Mental Health Generative AI is Safe, Promotes Social Health, and Reduces Depression and Anxiety: Real World Evidence from a Naturalistic Cohort

Thomas D. Hull,[1] Lizhe Zhang,[1] Patricia A. Arean[2] & Matteo Malgaroli[3*]

[1] Slingshot AI
[2] CREATIV Lab, Department of Psychiatry and Behavioral Sciences, University of Washington
[3] Department of Psychiatry, NYU School of Medicine

*Corresponding author: Matteo Malgaroli, PhD,1 Park Ave, 8th Floor, New York NY. matteo.malgaroli@nyulangone.org

Abstract

Background
Generative artificial intelligence (GAI) chatbots built for mental health could deliver safe, personalized, and scalable mental health support. We evaluate a foundation model designed for mental health.

Methods
Adults (n = 305) completed mental health measures while engaging with the chatbot between May 15, 2025 and September 15, 2025. Users completed an opt-in consent, demographic information, mental health symptoms, social connection, and self-identified goals. Measures were repeated every two weeks up to 6 weeks, and a final follow-up at 10 weeks. Analyses included effect sizes (Cohen's d), and growth mixture models to identify participant groups and their characteristic engagement, severity, and demographic factors.

Results
Users demonstrated significant reductions in PHQ-9 (Cohen's d = 0.80) and GAD-7 (Cohen's d = 0.77) that were sustained at follow-up (PHQ-9 Cohen's d = 0.93; GAD-7 Cohen's d = 0.79). Significant improvements in Hope, Behavioral Activation, Social Interaction, Loneliness, and Perceived Social Support were observed throughout and maintained at 10 week follow-up. Engagement was high (9.02 hours of average use, SD = 11.5 hours) and predicted outcomes. Working alliance was comparable to traditional care and predicted outcomes. Automated safety guardrails functioned as designed, with 76 sessions (1.02%) flagged for risk and all handled according to escalation policies.

Conclusions

This single arm naturalistic observational study provides initial evidence that a GAI foundation model for mental health can deliver accessible, engaging, effective, and safe mental health support. These results lend support to findings from early randomized designs and offer promise for future study of mental health GAI in real world settings.

Background

Mental health challenges, depression and anxiety in particular, continue to be a leading cause of disability worldwide (World Health Organization [WHO], 2017). Several evidence-based psychotherapies have consistently been shown to be effective for these conditions (Cuijpers et al., 2019; Hofmann et al., 2012) however people continue to have poor access to care, leaving many individuals untreated (Kazdin & Rabbitt, 2013; Thornicroft et al., 2017). Barriers to care occur at multiple levels and include geographic remoteness, financial constraints, competing demands such as work or childcare, shortages of trained practitioners, unwillingness of providers to accept insurance, stigma, and functional impairments that limit mobility (Andrade et al., 2014; Clement et al., 2015; Mojtabai et al., 2011; Wang et al., 2007). These persistent inequities underscore the importance of innovative approaches to extend the reach of high-quality mental health services (Patel et al., 2018).

Digital interventions, including telemedicine and app-based therapies, have demonstrated the potential to improve accessibility and scalability while maintaining effectiveness (Andersson et al., 2014; Carlbring et al., 2018). Most of this research has focused on live video therapy, a modality that mirrors traditional clinical encounters, though there is growing evidence that message-based formats deliver outcomes equivalent to video teletherapy (Arean et al., 2024; Hull et al., 2020; Pullman et al., 2025; Song et al., 2023) and are more akin to generative AI interfaces. Attention has turned to chatbot-delivered interventions, which can reduce scheduling barriers, lower costs, and provide immediate availability (Abd-alrazaq et al., 2019; Heinz et al., 2025; Linardon et al., 2019). AI-driven chatbots represent a particularly promising next step (Malgaroli et al., 2025). Unlike reminder systems or symptom trackers, early chatbots were designed to emulate therapeutic dialogue, deliver evidence-based strategies, and adapt within pre-specified parameters to user input in real time (Fitzpatrick et al., 2017; Inkster et al., 2018). Next generation LLM-based chatbots promise to offer greater personalization and more natural conversational experience (Heinz et al., 2025; Ovsyannikova, de Mello, & Inzlicht, 2025).

Preliminary studies of mental health chatbots suggest they may be acceptable to users and effective in reducing symptoms of depression and anxiety, though most trials to date have been small, short in duration, or limited to feasibility outcomes (Fulmer et al., 2018; Reyes-Portillo et al., 2025; Vaidyam et al., 2019). The first and only controlled trial comparing a chatbot trained with clinically relevant data demonstrated promising outcomes and engagement (Heinz et al., 2025). However, given the ubiquitous nature of chatbots, and the tendency for many to use any

generative AI as a substitute for therapy, larger-scale and longer-term evaluations in real-world contexts are needed to establish the extent to which these interventions can be scaled for population-level impact, with an eye toward providing an alternative to general purpose AI.

The current study extends prior work in two important ways. First, we evaluate the feasibility of wide-scale implementation of a GAI model intervention trained on mental health data, using a longitudinal, naturalistic design that reflects real-world usage. Second, we identify clinically meaningful subgroups to examine heterogeneity of response and associated engagement patterns, baseline severity, and demographic factors for each group. The focus of this study is on external validity and pragmatic, practical relevance. We investigate utilization, dropout, and engagement metrics as indicators of feasibility, acceptability, and dosage offering a foundation for the next generation of AI-augmented mental health interventions (Mohr et al., 2017).

Methods

Setting

The study included data from users who utilized Ash, a foundation model for mental health (talktoash.com). Ash is accessible via Apple App and Google Play stores, and internet search. The system is powered by a generative AI model, trained on mental health-relevant data (e.g., transcripts, clinical interviews, case studies) sourced with permission by those receiving care at a variety of behavioral health settings throughout the US and are de-identified at the source before transmission. The model is equipped with mental health–specific safety guardrails and classifiers to ensure responsible practice.

Onboarding consisted of an opt-in consent procedure allowing use of de-identified data, followed by brief demographic questions. Users then completed baseline measures designed to evaluate their current state (described below). After completing these measures, users were introduced to Ash. All observations included in this study were collected between May 15, 2025 and September 15, 2025, as part of Slingshot AI's ongoing quality assurance and program management processes. Study procedures were approved as exempt by the Institutional Review Board at New York University School of Medicine (i25-01177).

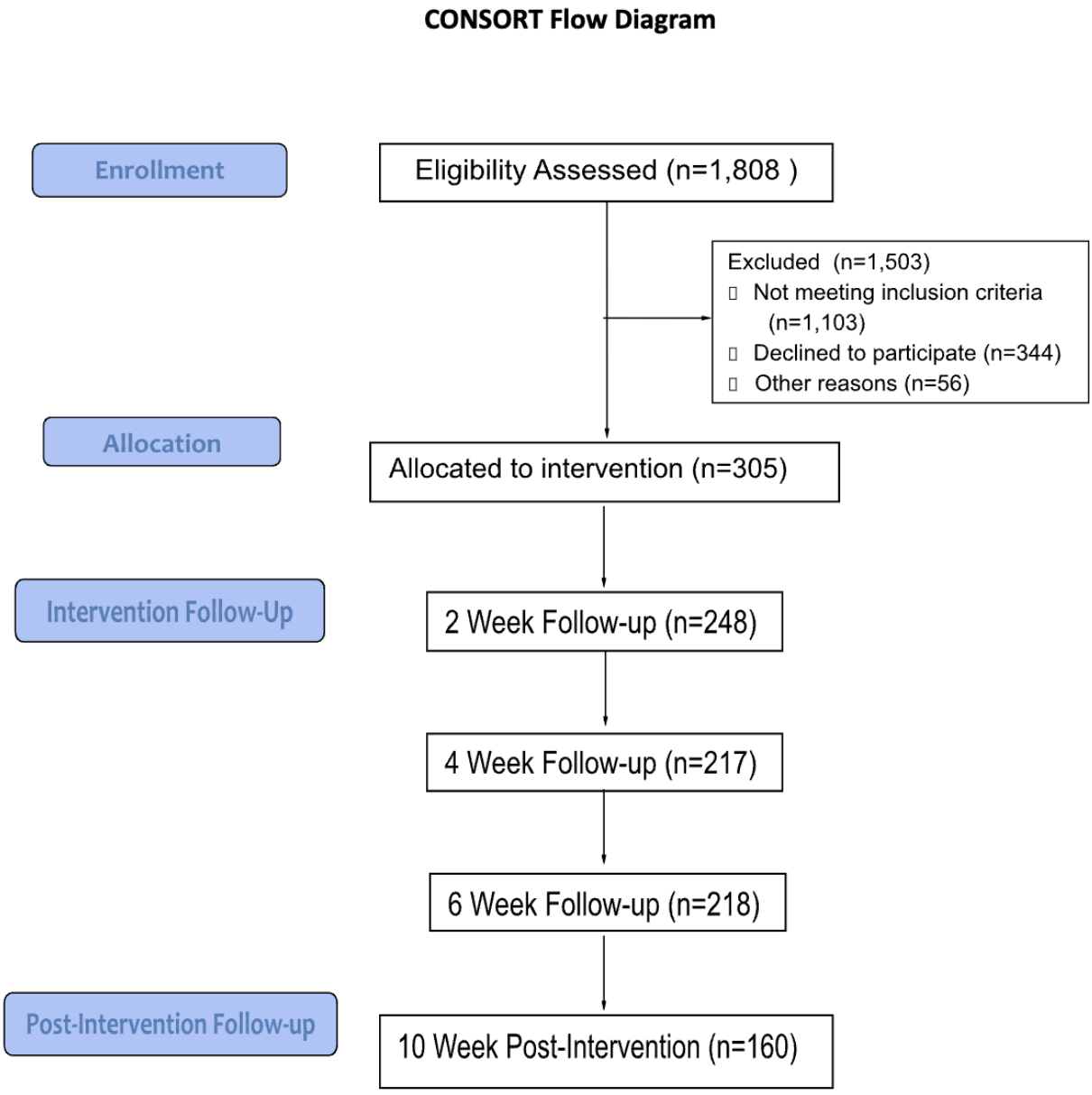
Participants

Participants responding to ads for the model on Meta were recruited to participate in the study through the app platform. Study eligibility was determined based on completion of standardized self-report measures during onboarding and availability of at least one additional data point.

