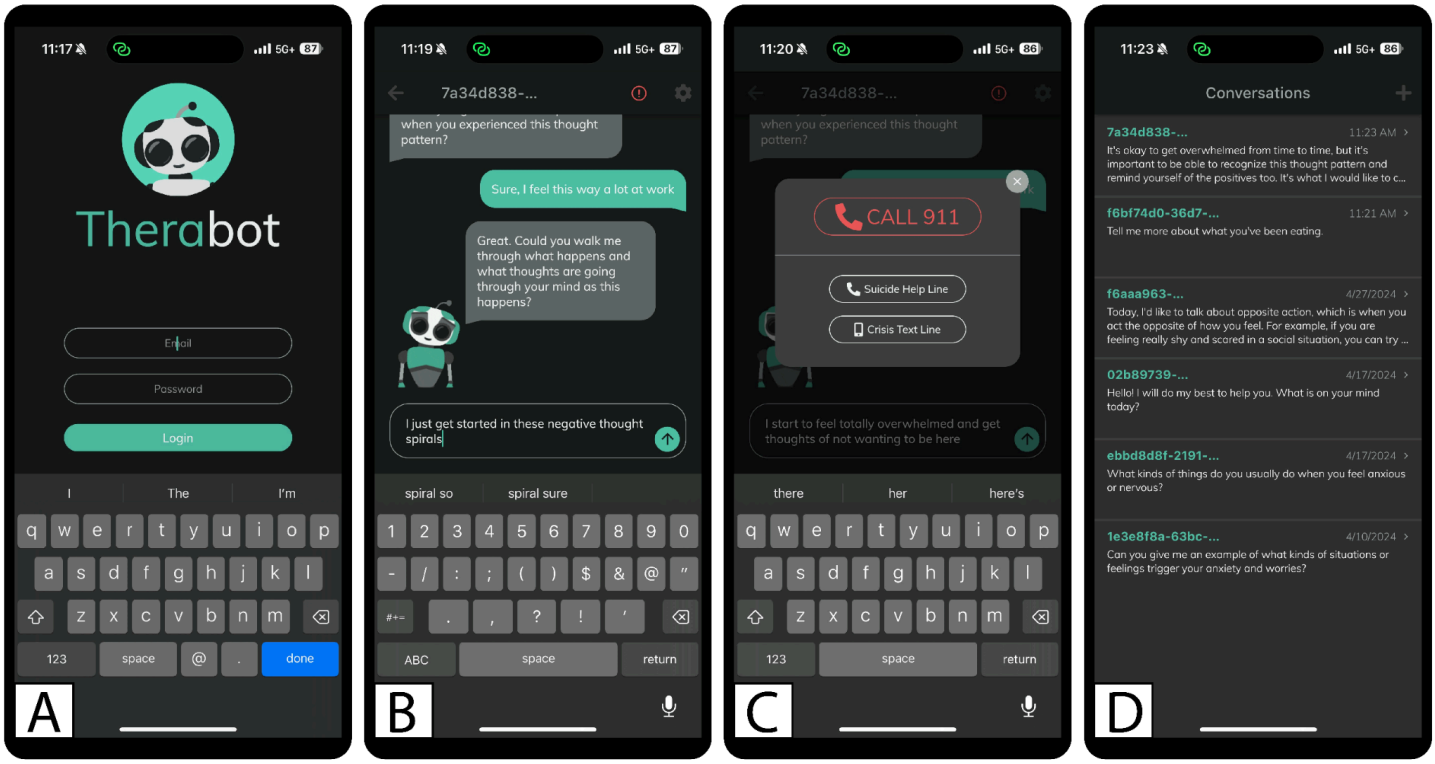

Figure 2. *Key design features of the Therabot application: (A) Therabot login screen; (B) main chat interface; (C) emergency module deployed in response to model detection of high-risk content (e.g., suicidal ideation); (D) conversation thread interface for users to initiate a thread or return to a prior thread.*

Measures

Primary outcome measures were administered at baseline, post-intervention (4 weeks), and follow-up (8 weeks) and included the Patient Health Questionnaire-9[27] (PHQ-9), the Generalized Anxiety Disorder Questionnaire for DSM-IV[28] (GAD-Q-IV), and the Stanford-Washington University Eating Disorder[29] (SWED), as measures of depression, anxiety, and weight and shape concerns (WSC), respectively. Secondary outcomes included therapeutic alliance (Working Alliance

Inventory-Short Revised[30]; WAI-SR), engagement with *Therabot* (number of messages sent), and satisfaction with *Therabot* (self-developed survey). Additional details on the primary and secondary outcomes are presented in Supplementary Materials 1.

Sample Size Justification

Based on the comorbidity between diagnoses, we expected to have 100 persons in each analysis. A Monte-Carlo simulation study was used to estimate the statistical power for the differential change in treatment response. For each simulated dataset (N = 80-150), we generated ordinal data such that the treatment group showed differential changes (d = 0.3-0.5) over time. We assumed 10% missing data. We fit Cumulative Link Mixed Models (CLMMs, see statistical methods below), with individual differences as random intercepts. The interaction effects between time and randomization group determined power, with the proportion of times the interaction terms were significant representing power. Results suggested that we had greater than 90% power to detect differential response of 0.3 and 0.5

Randomization

Randomization was performed using a computer generated random sequence with a fixed block size of 6. The randomization sequence was generated prior to the trial start and the assignment process was fully automated. Once group assignment occurred, neither researchers, nor participants were blinded to their group membership.

