Review

Generative AI Mental Health Chatbots as Therapeutic Tools: Systematic Review and Meta-Analysis of Their Role in Reducing Mental Health Issues

Qiyang Zhang[1], PhD; Renwen Zhang[2], PhD; Yiying Xiong[3], PhD; Yuan Sui[3], MSEd; Chang Tong[3], MSEd; Fu-Hung Lin[3], MSEd

[1]Department of Educational Advancement, Duke-NUS Medical School, Singapore, Singapore
[2]Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore
[3]School of Education, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:

Qiyang Zhang, PhD
Department of Educational Advancement
Duke-NUS Medical School
8 College Road
Singapore 169857
Singapore
Phone: 65 66012186
Email: qiyang39@duke-nus.edu.sg

Abstract

Background: In recent years, artificial intelligence (AI) has driven the rapid development of AI mental health chatbots. Most current reviews investigated the effectiveness of rule-based or retrieval-based chatbots. To date, there is no comprehensive review that systematically synthesizes the effect of generative AI (GenAI) chatbot's impact on mental health.

Objective: This review aims to (1) narratively synthesize existing GenAI mental health chatbots' technical features, treatment and research designs, and sample characteristics through a systematic review of quantitative studies and (2) quantify the effectiveness and key moderators of these rigorously designed trials on GenAI mental health chatbots through a meta-analysis of only randomized controlled trials (RCTs).

Methods: The search strategy includes 11 database searching, backward citation tracking, and a manual ad hoc search to update literature. This thorough literature search, completed in March 2025, returned 5555 records for screening. The systematic review included studies that (1) used generative or hybrid (rule/retrieval-based and generative) AI-based chatbots to deliver interventions and (2) quantitatively measured mental health-related outcomes. The meta-analysis has additional inclusion criteria: (1) studies must be RCTs, (2) must measure negative mental health issues, (3) the comparison group must not have chatbot features, and (4) must provide enough statistics for effect size calculation. We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist and registered the protocol retrospectively during the revision process (September 18, 2025). In meta-regression, data were synthesized in R software using a random-effects model.

Results: The narrative synthesis of 26 studies revealed that (1) GenAI chatbot interventions mostly took place in non-WEIRD countries (non-Western, Educated, Industrialized, Rich, and Democratic) and (2) there is a lack of studies focusing on young children and older adults. The meta-analysis of 14 RCTs showed a statistically significant effect (effect size [ES]=0.30, *P*=.047, N=6314, 95% CI 0.004, 0.59, 95% prediction interval [PI] –0.85, 1.67), which means that GenAI chatbots are, on average, effective in reducing negative mental health issues, such as depression, anxiety, among others. We found that social-oriented chatbots (ie, those that mainly provide social interactions) are more effective than task-oriented programs (ie, those that assist with specific tasks). Risk of bias in the nonrandomized studies and RCTs was assessed using Cochrane ROBINS-I (Risk Of Bias In Non-randomised Studies – of Interventions) and RoB2 (revised Cochrane risk-of-bias tool for randomized trials), respectively, indicating a moderate amount of risk. One main limitation of this meta-analysis is the small number of studies (n=14) included.

Conclusions: By identifying research gaps, we suggest that future researchers investigate user groups such as adolescents and older adults, outcomes other than depression and anxiety, cultural adaptations in non-WEIRD countries, ways to streamline chatbots in usual care practices, and explore applications in diverse settings. More importantly, we cannot ignore GenAI chatbots' risks while acknowledging their promise. This review also emphasized several ethical implications.

Trial Registration: Open Science Framework 9DAJ7; https://osf.io/9daj7

J Med Internet Res 2025;27:e78238; doi: 10.2196/78238

Keywords: generative artificial intelligence; AI chatbot; conversational agents; mental health; meta-analysis; systematic review; artificial intelligence

Introduction

Background of Generative Artificial Intelligence (GenAI) Mental Health Chatbots

Globally, one in every eight individuals is affected by a mental health issue [1], which can significantly impair people's physical health [2] and wellness [3]. Mental health conditions impose an estimated $1.9 trillion [4] economic burden worldwide. Despite the high social cost and pressing need for treatment, access to mental health services remains severely limited [5]. In fact, over 70% of people with mental disorders receive no treatment from professionals due to stigma, counselors' shortage, and under-resourced care infrastructures [6]. The COVID-19 pandemic has further exacerbated these challenges, highlighting the urgent need for scalable, accessible, and cost-effective interventions [7].

Recent advances in artificial intelligence (AI) have driven the rapid development of mental health chatbots [8,9]. These AI chatbot-based interventions offer 24/7 support, enhanced self-management [10], reduced stigma, and appeal to digital-native users [11]. Multimedia Appendix 1 presents an overview of existing AI mental health chatbots. A number of reviews showed promising effects of AI chatbot interventions on reducing mental health distress and improving quality of life [10,12-16] (see Multimedia Appendix 2). However, most of these interventions relied on retrieval-based chatbots, which use predefined responses or static databases and often result in rigid, repetitive interactions [17-19].

In contrast, GenAI chatbots, powered by large language models (LLMs), such as GPT models, generate novel responses in real time. By tailoring replies to the user's language, tone, and emotional content, they enable conversations that are more natural, personalized, and emotionally resonant. This capacity may strengthen user engagement [20], therapeutic alliance [21], and the sense of being understood [22]. A recent meta-analysis shows that GenAI chatbots outperform rule-based and retrieval-based chatbots in reducing depressive symptoms [13]. Emerging primary studies suggest that GenAI mental health chatbots can improve engagement and adherence by supporting between-session cognitive behavioral therapy tasks [23] and delivering positive psychology interventions, such as gratitude or self-reflection exercises [24]. Beyond structured therapy, companion chatbots such as Replika have been found to reduce loneliness [25] and produce outcomes comparable to mindfulness interventions among older adults [26]. Despite this huge potential, no review currently exists that has systematically synthesized the effect of GenAI chatbots' impact on mental health.

Aims of This Review

This review seeks to fill this gap by conducting a systematic review and meta-analysis of GenAI chatbot interventions for mental health. This paper aims to:

1. synthesize current GenAI mental health chatbots' technical and treatment features, research designs, and sample characteristics,
2. quantify the effectiveness of these interventions via a meta-analysis of randomized controlled trials (RCTs)
3. examine key moderators of intervention effectiveness, including chatbot design, population characteristics, intervention context, and outcome types.

Methods

Search Strategy

To ensure comprehensive coverage of the literature, the first author (QZ) implemented a thorough search strategy that included database queries, updated manual searches, and backward citation tracking. Using a predefined set of keywords, the author systematically searched 11 databases, including Scopus, Embase, Web of Science, APA PsycInfo, Child Development & Adolescent Studies, ERIC, ACM Digital Library, CINAHL, MEDLINE, PsyArXiv, and OpenDissertations. We developed a predefined set of keywords on "method," "generative AI chatbot," and "mental health" (keywords detailed in Multimedia Appendix 3 and search details can be found in Multimedia Appendix 4). Furthermore, through a tool called *CitationChaser*, we conducted backward citation tracking for 9 similar reviews listed in Multimedia Appendix 2. The data search was completed on November 1, 2024. In order to update our search, on March 5, 2025, we conducted another round of manual ad hoc search to seek newly published studies. Together, 5555 records were identified from the combined search strategies.

Inclusion Criteria for Systematic Review

1. Studies should have used generative or hybrid (rule or retrieval-based and generative) AI-based conversational agent or chatbot to deliver interventions. For example, rule-based chatbots that formulate responses to user queries through a predetermined set of rules without employing any AI algorithms or techniques were excluded (eg [27]).

2. Studies must quantitatively measure mental health-related outcomes, including both positive and negative constructs.
3. These must be primary studies. We excluded review papers.
4. Text must be available on the internet or written in English.
5. Studies must be published on or after January 1, 2014, as the modern generative AI chatbot era began around 2015 with neural conversational models [28].

Inclusion Criteria for Meta-Analysis

Apart from the above criteria, eligible studies for meta-analysis must meet additional inclusion criteria:

1. The study must quantitatively measure negative mental health issues in the outcome, such as depression, anxiety, psychological distress, stress, etc. We excluded well-being, happiness, positive emotions, etc. due to the scope of this study. The study by Vowels et al [29] was excluded because of its focus on positive well-being only. We also excluded usability studies that measure outcomes using scales such as System Usability Scale.
2. Studies must be RCTs. We excluded single-arm studies. The study by Zheng [30] was excluded because the study design is not RCT.
3. The comparison group must not have chatbot features since the treatment group includes chatbot. We excluded studies where both control and treatment had chatbot features. For example, the study by Liu et al [24] was excluded because all the control groups used chatbots.
4. Studies must have sufficient data provided to calculate effect sizes. This means that studies must either directly provide effect sizes in Cohen *d* or Hedges *g* or they must provide pre- and postintervention mean and SD for both treatment and control groups. The study by Maples et al [31] was excluded because of insufficient data.

Screening

In conducting the screening process, we used Covidence software owing to its robust full-text review features and the provision of complimentary licenses available via our affiliated institutions [32]. Deduplication was handled both manually through Zotero (Corporation for Digital Scholarship) and through Covidence software. The screening of titles and abstracts, as well as the full-text review, was executed by using a double-blinded methodology to guarantee impartial evaluations. Four authors participated in the screening stage (YS, FL, CT, QZ). To resolve any conflicts, we held weekly meetings and reached a 100% consensus.

Data Extraction and Narrative Synthesis Approach

Before data extraction, a Microsoft Excel coding framework was developed a priori. Apart from straightforward variables, intervention *duration* was extracted as the total number of weeks the intervention was delivered. If *duration* was reported as days or months, we transformed the variable into weeks using 1 week=7 days=0.25 months. To represent sex, we extracted female percentage among participants and coded *Fifty percent female* as 1 if at least 50% of the sample identified as female and 0 otherwise. The *age* variable was split into early adulthood (18–30 y old), middle adulthood (30–50 y old), and late adulthood (more than 50 y old) based on mean age. Following the framework established by Beyebach et al [33], we extracted and categorized countries as WEIRD or non-WEIRD. Studies were coded as WEIRD if they were conducted in countries that fit this classification, while those that did not meet these criteria were labeled as non-WEIRD. Studies were coded as *customized* if the chatbot allows users to customize the user interface, such as changing the app’s background color, etc. Conversely, studies were coded as *non-customized* if the chatbot did not mention *customization* in the study. Studies were coded as *clinical* if participants were recruited from healthcare or clinical service settings (eg, hospitals, outpatient clinics, or counseling centers), regardless of whether their condition was primarily mental or physical. These participants typically had documented health concerns or were receiving clinical care at baseline. By contrast, studies that recruited participants from schools, universities, or the general community without requiring a diagnosed condition were coded as *non-clinical*. *Outcome* measures were coded according to the primary mental health construct assessed, including depression, anxiety, and stress.