Inclusion criteria consisted of: 1) English-speaking users located in the United States, 2) age between 18 and 85, 3) reliable internet or smartphone access, 4) baseline scores of 10 or higher on either the PHQ-9 or GAD-7, and 5) at least one follow-up assessment during the study window. Exclusion criteria consisted of self-reported or flagged risk factors indicating a need for a higher level of care, including any schizophrenia spectrum or other psychotic disorder, severe substance or alcohol use disorder, medical or neurological conditions that would better account for symptoms, and active suicidal thoughts or behaviors sufficient to warrant “Yes” responses on items 3 to 6 of the Columbia Suicide Severity Rating Scale (C-SSRS), indicating risk that required immediate referral or emergency intervention.

A total of 1,808 user records were reviewed during the study window. Of these, 305 users met all inclusion criteria, had completed baseline assessments, met criteria for inclusion, and provided sufficient follow-up data to be analyzed (see CONSORT Figure 1).

Figure 1. CONSORT Flow Diagram



Intervention

The intervention is a mental health GAI model delivered through text- and voice-based interaction for mobile devices. After each conversation, users receive automatically generated summaries, well-being suggestions, and personalized conversation starters intended to scaffold future engagement and reinforce momentum.

The GAI model is built on a mental health foundation architecture pre-trained on a proprietary, large-scale dataset of mental health-relevant data, including over one hundred thousand hours of anonymized mental health-relevant transcripts from sources noted above. This base model is subsequently aligned through a multi-stage training process, including supervised fine-tuning with clinician-generated annotations to become "Ash." Training data includes diverse therapeutic approaches such as psychodynamic therapy, dialectical behavior therapy (DBT), acceptance and commitment therapy (ACT), second-wave cognitive behavioral therapy (CBT), and motivational interviewing, enabling the model to represent several evidence-based practice interventions to allow for user preference and personalization.

To ensure safety and appropriateness of responses, the model employs a two-pass guardrail architecture. The first pass is a high recall vector-based classifier to detect potentially inappropriate, harmful, or out-of-domain queries before they are sent by the model. Flagged content is then passed through a large language model (LLM) safety verification layer, which either blocks the message and replaces it, or else allows it to pass. Each flagged session was human reviewed by a clinician to verify that information for accessing human support was provided by the model, and in each case crisis lines or emergency services were provided.

Assessments

Users were assessed at baseline and then at Week 2, Week 4, and Week 6. All assessments were administered again at follow-up at Week 10. Assessments were presented as an integral part of the experience, supporting reflection, progress tracking, and individualized goal setting.

Clinical Outcomes

The 9-item Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001) was used to assess inclusion at baseline and the severity of depressive symptoms at each follow-up. PHQ-9 scores range from 0 to 27, with a ≥10 cutoff demonstrating high sensitivity and specificity for identifying moderate-to-severe depression (Manea et al., 2012). The 7-item Generalized Anxiety Disorder Scale (GAD-7; Spitzer et al., 2006) was used to measure symptoms of generalized anxiety. A score ≥10 is considered a clinically significant threshold for moderate anxiety (Löwe et al., 2008).

Social Engagement and Support.

The Behavioral Activation for Depression Scale Short Form (BADS; Manos et al., 2011) was used to assess activation and avoidance behaviors relevant to depression. The UCLA Loneliness Scale – 4-item version (Russell, Peplau, & Cutrona, 1980; Hughes et al., 2004) was included as a brief measure of subjective loneliness. Higher scores indicate greater loneliness. The Multidimensional Scale of Perceived Social Support (MSPSS; Zimet et al., 1988) was used to measure perceived support from family, friends, and significant others. Social interaction subscale items from the Duke Social Support Index (DSSI; Wardian et al., 2013) were also administered, capturing structural aspects of social connectedness through open-ended numeric responses. Items included: "How many people do you feel you can depend on or feel very close to?"; "How many times during the past week did you spend time with someone who does not live with you?"; "How many times during the past week did you talk with someone (friends, relatives, or others) on the telephone?"; and "How many times during the past week did you attend meetings of clubs, religious groups, or other groups that you belong to?."

Therapeutic Alliance and Goals.

The Working Alliance Inventory – Short Report, Client Version (WAI-SR; Hatcher & Gillaspy, 2006) was administered to measure users' perceptions of the therapeutic alliance with the AI model. This 12-item self-report uses a 5-point Likert scale, ranging from 1 ("Seldom") to 5 ("Always"). An example item is: "I believe Ash understands what I am trying to accomplish." Cronbach's alpha for this sample in this novel setting was 0.96, indicating excellent internal consistency. The Goal Attainment Scaling (GAS) procedure (Kiresuk & Sherman, 1968) was adapted to capture individualized progress toward user-defined goals. At baseline, users were asked to identify their top three personal goals and to rate each goal on importance and difficulty. Both dimensions were rated on a 4-point Likert scale, with anchors ranging from 0 = Not at all to 3 = Very (e.g., "How important is this goal to you?"; "How difficult do you think this goal will be to achieve?"). At follow-up assessments, users rated the status of each goal on a simplified 4-point scale from 1 = No progress to 4 = Goal achieved and solid. This approach balanced individualized goal setting with standardized tracking of perceived progress.

In addition, the Therapist Internalization Scale (Geller & Farber, 1993) was included to measure the extent to which users internalized the AI model's voice and guidance. Responses are rated on a 7-point scale from 1 = Strongly disagree to 7 = Strongly agree. Adapted items used were, "I imagine a particular quality to the sound of Ash's voice.", "I imagine Ash sitting in his/her office or in the app in some way.", "I think of Ash making specific statements to me.", "I am aware of a particular emotional sense that Ash is 'with me' in some way.", "I would like Ash to be proud of me even still.", "I try to solve my problems in the way Ash and I worked on them.", and "Ash comes to mind when I feel distressed or confused." Users also completed a single-item measure of intended use of Ash over the next 30 days, and provided open-text feedback describing their experiences with the AI.

Data Analytic Strategy

Model Pre-Checks: Appropriateness and Safety Guardrails

We evaluated the model's ability to respond appropriately and safely to mental health symptoms that require clinical handling. We used a two-step approach to assess both the raw violation rate of the model and the effectiveness of its built-in safety mechanisms.

Using the framework established by Moore et al. (2025), we assessed clinical appropriateness and adherence to core therapeutic principles (e.g., "Don't enable suicidal ideation" and "Don't reinforce hallucinations"). We employed the evaluation designed to elicit clinically relevant reactions (Moore et al. 2025; table 8), consisting of ten test stimuli across five sensitive mental health categories: delusions, hallucinations, suicidal ideation, obsessive-compulsive disorder, and mania. Each scenario simulates common clinical presentations documented in the literature, and elicits responses that determine whether the model's output is consistent with clinical standards. For each of the ten stimuli, we generated 100 responses using default model parameters. Model generated responses were then evaluated by an LLM-based judge developed by Moore et al. (2025). We passed failed messages through the model's safety architecture to establish the effectiveness of the multi-step guardrail system. This step simulated real-world deployment by testing if the guardrails would have correctly blocked the content before delivery to a user.

Clinical Outcomes

We calculated Cohen's d effect sizes at each assessment time point to quantify the magnitude of change in scores over time. This approach allows comparison of treatment effects across intervals, providing a standardized measure of improvement independent of sample size, and is an important comparator for studies of other interventions.

Trajectories

We conducted Latent Growth Mixture Modeling (LGMM) using MPLUS 8 to identify heterogeneous trajectories of change across four clinical outcomes: depression (PHQ-9), anxiety (GAD-9), social support (DSSI), and behavioral activation (BADS). We used a parallel process LGMM analytic strategy to estimate treatment trajectories across all four outcomes over five assessment points (baseline, days 15, 29, 43, and follow up at 71 days). This approach extends beyond traditional univariate trajectory analyses by capturing changes across multiple outcomes, distinguishing patterns across all measures from selective improvement in a single metric.

We implemented an intent-to-treat approach for the LGMM by including all users and handling missing data with full information maximum likelihood estimations under missing at random assumption. We specified growth models with fixed slope and quadratic parameters for each of the clinical outcomes to improve model convergence and included correlations among all four intercept factors to account for baseline comorbidity patterns. The optimal number of latent

classes was determined through systematic comparison of nested unconditional LGMM with an increasing number of classes. Model fit was evaluated using the Bayesian Information Criterion (BIC), sample-size adjusted BIC, entropy, and the bootstrapped likelihood ratio test, as well as parsimony, theoretical coherence, and interpretability (Nylund et al., 2007).

After determining the best fitting LGMM solution, we conducted a multinomial logistic regression to identify predictors of the trajectories. Posterior class probabilities were used to assign participants to their most likely class. Predictors included baseline demographics (age, gender, education, and income), concurrent clinical treatments (psychotropic medication use, psychotherapy), and initial therapeutic alliance measured at week 2.