Statistical Methods

To examine the effectiveness of Therabot, relative to the waitlist control group, we examined the effect of time and treatment assignment on depression, anxiety, and WSC among persons at a clinical level of MDD, GAD, and CHR-FED at baseline. To address the ordinal, non-equidistant nature in the response categories of our symptom measures, and eliminate potential distortion of effect size estimates and inflated error rates, we use Cumulative Link Mixed Models (CLMMs) to analyze the effects of both time and randomization as fixed effects, and random individual differences in the outcome. The logit link function was used, which was well-suited for proportional odds models, and thresholds were assumed symmetric around the latent mean of zero. Formally, the model is described as follows:

$$(1)\quad \text{logit}(P(Y \leq j \mid \text{Time}, \text{Group}, \text{ID})) = \alpha_j - (\beta_1 \times \text{Time} + \beta_2 \times \text{Group} + \beta_3 \times \text{Time} \times \text{Group} + u_{ID})$$

Here, Y represents the ordinal outcome (MDD, GAD, or WCS score), *j* denotes the threshold category, α_j are the threshold parameters, and β_1, β_2, and β_3 are the fixed effect coefficients for time, group, and their interaction, respectively. Here, u_{ID} signifies the random intercepts for participants. Odds ratios were calculated from these estimates, and effect sizes were calculated using $d = \log(OR) * \text{sqrt}(3)/\pi$.[31]

Compensation

Participants were compensated $25 USD for each of three assessments completed.

Trial Registration and Ethical Considerations

All participants provided informed written consent prior to their participation. The Dartmouth-Hitchcock Institutional Review Board approved the research protocol. The trial was preregistered through ClinicalTrials.gov NCT06013137.

Results

Participants

Participants included 210 adults, randomized to intervention and WLC. By the 4-week assessment, four participants had withdrawn from the study and three were lost to follow-up; six participants had opted to discontinue the Therabot intervention. By the 8-week assessment, one additional participant was lost to follow-up. Detailed aggregate

demographics are displayed in Table 1. Figure 1 displays the flow of participants from screening to analysis. Recruitment occurred from March 15-31, 2024, ending when we reached our recruitment target.

Characteristic	Waitlist Control (*n*=104)	Intervention (*n*=106)	Overall (*N*=210)	p-value
Mean age (s.d.), years	33.63 (10.56)	34.09 (11.41)	33.86 (10.97)	0.757
Gender, n (%)				
Man	37 (35.58)	41 (38.68)	78 (37.14)	0.602
Woman	62 (59.62)	63 (59.43)	125 (59.52)	
Nonbinary	4 (3.85)	2 (1.89)	6 (2.86)	
Other	1 (0.96)	0 (0.00)	1 (0.48)	
Transgender, n (%)				
Yes	5 (4.81)	3 (2.83)	8 (3.81)	0.698
No	99 (95.19)	103 (97.17)	202 (96.19)	
Sexual Orientation, n (%)				
Heterosexual	82 (78.85)	84 (79.25)	166 (79.05)	0.999
Homosexual/Gay	3 (2.88)	9 (8.49)	12 (5.71)	
Bisexual	11 (10.58)	10 (9.43)	21 (10.00)	
Pansexual	3 (2.88)	1 (0.94)	4 (1.90)	
Asexual	0 (0.00)	1 (0.94)	1 (0.48)	
Bicurious	1 (0.96)	0 (0.00)	1 (0.48)	
Other	4 (3.85)	1 (0.94)	5 (2.38)	
Race and Ethnicity, n (%)				
Non-Hispanic White	56 (53.85)	56 (52.83)	112 (53.33)	0.925
Hispanic White	7 (6.73)	9 (8.49)	16 (7.62)	
Black	28 (26.92)	26 (24.53)	54 (25.71)	
American Indian	0 (0)	1 (0.94)	1 (0.48)	
Asian	5 (4.81)	6 (5.66)	11 (5.24)	
Multiple / Other	8 (7.69)	8 (7.55)	16 (7.62)	
Highest Level of Education, n (%)				
High School	4 (3.85)	7 (6.60)	11(5.24)	0.837
Some College	14 (13.46)	19 (17.92)	33 (15.71)	
Associates Degree	23 (22.12)	17 (16.04)	40 (19.05)	
Bachelor's Degree	45 (43.27)	44 (41.51)	89 (42.38)	
Master's Degree	15 (14.42)	16 (15.09)	31 (14.76)	
Doctoral Degree	3 (2.88)	3 (2.83)	6 (2.86)	
Current Student, n (%)				
No	78 (75.00)	77 (72.64)	155 (73.81)	0.767
Yes; Part Time	12 (11.54)	11 (10.38)	23 (10.95)	
Yes; Full Time	14 (13.46)	18 (16.98)	32 (15.24)	

Table 1. *Baseline characteristics of the study sample (N=210), divided into intervention group (n=106) and waitlist control group (n=104). Results of a between-group statistical comparison are displayed in the far right column. The Chi-squared test was used for categorical variables, and the Welch two sample t-test was used for continuous variables, with a two-tailed threshold of p=0.05 for significance.*