Given the heterogeneity of study designs, interventions, and outcome measures across the included studies, a narrative synthesis was conducted to summarize and compare study characteristics systematically. Following the narrative synthesis guidance from Popay et al [34], we structured the synthesis around four analytical dimensions: (1) technical features of the chatbot systems (eg, AI architecture, delivery platform, modality, customization, and embodiment), (2) treatment features (eg, theoretical frameworks, intervention duration, target outcomes, and presence of human guidance), (3) research design characteristics (eg, study type, publication year, and methods), and (4) sample characteristics (eg, country, participant demographics, population type, and recruitment setting). Each study’s information was extracted by two out of three authors independently (YS, FL, CT). Conflicts were discussed and resolved through weekly team discussions with the first author (QZ) until full agreement was reached.

Analytic Plan

We used a random-effects model using the *metafor* package [35] in R statistical software (version 4.5.1, R Foundation for Statistical Computing). For weighted mean effect sizes, we assigned weights to each study based on inverse variance [36]. Several studies contributed multiple outcomes (eg, depression, anxiety, stress), which are statistically dependent because they are measured using the same participants. Our primary analysis used a multilevel random-effects meta-analysis with random intercepts at the study level and the outcome-within-study level [37]. The model was fit with restricted maximum likelihood (REML). To account for small-sample uncertainty in multilevel models, we used t-based inference with Satterthwaite-adjusted degrees of freedom by setting tdist=TRUE in rma.mv(). This approach provides an effect equivalent to the Hartung–Knapp–Sidik–Jonkman (HKSJ) adjustment and is particularly suitable for multivariate meta-regressions. We constructed a block-diagonal sampling variance-covariance matrix V with an assumed within-study correlation *r*=0.80. As a robustness check, we computed cluster-robust (CR2) standard errors with Satterthwaite degrees of freedom, clustering on study.

In addition to multivariate meta-analysis, as a sensitivity analysis, we fitted a univariate model using the HKSJ adjustment method. Since only 14 studies were included in the meta-analysis, we adopted HKSJ as it is recommended for meta-analysis with few studies [38]. In addition, the HKSJ method was found to outperform the standard DerSimonian-Laird method [39,40]. We fitted a random-effects model using the Sidik–Jonkman (SJ) estimator for the between-study variance (τ^2), which is robust to outliers and performs well when heterogeneity is substantial. We aggregated the dataset to a single effect size per study–outcome pair. Specifically, for each study-outcome pair, we computed the mean of the reported effect sizes and the mean of their reported sampling variances. For inference on the pooled effect, we applied the HKSJ adjustment.

We assessed residual dispersion with Cochran Q and its *P* value from the fitted model. For completeness, we computed I^2 as the proportion of total variability attributable to heterogeneity; however, following Borenstein et al [37], we emphasize that I^2 is not an absolute measure of heterogeneity and does not indicate the magnitude of variation in true effects across settings. We therefore also reported τ^2. Given that treatment effects over different settings may vary, we also reported 95% prediction intervals (PIs) for the true effect [41].

In line with open science principles, the complete dataset and R code are publicly available [42]. We registered the protocol retrospectively during the revision process (September 18, 2025), with Open Science Framework (10.17605/OSF.IO/9DAJ7) [43]. We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist. As for missing data, we either inferred from other relevant information or we reported them as NA in tables. When studies only provided post-test means and SDs, we assumed baseline equivalence and calculated Hedges *g*. When key statistics for calculating effect sizes were missing, we had to drop the study.

To evaluate publication bias, selection modeling was used. This approach used a weight function model created by Vevea and Woods [44], implemented through the *weightr* package. To assess the risk of bias for all 26 studies included in the systematic review, we adopted two Cochrane tools. For the nonrandomized studies included in the systematic reviews, we adopted the Cochrane ROBINS-I (Risk Of Bias In Non-randomised Studies – of Interventions) tool [45]. For the RCTs included in both the systematic reviews and the meta-analysis, we used the Cochrane RoB2 (revised Cochrane risk-of-bias tool for randomized trials) [46]. Two authors coded (CT, FL) independently, and a third author (QZ) resolved discrepancies.

Moderators

Considering the need for balanced moderator categories, theoretical and practical importance, small sample size (n=14), and the need for a degree of freedom to be larger than four to ensure enough statistical power [47], we tested three moderators only. Same as the intervention effects, for the primary analysis, we applied a multilevel random-effects meta-analysis. In addition, we conducted one-moderator random-effects meta-regressions for each candidate moderator separately to avoid over-parameterization given the small number of trials. For each model, we used the SJ estimator for between-study variance and HKSJ inference for the pooled effects and moderator coefficients.

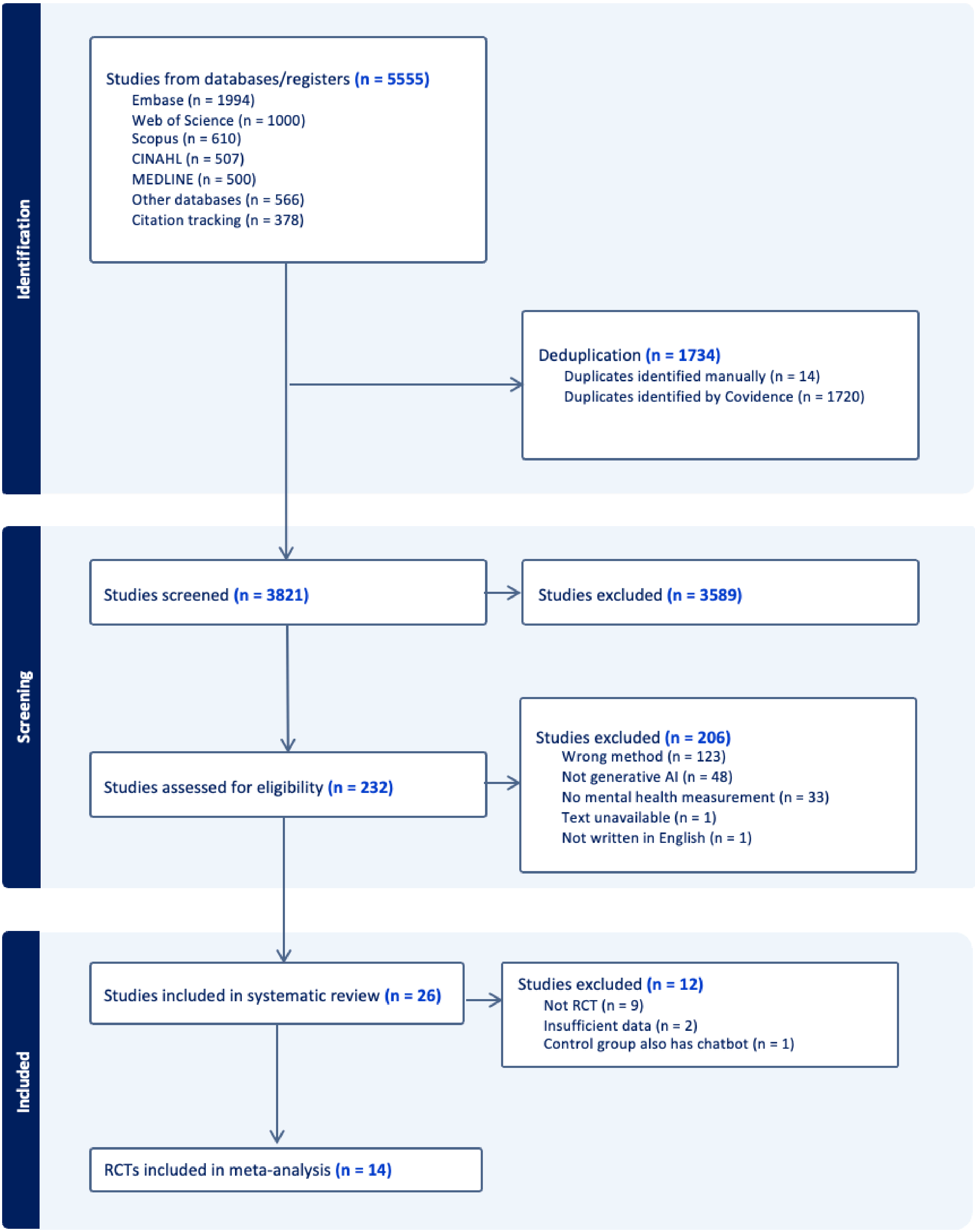
Studies were coded as *human assistance* if they included a preparatory session where a human introduced the chatbot or if human guidance was provided during chatbot use. One study, despite not explicitly mentioning human assistance, was also coded as *human assistance* based on an image showing a human assisting an older participant with ChatGPT [26]. Studies without any form of human involvement were coded as *self-guided*. Studies were coded based on social function as either *task-oriented* or *social-oriented*. *Task-oriented* studies were those where the chatbot's primary function was to assist with specific tasks, such as providing information, completing exercises, or helping with specific skills (eg, learning or mental health interventions). *Social-oriented* studies were those where the chatbot's primary function was to provide social interaction, emotional support, or companionship, without a specific focus on task completion or learning. Control group type was coded as *active* if participants received an alternative intervention, such as bibliotherapy, psychoeducation, routine care, or continued school-based support. Studies were coded as *passive* if participants received no intervention, such as waitlist control groups.

Results

Screening Procedures

During the title and abstract screening phase and the full-text review phase, Cohen ϰ values were 0.5 and 0.6, indicating fair and substantial agreement, respectively. From 5555 records across databases, 26 studies met inclusion criteria for narrative synthesis. Of these, 14 RCTs (19 treatment-control comparisons, N=6314) provided sufficient data for meta-analysis (see Figure 1 for data selection process). Among the 26 studies, 1 study measured only positive well-being instead of mental health issues [29], 1 study had chatbot designs in the control group [24], 3 studies did not report sufficient data for calculating pooled effect size, and 8 studies were not randomized trials, leaving a total of 14 RCTs eligible for meta-analysis to estimate the effectiveness of GenAI chatbot on mental health issues.

Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) diagram. RCT: randomized controlled trials.