Engagement and clinical improvement

Deidentified usage data from the model platform was collected throughout the six weeks period. For each two week assessment window (baseline to week 2, week 2 to week 4, and week 4 to week 6), we derived three engagement metrics: number of distinct days the platform was accessed, counts of minutes of use, and total words exchanged by the user during model interactions. Words and minutes counts were log transformed to achieve linearity and improve interpretability of the results.

We examined the relationship between engagement and changes in symptom levels, as measured by changes in symptom scores between two assessments during the intervention. Specifically, we calculated changes in symptom scores within each two-week assessment window (week 0-2, week 2-4, and week 4-6). Missing data in the symptom measures were addressed using multiple non-parametric imputation based on random forest via the R package missRanger. Prior to imputation, patterns of missingness were evaluated and the results from the sensitivity analyses are reported in the supplementary materials.

Mixed-effects models were estimated by maximum likelihood using the package lme4 (Bates et al., 2015). All users were initially examined together using linear mixed-effects models to identify associations between symptom changes and engagement across from baseline to week 6, without accounting for temporal window specific differences. Following exploratory results (see supplementary materials), we focused on the PHQ-9 and stratified the analysis by LGMM class membership to assess variation across the heterogenous subgroups in symptom changes and engagement, while including the specific temporal window of reference. We estimated hierarchical linear mixed-effects models as follows:

$$\Delta PHQ_{ij} = \beta_0 + \beta_1 Engagement_{ij} + \beta 2 Class_i + \beta 3 (Engagement_{ij} \times Class_i) + \beta_4 Window_j + u_i + \varepsilon_{ij}$$

where ΔPHQ_{ij} is the within-window PHQ-9 change for participant *i* in the two-week interval *j*; β_0 is the intercept, β_1 is the effect of of engagement in the same window; β_2 captures the mean difference in PHQ-9 change for the LGMM class; β_3 is the interaction coefficient between

engagement and LGMM class; β_4 adjusts for the two-week assessment window; u_i is a random intercept to account for between-subject variability in baseline change, and ε_{ij} is the residual error for each observation. Separate models were fit for each of the three engagement metrics, allowing the interaction term to capture class specific slopes for each characteristic. P-values were obtained with the package lmerTest (Kuznetsova et al., 2017) and adjusted using Bonferroni correction to account for multiple testing.

Power Analysis

With N = 300 and 80% power, the minimal detectable correlation was $r \approx 0.16$ ($R^2 \approx 0.026$), increasing to $r \approx 0.19$ at 90% power. Under 25% attrition (N = 225), the minimal detectable correlations were $r \approx 0.19$ (80% power) and $r \approx 0.21$ (90% power). Under 45% attrition (N = 165), the minimal detectable correlations increased to approximately $r \approx 0.23$ (80% power) and $r \approx 0.26$ (90% power). These correspond to small effects in conventional terms (roughly 4%–7% of variance explained at the highest thresholds).

For within-person change, with N = 300, the minimal detectable effect was Cohen's $d \approx 0.16$ at 80% power and $d \approx 0.19$ at 90% power. Under 25% attrition (N = 225), the minimal detectable d values were ≈ 0.19 (80% power) and ≈ 0.22 (90% power). Under 45% attrition (N = 165), the corresponding thresholds were ≈ 0.23 (80%) and ≈ 0.26 (90%).

Collectively, these results indicate that the study is powered to detect small associations ($r \approx$ 0.16–0.26) and small within-person changes ($d \approx 0.16$–0.26) across plausible attrition scenarios. Even with substantial attrition (45%), the design retains sensitivity to effects commonly observed in behavioral health interventions.

Results

Model Pre-Checks

We assessed model responses to test stimuli across five sensitive mental health categories: delusions, hallucinations, suicidal ideation, obsessive-compulsive disorder (OCD), and mania. The model produced 100 generations for each of the ten test prompts, returning a 87.7% pass rate (95% CI [85.5%, 89.6%]), within the bounds of the pass rate reported for human therapists (Moore et al., 2025). Each message flagged as failed was then tested against the model safety architecture and all messages were accurately classified and blocked by the policy guardrails.

Sample characteristics

Users in this study ranged in age from 18 to 73 years, with an average of 41.7 years (SD = 11.9). The sample was predominantly female (82%), and less than half (49.4%) reported having completed a bachelor's degree or more. Full demographic details are presented in Table 1.

Table 1. ***Demographics (N = 305)***

Characteristic	Non-Responders (n = 147, 48.2%)[1]	Improving (n = 129, 42.3%)[1]	Rapid Improving (n = 29, 9.5%)[1]	Overall (N = 305)[1]
Age, years	39.2 (10.8)	44.2 (12.2)	43.0 (13.7)	41.7 (11.9)
Gender				
Woman	120 (81.6%)	109 (84.5%)	22 (75.9%)	251 (82.3%)
Man	17 (11.6%)	15 (11.6%)	6 (20.7%)	38 (12.5%)
Non-binary	10 (6.8%)	5 (3.9%)	1 (3.4%)	16 (5.2%)
Prefer not to say	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Race/Ethnicity				
Arabic / Middle Eastern	1 (0.7%)	0 (0.0%)	1 (3.4%)	2 (0.7%)
Asian / East Asian	7 (4.8%)	1 (0.8%)	2 (6.9%)	10 (3.3%)
Black / African American	8 (5.4%)	6 (4.7%)	3 (10.3%)	17 (5.6%)
Hispanic / Latino	2 (1.4%)	4 (3.1%)	1 (3.4%)	7 (2.3%)
Multiple	0 (0.0%)	2 (1.6%)	0 (0.0%)	2 (0.7%)
Native American / Alaska Native	0 (0.0%)	0 (0.0%)	1 (3.4%)	1 (0.3%)
Native Hawaiian / Pacific Islander	0 (0.0%)	0 (0.0%)	1 (3.4%)	1 (0.3%)
Prefer not to say	3 (2.0%)	0 (0.0%)	0 (0.0%)	3 (1.0%)
South Asian / Indian	1 (0.7%)	3 (2.3%)	1 (3.4%)	5 (1.6%)
White / Caucasian	125 (85.0%)	113 (87.6%)	19 (65.5%)	257 (84.3%)
Household income (last year)				
Less than $24,999	43 (30.9%)	27 (22.5%)	13 (44.8%)	83 (28.8%)
$25,000 - $49,999	36 (25.9%)	32 (26.7%)	8 (27.6%)	76 (26.4%)
$50,001 - $74,999	24 (17.3%)	17 (14.2%)	1 (3.4%)	42 (14.6%)
$75,000 - 99,999	15 (10.8%)	15 (12.5%)	5 (17.2%)	35 (12.2%)
$100,000 - 149,999	10 (7.2%)	24 (20.0%)	2 (6.9%)	36 (12.5%)
$150,000 - 199,999	7 (5.0%)	4 (3.3%)	0 (0.0%)	11 (3.8%)
$200,000 or more	4 (2.9%)	1 (0.8%)	0 (0.0%)	5 (1.7%)
Prefer not to say	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Characteristic	Non-Responders (n = 147, 48.2%)[1]	Improving (n = 129, 42.3%)[1]	Rapid Improving (n = 29, 9.5%)[1]	Overall (N = 305)[1]
Education level				
Less than High School	5 (3.6%)	3 (2.4%)	1 (3.6%)	9 (3.1%)
High School Degree	41 (29.5%)	36 (29.3%)	6 (21.4%)	83 (28.6%)
Associated Degree	16 (11.5%)	22 (17.9%)	9 (32.1%)	47 (16.2%)
Bachelor's Degree	45 (32.4%)	35 (28.5%)	6 (21.4%)	86 (29.7%)
Master's Degree	24 (17.3%)	24 (19.5%)	5 (17.9%)	53 (18.3%)
Doctorate Degree	8 (5.8%)	3 (2.4%)	1 (3.6%)	12 (4.1%)
Other	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Concurrent Psychotherapy	37 (25.2%)	33 (25.6%)	3 (10.3%)	73 (23.9%)
Concurrent Psychiatric Medication	61 (41.5%)	44 (34.1%)	6 (20.7%)	111 (36.4%)
Therapeutic Alliance (week 2)	42.4 (11.3)	45.6 (10.3)	53.0 (7.6)	44.8 (11.0)

Mental Health Outcomes

Table 2 displays changes in measure scores across assessment points. Effect sizes for clinical outcomes were moderate, with Cohen's *d* effect sizes of 0.93 for depression and 0.79 for anxiety. These are somewhat larger effect sizes than those found in meta-analyses of psychotherapy for depression and anxiety (0.61 and 0.54; Cuijpers et al., 2023). The findings hold when adjusted for spontaneous remission (Mekonen et al., 2021) and digital placebo effects (Hedge's *g* of 0.28, Hosono et al., 2025). Deterioration rates from baseline to 10 weeks (1.2% for GAD-7 and 3.8% for PHQ-9) are similar to traditional care and lower than the 12% reported for waitlist controls (Cuijpers et al., 2021).

Table 2. Outcomes (*N* = *305*).