Primary Outcomes

Major Depressive Disorder

There were significant interactions between randomization and pre-post time(β = -1.533, SE = 0.404, Z = -3.797, p < .001, OR = 0.216, d = -0.845, $\Delta_{median,\ Therabot}$ = -6 pts, $\Delta_{median,\ Control}$ = -2 pts) and randomization and pre-follow-up time (β = -1.639, SE = 0.410, Z = -3.996, p < .001, OR = 0.194, d = -0.903, $\Delta_{median,\ Therabot}$ = -8 pts, $\Delta_{median,\ Control}$ = -4 pts), suggesting a large and differential response in the treatment group compared to the control group at post and follow-up (Figure 3, Row 1).

Generalized Anxiety Disorder

There were significant interactions in change from pre to post time and randomization (β = -1.523, SE = 0.424, Z = -3.597, p < .001, OR = 0.218, d = -0.840, $\Delta_{median,\ Therabot}$ = -2 pts, $\Delta_{median,\ Control}$ = 0 pts), and pre to follow-up time (β = -1.441, SE = 0.432, Z = -3.339, p < .001, OR = 0.237, d = -0.794, $\Delta_{median\ ,Therabot}$ = -3 pts, $\Delta_{median,\ Control}$ = -1 pts), indicating a large and differential response in the treatment group relative to the control group (Figure 3, Row 2).

Weight and Shape Concerns

There was a significant interaction between treatment and pre-post (β = -1.485, SE = 0.513, Z = -2.893, p = .004, OR = 0.227, d = -0.819, $\Delta_{median,\ Therabot}$ = -8 pts, median change in control group = -2 pts) and pre-follow-up time (β = -1.137, SE = 0.514, Z = -2.215, p = .027, OR = 0.321, d = -0.627, $\Delta_{median,\ Therabot}$ = -12 pts, $\Delta_{median,\ Control}$ = -5 pts), suggesting a large and differential response in the treatment group relative to a control group at post and follow-up (Figure 3, Row 3).