Identification
Studies from databases/registers (n = 5555)
Embase (n = 1994)
Web of Science (n = 1000)
Scopus (n = 610)
CINAHL (n = 507)
MEDLINE (n = 500)
Other databases (n = 566)
Citation tracking (n = 378)
Deduplication (n = 1734)
Duplicates identified manually (n = 14)
Duplicates identified by Covidence (n = 1720)
Screening
Studies screened (n = 3821)
Studies excluded (n = 3589)
Studies assessed for eligibility (n = 232)
Studies excluded (n = 206)
Wrong method (n = 123)
Not generative AI (n = 48)
No mental health measurement (n = 33)
Text unavailable (n = 1)
Not written in English (n = 1)
Included
Studies included in systematic review (n = 26)
Studies excluded (n = 12)
Not RCT (n = 9)
Insufficient data (n = 2)
Control group also has chatbot (n = 1)
RCTs included in meta-analysis (n = 14)

Narrative Synthesis

Table 1 presents selected major characteristics of the 26 studies included in the systematic review. Below, we provide a descriptive analysis of the interventions' technical features, treatment features, research methods, and sample characteristics.

Table 1. Characteristics of the 26 studies included in the systematic review (Only 14 were included in the meta-analysis).

Study	Included in meta-analysis?	GenAI[a] chatbot name	Response generation approach	AI[b] technique	Customized	Interaction mode	Human assistance	Implement/ integrate	Research design	Targeted outcomes	Age, mean[c]	Sample size	Female	Clinical/ nonclinical Populations	Duration	Number of sessions/ modules	Country
Al Mazroui & Alzyoudi, 2024 [48]	No, not RCT[d]	ChatGPT	Generative	LLM[e]	1	Text	0	No	Mixed methods	Loneliness	65.3; 60-73	20	7	Nonclinical	2 weeks	3	United States
Liu Auren et al, 2024 [25]	No, not RCT	Replika	Generative	LSTM[f]	1	Text	0	No	Survey	Loneliness	18-73	404	241	Nonclinical	<15 minutes, 15-30 minutes, 10 minutes-1hour, 1-2 hours, 2-4 hours, 4 hours+ranging from daily to once a month or less	NA[g]	United States
Carl et al, 2024 [49]	No, not RCT	OpenAI GPT-4	Generative	LLM, NLP[h], GPT	1	Text and voice	0	Medical Q&A integrated into urological counseling sessions	Pre-post intervention design	Anxiety	60.58; 18-96	292	212	Urological patients	NA	3	Germany
Chen et al, 2025 [50]	Yes	No name	Hybrid	NLP, LLM	0	Text	0	Implemented AI chatbot to compare its effectiveness with a nurse hotline in reducing anxiety and depression among school parents in Hong Kong	RCT	Depression, anxiety	18-60	124	NA	Nonclinical	2 weeks	2	China (Hong Kong)
Wang et al, 2024 [51]	Yes	Typebot+ EAP Talk mode 1	Generative	LLM	1	Text, voice, and video	1	An online general English-speaking course at a language institution in southeastern China	RCT	Anxiety	21; 19-23	99	NA	Nonclinical	6 weeks	12	China
Çakmak, 2022 [52]	No, not RCT	Replika	Generative	neural network machine learning model and scripted dialogue content	1	Text and voice	0	Conducted Oral Communication Skill II course for freshmen, 2 hours per week, as a follow-up to Oral Communication Skills I.	Mixed methods	Anxiety	18-23	90	57	Nonclinical	12 weeks	NA	Turkey
Gan et al, 2025 [53]	Yes	ChatGPT 4.0	Generative	GPT	1	Text	0	Used ChatGPT 4.0 to provide standardized responses during knee arthroplasty consent process, with physicians interpreting and contextualizing information.	RCT	Anxiety, depression, pain	72.71; 60-80	55	43	Patients with knee osteoarthritis	2 weeks	2	China

Study	Included in meta-analysis?	GenAI[a] chatbot name	Response generation approach	AI[b] technique	Customized	Interaction mode	Human assistance	Implement/integrate	Research design	Targeted outcomes	Age, mean[c]	Sample size	Female	Clinical/nonclinical Populations	Duration	Number of sessions/modules	Country
He et al, 2022 [54]	Yes	XiaoE	Generative	NLP, deep learning	0	Text, voice, and images	0	No	Single-Blind, Three-Arm RCT	Depressive symptoms	18.78; 17-34	148	55	Young adults with depressive symptoms	1 week	25.54	China
Heinz et al, 2024 [55]	Yes	Therabot	Generative	LLM	1	Text	0	No	RCT	Depression, anxiety, and high-risk eating disorders	33.86	210	125	Patients with depression, anxiety, or high risk of eating disorder	8 weeks	NA	United States
Habicht et al, 2024 [56]	No, not RCT	Limbic care	Generative	LLM	1	Text	0	Implemented Limbic Care AI tool in NHS Talking Therapies for Anxiety and Depression, supporting CBT exercises between sessions.	Observational study	Anxiety and depression	40.4; 18 and above	244	169	Patients with depression or anxiety receiving CBT	3 months	7	United Kingdom
Kimani et al, 2019 [57]	No, insufficient data	Angela	Hybrid	NLG[i]	1	Text, voice, and videos	1	No	Mixed methods	Confidence and anxiety	23; 18-30	28	22	Nonclinical	NA	2	United States
Liu IV et al, 2024 [24]	No, all the control groups used chatbots	GPT-3.5 Turbo and Baidu UNIT Platform Chatbot	Generative & Retrieval	NLP, GPT-3.5	1	Text	1	No	RCT	Anxiety, the satisfaction with life, positive and negative affect, psychological well-being	18-55	Total number: 154, sub study 1: 207, sub study 2: 70, sub study 3: 50	NA;NA;29	Nonclinical	Sub study 1: 6 days; sub study 2: 6 days; sub study 3: 2 weeks	NA	China
Liu Ivan et al, 2022 [58]	Yes	Philobot	Hybrid	NLP, sentiment analysis, BERT[j], deep learning, rule-based retrieval, and structured decision trees	1	Text and voice	0	No	RCT Pilot Study	Resilience, happiness, positive & negative affect, depression, anxiety, mental disorder, loneliness	21.8	79	NA	Nonclinical	4 days	4	China
Drouin et al, 2022 [59]	Yes	Replika	Generative	LSTM	1	Text	0	No	3 groups experiment study	Positive & negative affect, positive & negative emotion	19.82; 18-38	417	297	Nonclinical	20 minutes	NA	United States
Ali et al, 2024 [60]	Yes	ChatGPT, Gemini, and Perplexity	Generative	LLM	1	Text	0	No	RCT	Anxiety	18 and above	92	48	Nonclinical	4 weeks	4	Pakistan
Maples et al, 2024 [31]	No, no data	Replika	Generative	LLM, NLP, GPT	1	Text, voice, images, and videos	0	No	Cross-sectional survey study	Anxiety, social support, self-awareness, self-harm, and suicide prevention	18 and above	1006	NA	Nonclinical	NA	NA	75% were US-based, 25% were international

Study	Included in meta-analysis?	GenAI[a] chatbot name	Response generation approach	AI[b] technique	Customized	Interaction mode	Human assistance	Implement/ integrate	Research design	Targeted outcomes	Age, mean[c]	Sample size	Female	Clinical/ nonclinical Populations	Duration	Number of sessions/ modules	Country
McFadyen et al, 2024 [61]	Yes	Limbic Care	Generative	LLM	1	Text and images	0	No	RCT	Anxiety, depression	36.84	540	NA	Patients with anxiety/ depression	6 weeks	NA	UK-developed and US participants
Hu et al, 2024 [62]	No, not RCT	MyAI	Generative	LLM, GPT	1	Text	0	No	Within-subject experimental design	Negative affect, stress, social support, loneliness	21; 18-29	150	118	Nonclinical	3 weeks	2	Singapore
Romanovskyi et al, 2021 [63]	Yes	Elomia	Generative	LLM: RoBERTa(NER)	0	Text	0	No	RCT	Depression, anxiety	21; 19-23	82	39	Nonclinical	4 weeks	NA	Ukraine
Sabour et al, 2023 [64]	Yes	Emohaa (ES-Bot+CBT[k]-Bot)	Hybrid	NLP, LLM	0	Text	1	No	RCT	Depression, anxiety, positive & negative affect, insomnia	30.9	247	190	Nonclinical	3 weeks	7	China
Zheng, 2024 [30]	No, not RCT	Reading Bot	pretrained LLM	LLM	1	Text	1	Used AI chatbot to clarify reading confusion in junior secondary EFL class.	Quasi-experimental pre-test/post-test.	Anxiety	12.845; 11-14	84	46	Nonclinical	3 hours 45 minutes	5	China
Vowels et al, 2024 [29]	No, positive outcomes	Amanda	Generative	LLM, GPT	1	Text	0	No	RCT	Well-being, distress, hopefulness, confidence, satisfaction, self-demand/partner-withdraw, partner-demand/self-withdraw	36.6	258	169	Nonclinical	NA	1	United Kingdom
Wang & Farb, 2024 [65]	Yes	No name	Generative	LLM	1	Text	0	No	RCT	Wellness, mindfulness	18.86; 17-40	114	83	Nonclinical	1 week	7	Canada
Wang & Li, 2024 [26]	No, not RCT	ChatGPT-3.0	Generative	GPT	1	Text and voice	1	No	Controlled design	Loneliness, depression, and life satisfaction	80.4	12	1	Nonclinical	8 weeks	8	China
Yahagi et al, 2024 [66]	Yes	ChatGPT-3.5	Generative	GPT	0	Text	0	Patients interacted with ChatGPT for preoperative information about Anesthesia	RCT	Anxiety	57.5	85	42	Nonclinical	4 weeks	4	Japan
Zheng et al, 2025 [67]	Yes	No name	Generative	LLM, GPT	1	Text and voice	1	Adapted Wang's (2014) four-stage model for English speaking, integrating LLM functionality	RCT	Anxiety	18.66	83	57	Nonclinical	5-7 days	1	China

Study	Included in meta-analysis?	GenAI[a] chatbot name	Response generation approach	AI[b] technique	Customized	Interaction mode	Human assistance	Implement/integrate	Research design	Targeted outcomes	Age, mean[c]	Sample size	Female	Clinical/nonclinical Populations	Duration	Number of sessions/modules	Country
								into the experimental group									

[a] GenAI: generative artificial intelligence.

[b] AI: artificial intelligence.

[c] Range is provided, if available.

[d] RCT: randomized controlled trial.

[e] LLM: large language model.

[f] LSTM: long short-term memory.

[g] NA: not available.

[h] NLP: natural language processing.

[i] NLG: natural language generation.

[j] BERT: bidirectional encoder representations from transformers.

[k] CBT: cognitive behavioral therapy.

