











ORIGINAL ARTICLE

# Randomized Trial of a Generative AI Chatbot for Mental Health Treatment

Michael V. Heinz , M.D.,<sup>1,2</sup> Daniel M. Mackin , Ph.D.,<sup>1,2</sup> Brianna M. Trudeau , B.A.,<sup>1</sup> Sukanya Bhattacharya , B.A.,<sup>1</sup> Yinzhou Wang , M.S.,<sup>1</sup> Haley A. Banta ,<sup>1</sup> Abi D. Jewett , B.A.,<sup>1</sup> Abigail J. Salzhauer , B.A.,<sup>1</sup> Tess Z. Griffin , Ph.D.,<sup>1</sup> and Nicholas C. Jacobson , Ph.D.<sup>1,2,3,4</sup>

Received: August 11, 2024; Revised: November 18, 2024; Accepted: February 2, 2025; Published March 27, 2025

## Abstract

**BACKGROUND** Generative artificial intelligence (Gen-AI) chatbots hold promise for building highly personalized, effective mental health treatments at scale, while also addressing user engagement and retention issues common among digital therapeutics. We present a randomized controlled trial (RCT) testing an expert-fine-tuned Gen-AI-powered chatbot, Therabot, for mental health treatment.

**METHODS** We conducted a national, randomized controlled trial of adults (N=210) with clinically significant symptoms of major depressive disorder (MDD), generalized anxiety disorder (GAD), or at clinically high risk for feeding and eating disorders (CHR-FED). Participants were randomly assigned to a 4-week Therabot intervention (N=106) or waitlist control (WLC; N=104). WLC participants received no app access during the study period but gained access after its conclusion (8 weeks). Participants were stratified into one of three groups based on mental health screening results: those with clinically significant symptoms of MDD, GAD, or CHR-FED. Primary outcomes were symptom changes from baseline to postintervention (4 weeks) and to follow-up (8 weeks). Secondary outcomes included user engagement, acceptability, and therapeutic alliance (i.e., the collaborative patient and therapist relationship). Cumulative-link mixed models examined differential changes. Cohen's d effect sizes were unbounded and calculated based on the log-odds ratio, representing differential change between groups.

**RESULTS** Therabot users showed significantly greater reductions in symptoms of MDD (mean changes: -6.13 [standard deviation {SD}=6.12] vs. -2.63 [6.03] at 4 weeks; -7.93 [5.97] vs. -4.22 [5.94] at 8 weeks; d=0.845-0.903), GAD (mean changes: -2.32 [3.55] vs. -0.13 [4.00] at 4 weeks; -3.18 [3.59] vs. -1.11 [4.00] at 8 weeks; d=0.794-0.840), and CHR-FED (mean changes: -9.83 [14.37] vs. -1.66 [14.29] at 4 weeks; -10.23 [14.70] vs. -3.70 [14.65] at 8 weeks; d=0.627-0.819) relative to controls at postintervention and follow-up. Therabot was well utilized (average use >6 hours), and participants rated the therapeutic alliance as comparable to that of human therapists.

**CONCLUSIONS** This is the first RCT demonstrating the effectiveness of a fully Gen-AI therapy chatbot for treating clinical-level mental health symptoms. The results were promising for MDD, GAD, and CHR-FED symptoms. Therabot was well utilized and received high user ratings. Fine-tuned Gen-AI chatbots offer a feasible approach to delivering personalized mental

*The author affiliations are listed at the end of the article.*

*Dr. Heinz can be contacted at [michael.v.heinz@dartmouth.edu](mailto:michael.v.heinz@dartmouth.edu) or at the Center for Technology and Behavioral Health, Dartmouth College, 46 Centerra Pkwy, Lebanon, NH 03766.*

health interventions at scale, although further research with larger clinical samples is needed to confirm their effectiveness and generalizability. (Funded by Dartmouth College; ClinicalTrials.gov number, [NCT06013137](#).)

## Introduction

**T**he prevalence and burden of mental health disorders have increased significantly over the past three decades.<sup>1</sup> Despite the adverse impact of these disorders,<sup>2</sup> mental health infrastructure is inadequately resourced to meet the current and growing demand for care.<sup>3-5</sup> Although empirically validated psychosocial treatments exist,<sup>6-8</sup> they are resource intensive, and limited in scalability and accessibility, leading to fewer than half of the people with a mental health disorder receiving care.<sup>5,9</sup> Digital therapeutics (DTx) — automated, evidence-based software for the treatment or diagnosis of medical conditions<sup>10</sup> — offer a solution to bridge this gap.

While DTx aim to improve the accessibility and scalability of evidence-based mental health interventions,<sup>11</sup> these approaches have been plagued by attrition and low rates of engagement.<sup>12</sup> Within established psychotherapies, there is evidence for the benefit of nonspecific factors, such as therapeutic alliance (i.e., the collaborative relationship between patient and therapist), empathy, and shared goals.<sup>13</sup> The relative lack of personalization and alliance in DTx, compared with human-delivered psychosocial interventions, is likely to contribute to reduced engagement.<sup>14</sup> These nonspecific factors have been difficult to emulate via automated technologies and may be fundamentally different, or unachievable, in automated software.<sup>15</sup>

Artificial intelligence (AI) represents a promising direction for improving DTx. Chatbots hold particular promise given their capacity to imitate human conversation and dialogue, long known to be integral parts of psychotherapeutic treatments — that is, the talking cure.<sup>16</sup> In fact, mental health and wellness chatbots are not a new phenomenon, with ELIZA,<sup>17</sup> an early rule-based chatbot, used to emulate a Rogerian therapist. The study of chatbots for mental health remains nascent, with evaluation limited to exclusively rule-based conversational agents to date. Although such chatbots (e.g., Woebot) have shown benefits in clinical trials,<sup>18</sup> and in some cases a capacity to promote a therapeutic alliance,<sup>19</sup> they are inherently limited by their reliance on an explicitly programmed decision trees and restricted inputs.

Recent advances in computing and machine learning now allow for sophisticated systems capable of learning, adapting, and understanding context in natural language, removing the necessity for explicit programming. Pushing these bounds even further in the language domain, the advent of generative AI (Gen-AI), recently popularized by ChatGPT,<sup>20</sup> has enabled the 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,<sup>21</sup> rather than software intended for the diagnosis and treatment of mental health disorders. Although some Gen-AI-powered chatbots have shown both broad appeal and the capacity to form human-like bonds,<sup>22</sup> they are not intended and have not been evaluated for mental health treatment. The non-deterministic and open-ended nature of Gen-AI, enabling the possibility of harmful responses, has paused 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 potential benefit from the use of therapeutic Gen-AI-powered chatbots. Paired with existing frameworks for DTx, Gen-AI chatbots have unprecedented potential to address existing problems with engagement while powering the development of new and personalized interventions. Although the literature supports the effectiveness of cognitive behavioral therapy (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 health care, to our knowledge, no prior RCTs have investigated the effectiveness and safety of a Gen-AI chatbot for the treatment of mental health symptoms. Beginning in 2019, we started developing Therabot, a Gen-AI chatbot trained using expertly written therapist-patient dialogues based on third-wave CBT,<sup>23</sup> integrating empirically grounded contextual and functional approaches to mental health problems. Developed with over 100,000 human hours comprising software development, training dialogue creation, and refinement, Therabot is designed to augment and enhance conventional mental health treatment services by delivering personalized, evidenced-based mental health interventions at scale.