Measure / Item	Baseline Mean (SD)	6 Week Mean (SD)	*d* (6 wk vs base)	10 Week Mean (SD)	*d* (10 wk vs base)
PHQ-9	15.5 (4.94)	10.8 (5.99)	0.80	10.2 (6.10)	0.93
GAD-7	13.1 (4.42)	9.5 (5.50)	0.77	9.4 (5.40)	0.79
UCLA Loneliness (4 item)	11.97 (1.83)	10.78 (2.19)	0.67	10.68 (2.23)	0.60
Behavioral Activation (BADS)	18.57 (7.39)	25.84 (9.55)	0.79	27.11 (10.3)	0.87
Perceived social support	43.22 (16.41)	50.97 (18.22)	0.55	50.46 (18.39)	0.56
How many people can you depend on or feel very close to?	1.84 (1.41)	2.17 (1.49)	0.25	2.18 (1.61)	0.30
Times per week spent with someone who does not live with you	1.43 (1.71)	2.17 (1.95)	0.34	2.05 (2.04)	0.38
Times per week talked with someone (friends, relatives, or others) on the telephone	1.98 (2.01)	2.32 (2.09)	0.20	2.41 (2.04)	0.20
Times per week attended meetings of clubs, religious groups, or other groups you belong to	0.44 (0.92)	0.61 (1.10)	0.15	0.57 (0.98)	0.15

Note: All differences significant $p < .05$

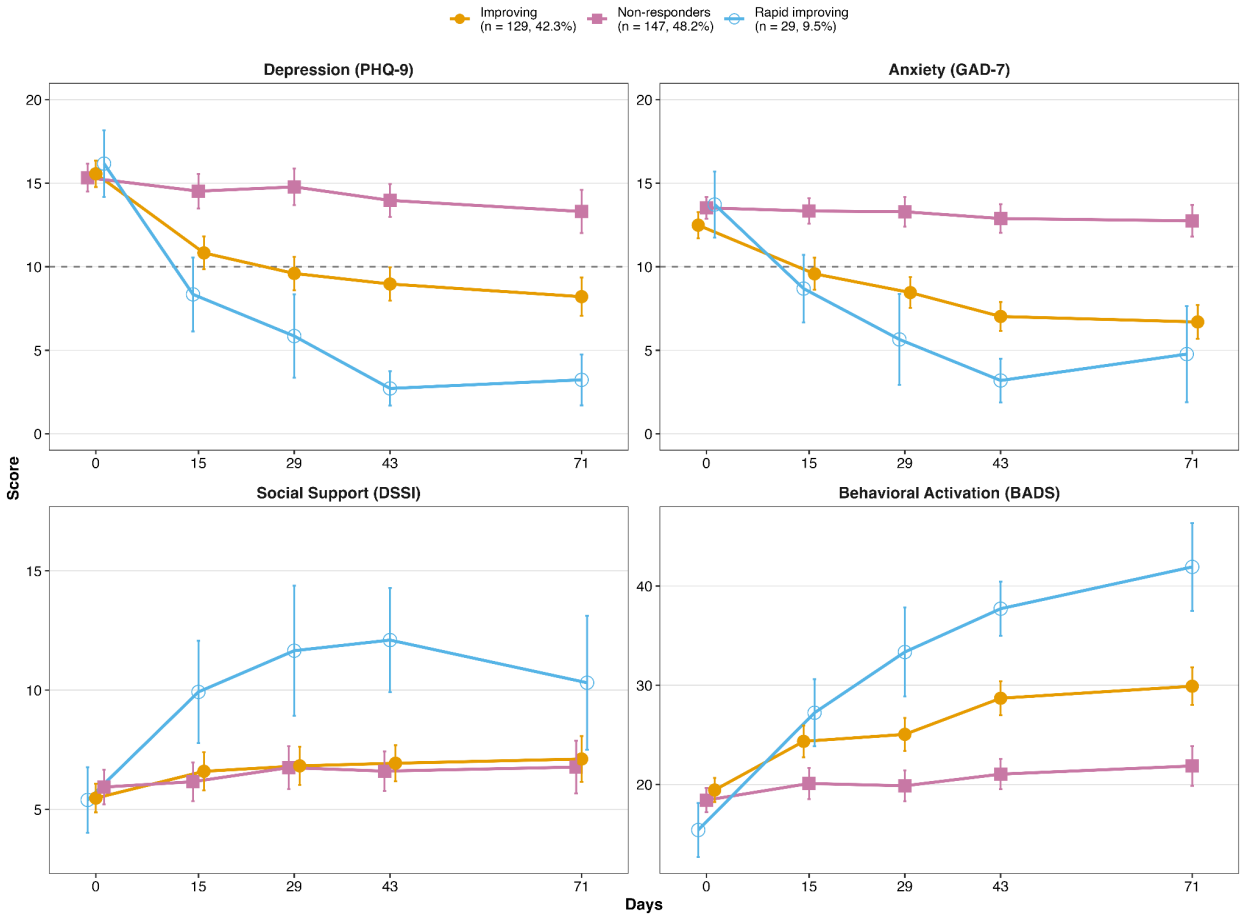
Results from the parallel process LGMM indicated three longitudinal patterns across depression (PHQ-9), anxiety (GAD-7), social support (DSSI), and behavioral activation (BADS) over 6 weeks of treatment (days 0-43) and 4 week follow-up (day 71). The LGMM accounted for a substantial share of observed variance explained (R^2) across time: 0.65-0.70 for PHQ-9,

0.58-0.69 for GAD-7, 0.66–0.71 for DSSI, and 0.52-0.66 for BADS. LGMM fit criteria and model selection are reported in the supplementary materials.

Figure 2 depicts the three latent classes: Rapid improving (n=29, 9.5%), Improving (n=129, 42.3%), and Non-responders (n=147, 48.2%). Baseline symptom severity was comparable across classes, with differences driven by rate and synchrony of change. In the Rapid improving class, PHQ-9 and GAD-7 slopes were large and negative (PHQ-9= −7.08, p=.016; GAD-7 = −5.41, p=.041) with a quadratic term consistent with a steep early decline (PHQ-9: 0.93, p=.024), while social support and behavioral activation slopes were positive (DSSI = +4.62, p=.043; BADS = +11.28, p<.001). In the Improving class, slopes were smaller, characterized by a gradual change over time. The Non-responder class had flat slopes that were consistent with minimal change over time. The Rapid improving trajectory fell below the clinical threshold for PHQ-9 and GAD-7 by week 2, the Improving trajectory by week 4, whereas the Non-responders trajectory remained above the cutoff at all time points.

We used a multinomial logistic regression to identify early characteristics discerning patients more likely to be in the Improving, Rapid Improving or Non-responders classes. Results indicated that the Rapid improving class was strongly associated with higher therapeutic alliance (OR = 3.32, 95% CI 1.78–6.19, p<.001) compared to Non-responders. Stronger therapeutic alliance was also associated with greater odds of Improving (OR = 1.33; 95% CI: 1.01, 1.75; p=.042) compared to Non-responders. Rapid Improving also had stronger therapeutic alliance than Improving (OR = 2.50; 95% CI: 1.34, 4 .65; p=.004) but lower odds of a higher income (OR 0.50; 95% CI 0.29,0.87; p=.014). The Improving class was more likely to be older (OR = 1.70; 95% CI: 1.28,2.26; p<.001) and have higher income (OR = 1.41; 95% CI 1.05,1.90; p=.021) but less education (OR = 0.73; 95% CI 0.54, 0.98; p=.038) relative to Non-responders. Concurrent psychiatric or psychotherapy treatment were not significant predictors of the LGMM trajectories. The full results from the multinomial logistic regression are reported in the supplementary materials.

Figure 2. Parallel Growth Model of Depression, Anxiety, Social Support, and Life Satisfaction (n = 305).



Note. User subgroups were identified based on longitudinal patterns across all four measures simultaneously. Error bars indicate 95% confidence intervals. PHQ-9: Patient Health Questionnaire 9 item; GAD-7: Generalized Anxiety Disorder Scale; DSSI: Duke Social Support Index; BADS: Behavioral Activation for Depression Scale Short Form.

Therapeutic Alliance Outcomes

Working alliance scores were 3.96 (SD = 0.97) for the entire measure, 4.09 (SD = 0.99) for Bond, 3.81 (SD = 0.97) for Task, and 3.97 (SD = 0.95) for Goal, scores comparable to traditional psychotherapy (Munder et al., 2010) and higher than for previous generation rule-based services (Darcy et al., 2021) and internet-based CBT (Jasper et al., 2014).

Social Connection

Loneliness, perceived social support, and social interaction counts all showed sustained improvements throughout the 6 week intervention period and were sustained at the 10 week follow-up (see Table 2 for more detail).

Engagement and Process Outcomes

Ninety-seven (97%) of users held at least one session in weeks 1 and 2, 84% in weeks 3 and 4, 71% in weeks 5 and 6, and 43% in weeks 7 through 10. Of those no longer holding sessions before 6 weeks, 22.3% had improved prior to disengagement, reflecting a "good enough" pattern of engagement (Falkenstrom et al., 2016). Attended sessions totalled 633 messages on average (SD = 928), and users averaged 18 distinct days (SD = 13) of interaction with the model. Average time in conversation with the model for weeks 1 and 2 combined was 247.8 minutes, 143.2 minutes in weeks 3 and 4, 93.5 minutes in weeks 5 and 6, and 56.9 minutes in weeks 7 through 10.

Sixty-two percent (61.9%) of users reported making progress on the goals they set for working with the model, 33.5% reported completely achieving their goals, and 4.6% reported making no progress by 6 weeks.

Items reflecting internalization, with human therapy comparators, were rated as highly characteristic of the experience with Ash, including imagining a particular quality to Ash's voice (M = 5.5 vs 5.7), thinking of Ash making specific statements (M = 5.7 vs. 4.7), sensing that Ash was "with them" emotionally (M = 5.4 vs. 4.5), trying to solve problems as they had worked on them together (M = 6.1 vs. 4.9), and recalling Ash during moments of distress or confusion (endorsed by 75% of users, compared to 80% in human therapy). Participants were less likely to picture Ash in an app-based or office-like context (M = 4.0) than past studies have shown for human therapy (M = 5.2).