Technical Features

Among the 26 included studies, 21 used purely GenAI and 5 used hybrid models combining generative and rule or retrieval-based approaches. Among the 11 studies utilizing GenAI, all systems were based on large language model (LLM) architectures, within which three studies specifically used GPT-series models. In addition, eight studies integrated LLM with other AI techniques, such as natural language processing (NLP). Two studies used long short-term memory (LSTM) models [25,59], one incorporated NLP with dynamic programming (DP) [54], one study used neural network machine learning model and scripted dialogue content [52], one incorporated NLP alongside GPT-3.5 [24], another integrated NLP, BERT, deep learning [58], and one used natural language generation (NLG) [57].

Most studies used a single AI chatbot, including ChatGPT (various versions, n=5), Replika (n=4), Limbic care (n=2), and one study each for Elomia [63], Philobot [58], MyAI [62], Virtual Coach Angela [57], Reading Bot [30], Emohaa [64], Amanda [29], XiaoE [54], and Therabot [55]. Three studies explored multiple AI chatbots, such as ChatGPT, Gemini, and Perplexity [60], or Typebot/ D-ID Agent and EAP Talk [51], or GPT-3.5 Turbo and Baidu UNIT [24] platform chatbot. Three studies did not specify a chatbot name.

Regarding delivery format, the majority (n=11) required the use of a smartphone only, 6 used only web-based platforms (including two web applications), 2 used both smartphone and web-based platforms and 3 studies did not specify the platform. As for interaction mode, all 26 studies used text-based interaction, with 11 studies incorporating voice features and six studies further including image-based interactions.

All studies used nonembodied AI chatbots, except for two that incorporated embodied virtual agents [51,57]. Furthermore, among the 26 studies, 21 implemented AI chatbots with customized features.

Treatment Features

The duration of chatbot intervention ranged from 20 minutes [59] to 3 months [56] (mean 3.63, SD 1.74 wk). Nine studies explicitly incorporated cognitive behavioral therapy principles, while one drew from positive psychology [58] and two from mindfulness interventions [26,65]. However, the rest of the studies (n=14) did not specify the guiding theoretical model.

Interventions targeted outcomes varied, including mental health concerns (eg, depression, anxiety, insomnia, stress), social well-being (eg, loneliness, social support), school-, language-, or test-related anxiety, and anxiety related to medical procedure (eg, preoperative anxiety, hospital anxiety).

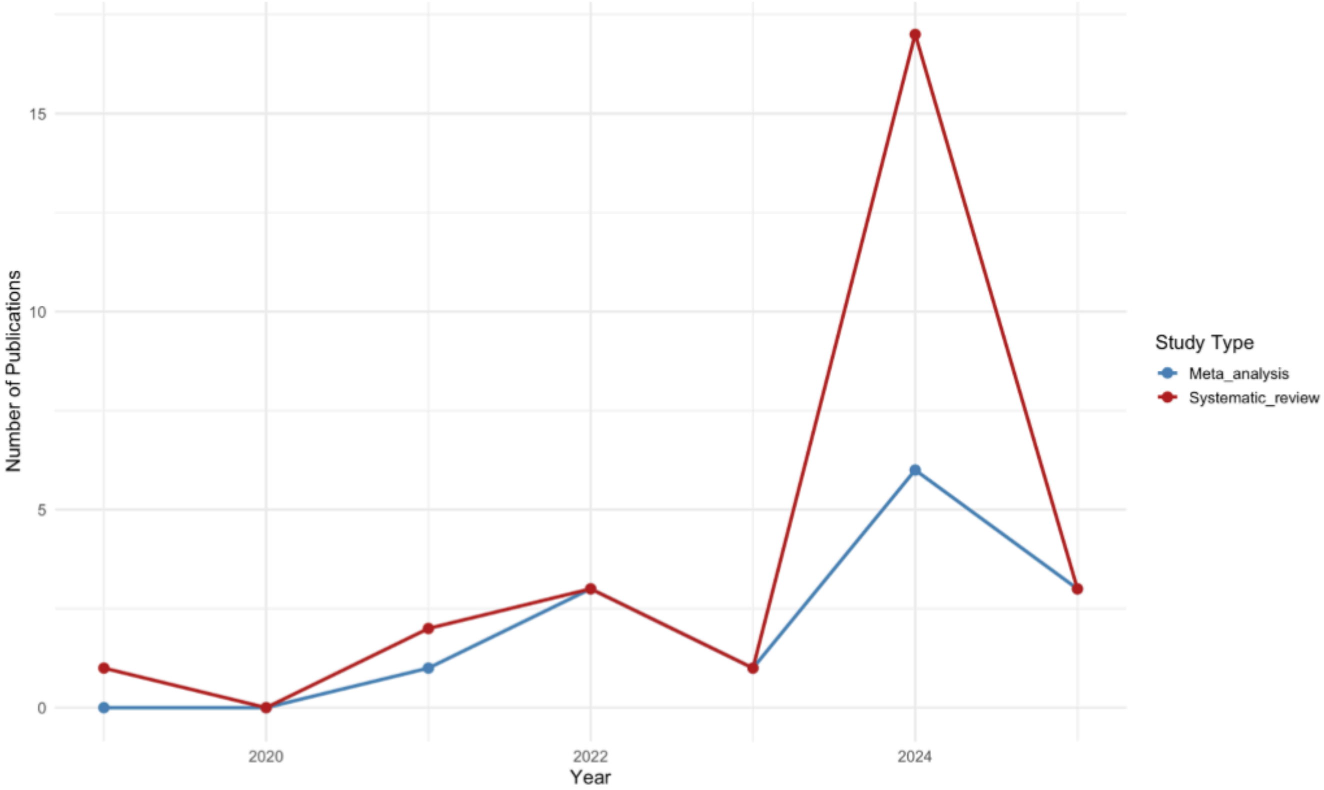
Most studies (n=18) incorporated some form of human support alongside chatbot use, such as clinician facilitation [56] and teacher supervision [51]. In contrast, eight studies featured fully autonomous AI chatbots. It seems like AI chatbots can assist in-person counseling sessions but might not replace human therapy. For example, one study found that the GenAI support system significantly improved patient attendance and treatment outcomes [68]. However, a study also found that mental health chatbot alone did not outperform the traditional bibliotherapy method in developing [58] participants' resilience.

A few studies embedded chatbots into structured health care (n=5), including urological counseling for urological surgery [49], preoperative patient education [66], therapy support between sessions [23], knee arthroplasty patients' consent process [53], assistance to health care professionals in a conventional school nurse hotline [53]. Some studies embedded chatbots in educational settings (n=3), such as oral communication skills courses for university freshmen [52], junior secondary English as a Foreign Language classes' activities for clarifying reading confusion [30], and online general English-speaking courses [51].

Research design

Fifteen studies used RCTs, while the remainder used quasi-experimental (n=5), mixed-methods (n=3), survey (n=2) [26], and observational design (n=1) [23]. Publication years ranged from 2019 to early 2025 (March when the search concluded), with 17 studies published in 2024 alone, which resonates with AI's exponential growth. Figure 2 presents a plot of the number of studies published per year.

Figure 2. Number of studies published per year. The chart illustrates the number of systematic review and meta-analysis publications from 2019 to 2025. Data represent the total number of studies published each year, based on the inclusion criteria established for this review. Note that the literature search concluded in March 2025, which explains the drop in 2025.



Sample Characteristics

Collectively, a total of 5469 participants from 11 countries and regions were involved. Most were single-site studies, with 10 conducted in China, 6 in the United States, 3 in the United Kingdom [26,37,61], and 1 each in Turkey [52], Germany [49], Singapore [62], Pakistan [60], Ukraine [63], Canada [65], and Japan [66]. Out of the 10 studies conducted in China, 2 used English-language AI platforms because they targeted English learning among Chinese students [30,67], and the remaining 8 studies used Chinese. Based on Beyebach et al's [33] WEIRD versus non-WEIRD framework, 15 out of 26 studies (58%) were conducted in non-WEIRD countries.

Exactly half of the studies [55] (n=13, including one study recruited partially students and partially nonstudents) focused on student populations, while the remainder involved general adult participants (n=13). The participants' age ranged from 11 [30] to 96 [49] years. Meanwhile, 21 out of 26 studies (81%) from the systematic review focus on early- and middle-aged adults (18–50 y old). Only one study focused on adolescents exclusively [41], and three studies focused on older adults [48,53,66]. The studies' sample size ranges all the way from 12 [26] to 1006 [31] (mean 198.85, SD 210.68). Most studies (n=20) involved nonclinical participants, while six recruited clinical populations, including patients attending urological counseling for elective surgery [49] (n=1), people experiencing symptoms of depression or anxiety [16,23,61] (n=3), people with high-risk eating disorders and other mental health conditions (n=1) [55], or knee osteoarthritis patients (n=1) [53].

Intervention Effects and Sensitivity Analyses

Table 2 presents the descriptive statistics of the 14 RCT studies with 19 treatment-control pairs (n=6314) included in the meta-analysis. Figure 3 presents the forest plot of these 14 studies' effect sizes and outcomes.

Table 2. Descriptive statistics of the 13 studies included for meta-analysis (14 studies include 19 treatment-control pairs). We have 14 articles but 19 treatments. Therefore, for the sake of analysis, 18 was used for meta-analysis.