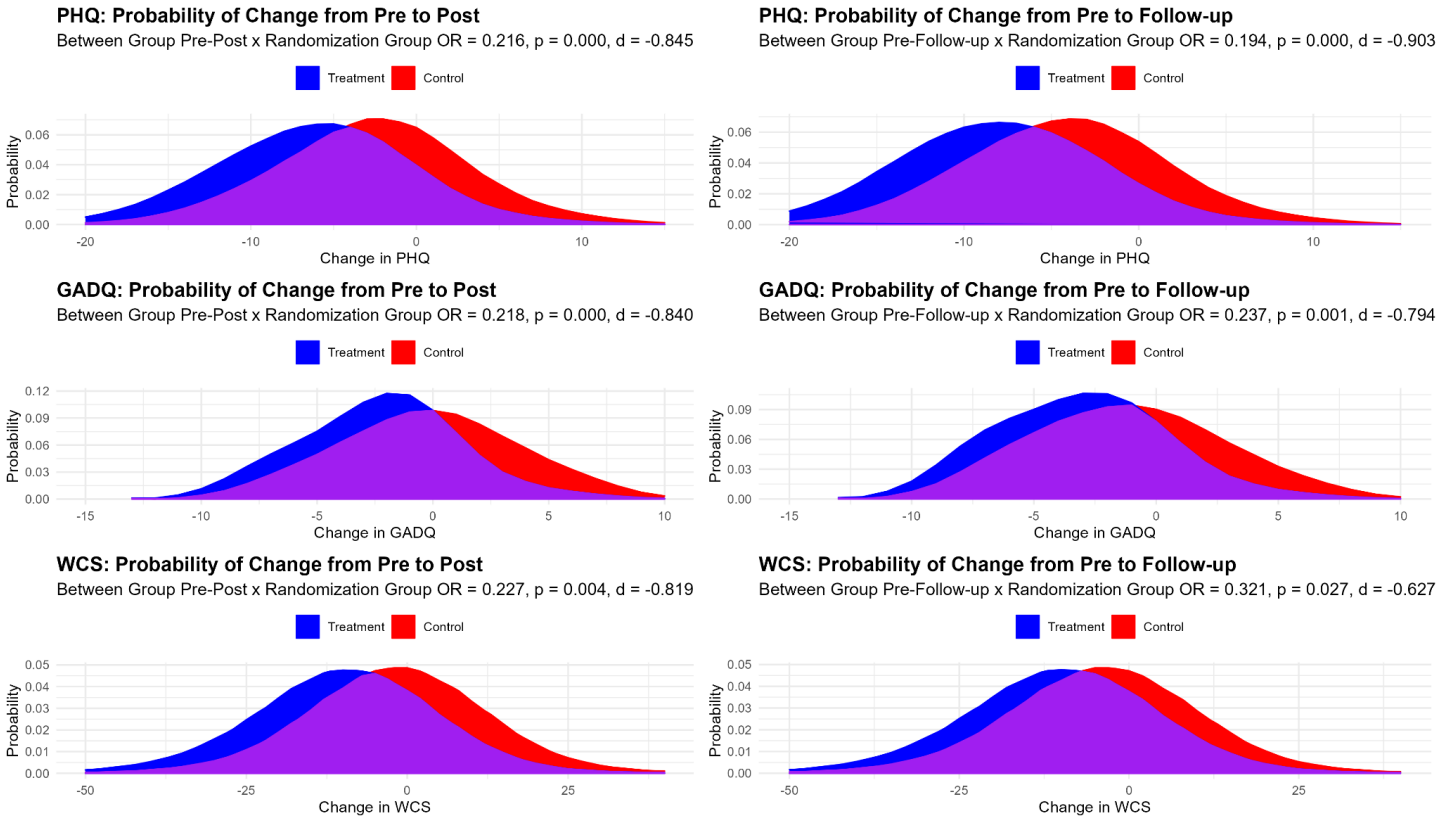

Figure 3. *Distributions representing smoothed probability of changes in clinical outcomes (depression, anxiety, weight/shape concerns, row-wise) at post (4 weeks, left column) and follow up (8 weeks, right column). Treatment group is visualized in blue and control in red.*

Secondary Outcomes

User Working Alliance

The total working alliance scores for Therabot are displayed in Figure 4. Compared to outpatient provider norms[30], results suggest strong working alliances with Therabot.

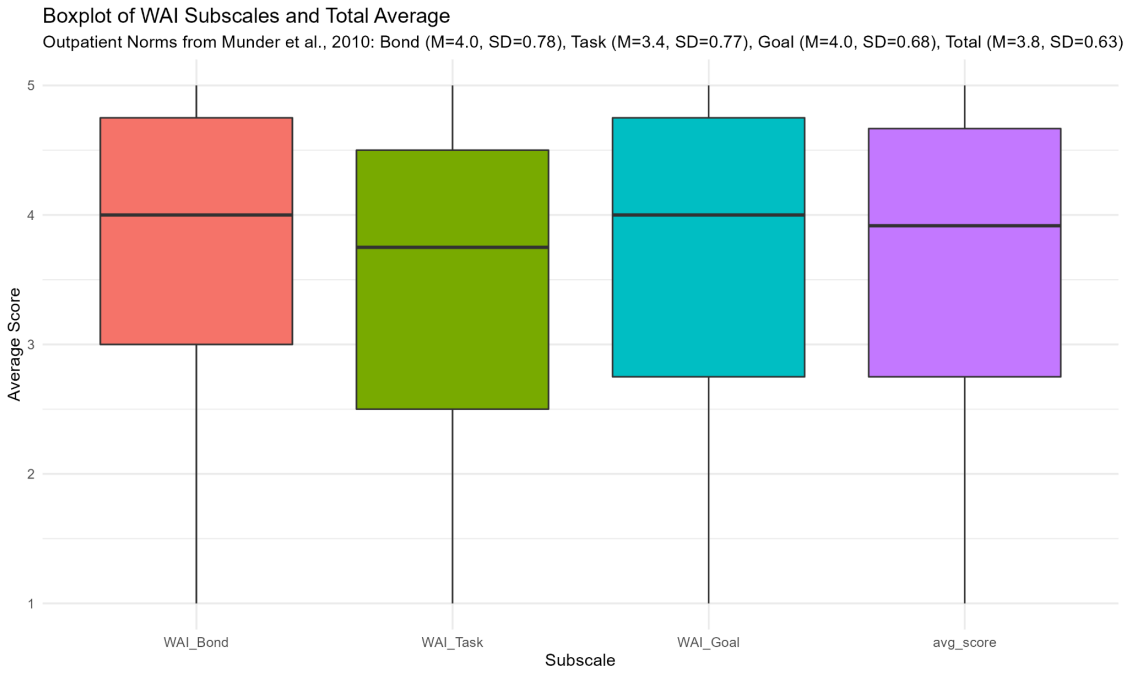

Figure 4. *Boxplot showing the aggregate subscale scores from the working alliance inventory including those interacting with the Therabot. The average of subscores is shown on the far right.*

User Satisfaction

User satisfaction was high, indicating that users found *Therabot* was both easy and intuitive to use, and liked the interface and overall design; users reported feeling better after interacting with *Therabot* and found sessions with *Therabot* helpful, reporting that *Therabot* behaved similarly to a real therapist. Overall, users were satisfied with *Therabot* and reported that they would use it on their own.