In this RCT, we examined Therabot's effectiveness for the treatment of major depressive disorder (MDD) symptoms, generalized anxiety disorder (GAD) symptoms, and clinically high-risk feeding and eating disorder (CHR-FED) symptoms in a large, nationally representative sample of

participants. We hypothesized that participants assigned to a 4-week intervention with Therabot would measurably improve in mental health symptoms across all clinical domains relative to patients assigned to the waitlist control (WLC) condition at both postintervention (4 weeks) and follow-up (8 weeks). Furthermore, we hypothesized 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 randomly assigned, WLC trial with a 1:1:1 allocation ratio across the MDD, GAD, and CHR-FED symptom groups. A Meta Ads campaign was used to recruit adults across the United States. Given its broad and diverse reach, samples gathered via Meta Ads have been shown to represent varied age, education, and gender groups.<sup>24</sup> Based on self-reported responses to a baseline questionnaire, with instruments detailed in the Supplementary Appendix, participants screening positive for MDD, GAD, or CHR-FED symptoms were stratified accordingly into one of the three groups and then randomly assigned to either the control or intervention group. We assumed participant identity to be truthful unless we detected irregularities in identity data. We automatically blocked identical email addresses, phone numbers, and IP addresses to prevent duplicates during sign-up. Furthermore, we used a custom-designed two-factor authentication system built into REDCap to ensure that participants indeed owned the phone number that they provided. We included a manual review in cases of uncertainty. Comorbidity was allowed, and outcomes were analyzed based on participants scoring within the respective groups at baseline, regardless of their primary presenting problem.

Participants randomly assigned to the intervention group were prompted daily to interact with Therabot during the treatment phase (first 4 weeks). During the subsequent postintervention follow-up phase (weeks 4–8), participants were not prompted but were permitted access to Therabot. Both groups, intervention, and WLC, received assessments at baseline, postintervention (4 weeks), and follow-up (8 weeks). Therabot access was disabled in the treatment group after 8 weeks; owing to a 72-hour grace period and variations across time zones, the actual access duration was potentially up to 60 days. After completing their final assessment, the WLC group was also provided full access

to Therabot. Study data were collected and managed using REDCap electronic data capture tools hosted at Dartmouth College.<sup>25,26</sup>

### MODEL TRAINING AND INFERENCE

We utilized transformer-based, decoder-only architectures. The system employed both Falcon-7B and LLaMA-2-70B models in tandem. The models were trained and deployed on AWS SageMaker. During training, we applied quantized low-rank adaptation (QLoRA) for efficient fine-tuning to optimize the models' generation of appropriate responses based on conversation history. For inference, the conversation history was used to prompt the fine-tuned models via SageMaker end points.

### PARTICIPANTS

The participants were required to be at least 18 years of age and to screen positive for clinical-level symptoms of at least one of the following: MDD, GAD, or CHR-FED. 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 the Supplementary Appendix).

### INTERVENTION

Therabot is a text-based multithread chat application for iOS and Android that can interact with participants regarding their mental health problems in natural language. The intervention utilizes a generative large language model (LLM) fine-tuned on expert-curated mental health dialogues. The dialogues were developed by our research team, including a board-certified psychiatrist and a clinical psychologist, and peer-reviewed using evidence-based (primarily CBT) modalities. Given the potential risks associated with Gen-AI, we added multiple guard rails, including a crisis classification model.

For the first 4 weeks, participants in the intervention group were prompted daily to engage with Therabot. During the subsequent follow-up phase (weeks 4–8), participants engaged with Therabot as frequently as they desired. 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 the most recent user message were combined and sent to the LLM. Common output included empathetic responses, validation, targeted interventions, or questions prompting

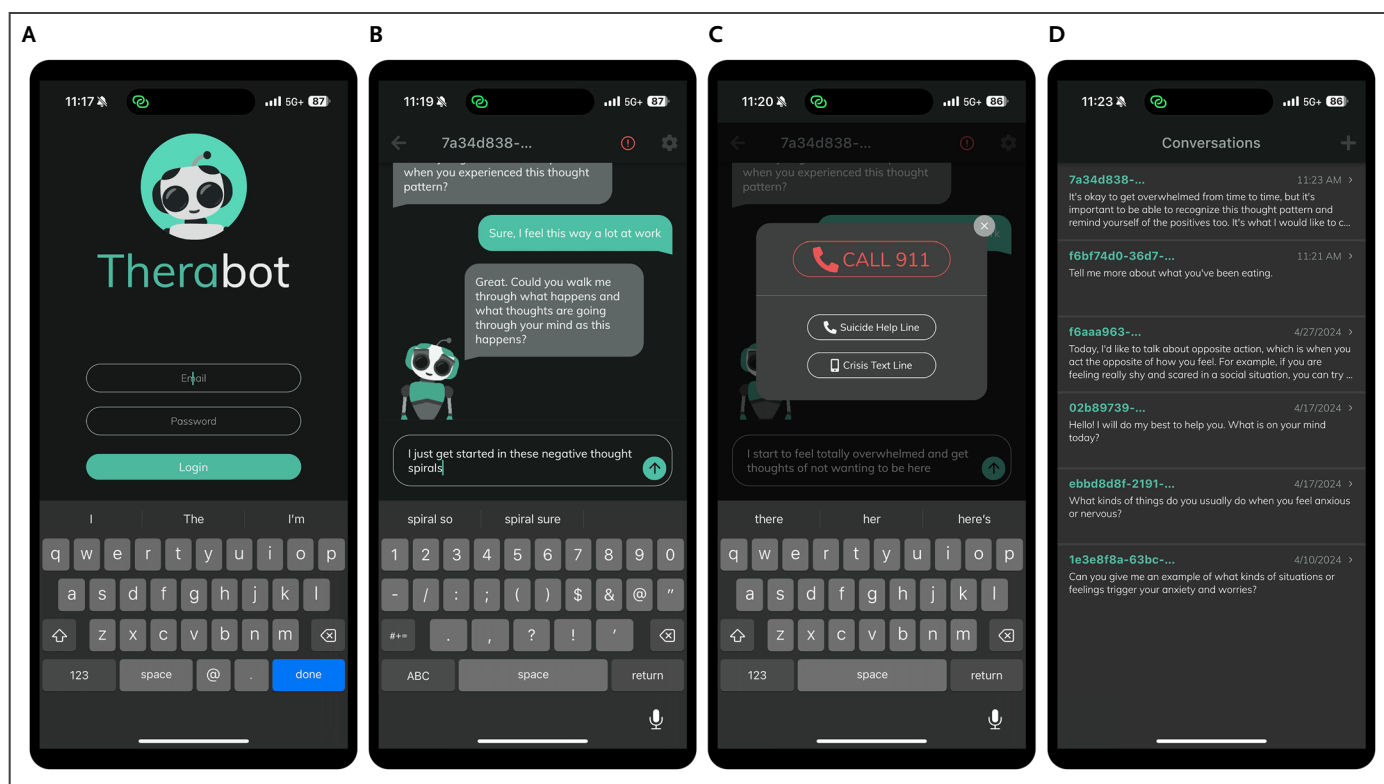


Figure 1. Key Design Features of the Therabot Application.

Panel A shows the Therabot login screen. Panel B shows the main chat interface. Panel C shows the emergency module deployed in response to model detection of high-risk content (e.g., suicidal ideation). Panel D shows the conversation thread interface for users to initiate a thread or return to a prior thread.

elaboration (see [Fig. 1](#)). All responses from Therabot were supervised by trained clinicians and researchers post-transmission. In the event of an inappropriate response from Therabot (e.g., providing medical advice), we contacted the participant to provide correction. In the event of a participant raising safety concerns (e.g., suicidal ideation), we contacted the participant to provide safety guidance and emergency resources.