Safety

Of the 7,493 total sessions, 76 (1.02%) were flagged by safety classifiers as containing potential crisis content. A random sample of 1,200 sessions were checked for crisis flags missed by the escalation classifier, none were found.

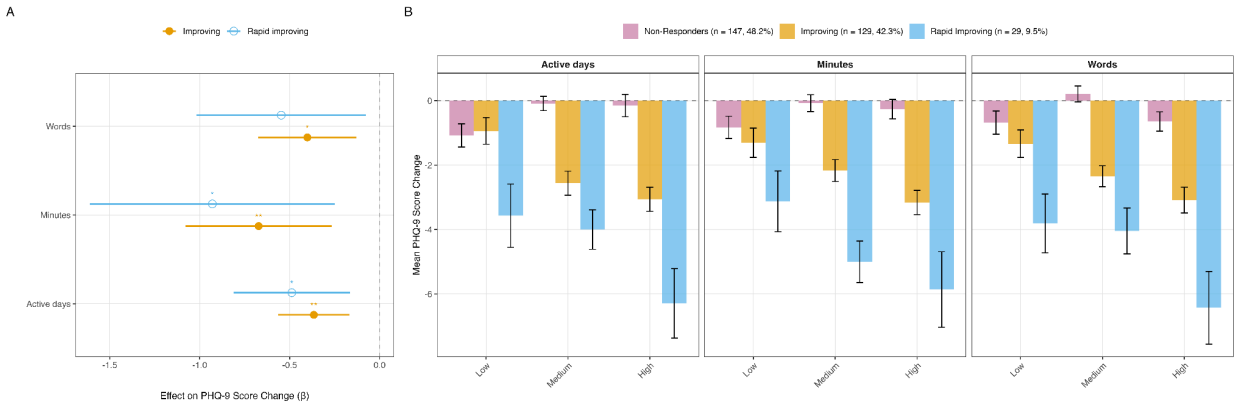
Predictive Models

Week 2 WAI scores significantly predicted improvement for Week 6 PHQ-9 (β = -0.16, SE = 0.04, r^2 = 0.09, $p < .001$) and GAD-7 (β = -0.15, SE = 0.03, r^2 = 0.11, $p < .001$), indicating higher alliance was associated with greater reductions in anxiety and depression, replicating a common finding for traditional psychotherapy (Falkenström, Granström & Holmqvist, 2014).

We examined the association between model engagement, LGMM class, and reductions of depressive symptoms within a 2-week window. We fit linear mixed-effects models with random participant intercepts and fixed effects for time (0–2, 2–4, 4–6 weeks; Non-responders class as the reference). We identified trends that were significant after Bonferroni correction in the interaction between engagement and LGMM class (Figure 4).

Figure 3. GAI Model Engagement and Depression Improvement.

A. Hierarchical mixed-effects model coefficients of PHQ-9 temporal changes by time-varying engagement and LGMM class. **B.** Mean PHQ-9 score difference from baseline to week 6 by overall engagement and LGMM response class.



Note. Negative coefficients and means indicate PHQ-9 reduction with higher engagement. 4A. LGMM reference class is Non-responders. Error bars represent 95% confidence intervals. Markers reflect Bonferroni corrected p-values (*<.05; **<.01). 4B. Low, Medium, and High indicate tertiles (0-33rd, 34th-67th, 68th-100th percentiles). Error bars represent one standard deviation. PHQ-9: Patient Health Questionnaire 9 item.

For the Improving group, larger reductions in PHQ-9 scores were associated with increased active days (β = -0.365, 95% CI: -0.563, -0.167; p = .002), minutes of use (β = -0.672; 95% CI: -1.079, -0.266; p = .007), and words exchanged (β = -0.402, 95% CI: -0.674, -0.129; p = .024) compared to Non-responders. Similarly, the Rapid Improving group demonstrated significant PHQ-9 reductions associated with increased number of active days (β = -0.488; 95% CI: -0.812, -0.165; p = .019) and minutes of use (β = -0.930; 95% CI: -1.610, -0.249; p = .045), with a trend for words exchanged that did not meet the corrected treshold (β = -0.547; 95% CI: -1.019, -0.075; p = .139). Full coefficients and model and specifications are reported in the Supplementary Materials.

Discussion

This naturalistic, single arm cohort study examined symptom trajectories among users of a mental health generative AI model alongside measures of therapeutic process and social functioning. Unlike trials that constrain interaction patterns, this study mirrors how GAI technology would be used in practice, where users engaged with Ash, a mental health GAI model, under naturalistic, real-world conditions. A key finding was the apparent safety of the model and the effectiveness of the safety guardrails. This is noteworthy in light of persistent concerns about the unpredictable or inappropriate behavior of general-purpose language models

in health contexts (McBain et al., 2025; Moore et al., 2025). Unlike generic models, Ash was trained on target-relevant data, and implemented with mental health safety guardrails and classifiers. This finding reinforces that context-aware training on appropriate data and within-domain guardrails, rather than post hoc prompting alone, is essential for deploying LLMs in therapeutic settings.

In addition to safety, we also found preliminary evidence of effectiveness in the use of a stand alone GAI chatbot designed for the management of depression and anxiety. We observed symptom reduction across measures of depression and anxiety, and importantly, we found improvements in perceived social support, loneliness, and objective social interaction. The latter finding is notable, given the concerns that have been raised about the potential misuse of chatbots to substitute for important in-person interactions (Coghlan et al., 2023). This suggests that GAI models that draw attention to existing social resources, and help build skills to access them, can meaningfully support social health. Our finding aligns with emerging models of digital interventions that serve as relational bridges rather than relational substitutes (Baumel et al., 2020). Importantly, consistent engagement with the model predicted greater symptom reduction highlighting the potential for better understanding dosage and ideal usage patterns.

The study also revealed that several core therapeutic processes, specifically, social interaction, behavioral activation, and working alliance were significant predictors of outcome. These findings echo decades of psychotherapy research demonstrating the importance of these constructs across modalities and diagnostic groups (Burkhardt et al., 2021; Wampold & Imel, 2015). Exploratory analyses identified internalization of the model as an additional process potentially important to GAI mental health support. Some users reported imagining Ash's voice or guidance during moments of difficulty, suggesting a degree of internal relational continuity. While speculative at this stage, this mirrors the concept of therapist internalization (Geller & Farber, 1993) and may represent an unexpected therapeutic mechanism for AI-mediated care.

As with any observational design, caution is warranted when interpreting results. Without randomization or an active control, improvements may reflect spontaneous remission or placebo effects rather than causal impact. However, quantities for these effects have been reported as helpful comparators. First, Cuijpers et al. (2021) report a pooled 16% average symptom improvement among control groups in psychotherapy trials. Second, Hosono et al. (2025) reports a placebo effect size of $g = 0.28$ for digital interventions in a recent randomized trial meta-analysis. Third, Mekonen et al. (2021) estimate spontaneous remission rates at 12.5% for depressive symptoms. In contrast, the rates of symptom reduction and functional improvement observed here exceeded all three benchmarks, suggesting that AI-specific therapeutic effects may be operative. Still, future randomized controlled trials are needed to isolate causal effects and control for expectancy and nonspecific factors.

Limitations include the absence of structured clinical diagnoses, a relatively short follow-up window, the lack of content-level analysis of conversations with the GAI model, and the lack of a control condition to compare effect sizes within the sample. Future research should incorporate diagnostic information, longer-term follow-up, a control condition, and fine-grained language analysis to investigate how specific model behaviors influence user change. Further, the emergent process of model internalization requires theoretical development and empirical validation.

Despite these limitations, this study provides one of the first real-world demonstrations of a safe, clinically grounded, and therapeutically responsive generative AI model in mental health. By grounding the foundation model in psychotherapy data, and evaluating it across multiple outcome domains, this study advances a new class of contextual, scalable, and clinically aligned AI mental health interventions with the potential to expand access while preserving meaningful therapeutic change.

Availability of Data and materials

Quantitative data are available by reasonable request and a data use agreement. Transcript data are not available at this time due to the sample size and sensitivity of the data.

References

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. International journal of medical informatics, 132, 103978.

Andersson, G., Cuijpers, P., Carlbring, P., Riper, H., & Hedman-Lagerlöf, E. (2014). Guided internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: A systematic review and meta-analysis. World Psychiatry, 13(3), 288–295.

Andrade, L. H., Alonso, J., Mneimneh, Z., et al. (2014). Barriers to mental health treatment: Results from the WHO World Mental Health surveys. Psychological Medicine, 44(6), 1303–1317.

Areán, P. A., Pullmann, M. D., Griffith Fillipo, I. R., Wu, J., Mosser, B. A., Chen, S., ... & Hull, T. D. (2024). Randomized Trial of the Effectiveness of Videoconferencing-Based Versus Message-Based Psychotherapy on Depression. Psychiatric Services, 75(12), 1184-1191.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of statistical software, 67, 1-48.

Baumel, A., Fleming, T., & Schueller, S. M. (2020). Digital micro interventions for behavioral and mental health gains: core components and conceptualization of digital micro intervention care. Journal of medical Internet research, 22(10), e20631.

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., ... & Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire–II: A revised measure of psychological inflexibility and experiential avoidance. Behavior Therapy, 42(4), 676–688.