Category	Level	Overall (%)
Study level		
Total treatments (n=19)		
WEIRD[a]		
No		11 (57.9)

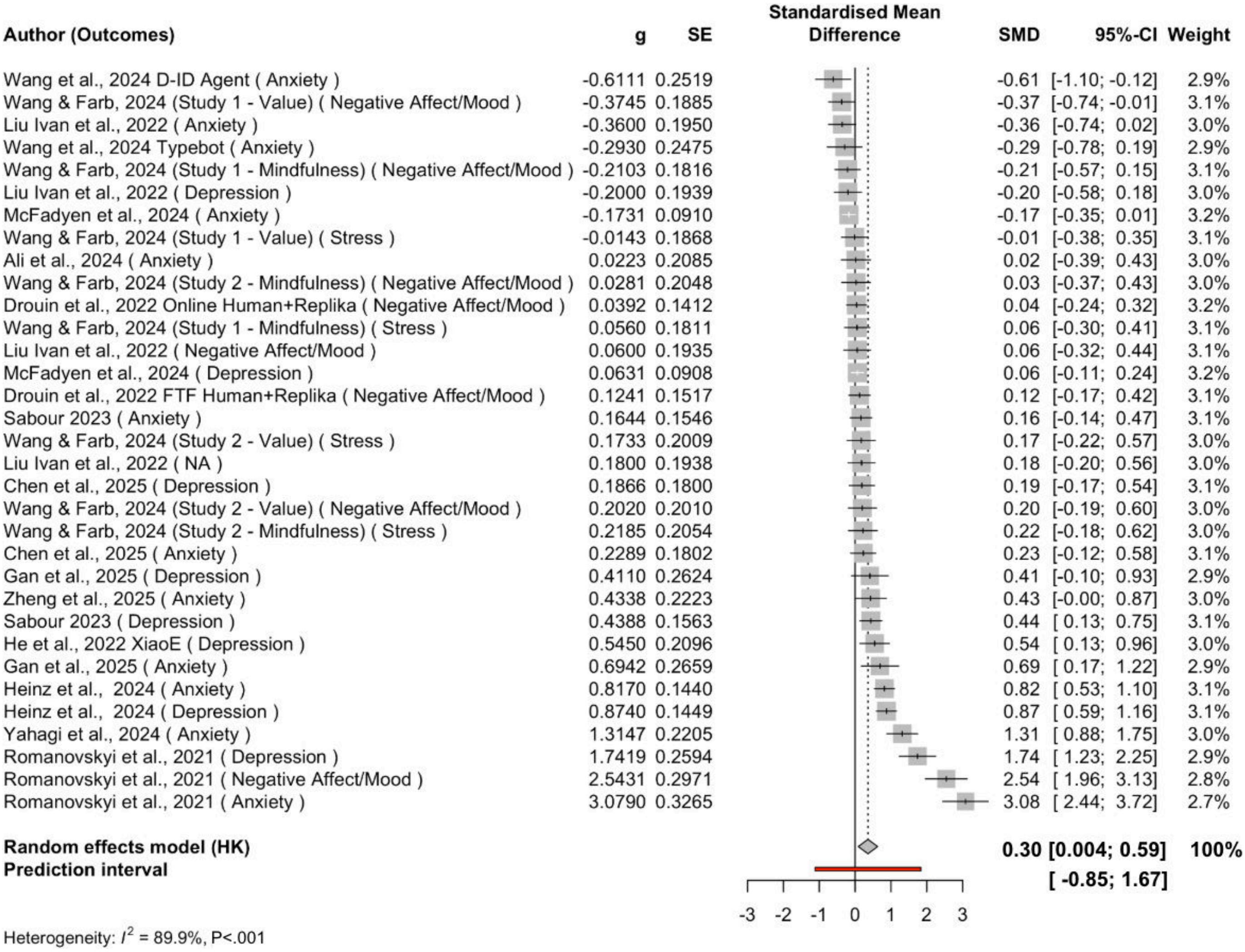
Category	Level	Overall (%)
Yes		8 (42.1)
Clinical populations		
No		14 (73.7)
Yes		5 (26.3)
Age[b]		
Early adulthood		12 (63.2)
Late adulthood		2 (10.5)
Middle adulthood		4 (21.1)
NA[c]		1 (5.3)
Sex		
Less than 50% female		9 (47.4)
More than 50% female		10 (52.6)
Customized		
No		6 (31.6)
Yes		13 (68.4)
Human assistance		
Purely self-guided program		15 (78.9)
With human assistance		4 (21.1)
Modality		
Hybrid		7 (36.8)
Text-based		12 (63.2)
Social function		
Social-oriented		7 (36.8)
Task-oriented		12 (63.2)
Outcome level		
Total effect sizes (n=44)		
Outcomes		
Anxiety		19 (43.2)
Depression		12 (27.3)
Loneliness		1 (2.3)
Negative mood or Affect		8 (18.2)
Stress		4 (9.1)
Clustered		
0		38 (86.4)
NA		6 (13.6)
Control group		
Active		28 (63.6)
Passive		16 (36.4)
Follow-up		
Without follow-up assessments		37 (84.1)
With follow-up assessments		7 (15.9)

[a] WEIRD is an acronym for Western, Educated, Industrialized, Rich, and Democratic following Beyebach et al’s [33] framework.

[b] Age was split into three categories: early adulthood (18–30 y old), middle adulthood (30–50 y old), and late adulthood (more than 50 y old) based on mean age.

[c] Not available.

Figure 3. Forest plot for all outcomes**.** Studies were organized from smaller SMDs to larger SMDs [26,50,51,53-55,58-61,63-67]. g: Hedges *g*; PI: predictive intervals; SE: standard error; SMD: standardized mean difference.



Author (Outcomes)	g	SE	SMD	95%-CI	Weight
Wang et al., 2024 D-ID Agent (Anxiety)	-0.6111	0.2519	-0.61	[-1.10; -0.12]	2.9%
Wang & Farb, 2024 (Study 1 - Value) (Negative Affect/Mood)	-0.3745	0.1885	-0.37	[-0.74; -0.01]	3.1%
Liu Ivan et al., 2022 (Anxiety)	-0.3600	0.1950	-0.36	[-0.74; 0.02]	3.0%
Wang et al., 2024 Typebot (Anxiety)	-0.2930	0.2475	-0.29	[-0.78; 0.19]	2.9%
Wang & Farb, 2024 (Study 1 - Mindfulness) (Negative Affect/Mood)	-0.2103	0.1816	-0.21	[-0.57; 0.15]	3.1%
Liu Ivan et al., 2022 (Depression)	-0.2000	0.1939	-0.20	[-0.58; 0.18]	3.0%
McFadyen et al., 2024 (Anxiety)	-0.1731	0.0910	-0.17	[-0.35; 0.01]	3.2%
Wang & Farb, 2024 (Study 1 - Value) (Stress)	-0.0143	0.1868	-0.01	[-0.38; 0.35]	3.1%
Ali et al., 2024 (Anxiety)	0.0223	0.2085	0.02	[-0.39; 0.43]	3.0%
Wang & Farb, 2024 (Study 2 - Mindfulness) (Negative Affect/Mood)	0.0281	0.2048	0.03	[-0.37; 0.43]	3.0%
Drouin et al., 2022 Online Human+Replika (Negative Affect/Mood)	0.0392	0.1412	0.04	[-0.24; 0.32]	3.2%
Wang & Farb, 2024 (Study 1 - Mindfulness) (Stress)	0.0560	0.1811	0.06	[-0.30; 0.41]	3.1%
Liu Ivan et al., 2022 (Negative Affect/Mood)	0.0600	0.1935	0.06	[-0.32; 0.44]	3.1%
McFadyen et al., 2024 (Depression)	0.0631	0.0908	0.06	[-0.11; 0.24]	3.2%
Drouin et al., 2022 FTF Human+Replika (Negative Affect/Mood)	0.1241	0.1517	0.12	[-0.17; 0.42]	3.1%
Sabour 2023 (Anxiety)	0.1644	0.1546	0.16	[-0.14; 0.47]	3.1%
Wang & Farb, 2024 (Study 2 - Value) (Stress)	0.1733	0.2009	0.17	[-0.22; 0.57]	3.0%
Liu Ivan et al., 2022 (NA)	0.1800	0.1938	0.18	[-0.20; 0.56]	3.0%
Chen et al., 2025 (Depression)	0.1866	0.1800	0.19	[-0.17; 0.54]	3.1%
Wang & Farb, 2024 (Study 2 - Value) (Negative Affect/Mood)	0.2020	0.2010	0.20	[-0.19; 0.60]	3.0%
Wang & Farb, 2024 (Study 2 - Mindfulness) (Stress)	0.2185	0.2054	0.22	[-0.18; 0.62]	3.0%
Chen et al., 2025 (Anxiety)	0.2289	0.1802	0.23	[-0.12; 0.58]	3.1%
Gan et al., 2025 (Depression)	0.4110	0.2624	0.41	[-0.10; 0.93]	2.9%
Zheng et al., 2025 (Anxiety)	0.4338	0.2223	0.43	[-0.00; 0.87]	3.0%
Sabour 2023 (Depression)	0.4388	0.1563	0.44	[0.13; 0.75]	3.1%
He et al., 2022 XiaoE (Depression)	0.5450	0.2096	0.54	[0.13; 0.96]	3.0%
Gan et al., 2025 (Anxiety)	0.6942	0.2659	0.69	[0.17; 1.22]	2.9%
Heinz et al., 2024 (Anxiety)	0.8170	0.1440	0.82	[0.53; 1.10]	3.1%
Heinz et al., 2024 (Depression)	0.8740	0.1449	0.87	[0.59; 1.16]	3.1%
Yahagi et al., 2024 (Anxiety)	1.3147	0.2205	1.31	[0.88; 1.75]	3.0%
Romanovskyi et al., 2021 (Depression)	1.7419	0.2594	1.74	[1.23; 2.25]	2.9%
Romanovskyi et al., 2021 (Negative Affect/Mood)	2.5431	0.2971	2.54	[1.96; 3.13]	2.8%
Romanovskyi et al., 2021 (Anxiety)	3.0790	0.3265	3.08	[2.44; 3.72]	2.7%
Random effects model (HK)			**0.30**	**[0.004; 0.59]**	**100%**
Prediction interval				**[-0.85; 1.67]**	

Heterogeneity: I^2 = 89.9%, P<.001

Table 3 presents the multilevel meta-analysis including 44 effects across 14 study clusters. The overall pooled effect was 0.30 (SE 0.14), *P*=.047, with a 95% CI (0.004, 0.59) and 95% PI (–0.85 to 1.67) under the REML estimator. This corresponds to a small-to-moderate positive effect of chatbot-based mental health interventions compared to control groups. The 95% PI indicated wide real-world variability in effects, suggesting that true effects across similar contexts could range from negligible to moderately large in magnitude. Between-study heterogeneity was substantial (σ^2=0.332, τ=0.576), with additional within-study residual variation at the outcome level (σ^2=0.039, τ=0.198). The residual heterogeneity test confirmed significant dispersion (Q [35]=230.41, *P*<.001).

Table 3. Multivariate random-effects meta-regression model results for models with and without moderators.