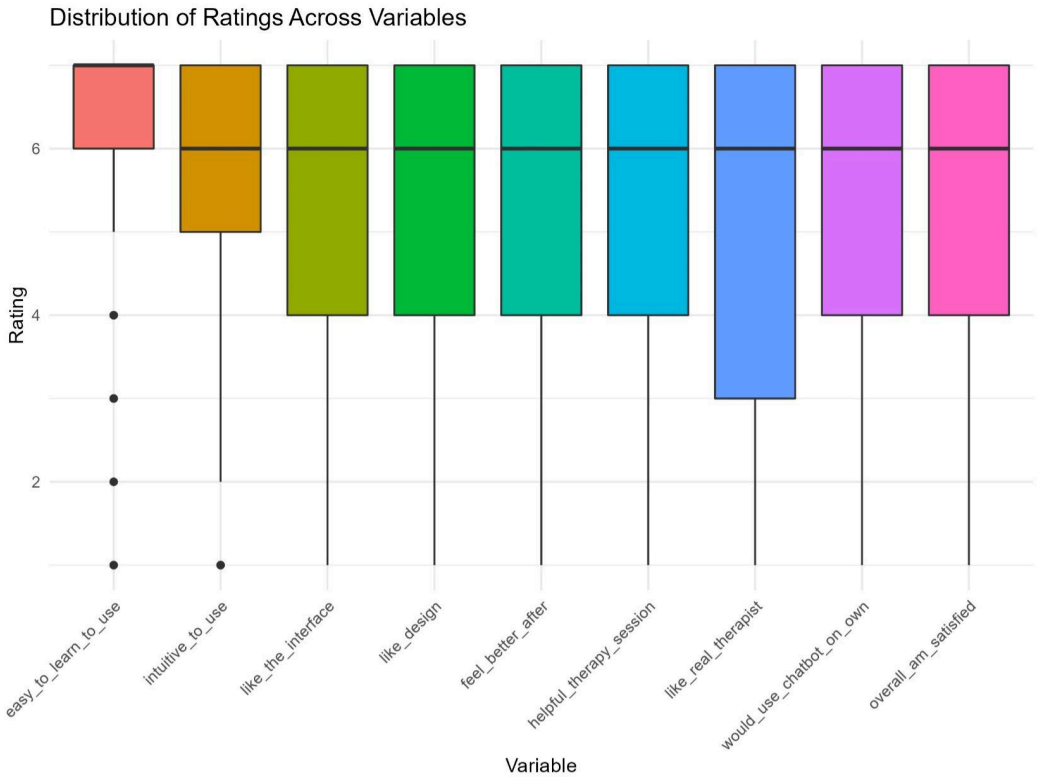

Figure 5. *Boxplots showing distribution of user ratings across variables.*

User Engagement

Of participants assigned to the *Therabot* group, 101 (95%) interacted with *Therabot*. The mean number of messages sent by participants was 260 (min = 1, max = 1,557), with the mean number of days interacting with *Therabot* being 24 days (min = 1, max = 60). The mean total amount of time participants interacted with *Therabot* was 6.18 hours across the course of the study. Participant engagement is depicted in Figure 6.

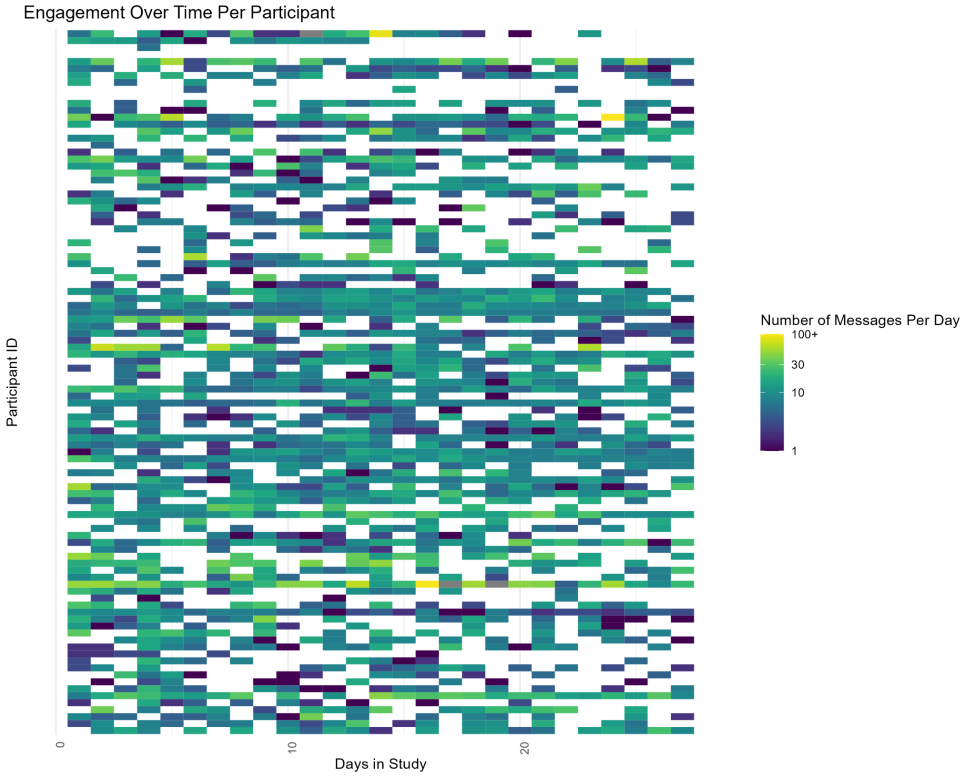

Figure 6. *Heat map representing user engagement across days in study (horizontal axis) by participant (vertical axis). Color represents the number of messages sent per day.*