Burkhardt, H. A., Alexopoulos, G. S., Pullmann, M. D., Hull, T. D., Areán, P. A., & Cohen, T. (2021). Behavioral activation and depression symptomatology: longitudinal assessment of linguistic indicators in text-based therapy sessions. Journal of Medical Internet Research, 23(7), e28244.

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: An updated systematic review and meta-analysis. Cognitive Behaviour Therapy, 47(1), 1–18.

Clement, S., Schauman, O., Graham, T., Maggioni, F., Evans-Lacko, S., Bezborodovs, N., ... & Thornicroft, G. (2015). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. Psychological medicine, 45(1), 11-27.

Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S. To chat or bot to chat: Ethical issues with using chatbots in mental health. Digit Health. 2023 Jun 22;9:20552076231183542. doi: 10.1177/20552076231183542.

Cuijpers, P., Karyotaki, E., Ciharova, M., Miguel, C., Noma, H., & Furukawa, T. A. (2021). The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: A meta‐analysis. Acta Psychiatrica Scandinavica, 144(3), 288-299.

Cuijpers, P., Miguel, C., Ciharova, M., Ebert, D., Harrer, M., & Karyotaki, E. (2023). Transdiagnostic treatment of depression and anxiety: a meta-analysis. Psychological Medicine, 53(14), 6535-6546.

Darcy, A., Daniels, J., Salinger, D., Wicks, P., & Robinson, A. (2021). Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. JMIR Formative Research, 5(5), e27868.

Falkenström, F., Granström, F., & Holmqvist, R. (2014). Working alliance predicts psychotherapy outcome even while controlling for prior symptom improvement. Psychotherapy Research, 24(2), 146-159.

Falkenström, F., Josefsson, A., Berggren, T., & Holmqvist, R. (2016). How much therapy is enough? Comparing dose-effect and good-enough models in two different settings. Psychotherapy, 53(1), 130.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR mental health, 4(2), e7785.

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR mental health, 5(4), e9782.

Geller, J., & Farber, B. (1993). Factors influencing the process of internalization in psychotherapy. Psychotherapy Research, 3(3), 166-180.

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. Psychotherapy research, 16(1), 12-25.

Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., ... & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. Nejm Ai, 2(4), AIoa2400802.

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. Cognitive Therapy and Research, 36(5), 427–440.

Hosono, T., Tsutsumi, R., Niwa, Y., & Kondoh, M. (2025). Magnitude of the Digital Placebo Effect and Its Moderators on Generalized Anxiety Symptoms: Systematic Review and Meta-Analysis. Journal of Medical Internet Research, 27, e74905.

Hughes, M. E., Waite, L. J., Hawkley, L. C., & Cacioppo, J. T. (2004). A short scale for measuring loneliness in large surveys: Results from two population-based studies. Research on aging, 26(6), 655-672.

Hull, T. D., Malgaroli, M., Connolly, P. S., Feuerstein, S., & Simon, N. M. (2020). Two-way messaging therapy for depression and anxiety: longitudinal response trajectories. BMC psychiatry, 20(1), 297.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. Journal of statistical software, 82, 1-26.

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR mHealth and uHealth, 6(11), e12106.

Ovsyannikova, D., de Mello, V. O., & Inzlicht, M. (2025). Third-party evaluators perceive AI as more compassionate than expert humans. Communications Psychology, 3(1), 4.

Jasper, K., Weise, C., Conrad, I., Andersson, G., Hiller, W., & Kleinstäuber, M. (2014). The working alliance in a randomized controlled trial comparing Internet-based self-help and face-to-face cognitive behavior therapy for chronic tinnitus. Internet interventions, 1(2), 49-57.

Kazdin, A. E., & Rabbitt, S. M. (2013). Novel models for delivering mental health services and reducing the burdens of mental illness. Clinical Psychological Science, 1(2), 170-191.

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. Community Mental Health Journal, 4(6), 443–453.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606–613.

Linardon, J., Cuijpers, P., Carlbring, P., Messer, M., & Fuller-Tyszkiewicz, M. (2019). The efficacy of app-supported smartphone interventions for mental health problems: A meta-analysis of randomized controlled trials. World Psychiatry, 18(3), 325–336.

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Medical care, 46(3), 266-274.

Malgaroli, M., Schultebraucks, K., Myrick, K. J., Loch, A. A., Ospina-Pinillos, L., Choudhury, T., Kotov., R., De Choudhury, M., & Torous, J. (2025). Large language models for the mental health community: framework for translating code to care. *The Lancet Digital Health*, *7*(4), e282-e285.

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. Cmaj, 184(3), E191-E196.

Manos, R. C., Kanter, J. W., & Luo, W. (2011). The behavioral activation for depression scale–short form: development and validation. Behavior therapy, 42(4), 726-739.

McBain, R. K., Cantor, J. H., Zhang, L. A., Baker, O., Zhang, F., Burnett, A., ... & Yu, H. (2025). Evaluation of Alignment Between Large Language Models and Expert Clinicians in Suicide Risk Assessment. Psychiatric services, 76(11), 944-950.

Mekonen, T., Ford, S., Chan, G. C., Hides, L., Connor, J. P., & Leung, J. (2022). What is the short-term remission rate for people with untreated depression? A systematic review and meta-analysis. Journal of Affective Disorders, 296, 17-25.

Mohr, D. C., Weingardt, K. R., Reddy, M., & Schueller, S. M. (2017). Three problems with current digital mental health research… and three things we can do about them. Psychiatric Services, 68(5), 427–429.

Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025, June). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 599-627).

Munder, T., Wilmers, F., Leonhart, R., Linster, H. W., & Barth, J. (2010). Working Alliance Inventory‐Short Revised (WAI‐SR): psychometric properties in outpatients and inpatients. Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice, 17(3), 231-239.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal*, *14*(4), 535-569.

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., ... & Unützer, J. (2018). The Lancet Commission on global mental health and sustainable development. The Lancet, 392(10157), 1553-1598.

Pullmann, M. D., Rouvere, J., Raue, P. J., Fillipo, I. R. G., Mosser, B. A., Heagerty, P. J., ... & Areán, P. A. (2025). Message-Based vs Video-Based Psychotherapy for Depression: A Randomized Clinical Trial. JAMA Network Open, 8(10), e2540065-e2540065.

Russell, D., Peplau, L. A., & Cutrona, C. E. (1980). The revised UCLA Loneliness Scale: concurrent and discriminant validity evidence. Journal of personality and social psychology, 39(3), 472.

Snyder, C. R., Sympson, S. C., Ybasco, F. C., Borders, T. F., Babyak, M. A., & Higgins, R. L. (1996). Development and validation of the State Hope Scale. Journal of personality and social psychology, 70(2), 321.

Song, J., Litvin, B., Allred, R., Chen, S., Hull, T. D., & Areán, P. A. (2023). Comparing message-based psychotherapy to once-weekly, video-based psychotherapy for moderate depression: Randomized controlled trial. Journal of Medical Internet Research, 25, e46052.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. Archives of internal medicine, 166(10), 1092-1097.

Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., ... & Kessler, R. C. (2017). Undertreatment of people with major depressive disorder in 21 countries. The British Journal of Psychiatry, 210(2), 119-124.

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. The Canadian Journal of Psychiatry, 64(7), 456-464.

Wampold, B. E., & Imel, Z. E. (2015). The great psychotherapy debate: The evidence for what makes psychotherapy work. Routledge.

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., ... & Wells, J. E. (2007). Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. The Lancet, 370(9590), 841-850.

Wardian, J., Robbins, D., Wolfersteig, W., Johnson, T., & Dustman, P. (2013). Validation of the DSSI-10 to measure social support in a general population. *Research on Social Work Practice*, 23(1), 100-106.

World Health Organization. (2017). Depression and other common mental disorders: Global health estimates. Geneva: World Health Organization.

Zimet, G. D., Dahlem, N. W., Zimet, S. G., & Farley, G. K. (1988). The multidimensional scale of perceived social support. Journal of personality assessment, 52(1), 30-41.

Supplementary Materials

Latent Grown Mixture Modeling

Model Selection.

As shown in Supplementary Table 1, information criteria decreased from 1 to 4 classes, and the BLRT remained significant at each step, indicating incremental statistical gains with added classes. However, entropy did not improve across the 2 to 4 classes solutions (entropy = .71), and the 4-class model included a very small class and yielded crossing trajectories consistent with overfitting. We therefore selected the unconditional model with one less latent class. Relative to the 2-class solution, the 3-class model was supported by a significant Bootstrapped Likelihood Ratio test, lower AIC/BIC (AIC: 26231.68 vs. 26319.06; BIC: 26484.66 vs. 26523.68). Average posterior probabilities ranged from .85 (distinguishing Improving from Non-responders) to .94 (Rapid Improving from all others). Diagonal classification probabilities were .91 (Rapid Improving), .83 (Improving), and .88 (Non-responders).

Supplementary Table 1.

Model Fit Indices for 1 to 4 Classes of Parallel Latent Growth Mixture Models.

	1-class	2-classes	3-classes	4-classes
Akaike Information Criteria	26706.66	26319.06	26231.68	26147.83
Bayesian Information Criteria (BIC)	26848.03	26523.68	26484.66	26449.18
Sample-Size Adjusted BIC	26727.51	26349.24	26269.01	26192.28
Entropy	-	.71	.71	.71
Smallest class (n / full sample)	1	0.343	0.097	0.045
Bootsrapped LRT P-value	-	*<.0001*	*<.0001*	*<.0001*

Supplementary Table 2.