## MEASURES

Primary outcome measures were administered to both groups at baseline, postintervention (4 weeks), and follow-up (8 weeks) and included the Patient Health Questionnaire 9<sup>27</sup> (PHQ-9), the Generalized Anxiety Disorder Questionnaire for the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (DSM-IV)<sup>28</sup> (GAD-Q-IV), and the Weight Concerns Scale (WCS) within the Stanford-Washington University Eating Disorder<sup>29</sup> (SWED), as measures of depression, anxiety, and weight concerns, respectively. The SWED was used in full as a

baseline risk stratification tool to ensure we represented participants with CHR-FED.

The PHQ-9 sum score ranges from 0 to 27, with a decrease of 5 or more shown to constitute a clinically meaningful change<sup>30</sup> and a score of 10 or more often used as a clinical threshold for MDD screening. The GAD-Q-IV, designed to assess GAD based on DSM-IV and DSM-5 criteria, can be scored via binary criterion method or a total sum score ranging from 0 to 13. The WCS score ranges from 0 to 100 and constitutes a mean of five Likert-style items, normalized from 0 to 100.<sup>29</sup> Although the PHQ-9 has established clinically meaningful change thresholds, the GAD-Q-IV and WCS have not; therefore, we use Cohen's d values for primary effect sizes, given their widespread use in evaluating psychiatric trials.

Secondary outcomes included therapeutic alliance (Working Alliance Inventory — Short Revised<sup>31</sup> [WAI-SR]), engagement with Therabot (number of messages sent), and satisfaction with Therabot (self-developed survey). Each domain of the WAI-SR (goal, task, and bond) is scored as the



mean of four Likert-style items, resulting in a score range of 1–5. Additional details on the primary and secondary outcomes are presented in the Supplementary Appendix.

### SAMPLE SIZE JUSTIFICATION

Based on the comorbidity between diagnoses, we expected to have 100 participants 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 randomly assigned group determined power, with the proportion of times the interaction terms were significant representing power. The results suggested that we had greater than 90% power to detect a differential response of 0.3 and 0.5

### RANDOM ASSIGNMENT

Random assignment was performed using a computer-generated random sequence with a fixed block size of six. The random assignment 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 weight concerns among participants at a clinical level of MDD, GAD, and CHR-FED at baseline. The participants were analyzed within the groups to which they were randomly assigned. Owing to the trivial percentage of missing outcome data, we used complete case analysis. To ensure complete case analysis provided an unbiased estimate, we also analyzed the data using multiple imputation for missing data handling. The results were nearly identical, with no changes to the interpretation, and are displayed in the Supplementary Appendix.

To address the ordinal, nonequidistant nature in the response categories of our symptom measures and eliminate potential distortion of effect size estimates and inflated error rates, we used CLMMs to analyze the effects of time and random assignment 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 to be symmetric around the latent mean of zero. Formally, the model is described as follows:

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

where  $Y$  represents the ordinal outcome (MDD, GAD, or WCS score),  $j$  denotes the threshold category,  $\alpha_j$  are the threshold parameters, and  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  are the fixed-effect coefficients for time, group, and their interaction, respectively. The coefficients of primary interest are the  $\beta_3$  terms representing the differential change between groups from baseline at each follow-up time point. Here,  $u_{\text{ID}}$  signifies the random intercepts for the participants. Odds ratios were calculated from these estimates, and effect sizes were calculated using  $d = \log(\text{OR}) \times \sqrt{3} / \pi$ .<sup>32</sup> All Cohen's  $d$  effect sizes were calculated between groups representing differential change. The effect sizes are unbounded. All treatment effects reported in our analyses are marginal effects derived from CLMMs. These models provide predicted probabilities for each possible outcome score under each condition (treatment/control) at each time point. To obtain interpretable statistics, we drew 10,000 samples from these model-predicted probability distributions, which provided stable estimates of means and standard deviations rather than using raw scores.<sup>33,34</sup> For example, with depression symptoms (PHQ-9), the model predicts the probability of each possible score (0–27) for both groups at baseline, postintervention and follow-up. The change scores were calculated by comparing these sampled distributions across time points. This approach allows us to honor the ordinal nature of our measures while also providing clinically meaningful statistics, with effect sizes computed directly from the model's log-odds ratios and supporting statistics (means, standard deviations [SDs]) derived from the sampling procedure. To control for multiple comparisons across our six primary end points (three outcomes measured at two time points), we implemented the Holm–Bonferroni sequential procedure for familywise error rate control. Adjusted  $P$  values are reported for all primary analyses (see [Table 1](#)).

Descriptive results for working alliance and user satisfaction were visualized using box plots. In each plot, the box represents the interquartile range (IQR), with the median presented as a horizontal line. Whiskers extend to the smallest and largest values within  $1.5 \times \text{IQR}$  from the hinges. Data beyond the whiskers are plotted as individual outliers.

Outcome and Time Point	Intervention			Control			Between-Group Differences in Change		
	M (SD)	ΔM (SD)	N	M (SD)	N	ΔM (SD)	β (SE)	d (95% CI)	Adjusted P Value†
MDD symptoms (as measured by PHQ-9)									
Baseline	15.63 (4.33)		73	15.91 (4.45)	69				
Postintervention (4 weeks)	9.50 (7.50)	−6.13 (6.12)	70	13.28 (7.50)	69	−2.63 (6.03)	−1.533 (0.404)	0.845 (0.409 to 1.282)	<0.001
Follow-up (8 weeks)	7.70 (7.38)	−7.93 (5.97)	70	11.69 (7.43)	69	−4.22 (5.94)	−1.639 (0.410)	0.903 (0.460 to 1.347)	<0.001
GAD symptoms (as measured by GAD-Q-IV)									
Baseline	10.43 (1.14)		60	10.42 (1.13)	56				
Postintervention (4 weeks)	8.11 (3.73)	−2.32 (3.55)	53	10.30 (4.16)	56	−0.13 (4.00)	−1.523 (0.424)	0.840 (0.382 to 1.298)	0.001
Follow-up (8 weeks)	7.24 (3.77)	−3.18 (3.59)	52	9.31 (4.16)	56	−1.11 (4.00)	−1.441 (0.432)	0.794 (0.328 to 1.261)	0.003
Weight concerns (as measured by WCS)									
Baseline	54.13 (11.38)		42	54.65 (16.41)	47				
Postintervention (4 weeks)	44.30 (18.33)	−9.83 (14.37)	42	52.99 (21.76)	47	−1.66 (14.29)	−1.485 (0.513)	0.819 (0.264 to 1.373)	0.008
Follow-up (8 weeks)	43.90 (18.60)	−10.23 (14.70)	42	50.95 (22.00)	47	−3.70 (14.65)	−1.137 (0.514)	0.627 (0.072 to 1.182)	0.027

\* Marginal effects were estimated using cumulative-link mixed models based on model-derived probabilities. Sample sizes analyzed at each time point are provided in the column immediately to the right of the mean. CI denotes confidence interval; GAD-Q-IV, Generalized Anxiety Disorder Questionnaire for the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition; MDD, major depressive disorder; PHQ-9, Patient Health Questionnaire 9; SD standard deviation; and WCS, Weight Concerns Scale.