Discussion

As the first RCT of its kind, our study supports the feasibility, acceptability, and effectiveness of a fine-tuned, fully GenAI-powered chatbot for treating mental health symptoms. Users demonstrated sustained engagement and rated their alliance with *Therabot* comparable to human therapists during the four-week trial. Critically, as compared to the WLC, *Therabot* users showed a greater reduction in depression, anxiety, and CHR-FED symptoms at post-intervention and at follow-up. We posit *Therabot's* success to be driven by three main factors. First, akin to effective rule-based conversational agents,[32] *Therabot* is rooted in evidence-based psychotherapies for anxiety[33], depression[34], and WSC[35]. Second, users had unrestricted access to *Therabot*, allowing for anytime-anywhere interactions. Notably, the ability to access therapeutic support when most needed, regardless of the time or location, may be one of the most significant advantages of DTx. Third, and unlike existing chatbots for mental health treatment, *Therabot* was powered by Gen-AI, allowing for natural, highly personalized, open-ended dialogue. Moreover, we argue that the Gen-AI approach promoted the therapeutic alliance, a critical nonspecific mediator of change in psychotherapy.[36] Although some evidence supports developing a therapeutic alliance with rule-based agents,[37] we see such a bond as inherently limited compared to that possible with Gen-AI-powered agents; Gen-AI provides greater capacity for personalized adaptation and more closely resembles human-human interaction. Our results suggest that, within four weeks, participants were able to develop a strong therapeutic relationship with *Therabot*, using it at consistently high rates.

Although existing companion Gen-AI chatbots can be highly engaging, they are not trained or evaluated for treatment of mental health disorders. Such chatbots may also be compromised by competing interests, such as user engagement or profit, which may be at odds with best practices for treating mental disorders.[19] Therefore, Gen-AI conversational agents tailored to integrate both evidence-based techniques and important nonspecific factors contributing

to psychotherapy outcomes represent a significant opportunity to provide scalable, on-demand, and effective mental health treatment. The nascency of Gen-AI and associated risks have likely contributed to the absence of a clinically-validated Gen-AI chatbot for mental mental health treatment. Indeed, the nondeterministic nature of Gen-AI models introduce the possibility of "hallucinations" and incorrect or potentially harmful content.[23] While human-delivered therapy is not immune to patient iatrogenesis,[38] such effects in Gen-AI models have the potential to impact more people and are less regulated.

We thus emphasize, first, the necessity of understanding Gen-AI's potential role and risks associated with mental health treatment and, second, the need for guardrails and close human supervision while testing such methods. All content was closely supervised for quality and safety in our trial, with rapid expert intervention available. This approach may continue to be necessary when testing similar future models to ensure safety. In addition, given the inscrutable "black box" nature of Gen-AI models, the inner processes are difficult or impossible to understand analytically. In this way, Gen-AI models are similar to human minds – intractable in complexity and predominantly studied by the data they produce – and may, thus, require extensive observation to obtain a reliable assessment.

Our results have important implications, forming the early foundational evidence for the use of fine-tuned Gen-AI powered chatbots in mental health treatment. *Therabot* shows promise as a means to scale evidence-based therapies in a way that maintains a high degree of personalization and engagement. Further, our approach enables novel translations of therapeutic techniques not possible in rule-based agents. Consider, for instance, detailed and personalized imaginal exposures prompted by Gen-AI agents. Interventions dependent upon the therapeutic alliance or specific patient-therapist interactions, may also benefit from the integration of Gen-AI chatbots.

Our study has notable strengths, including a nationally-recruited, demographically diverse, moderate sample size. Further, unlike many digital mental health studies,[39] *Therabot* ran on both Android and iOS devices, increasing generalizability. However, we also acknowledge several limitations. First, given our recruitment strategy, there was potential for selection bias towards younger, more technologically-minded participants who were open to AI. However, this may also resemble the most likely end users. Second, characteristic of WLC RCTs, there was potential for differential contact between the intervention and control group. We helped mitigate this by planning equivalent contact between groups whenever possible. Lastly, our follow-up period was limited to eight weeks; while this allowed for testing early effectiveness and safety, longer studies are needed to assess long-standing effectiveness.

Overall, results from the *Therabot* RCT – the first investigation of a Gen-AI conversational agent for mental health treatment – are highly promising. We found high engagement and acceptability of the intervention, as well as symptom decreases across disorders while maintaining a therapeutic alliance comparable to that of human therapists and their patients. Future work may extend the range of psychopathologies treated (e.g., obsessive-compulsive disorders), the settings in which Gen-AI chatbots are provided (e.g., emergency rooms), and the role of the Gen-AI chatbot (e.g., adjunctive to in-person psychotherapy). Our study provides key groundwork in the development of Gen-AI chatbots for mental health treatment.

Author Disclosures

MVH has worked as a paid consultant for Artisight and has received speaking fees related to his research. MVH and DM have worked as paid consultants for ChatNexus. NCJ has received a grant from Boehringer-Ingelheim. NCJ has edited a book through Academic Press and receives book royalties, and NCJ also receives speaking fees related to his research.