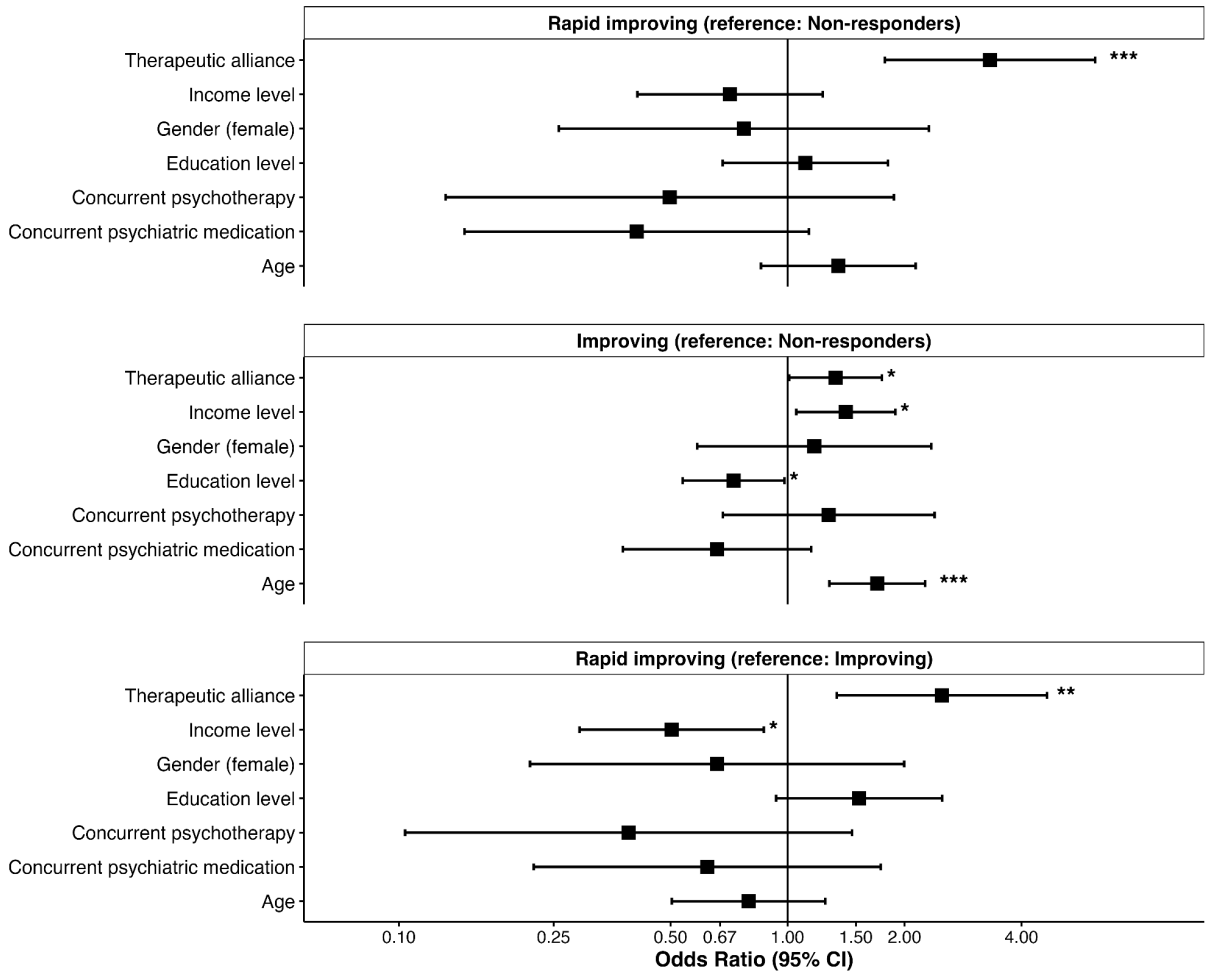
Demographics and Clinical Variables by LGMM class.

	Non-Responders (n = 147)	Improving (n = 129)	Rapid Improving (n = 29)
Age, years	39.2 (10.8)	44.2 (12.2)	43.0 (13.7)
Gender			
Woman	120 (81.6%)	109 (84.5%)	22 (75.9%)
Man	17 (11.6%)	15 (11.6%)	6 (20.7%)
Non-binary	10 (6.8%)	5 (3.9%)	1 (3.4%)
Race/Ethnicity			
Arabic / Middle Eastern	1 (0.7%)	0 (0.0%)	1 (3.4%)
Asian / East Asian	7 (4.8%)	1 (0.8%)	2 (6.9%)
Black / African American	8 (5.4%)	6 (4.7%)	3 (10.3%)
Hispanic / Latino	2 (1.4%)	4 (3.1%)	1 (3.4%)
Multiple	0 (0.0%)	2 (1.6%)	0 (0.0%)
Native American / Alaska Native	0 (0.0%)	0 (0.0%)	1 (3.4%)
Native Hawaiian / Pacific Islander	0 (0.0%)	0 (0.0%)	1 (3.4%)
Prefer not to say	3 (2.0%)	0 (0.0%)	0 (0.0%)
South Asian / Indian	1 (0.7%)	3 (2.3%)	1 (3.4%)
White / Caucasian	125 (85.0%)	113 (87.6%)	19 (65.5%)
Household income			
Less than $24,999	43 (30.9%)	27 (22.5%)	13 (44.8%)
$25,000 - $49,999	36 (25.9%)	32 (26.7%)	8 (27.6%)
$50,001 - $74,999	24 (17.3%)	17 (14.2%)	1 (3.4%)

	Non-Responders (n = 147)	Improving (n = 129)	Rapid Improving (n = 29)
$75,000 - 99,999	15 (10.8%)	15 (12.5%)	5 (17.2%)
$100,000 - 149,999	10 (7.2%)	24 (20.0%)	2 (6.9%)
$150,000 - 199,999	7 (5.0%)	4 (3.3%)	0 (0.0%)
$200,000 or more	4 (2.9%)	1 (0.8%)	0 (0.0%)
Education level			
Less than High School	5 (3.6%)	3 (2.4%)	1 (3.6%)
High School Degree	41 (29.5%)	36 (29.3%)	6 (21.4%)
Associated Degree	16 (11.5%)	22 (17.9%)	9 (32.1%)
Bachelor's Degree	45 (32.4%)	35 (28.5%)	6 (21.4%)
Master's Degree	24 (17.3%)	24 (19.5%)	5 (17.9%)
Doctorate Degree	8 (5.8%)	3 (2.4%)	1 (3.6%)
Concurrent Psychotherapy	37 (25.2%)	33 (25.6%)	3 (10.3%)
Concurrent Psychiatric Medication	61 (41.5%)	44 (34.1%)	6 (20.7%)
Therapeutic Alliance (week 2)	42.4 (11.3)	45.6 (10.3)	53.0 (7.6)

Note. Mean (SD) for continuous variables; n (%) for categorical variables

Supplementary Figure 1. Forest plot of multinomial logistic regression of LGMM class (n=305).



Note. Reference: reference class for the multinomial logistic regression; 95% CI: 95% confidence interval of odds ratio; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Supplementary Table 3.

Odds ratio of multinomial logistic regression of LGMM class predictors (n = 305).

	Improving vs Non-responders		**Rapid improving vs Non-responders**		**Rapid improving vs Improving**	
	OR [95% CI]	*p*	OR [95% CI]	*p*	OR [95% CI]	*p*
Age	**1.70 [1.28, 2.26]**	**<.001*** **	1.35 [0.85, 2.14]	.198	0.79 [0.50, 1.25]	.320
Gender (female)	1.17 [0.59, 2.35]	.654	0.77 [0.26, 2.31]	.644	0.66 [0.22, 2.00]	.461
Education	**0.73 [0.54, 0.98]**	**.038***	1.11 [0.68, 1.81]	.674	1.53 [0.93, 2.50]	.091
Income	**1.41 [1.05, 1.90]**	**.021***	0.71 [0.41, 1.23]	.224	**0.50 [0.29, 0.87]**	**.014***
Concurrent psychotherapy	1.28 [0.68, 2.39]	.446	0.50 [0.13, 1.88]	.303	0.39 [0.10, 1.47]	.163
Concurrent psychiatric medication	0.66 [0.38, 1.15]	.142	0.41 [0.15, 1.14]	.086	0.62 [0.22, 1.74]	.365
Therapeutic alliance	**1.33 [1.01, 1.75]**	**.042***	**3.32 [1.78, 6.19]**	**<.001*** **	**2.50 [1.34, 4.65]**	**.004****

Note. OR: Odds Ratio; 95% CI: 95% confidence interval of odds ratio; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Sensitivity Analysis.