Coefficient	SMD[a]	SE[b]	*t* test	*df*[c]	*P* value	95% PI[d]
Null model for all outcomes						
Intercept	0.30	0.14	2.13	17.9	.047[e]	–0.85, 1.67
Null model for depression						
Intercept	0.49	0.20	2.5	6.97	.04[e]	–0.51, 1.54
Null model for anxiety						
Intercept	0.43	0.28	1.56	11	.15	–1.08, 2.05
Null model for negative affect and mood						
Intercept	0.28	0.31	0.92	7	.39	–1.95, 2.52
Null model for stress						
Intercept	0.10	0.05	1.92	2.96	.15	–0.31, 0.51

Coefficient	SMD[a]	SE[b]	*t* test	*df*[c]	*P* value	95% PI[d]
Single predictor model with one moderator						
Social function: task-oriented (as compared to social-oriented)	–0.78	0.28	–2.76	12.45	.02[e]	—[f]
Single predictor model with one moderator						
Human assistance (as compared to self-guided)	–0.39	0.29	–1.34	4.63	.24	—
Single predictor model with one moderator						
Passive control group (as compared to active control group)	0.202	0.25	0.82	4.78	.45	—
Full model with three moderators						
Intercept	0.77	0.34	2.28	5.74	.06	—
Passive control group (as compared to active control group)	0.07	0.24	0.31	4.90	.77	—
Human assistance (as compared to self-guided)	–0.03	0.25	–0.12	5.812	.91	—
Social function: task-oriented (as compared to social-oriented)	–0.76	0.32	–2.38	11.26	.04[e]	—

[a] SMD: standardized mean difference.
[b] SE: standard error.
[c] df: degrees of freedom.
[d] PI: predictive intervals.
[e] *P*<.05.
[f] Not available.

As for sensitivity analysis, we performed two changes: (1) we changed the *r* specification between 0.2 to 0.8 and the result is robust, and (2) we applied univariate HKSJ-SJ random-effects meta-analysis, and the conclusion remains the same. Details of the univariate HKSJ-SJ meta-regression model results for models with and without moderators can be found in Multimedia Appendices 5 and 6.

Apart from weighted average effects across all included studies, we also conducted subgroup analyses by outcome (Table 3). The pooled effect for depression (k=12) was 0.49 (SE 0.20; *P*=.04, 95% CI 0.03, 0.96, 95% PI –0.51, 1.54), with high heterogeneity (Q[11]=68.23; *P*<.001; τ²=0.225, I² ≈ 90%). The pooled effect for anxiety (k=19) was 0.43 (SE 0.28; *P*=0.15, 95% CI –0.18, 1.03, 95% PI (–1.08, 2.051), also with substantial heterogeneity (Q[18]=142.63, *P*<.001; τ²=0.857). The pooled effect for negative affect or mood (k=8) was 0.28 (SE=0.31; *P*=.39, 95% CI –0.45, 1.02, 95% PI –1.95, 2.52), with very high heterogeneity (Q[7]=77.00, *P*<.001; τ²=0.664). The pooled effect for stress (k=4) was 0.10 (SE=0.05: *P*=.15, 95% CI –0.21, 0.41, 95% PI –0.31, 0.51, with negligible heterogeneity (Q[3]=0.90, *P*=.83; τ²=0). There is only one effect size on loneliness; therefore, we skipped subgroup analysis on this outcome. Figures 4–7 present forest plots for four subgroup outcomes.

Figure 4. Forest plot for depression outcomes. Note studies were organized from smaller SMDs to larger SMDs [50,53-55,58,61,63,64]. g: Hedges *g*; PI: predictive intervals; SE: standard error; SMD: standardized mean difference.

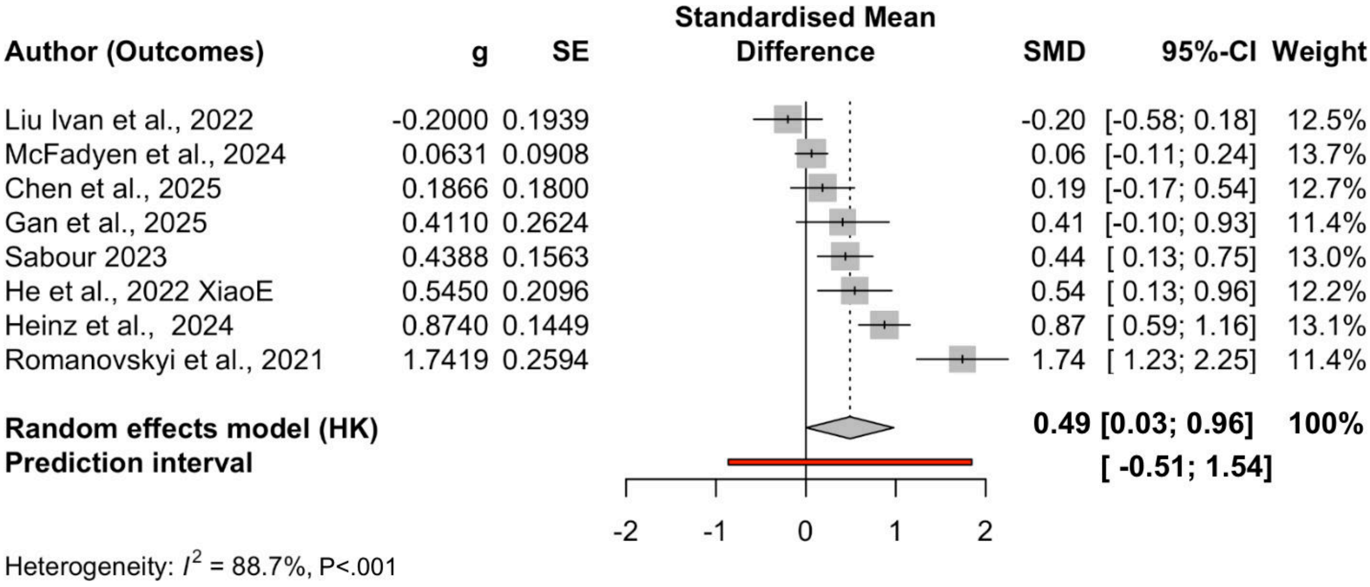
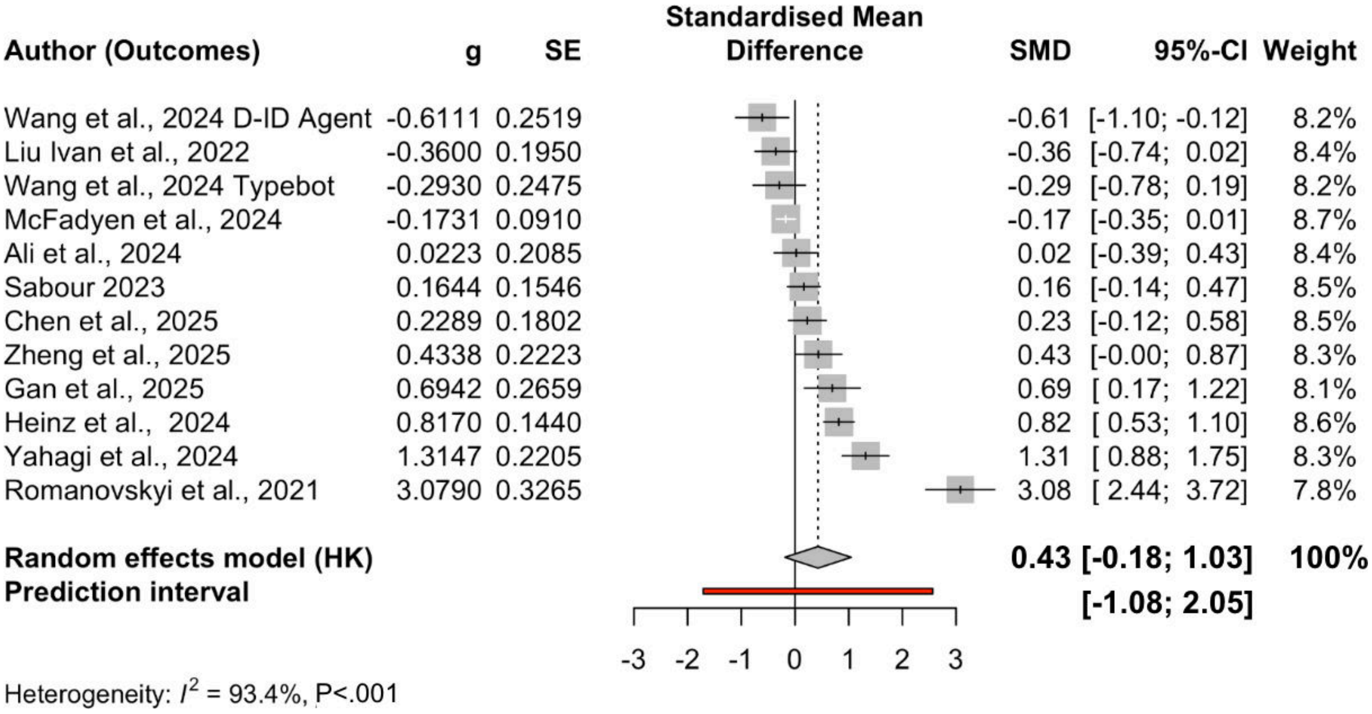


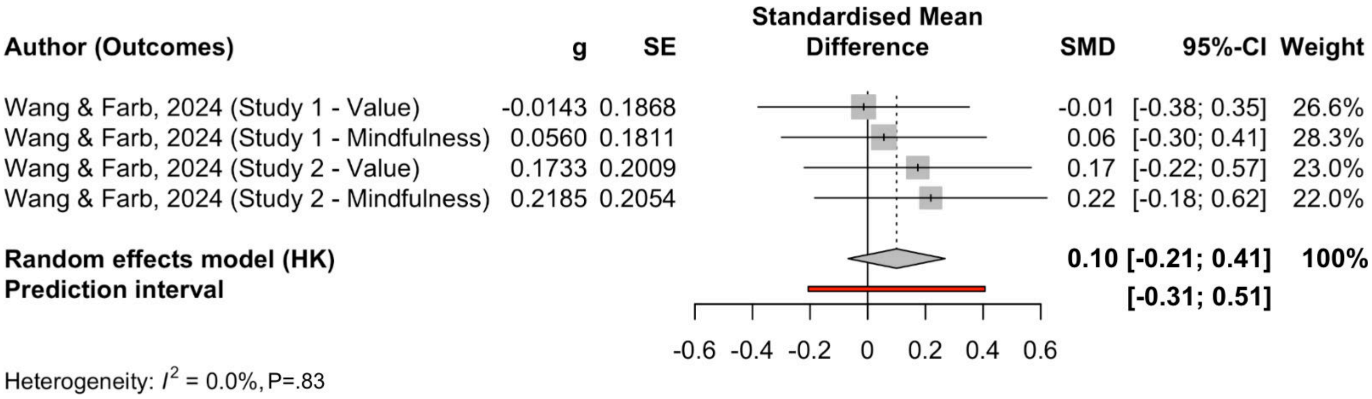
Figure 5. Forest plot for anxiety outcomes. Note studies were organized from smaller SMDs to larger SMDs [26,50,51,53,55,58,60,61,63,64,66,67]. g: Hedges *g*; PI: predictive intervals; SE: standard error; SMD: standardized mean difference.