† P values were adjusted using the Holm–Bonferroni method for multiple comparisons across six primary end points.

## COMPENSATION

Participants were compensated \$25 (U.S. dollars) for each of the three assessments completed.

## Results

### PARTICIPANTS

Participants included 210 adults, randomly assigned to intervention and WLC. By the 4-week assessment, four participants had withdrawn from the study (treatment group), and three were lost to follow-up (treatment group); six of these participants had opted to discontinue the Therabot intervention. By the 8-week assessment, one participant was lost to follow-up (treatment group). Detailed aggregate demographics are displayed in [Table 2](#). [Figure 2](#) displays the flow of participants from screening to analysis, including counts of participants meeting clinical screening thresholds for each group. The study spanned March through May 2024, with recruitment from March 15–31, 2024. Recruitment ended when the enrollment target was reached.

## PRIMARY OUTCOMES

### Major Depressive Disorder Symptoms

Participants receiving the Therabot intervention showed large and greater improvement in depression symptoms compared with the control participants across both time points (see [Table 1](#)), with both comparisons remaining significant after adjusting for multiple comparisons. The mean change (SD) in PHQ-9 score from baseline to postintervention was −6.13 (6.12) in the intervention group and −2.63 (6.03) in the control group. The change from baseline to follow-up was −7.93 (5.97) in the intervention group and −4.22 (5.94) in the control group ([Fig. 3](#), Row 1).

### Generalized Anxiety Disorder Symptoms

Similar patterns of improvement were observed for anxiety symptoms ([Table 1](#)), where participants receiving the Therabot intervention showed a large and differential response in anxiety symptoms across both postintervention and follow-up. The mean change (SD) in GAD-Q-IV score from baseline to postintervention was −2.32 (3.55) in the

Characteristic	Waitlist Control (n=104)	Intervention (n=106)	Overall (N=210)
Age (years) — mean (SD)	33.63 (10.56)	34.09 (11.41)	33.86 (10.97)
Gender — n (%)			
Male	37 (35.58)	41 (38.68)	78 (37.14)
Female	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)
No	99 (95.19)	103 (97.17)	202 (96.19)
Sexual orientation — n (%)			
Heterosexual	82 (78.85)	84 (79.25)	166 (79.05)
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 or Ethnicity — n (%)*			
Non-Hispanic White	56 (53.85)	56 (52.83)	112 (53.33)
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)
Some college	14 (13.46)	19 (17.92)	33 (15.71)
Associate's 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)
Yes — part time	12 (11.54)	11 (10.38)	23 (10.95)
Yes — full time	14 (13.46)	18 (16.98)	32 (15.24)
Current treatment, n (%)			
Medication	23 (22.12)	24 (22.64)	47 (22.38)
Psychotherapy	16 (15.38)	13 (12.26)	29 (13.81)
Medication and psychotherapy	10 (9.62)	7 (6.60)	17 (8.10)
No treatment	75 (72.12)	79 (74.53)	154 (73.33)

\*Race was reported by the participants.

intervention group and  $-0.13$  (4.00) in the control group. The change from baseline to follow-up was  $-3.18$  (3.59) in the intervention group and  $-1.11$  (4.00) in the control group ([Fig. 3](#), Row 2).

### Weight Concerns

The intervention group showed significantly greater improvements in weight concerns than the control group,

exhibiting a large differential response across both time points, with comparisons remaining significant after adjusting for multiple comparisons. The mean change (SD) in WCS score from baseline to postintervention was  $-9.83$  (14.37) in the intervention group and  $-1.66$  (14.29) in the control group. The change from baseline to follow-up was  $-10.23$  (14.70) in the intervention group and  $-3.70$  (14.65) in the control group. ([Fig. 3](#), Row 3).

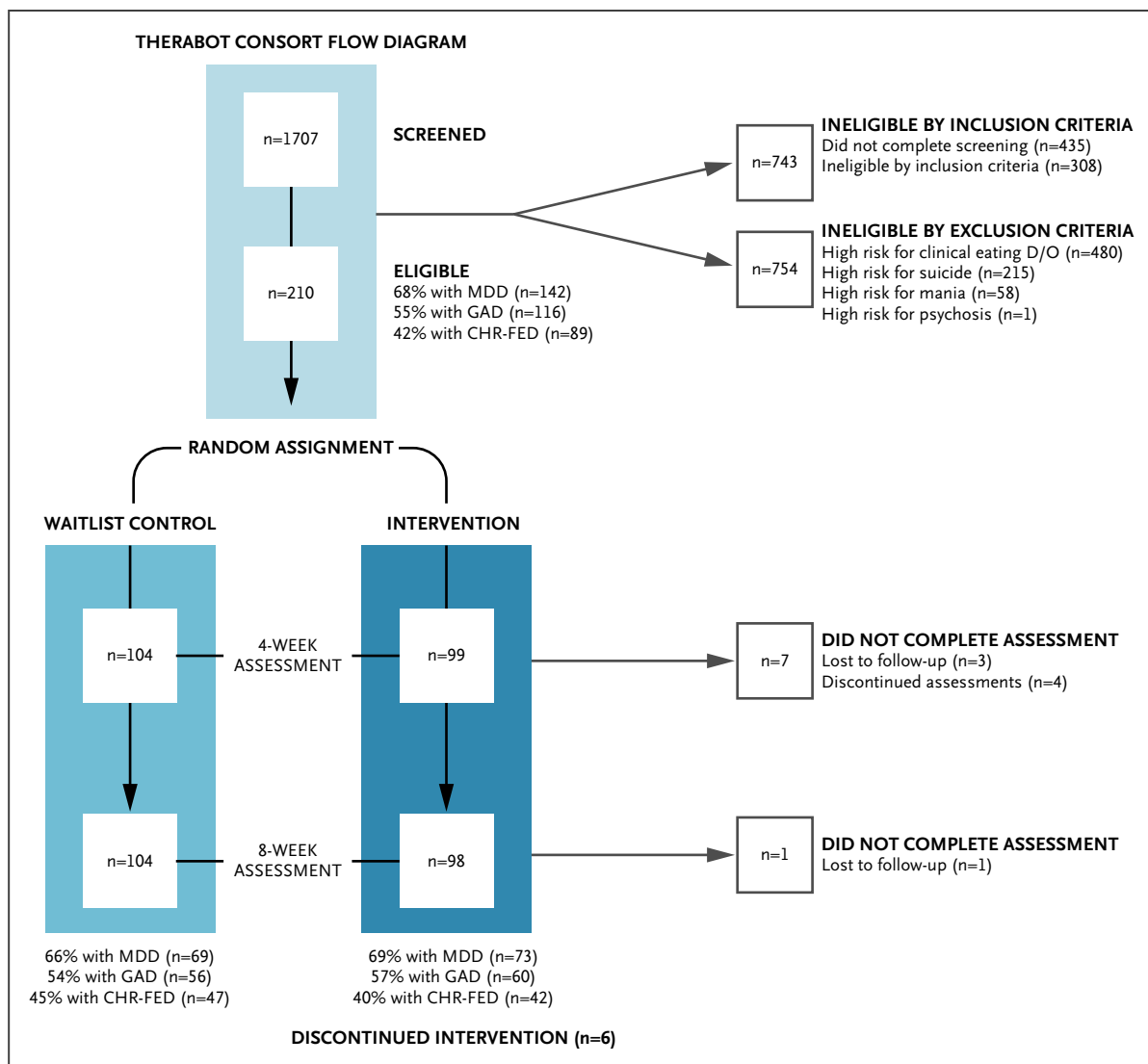


Figure 2. CONSORT Flow Diagram Showing Movement of Participants through the Study, with Associated Counts.