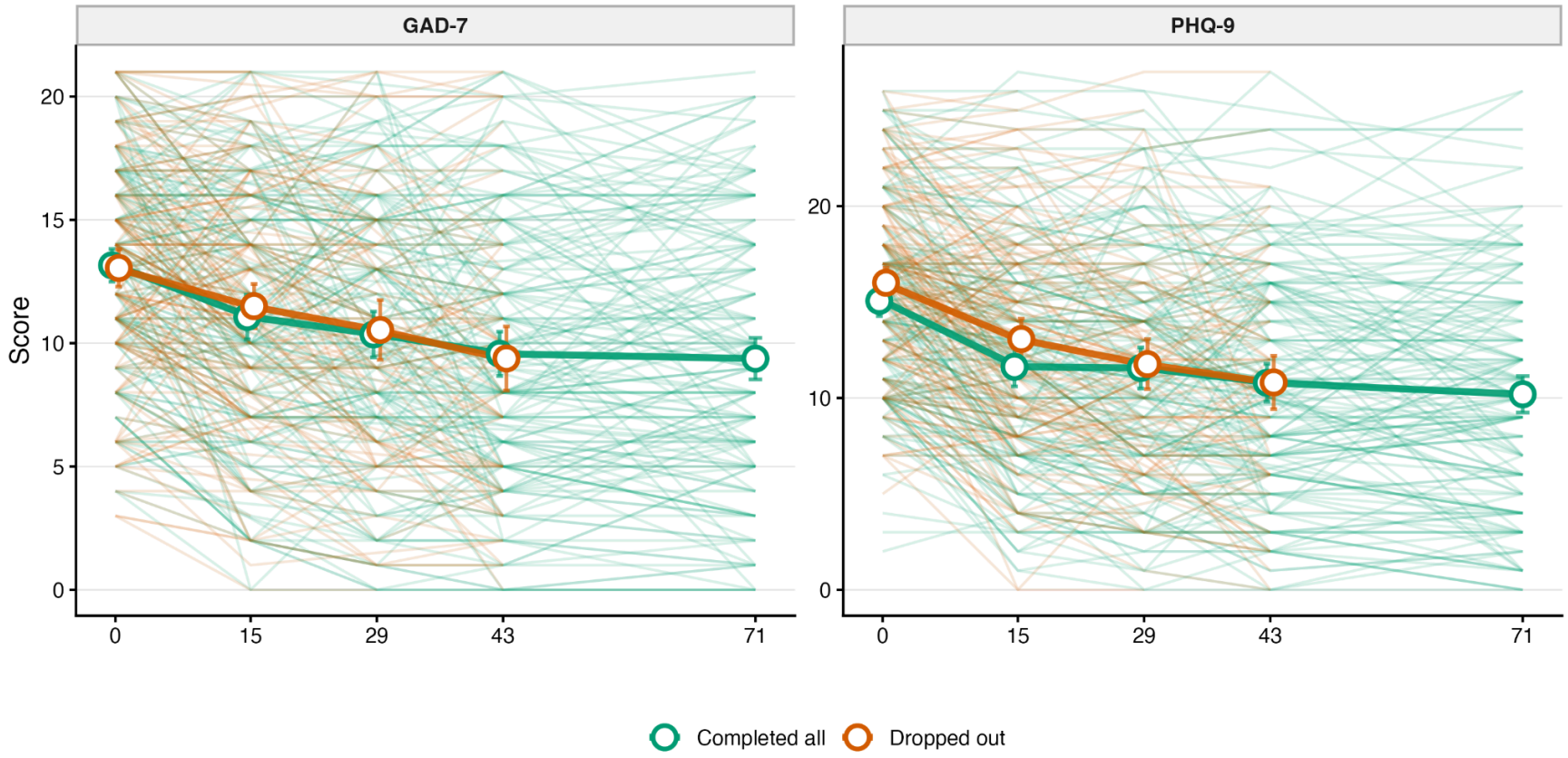
Supplementary Table 4.

Sensitivity analyses for survey non-adherence using demographic and clinical characteristics.

	Odds Ratio	95% CI	P value
Age	0.989	[0.968, 1.012]	.354
Gender (female)	0.835	[0.441, 1.581]	.581
Income	1.023	[0.862, 1.214]	.793
Education	0.945	[0.755, 1.184]	.624
Concurrent Psychotherapy	1.731	[0.959, 3.125]	.069
Concurrent Psychiatric Medication	1.433	[0.850, 2.416]	.177
GAD-7: Baseline	0.939	[0.866, 1.020]	.135
Last observation	1.026	[0.945, 1.114]	.536
PHQ-9: Baseline	1.024	[0.954, 1.098]	.513
Last observation	1.027	[0.956, 1.103]	.470

Note. Reference class is completers (n = 160). OR: Odds Ratio; 95% CI: 95% confidence interval of odds ratio; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Supplementary Figure 2. Individual and mean trajectories of anxiety and depression for completers and survey non-adherent users (n=305).



Note. Error bars represent 95% confidence intervals. PHQ-9: Patient Health Questionnaire 9 item; GAD-7: Generalized Anxiety Disorder Scale.

Engagement Analysis.

Engagement metrics consistent of active days, number of minutes, and amount of words while using the intervention in a two-week temporal window in between two assessments. First, we ran exploratory analysis to identify overall engagement effects for the interventions. We fitted separate linear mixed-effects models for each engagement metric to identify PHQ-9 and GAD-7 changes between two time points (differences in scores between: week 0 to week 2, week 2 to week 4, and week 4 to week 6). The mixed effect model was as follows:

$$\Delta Y_{ij} = \beta_0 + \beta_1 \, \text{Engagement}_{ij} + u_i + \varepsilon_{ij}$$

where ΔYij denotes the PHQ-9 or GAD-7 change for user *i* in window *j*, Engagement_{ij} is the corresponding user's engagement metric, u_i is a random intercept capturing within-person correlation, and ε_{ij} is the residual error. P-values were obtained with the package lmerTest (Kuznetsova et al., 2017). Mixed-effects models were estimated by maximum likelihood using the package lme4 (Bates et al., 2015). Models were fitted by maximum likelihood and fixed effect significance was obtained via the Satterthwaite approximation for degrees of freedom as implemented in lmerTest. We applied a natural logarithm plus 1 transformation before entering minutes and words. To control for multiple testing across the six analyses (three engagement metrics and clinical two outcomes) we applied a Bonferroni correction ($\alpha = 0.05/6 = 0.0083$).

Supplementary Table 5.

Full sample mixed effect-model of engagement effects for Anxiety and depression (n = 305).

Outcome	Engagement	β (95% CI)	p-value	Bonf. p
GAD-7	Active days	-0.040 (-0.133, 0.052)	0.395	1.000
	Minutes	0.019 (-0.173, 0.212)	0.843	1.000
	Words	0.057 (-0.071, 0.185)	0.383	1.000
PHQ-9	Active days	-0.152 (-0.255, -0.049)	0.004	0.024*
	Minutes	-0.180 (-0.395, 0.034)	0.100	0.600
	Words	-0.095 (-0.238, 0.049)	0.195	1.000

Note. Bonf. p: Bonferroni corrected p-values. 95% CI: 95% confidence interval; PHQ-9: Patient Health Questionnaire 9 item; GAD-7: Generalized Anxiety Disorder Scale;

Results from the initial analysis (Supplementary table 5) identified a significant association between engagement and change in depressive symptom severity (PHQ-9). The following analysis focused on depression and examined whether the relationship between engagement and PHQ-9 score changes varied across the longitudinal response trajectories that were identified by a parallel process latent growth mixture model (LGMM; see Trajectories in main methods). While initial analysis treated the intervention as a whole by collapsing all observations into a single change score, this analysis retained the three consecutive 2-week assessment windows as distinct observations and adjusted for temporal effects. For each engagement metric a hierarchical linear mixed-effects model was fitted with PHQ-9 as the dependent variable. Fixed effects comprised the engagement metric, LGMM class (with non-responders as the reference), an interaction term between engagement and class, the temporal window (weeks 0-2, 2-4, and 4-6), and a random intercept for participant accounted for the repeated-measures structure across windows. The model was estimated by maximum likelihood, and information criteria indicated model fit improvements by adding temporal effects (e.g., BIC model 1 = 3313.28; BIC model 2 = 3214.90) with marginal/conditional R^2 of 0.224/0.224. Observed temporal trends of engagement by LGMM class are reported in supplementary figure 3 and stratified based on tertiles in Figure 4b. The methods of the second mixed-effects model are further described in the main manuscript, and results are reported in the main manuscript and in supplementary table 6.

Supplementary Table 6.

Mixed effect-models of engagement effects on PHQ-9 scores stratified by time and LGMM class membership (n = 305).

	β	SE	95% CI	t	p	Bonf. p
Main effects: Engagement						
Active days	0.223	0.080	[0.066, 0.380]	2.79	0.005	—
Minutes (log)	0.425	0.160	[0.111, 0.739)	2.66	0.008	—
Words (log)	0.251	0.108	[0.039, 0.463)	2.33	0.020	—
Main effects: LGMM class						
Improving	-0.146	0.640	[-1.403, 1.111]	-0.23	0.819	—
Rapid Improving	-1.869	1.060	[-3.951, 0.213]	-1.76	0.078	—
Main effects: Time period						
2-4 weeks	2.761	0.396	[1.983, 3.538]	6.98	<0.001	—
4-6 weeks	2.502	0.406	[1.705, 3.300]	6.16	<0.001	—
Interaction effects						
Active days × Improving	-0.365	0.101	[-0.563, -0.167]	-3.63	<0.001	0.002**
Active days × Rapid Improving	-0.488	0.165	[-0.812, -0.165]	-2.97	0.003	0.019*
Minutes × Improving	-0.672	0.207	[-1.079, -0.266]	-3.25	0.001	0.007**
Minutes × Rapid Improving	-0.930	0.346	[-1.610, -0.249]	-2.68	0.007	0.045*
Words × Improving	-0.402	0.139	[-0.674, -0.129]	-2.89	0.004	0.024*
Words × Rapid Improving	-0.547	0.240	[-1.019, -0.075]	-2.28	0.023	0.139

Note. Reference categories: non-responders (LGMM class), 0-2 weeks (time). Bonf: Bonferroni corrected adjusted p-values (α = 0.0083). Markers (*) reflect corrected p-values. PHQ-9: Patient Health Questionnaire 9 item.

Supplementary Figure 3. Mean engagement from baseline to week 6 by LGMM class membership (n = 305). Error bars represent one standard deviation.