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Psychiatry* **9**, 137–150 (2022).
2. CLONINGER, C. R. The science of well-being: an integrated approach to mental health and its disorders. *World Psychiatry* **5**, 71–76 (2006).
3. National Center for & Health Workforce Analysis. Behavioral Health Workforce Brief (2023). (2023).
4. Health Resource and Service Administration. Workforce Projections. *Health Resource and Service Administration* https://data.hrsa.gov/topics/health-workforce/workforce-projections.
5. Kohn, R. *et al.* Mental health in the Americas: an overview of the treatment gap. *Rev. Panam. Salud Pública* **42**, (2018).
6. Monteleone, A. M. *et al.* Treatment of eating disorders: A systematic meta-review of meta-analyses and network meta-analyses. *Neurosci. Biobehav. Rev.* **142**, 104857 (2022).
7. Cuijpers, P. *et al.* Psychotherapy for Depression Across Different Age Groups: A Systematic Review and Meta-analysis. *JAMA Psychiatry* **77**, 694 (2020).
8. Van Dis, E. A. M. *et al.* Long-term Outcomes of Cognitive Behavioral Therapy for Anxiety-Related Disorders: A Systematic Review and Meta-analysis. *JAMA Psychiatry* **77**, 265 (2020).
9. Coombs, N. C., Meriwether, W. E., Caringi, J. & Newcomer, S. R. Barriers to healthcare access among U.S. adults with mental health challenges: A population-based study. *SSM - Popul. Health* **15**, 100847 (2021).
10. Fürstenau, D., Gersch, M. & Schreiter, S. Digital Therapeutics (DTx). *Bus. Inf. Syst. Eng.* **65**, 349–360 (2023).
11. Wang, C., Lee, C. & Shin, H. Digital therapeutics from bench to bedside. *Npj Digit. Med.* **6**, 38 (2023).
12. Nwosu, A., Boardman, S., Husain, M. M. & Doraiswamy, P. M. Digital therapeutics for mental health: Is attrition the Achilles heel? *Front. Psychiatry* **13**, (2022).
13. Huibers, M. & Cuijpers, P. Common (Nonspecific) Factors in Psychotherapy. in 1–6 (2015). doi:10.1002/9781118625392.wbecp272.
14. Henson, P., Peck, P. & Torous, J. Considering the Therapeutic Alliance in Digital Mental Health Interventions. *Harv. Rev. Psychiatry* **27**, 268–273 (2019).
15. Tong, F., Lederman, R., D'Alfonso, S., Berry, K. & Bucci, S. Digital Therapeutic Alliance With Fully Automated Mental Health Smartphone Apps: A Narrative Review. *Front. Psychiatry* **13**, 819623 (2022).
16. Xu, Y. *et al.* Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* **2**, 100179 (2021).
17. Breuer, J. & Freud, S. *Studies on Hysteria*. xxxi, 335 (Basic Books, Oxford, England, 1957).
18. Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966).
19. Haque, M. D. R. & Rubya, S. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. *JMIR MHealth UHealth* **11**, e44838 (2023).
20. Fitzpatrick, K. K., Darcy, A. & Vierhile, M. 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 Ment. Health* **4**, e19 (2017).
21. Darcy, A., Daniels, J., Salinger, D., Wicks, P. & Robinson, A. Evidence of Human-Level Bonds Established With a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study. *JMIR Form. Res.* **5**, e27868 (2021).
22. Gill, S. S. & Kaur, R. ChatGPT: Vision and challenges. *Internet Things Cyber-Phys. Syst.* **3**, 262–271 (2023).
23. De Freitas, J. & Cohen, I. G. The health risks of generative AI-based wellness apps. *Nat. Med.* **30**, 1269–1275 (2024).
24. Pentina, I., Hancock, T. & Xie, T. Exploring relationship development with social chatbots: A mixed-method study of replika. *Comput. Hum. Behav.* **140**, 107600 (2023).
25. Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
26. Harris, P. A. *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).
27. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613 (2001).
28. Newman, M. G. *et al.* Preliminary reliability and validity of the generalized anxiety disorder questionnaire-IV: A revised self-report diagnostic measure of generalized anxiety disorder. *Behav. Ther.* **33**, 215–233 (2002).
29. Graham, A. K. *et al.* A screening tool for detecting eating disorder risk and diagnostic symptoms among college-age women. *J. Am. Coll. Health* **67**, 357–366 (2019).
30. Munder, T., Wilmers, F., Leonhart, R., Linster, H. W. & Barth, J. Working Alliance Inventory‐Short Revised