As comorbidity is the norm rather than the exception, data were analyzed based on each participant's pathology group membership at baseline. CHR-FED denotes clinically high risk for feeding and eating disorders; CONSORT, Consolidated Standards of Reporting Trials; D/O, disorder; GAD, generalized anxiety disorder; and MDD, major depressive disorder.

## SECONDARY OUTCOMES

### User Working Alliance

The user working alliance, measured via the WAI-SR and offered to all participants in the treatment group who interacted at least once with Therabot at 4 weeks, was completed by 96 participants. The overall mean WAI score (SD) was 3.59 (1.27); the mean (SD) score for Bond was 3.71 (1.28); the mean score for Task was 3.47 (1.30); and the

mean score for Goal was 3.59 (1.35). Box plots for working alliance scores are displayed in [Figure 4](#). Participants, on average, reported a therapeutic alliance comparable to norms reported in an outpatient psychotherapy sample.<sup>31</sup>

### User Satisfaction

The user satisfaction survey was offered to all those in the treatment group who used Therabot, with 96 participants completing the survey. Participants rated their experience



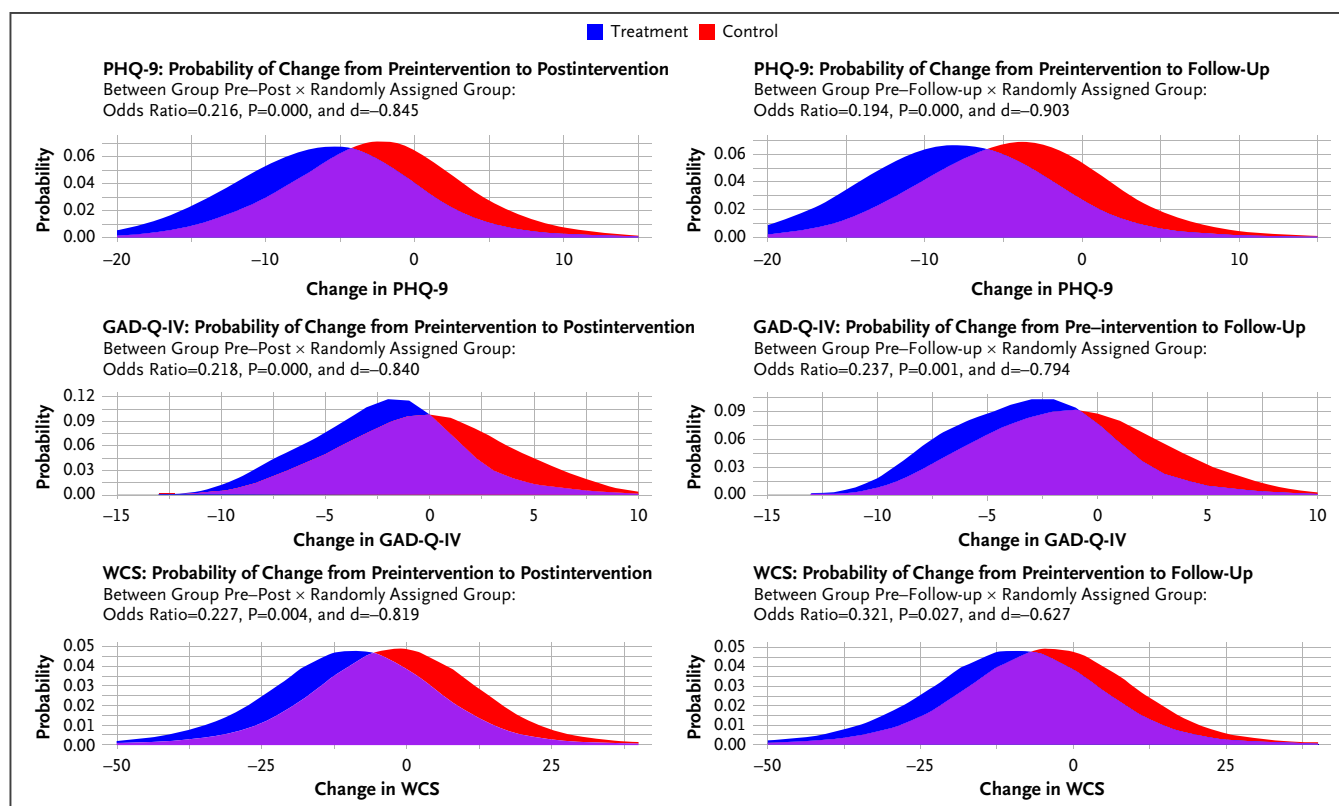


Figure 3. Distributions Representing Smoothed Probability of Changes in Clinical Outcomes (Depression, Anxiety, Weight Concerns, Row-Wise) Postintervention (4 Weeks, Left Column) and at Follow-Up (8 Weeks, Right Column).

The probabilities shown in these plots are derived directly from the CLMMs through predicted probabilities for each possible change score under treatment and control conditions. The treatment group is visualized in blue, and the control group is visualized in red. CLMM denotes cumulative-link mixed model; GAD-Q-IV, Generalized Anxiety Disorder Questionnaire for the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition; PHQ-9, Patient Health Questionnaire 9; and WCS, Weight Concerns Scale.

on a seven-point Likert scale, with seven being the highest. Users rated Therabot as easy to learn to use (6.42, SD=1.18) and intuitive (5.58, SD=1.58). On average, users liked the interface (5.46, SD=1.93) and design (5.53, SD=1.98). Users reported feeling better after interaction (5.39, SD=1.84) and found the Therabot sessions helpful (5.44, SD=1.82). Users reported that they would use Therabot on their own (5.12, SD=2.02) and rated Therabot as similar to a real therapist (4.90, SD=2.21). On average, overall satisfaction was 5.30 (SD=1.89). Box plots for user satisfaction ratings are shown in [Figure 5](#).

### User Engagement

Of participants randomly assigned to the Therabot group, 101 (95%) interacted with Therabot. The mean number of messages sent by participants was 260 (min.=1, max.=1557), 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 over the 4-week treatment phase is depicted in [Figure 6](#). Participant engagement over the full study period, including follow-up, is depicted in the Supplementary Appendix. Post-transmission staff intervention was required 15 times for participant safety concerns (e.g., expressions of suicidal ideation) and 13 times to correct inappropriate responses provided by Therabot (e.g., providing medical advice).