Author (Outcomes)	g	SE	SMD	95%-CI	Weight
Wang et al., 2024 D-ID Agent	-0.6111	0.2519	-0.61	[-1.10; -0.12]	8.2%
Liu Ivan et al., 2022	-0.3600	0.1950	-0.36	[-0.74; 0.02]	8.4%
Wang et al., 2024 Typebot	-0.2930	0.2475	-0.29	[-0.78; 0.19]	8.2%
McFadyen et al., 2024	-0.1731	0.0910	-0.17	[-0.35; 0.01]	8.7%
Ali et al., 2024	0.0223	0.2085	0.02	[-0.39; 0.43]	8.4%
Sabour 2023	0.1644	0.1546	0.16	[-0.14; 0.47]	8.5%
Chen et al., 2025	0.2289	0.1802	0.23	[-0.12; 0.58]	8.5%
Zheng et al., 2025	0.4338	0.2223	0.43	[-0.00; 0.87]	8.3%
Gan et al., 2025	0.6942	0.2659	0.69	[0.17; 1.22]	8.1%
Heinz et al., 2024	0.8170	0.1440	0.82	[0.53; 1.10]	8.6%
Yahagi et al., 2024	1.3147	0.2205	1.31	[0.88; 1.75]	8.3%
Romanovskyi et al., 2021	3.0790	0.3265	3.08	[2.44; 3.72]	7.8%
Random effects model (HK)			**0.43**	**[-0.18; 1.03]**	**100%**
Prediction interval				**[-1.08; 2.05]**	

Heterogeneity: I^2 = 93.4%, P<.001

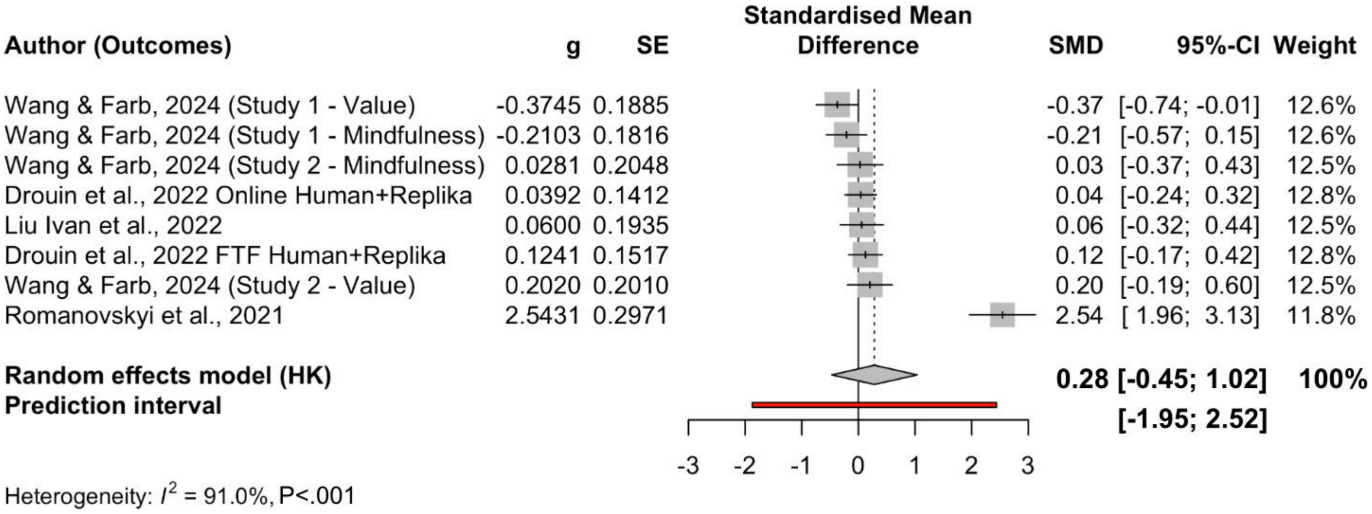
Figure 6. Forest plot for stress outcomes. Note studies were organized from smaller SMDs to larger SMDs [65]. g: Hedges *g*; PI: predictive intervals; SE: standard error; SMD: standardized mean difference.



Author (Outcomes)	g	SE	SMD	95%-CI	Weight
Wang & Farb, 2024 (Study 1 - Value)	-0.0143	0.1868	-0.01	[-0.38; 0.35]	26.6%
Wang & Farb, 2024 (Study 1 - Mindfulness)	0.0560	0.1811	0.06	[-0.30; 0.41]	28.3%
Wang & Farb, 2024 (Study 2 - Value)	0.1733	0.2009	0.17	[-0.22; 0.57]	23.0%
Wang & Farb, 2024 (Study 2 - Mindfulness)	0.2185	0.2054	0.22	[-0.18; 0.62]	22.0%
Random effects model (HK)			**0.10**	**[-0.21; 0.41]**	**100%**
Prediction interval				**[-0.31; 0.51]**	

Heterogeneity: I^2 = 0.0%, P=.83

Figure 7. Forest plot for negative affect and mood outcomes. Studies were organized from smaller SMDs to larger SMDs [58,59,63,65]. g: Hedges *g*; PI: predictive intervals; SE: standard error; SMD: standardized mean difference.

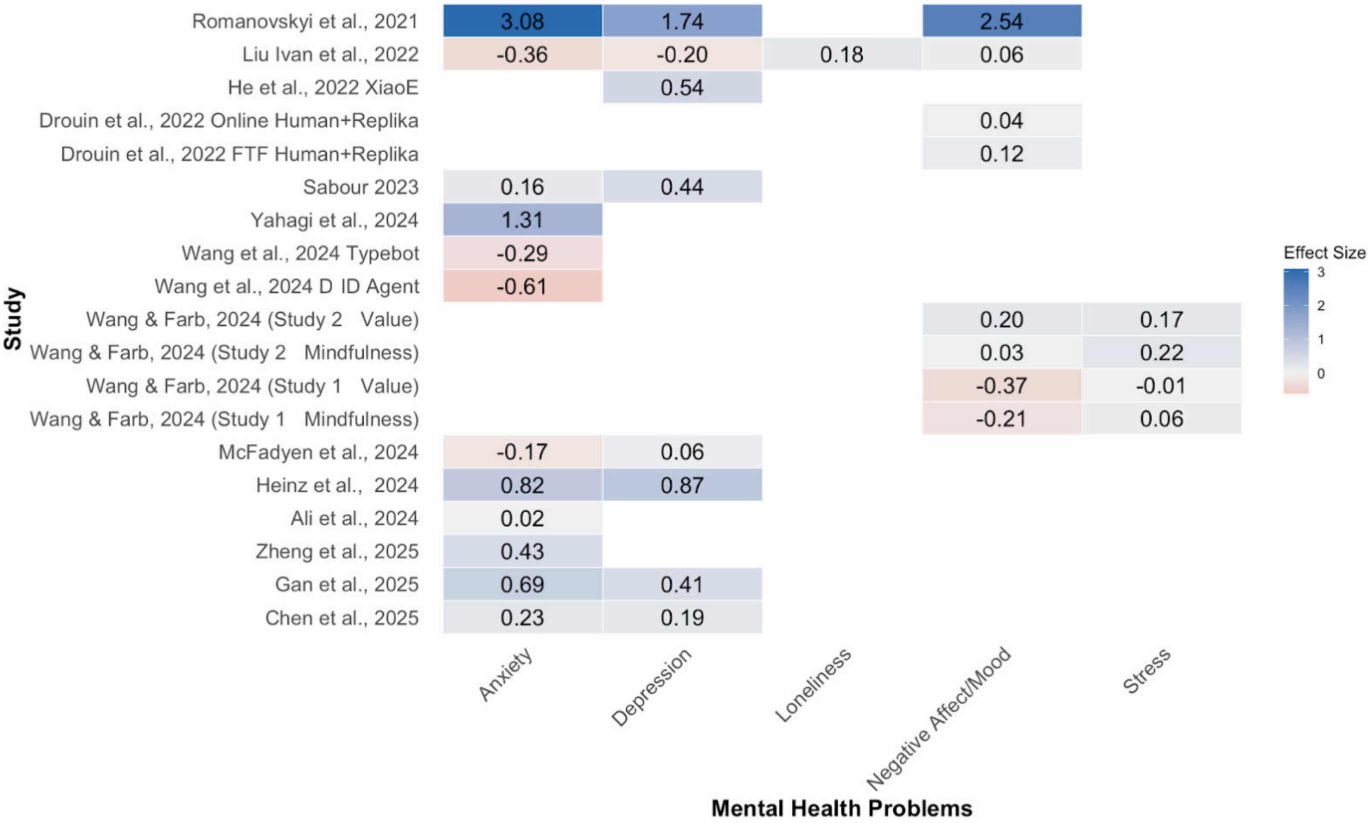


Author (Outcomes)	g	SE	SMD	95%-CI	Weight
Wang & Farb, 2024 (Study 1 - Value)	-0.3745	0.1885	-0.37	[-0.74; -0.01]	12.6%
Wang & Farb, 2024 (Study 1 - Mindfulness)	-0.2103	0.1816	-0.21	[-0.57; 0.15]	12.6%
Wang & Farb, 2024 (Study 2 - Mindfulness)	0.0281	0.2048	0.03	[-0.37; 0.43]	12.5%
Drouin et al., 2022 Online Human+Replika	0.0392	0.1412	0.04	[-0.24; 0.32]	12.8%
Liu Ivan et al., 2022	0.0600	0.1935	0.06	[-0.32; 0.44]	12.5%
Drouin et al., 2022 FTF Human+Replika	0.1241	0.1517	0.12	[-0.17; 0.42]	12.8%
Wang & Farb, 2024 (Study 2 - Value)	0.2020	0.2010	0.20	[-0.19; 0.60]	12.5%
Romanovskyi et al., 2021	2.5431	0.2971	2.54	[1.96; 3.13]	11.8%
Random effects model (HK)			**0.28**	**[-0.45; 1.02]**	**100%**
Prediction interval				**[-1.95; 2.52]**	

Heterogeneity: I^2 = 91.0%, P<.001

Among these, only the depression subgroup showed a statistically significant positive effect. The wide PIs, especially for anxiety and negative affect or mood, indicate that true effects in new, similar settings may range from negligible or unfavorable to moderately beneficial. Readers must be cautious since some subgroups have a small number of studies. Figure 8 complements the subgroup findings with a heatmap of each study's outcome and effect size.