(WAI‐SR): psychometric properties in outpatients and inpatients. *Clin. Psychol. Psychother.* **17**, 231–239 (2010).
31. Sánchez-Meca, J., Marín-Martínez, F. & Chacón-Moscoso, S. Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis. *Psychol. Methods* **8**, 448–467 (2003).
32. Fitzpatrick, K. K., Darcy, A. & Vierhile, M. 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 Ment. Health* **4**, e19 (2017).
33. Covin, R., Ouimet, A. J., Seeds, P. M. & Dozois, D. J. A. A meta-analysis of CBT for pathological worry among clients with GAD. *J. Anxiety Disord.* **22**, 108–116 (2008).
34. Lepping, P. *et al.* Clinical relevance of findings in trials of CBT for depression. *Eur. Psychiatry* **45**, 207–211 (2017).
35. Jarry, J. L. & Ip, K. The effectiveness of stand-alone cognitive-behavioural therapy for body image: A meta-analysis. *Body Image* **2**, 317–331 (2005).
36. Baier, A. L., Kline, A. C. & Feeny, N. C. Therapeutic alliance as a mediator of change: A systematic review and evaluation of research. *Clin. Psychol. Rev.* **82**, 101921 (2020).
37. Darcy, A., Daniels, J., Salinger, D., Wicks, P. & Robinson, A. Evidence of Human-Level Bonds Established With a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study. *JMIR Form. Res.* **5**, e27868 (2021).
38. Boisvert, C. M. & Faust, D. Iatrogenic Symptoms in Psychotherapy. *Am. J. Psychother.* **56**, 244–259 (2002).
39. Bryan, A. C. *et al.* Behind the Screen: A Narrative Review on the Translational Capacity of Passive Sensing for Mental Health Assessment. *Biomed. Mater. Devices* (2024) doi:10.1007/s44174-023-00150-4.
40. Bjureberg, J. *et al.* Columbia-Suicide Severity Rating Scale Screen Version: initial screening for suicide risk in a psychiatric emergency department. *Psychol. Med.* **52**, 1–9 (2021).
41. Hirschfeld, R. M. A. The Mood Disorder Questionnaire: A Simple, Patient-Rated Screening Instrument for Bipolar Disorder. *Prim. Care Companion CNS Disord.* **4**, (2002).
42. Degenhardt, L., Hall, W., Korten, A. & Jablensky, A. Use of a brief screening instrument for psychosis: Results of an ROC analysis.
43. Kocalevent, R.-D., Hinz, A. & Brähler, E. Standardization of the depression screener Patient Health Questionnaire (PHQ-9) in the general population. *Gen. Hosp. Psychiatry* **35**, 551–555 (2013).

Supplemental Materials 1

Screening Measures

Suicidality, Bipolar Disorder, and Psychosis

The Columbia Suicide Severity Rating Scale - Screen Version[40] (CSSRS-SV), Mood Disorder Questionnaire[41] (MDQ) and Brief Psychosis Screen[42](BPS) were used to assess symptoms of suicidality, mania, and psychosis, respectively. Individuals were excluded from participation in the study if they endorsed items placing them at moderate risk or greater for suicide according to the CSSRS-SV, a history of mania with associated impairment or a bipolar disorder diagnosis based upon the MDQ, or more than two psychotic symptom items on the BPS.

Depression

Depression symptoms were assessed using the Patient Health Questionnaire-9[27] (PHQ-9) at the baseline, post-intervention, and follow-up time points. The PHQ-9 is a widely used, 9-item self-report measure of depressive symptoms over the past two weeks, with items ranging from "0" (not at all) to "3" (nearly every day). The established cut point of 10[43] was used to characterize a participant as qualifying for enrollment into the depressed treatment group. The PHQ-9 demonstrated excellent reliability in the current study ($\alpha_{baseline} = 0.858$; $\alpha_{post\text{-}intervention} = 0.897$; $\alpha_{follow\text{-}up} = 0.890$).

Anxiety

The Generalized Anxiety Disorder Questionnaire for DSM-IV[28] (GAD-Q-IV) was used to assess symptoms of anxiety at the baseline, post-intervention, and follow-up time points. The GAD-Q-IV is a 9-item, self-reported diagnostic measure of generalized anxiety that assesses each of the diagnostic criteria of GAD. The weighted scoring system suggested by Newman and colleagues (2002) was used to create an anxiety severity score ranging from 0 to 13, with higher scores reflecting greater anxiety. As suggested by the authors, scores of 5.7 or greater were used as the threshold to determine eligibility for enrollment into the anxiety treatment group. The reliability of the GAD-Q-IV was good in the current study ($\alpha_{baseline} = 0.932$; $\alpha_{post\text{-}intervention} = 0.908$; $\alpha_{follow\text{-}up} = 0.917$).

Weight and Shape Concerns

Weight and shape concerns (WSC) were assessed using the 18-item Stanford-Washington University Eating Disorder[29] (SWED) at the baseline, post-intervention, and follow-up time points. Responses are used to sort participants into high vs. low eating disorder risk groups, while also providing continuous measures of WSC, frequencies of eating disorder behaviors in the past 3 months, and eating disorder-related impairment rated on a scale ranging from 1 (never) to 5 (always). Eating disorder risk classification was used to determine eligibility for enrollment into the eating disorder risk treatment group. The SWED demonstrated good internal consistency in the present study ($\alpha_{baseline} = 0.782$; $\alpha_{post\text{-}intervention} = 0.786$; $\alpha_{follow\text{-}up} = 0.794$).

Secondary Measures

Therapeutic Alliance

The Working Alliance Inventory-Short Revised[30] (WAI-SR) was used to measure the therapeutic alliance between Therabot and the participant at the post-intervention and follow-up assessments. The WAI-SR consists of 12 self-report items ranging from 1 (seldom) to 5 (always) that measure agreement on the tasks of therapy, agreement on the goals of therapy, and development of an affective bond. Internal consistency was excellent for the WAI-SR ($\alpha_{post\text{-}intervention} = 0.978$).

Satisfaction with Therabot

Satisfaction with Therabot was assessed at the post-intervention and follow-up assessment using an 11-item measure developed for the current study. Nine of the items were assessed on a 7-point likert-type scale ranging from strongly

disagree (1) to strongly agree (7). Two additional qualitative, open-ended items were also included to obtain feedback on the positive and negative aspects of participants' experience with Therabot.