### 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

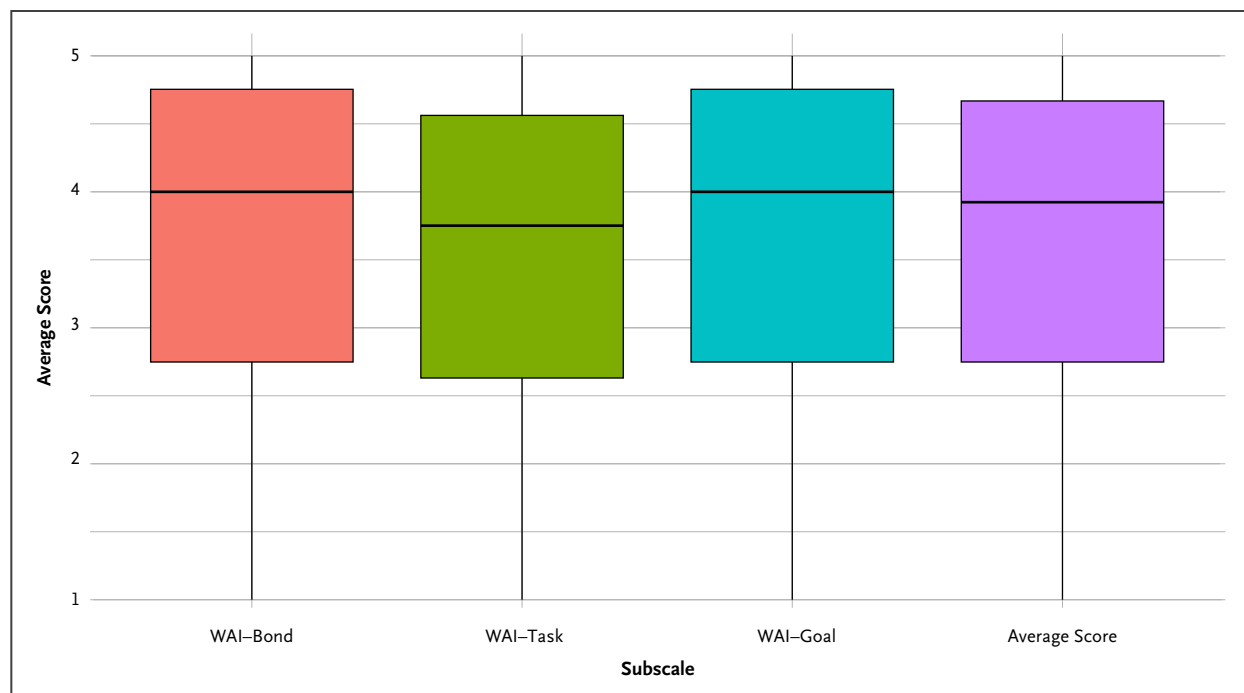


Figure 4. Box Plots of Aggregate Working Alliance Inventory Subscale Scores with Average Subscale Score (Right).

Box plots display the median (line), interquartile range (IQR) (box), and values within  $1.5 \times \text{IQR}$  (whiskers). For comparison, the outpatient norms from Munder et al.<sup>31</sup> are as follows: Bond ( $M=4.0$ ,  $SD=0.78$ ), Task ( $M=3.4$ ,  $SD=0.77$ ), Goal ( $M=4.0$ ,  $SD=0.68$ ), and Average ( $M=3.8$ ,  $SD=0.63$ ). IQR denotes interquartile range; SD, standard deviation; and WAI, Working Alliance Inventory.

symptoms. Users demonstrated sustained engagement and rated their alliance with Therabot as comparable to human therapists during the 4-week trial. Critically, as compared with the WLC, Therabot users showed a greater reduction in depression, anxiety, and CHR-FED symptoms at postintervention (4 weeks) and at follow-up (8 weeks). We posit that Therabot's success is driven by three main factors. First, akin to effective rule-based conversational agents,<sup>18</sup> Therabot is rooted in evidence-based psychotherapies for anxiety,<sup>35</sup> depression,<sup>36</sup> and weight concerns.<sup>37</sup> Second, users had unrestricted access to Therabot, allowing for any time-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, 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.<sup>38</sup> Although some evidence supports developing a therapeutic alliance with rule-based agents,<sup>19</sup> we

see such a bond as inherently limited compared with 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 4 weeks, participants were able to develop a working alliance comparable to that shown in an outpatient psychotherapy sample,<sup>32</sup> with use at consistently high rates.

Although existing companion Gen-AI chatbots can be highly engaging, they are not trained or evaluated for treating clinical-level mental health symptoms. Such chatbots may also be compromised by competing interests, such as user engagement or profit, which may be at odds with best practices for treatment.<sup>39</sup> 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 health treatment. Indeed,

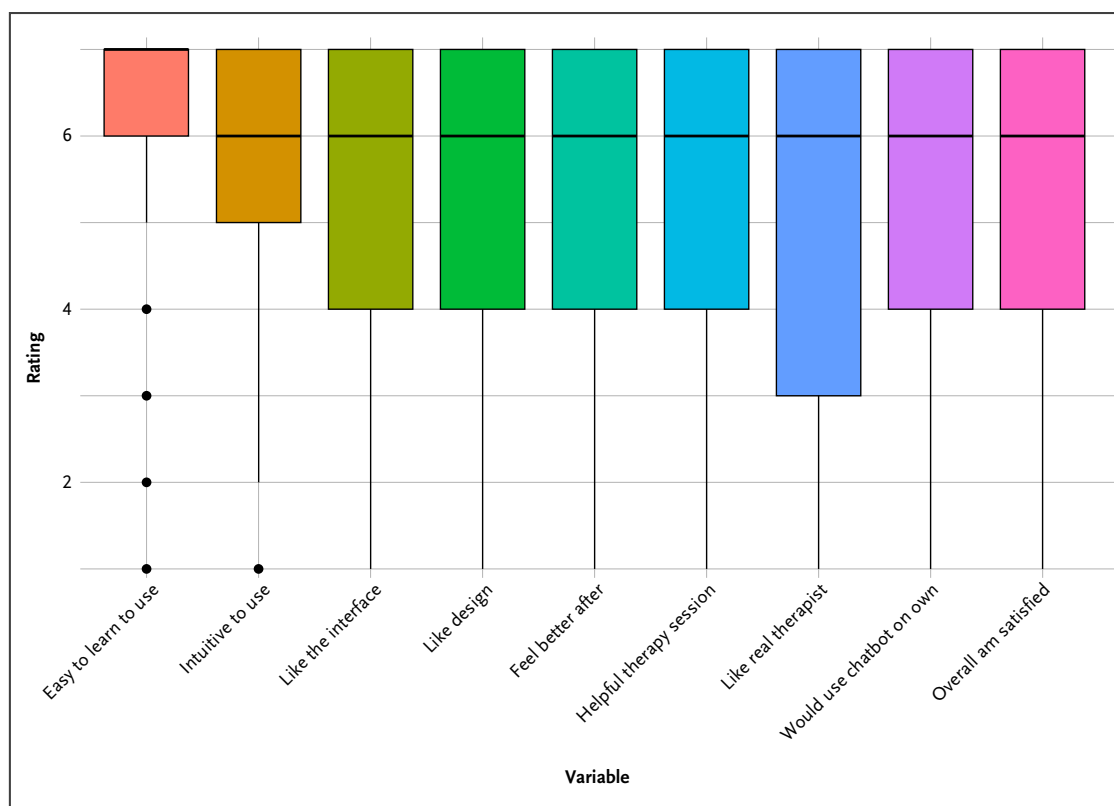


Figure 5. Box Plots Showing the Distribution of User Ratings across Satisfaction Variables.

Box plots display the median (line), interquartile range (IQR) (box), and values within  $1.5 \times \text{IQR}$  (whiskers). Points beyond the whiskers represent outliers. IQR denotes interquartile range

the nondeterministic nature of Gen-AI models introduces the possibility of hallucinations and incorrect or potentially harmful content.<sup>22</sup> While human-delivered therapy is not immune to patient iatrogenesis,<sup>40</sup> such effects in Gen-AI models have the potential to impact more people and are less regulated.

We emphasize the need to understand Gen-AI's potential role and risks associated with mental health treatment and also 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 thus may 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. Furthermore, our approach enables novel translations of therapeutic techniques that are not possible in rule-based agents. Consider, for instance, detailed and personalized imaginal exposures prompted by Gen-AI agents. Interventions dependent on 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. Furthermore, unlike many digital mental health studies,<sup>41</sup> 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 toward younger, more technologically minded participants who were open to AI.

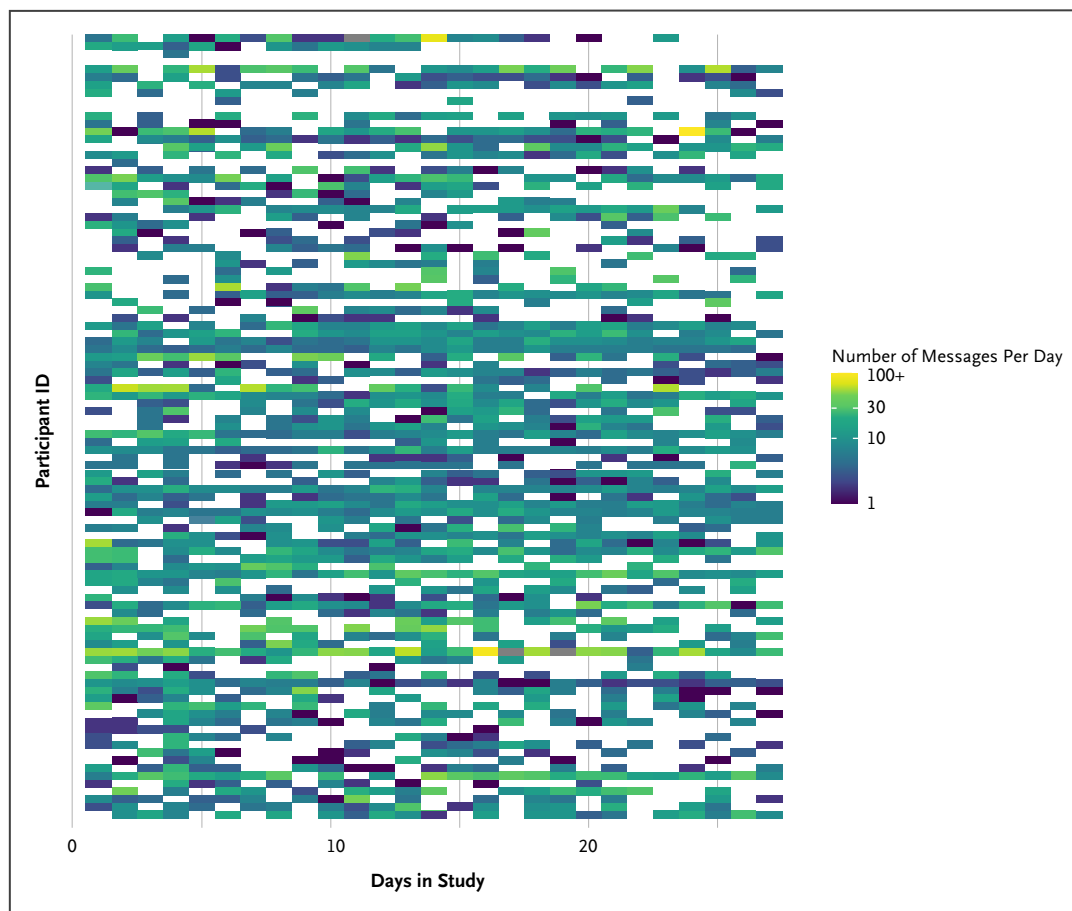


Figure 6. Heat Map Representing User Engagement across Days in the Study (x-Axis), by Participant (y-Axis)

The color represents the number of messages sent per day.

However, this may also resemble the most likely end users. Relatedly, up-front disclosure of a monetary incentive may have introduced selection bias, favoring participants who were motivated financially. Second, the characteristic of WLC RCTs is that there was potential for differential contact between the intervention and control group. We helped mitigate this by planning equivalent contact between groups whenever possible. Last, our follow-up period was limited to 4 weeks postintervention (weeks 4–8). While this allowed for testing early effectiveness and safety, longer studies are needed to assess the durability of Therabot’s effectiveness.

Overall, results from the Therabot RCT are highly promising. We found high engagement and acceptability of the intervention, as well as symptom decreases 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). Furthermore, future work should build on our descriptive user engagement metrics to define clinical thresholds pertaining to minimal doses of Gen-AI-driven therapies. Our study provides key groundwork for the development of Gen-AI chatbots for mental health treatment.

## Disclosures

Author disclosures and other supplementary materials are available at [ai.nejm.org](https://ai.nejm.org).

Supported by Dartmouth College.

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 Matong and Dr. Jianjun Xu, who provided valuable technical support for Therabot.

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 as [NCT06013137](https://clinicaltrials.gov/study/NCT06013137).

## Author Affiliations

<sup>1</sup>Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH

<sup>2</sup>Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Hanover, NH

<sup>3</sup>Quantitative Biomedical Sciences Program, Dartmouth College, Hanover, NH

<sup>4</sup>Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH

## References

- 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* 2022;9:137-150. DOI: [10.1016/S2215-0366\(21\)00395-3](https://doi.org/10.1016/S2215-0366(21)00395-3).
- Cloninger CR. The science of well-being: an integrated approach to mental health and its disorders. *World Psychiatry* 2006;5:71-76.
- National Center for Health Workforce Analysis. Behavioral health workforce, 2023. Rockville, MD: Bureau of Health Workforce, December, 2023.
- Health Resource & Service Administration. Workforce projections. Rockville, MD: Bureau of Health Workforce, 2024 (<https://data.hrsa.gov/topics/health-workforce/workforce-projections>).
- Kohn R, Ali AA, Puac-Polanco V, et al. Mental health in the Americas: an overview of the treatment gap. *Rev Panam Salud Pública* 2018;42:e165. DOI: [10.26633/RPSP.2018.165](https://doi.org/10.26633/RPSP.2018.165).
- Monteleone AM, Pellegrino F, Croatto G, et al. Treatment of eating disorders: a systematic meta-review of meta-analyses and network meta-analyses. *Neurosci Biobehav Rev* 2022;142:104857. DOI: [10.1016/j.neubiorev.2022.104857](https://doi.org/10.1016/j.neubiorev.2022.104857).
- Cuijpers P, Karyotaki E, Eckshtain D, et al. Psychotherapy for depression across different age groups: a systematic review and meta-analysis. *JAMA Psychiatry* 2020;77:694. DOI: [10.1001/jamapsychiatry.2020.0164](https://doi.org/10.1001/jamapsychiatry.2020.0164).
- Van Dis EAM, Van Veen SC, Hageraars MA, et al. Long-term outcomes of cognitive behavioral therapy for anxiety-related disorders: a systematic review and meta-analysis. *JAMA Psychiatry* 2020;77:265. DOI: [10.1001/jamapsychiatry.2019.3986](https://doi.org/10.1001/jamapsychiatry.2019.3986).
- Coombs NC, Meriwether WE, Caringi J, Newcomer SR. Barriers to healthcare access among U.S. adults with mental health challenges: a population-based study. *SSM Popul Health* 2021;15:100847. DOI: [10.1016/j.ssmph.2021.100847](https://doi.org/10.1016/j.ssmph.2021.100847).
- Fürstenau D, Gersch M, Schreiter S. Digital therapeutics (DTx). *Bus Inf Syst Eng* 2023;65:349-360. DOI: [10.1007/s12599-023-00804-z](https://doi.org/10.1007/s12599-023-00804-z).
- Wang C, Lee C, Shin H. Digital therapeutics from bench to bedside. *NPJ Digit Med* 2023;6:38. DOI: [10.1038/s41746-023-00777-z](https://doi.org/10.1038/s41746-023-00777-z).
- Nwosu A, Boardman S, Husain MM, Doraiswamy PM. Digital therapeutics for mental health: is attrition the Achilles heel? *Front Psychiatry* 2022;13:900615. DOI: [10.3389/fpsy.2022.900615](https://doi.org/10.3389/fpsy.2022.900615).
- Huibers M, Cuijpers P. Common (nonspecific) factors in psychotherapy. 2015:1-6. In: Cautina RL, Lilienfeld SO, eds. *The encyclopedia of clinical psychology*. 2nd Edn. Hoboken, NJ: Wiley-Blackwell. DOI: [10.1002/9781118625392.wbecp272](https://doi.org/10.1002/9781118625392.wbecp272)
- Henson P, Peck P, Torous J. Considering the therapeutic alliance in digital mental health interventions. *Harv Rev Psychiatry* 2019;27:268-273. DOI: [10.1097/HRP.0000000000000224](https://doi.org/10.1097/HRP.0000000000000224).
- 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* 2022;13:819623. DOI: [10.3389/fpsy.2022.819623](https://doi.org/10.3389/fpsy.2022.819623).
- Breuer J, Freud S. *Studies on hysteria*. Oxford: Basic Books, 1957.
- Weizenbaum J. ELIZA — a computer program for the study of natural language communication between man and machine. *Commun ACM* 1966;9:36-45. DOI: [10.1145/365153.365168](https://doi.org/10.1145/365153.365168).
- Fitzpatrick KK, 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* 2017;4:e19. DOI: [10.2196/mental.7785](https://doi.org/10.2196/mental.7785).



19. 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* 2021;5:e27868. DOI: [10.2196/27868](https://doi.org/10.2196/27868).
20. Gill SS, Kaur R. ChatGPT: vision and challenges. *Internet Things Cyber-Phys Syst* 2023;3:262-271. DOI: [10.1016/j.iotcps.2023.05.004](https://doi.org/10.1016/j.iotcps.2023.05.004).
21. De Freitas J, Cohen IG. The health risks of generative AI-based wellness apps. *Nat Med* 2024;30:1269-1275. DOI: [10.1038/s41591-024-02943-6](https://doi.org/10.1038/s41591-024-02943-6).
22. Pentina I, Hancock T, Xie T. Exploring relationship development with social chatbots: a mixed-method study of replika. *Comput Hum Behav* 2023;140:107600. DOI: [10.1016/j.chb.2022.107600](https://doi.org/10.1016/j.chb.2022.107600).
23. Hayes SC, Hofmann SG. "Third-wave" cognitive and behavioral therapies and the emergence of a process-based approach to intervention in psychiatry. *World Psychiatry* 2021;20:363-375. DOI: [10.1002/wps.20884](https://doi.org/10.1002/wps.20884).
24. Neundorff A, Öztürk A. How to improve representativeness and cost-effectiveness in samples recruited through meta: a comparison of advertisement tools. *PLoS One* 2023;18:e0281243. DOI: [10.1371/journal.pone.0281243](https://doi.org/10.1371/journal.pone.0281243).
25. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform* 2019;95:103208. DOI: [10.1016/j.jbi.2019.103208](https://doi.org/10.1016/j.jbi.2019.103208).
26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377-381. DOI: [10.1016/j.jbi.2008.08.010](https://doi.org/10.1016/j.jbi.2008.08.010).
27. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606-613. DOI: [10.1046/j.1525-1497.2001.016009606.x](https://doi.org/10.1046/j.1525-1497.2001.016009606.x).
28. Newman MG, Zuellig AR, Kachin KE, 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* 2002;33:215-233. DOI: [10.1016/S0005-7894\(02\)80026-0](https://doi.org/10.1016/S0005-7894(02)80026-0).
29. Graham AK, Trockel M, Weisman H, et al. A screening tool for detecting eating disorder risk and diagnostic symptoms among college-age women. *J Am Coll Health J ACH* 2019;67:357-366. DOI: [10.1080/07448481.2018.1483936](https://doi.org/10.1080/07448481.2018.1483936).
30. McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. *J Affect Disord* 2010;127:122-129. DOI: [10.1016/j.jad.2010.04.030](https://doi.org/10.1016/j.jad.2010.04.030).
31. Munder T, Wilmers F, Leonhart R, Linster HW, Barth J. Working alliance inventory-short revised (WAI-SR): psychometric properties in outpatients and inpatients. *Clin Psychol Psychother* 2010;17:231-239. DOI: [10.1002/cpp.658](https://doi.org/10.1002/cpp.658).
32. Sánchez-Meca J, Marín-Martínez F, Chacón-Moscoso S. Effect-size indices for dichotomized outcomes in meta-analysis. *Psychol Methods* 2003;8:448-467. DOI: [10.1037/1082-989X.8.4.448](https://doi.org/10.1037/1082-989X.8.4.448).
33. Christensen RHB. Cumulative link models for ordinal regression with the R package ordinal. 2018 ([https://cran.r-project.org/web/packages/ordinal/vignettes/clm\\_article.pdf](https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf)).
34. UCLA: Statistical Consulting Group. Ordinal logistic regression. R data analysis examples. 2024. (<https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/>).
35. Covin R, Ouimet AJ, Seeds PM, Dozois DJA. A meta-analysis of CBT for pathological worry among clients with GAD. *J Anxiety Disord* 2008;22:108-116. DOI: [10.1016/j.janxdis.2007.01.002](https://doi.org/10.1016/j.janxdis.2007.01.002).
36. Lepping P, Whittington R, Sambhi RS, et al. Clinical relevance of findings in trials of CBT for depression. *Eur Psychiatry* 2017;45:207-211. DOI: [10.1016/j.eurpsy.2017.07.003](https://doi.org/10.1016/j.eurpsy.2017.07.003).
37. Jarry JL, Ip K. The effectiveness of stand-alone cognitive-behavioural therapy for body image: a meta-analysis. *Body Image* 2005;2:317-331. DOI: [10.1016/j.bodyim.2005.10.001](https://doi.org/10.1016/j.bodyim.2005.10.001).
38. Baier AL, Kline AC, Feeny NC. Therapeutic alliance as a mediator of change: a systematic review and evaluation of research. *Clin Psychol Rev* 2020;82:101921. DOI: [10.1016/j.cpr.2020.101921](https://doi.org/10.1016/j.cpr.2020.101921).
39. Haque MDR, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. *JMIR MHealth UHealth* 2023;11:e44838. DOI: [10.2196/44838](https://doi.org/10.2196/44838).
40. Boisvert CM, Faust D. Iatrogenic symptoms in psychotherapy. *Am J Psychother* 2002;56:244-259. DOI: [10.1176/appi.psychotherapy.2002.56.2.244](https://doi.org/10.1176/appi.psychotherapy.2002.56.2.244).
41. Bryan AC, Heinz MV, Salzhauer AJ, Price GD, Tlachac ML, Jacobson NC. Behind the screen: a narrative review on the translational capacity of passive sensing for mental health assessment. *Biomed Mater Devices* 2024;2:778-810. DOI: [10.1007/s44174-023-00150-4](https://doi.org/10.1007/s44174-023-00150-4).