Figure 8. Heatmap of each study's outcome and effect size**.** Note the studies were organized chronologically. For effect sizes, darker color means higher absolute values, with blue indicating positive effect sizes and red indicating negative effect sizes [50,51,53-55,58-61,63-67].

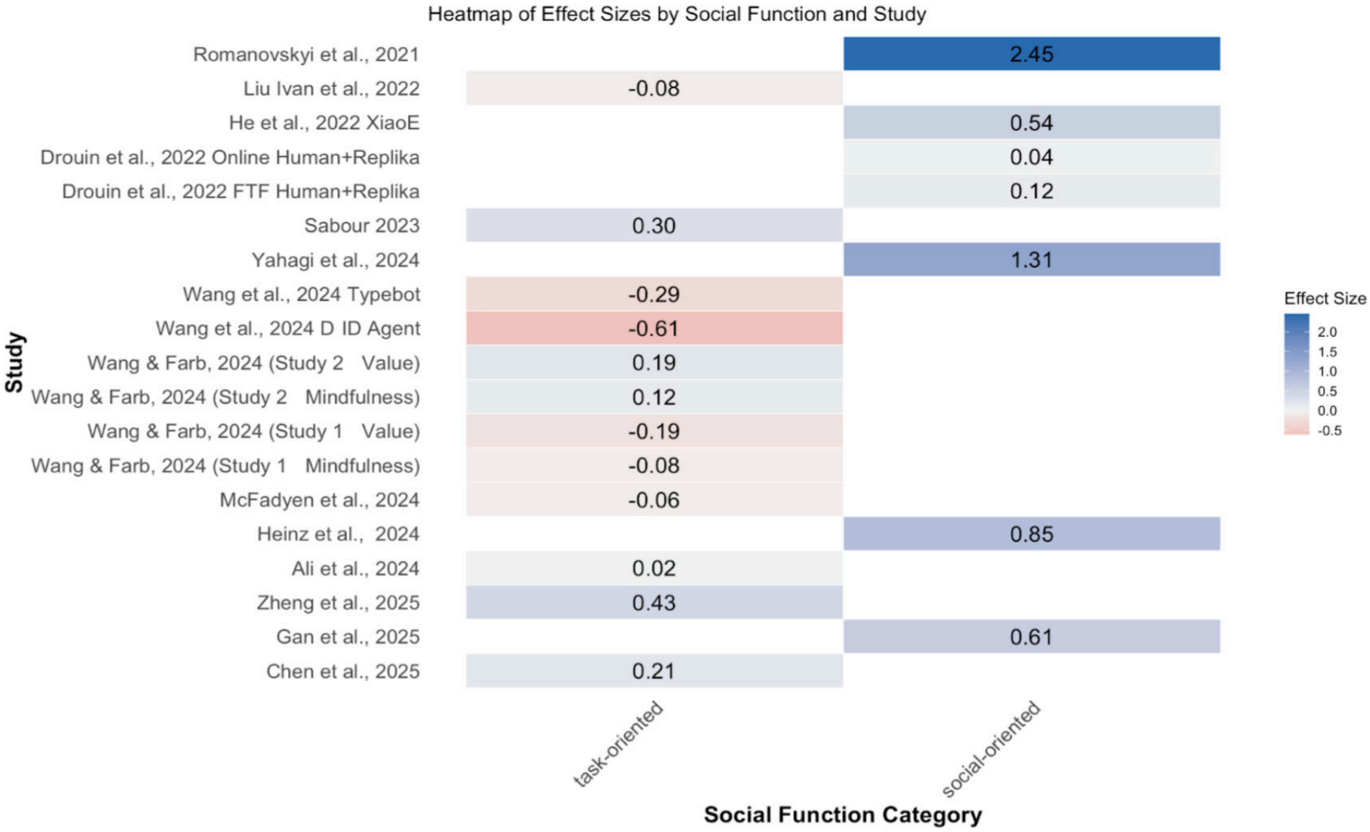


Moderator Analysis

The combined moderator model (Active/Passive+Human Assistance+Social Function) was significant overall (F(3, 36)=3.11, *P*=.04). Within this model, social function is a strong predictor (SMD=–0.76, *P*=.04), whereas human assistance (SMD=–0.03, *P*=.91) and active versus passive control group (SMD=0.07, *P*=.77) did not influence the effect. Specifically, task-oriented chatbots showed smaller effects (SMD=0.007, SE=0.06, *P*=.91) compared to socially oriented chatbots (SMD=0.77, SE=0.34, *P*=.06).

The sensitivity analyses using one-moderator univariate random-effects models with SJ τ^2 and HKSJ confirmed the results from the multivariate model that task-oriented chatbots are less effective as compared to social-oriented chatbots. Figure 9 presents a heatmap of each study's effect size by social function and study, which clearly shows that social-oriented chatbots are consistently found with more positive effect sizes across different studies.

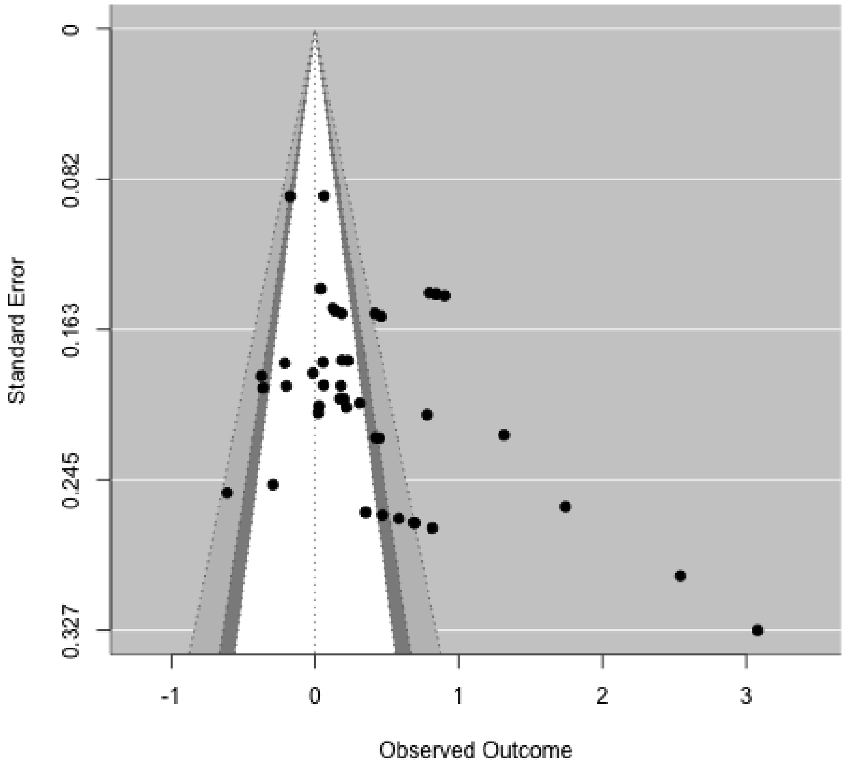
Figure 9. Heatmap of each study's effect size by social function and study**.** The studies were organized chronologically. For effect sizes, darker color means higher absolute values, with blue indicating positive effect sizes and red indicating negative effect sizes [50,51,53-55,58-61,63-67].



Selection Bias

Figure 10 presents a funnel plot. Visual assessment reveals that there is asymmetry. We then applied Vevea and Woods' [44] weight-function model to assess potential publication bias. The unadjusted model (k=44) estimated a pooled effect size of g=0.41 (SE=0.10, z=4.24; *P*<.001, 95% CI 0.22, 0.60). After adjustment for potential selection bias, the two-step model (*P* value cutpoints=.025, .50, 1) yielded a smaller and nonsignificant change in the pooled effect (g=0.54, SE=0.26, z=2.10; *P*=.04, 95% CI 0.04, 1.04). The estimated weights indicated that studies with *P*<.025* were about 3.54 times more likely to be included than those with larger *P* values, suggesting moderate publication bias favoring statistically significant findings. The likelihood ratio test comparing adjusted and unadjusted models was significant (χ^2_2=11.25, *P*=.004), confirming evidence of selection bias.

Figure 10. Funnel plot.



Risk of Bias Analysis

Table 4 presents the risk of bias assessment from the Cochrane tool, ROBINS-I for nonrandomized studies included in systematic review but excluded from meta-analysis. This tool assesses the risk of bias in nonrandomized studies across seven domains: (1) confounding, or whether external factors influenced outcomes; (2) selection of participants, assessing if inclusion or exclusion introduced bias; (3) classification of interventions, evaluating accurate group assignment; (4) deviations from intended interventions, considering adherence and co-interventions; (5) missing data, addressing loss and its impact; (6) measurement of outcomes, assessing objectivity and consistency; and (7) selection of the reported result, evaluating selective outcome reporting. Overall, the risk of bias ranges from moderate to serious, and 66.67% (8/12) studies were ranked as serious risk.

Table 4. Risk of bias using ROBINS-I (Risk Of Bias In Non-randomised Studies – of Interventions) for nonrandomized studies.

Study	Domain 1 confounding	Domain 2 classification of interventions	Domain 3 selection of participants	Domain 4 deviations from intended interventions	Domain 5 missing data	Domain 6 measurement of outcomes	Domain 7 selection of the reported result	Overall risk of bias
Al Mazroui & Alzyoudi, 2024 [48]	S[a]	L[b]	S	M[c]	M	M	M	S
Çakmak, 2022 [52]	S	M	M	M	M	M	M	S
Carl et al, 2024 [49]	L	L	M	M	L	M	L	M
Habicht et al, 2024 [56]	S	M	M	M	L	M	L	S
Hu et al, 2024 [62]	L	L	L	M	L	M	L	M
Kimani et al, 2019 [57]	L	L	L	L	L	M	L	M
Liu Auren et al, 2024 [25]	L	L	L	M	L	M	L	M
Liu IV et al, 2024 [24]	L	L	M	M	M	M	M	S

Study	Domain 1 confounding	Domain 2 classification of interventions	Domain 3 selection of participants	Domain 4 deviations from intended interventions	Domain 5 missing data	Domain 6 measurement of outcomes	Domain 7 selection of the reported result	Overall risk of bias
Maples et al, 2024 [31]	S	L	M	L	M	M	M	S
Vowels et al, 2024 [29]	L	L	L	L	L	M	M	S
Wang and Li, 2024 [26]	S	L	M	M	S	M	L	S
Zheng, 2024 [24]	S	L	L	M	L	M	L	S

[a]S: serious risk.
[b]L: low risk.
[c]M: moderate risk.

Table 5 presents the Cochrane RoB2 for the RCTs included in meta-analysis. This tool assesses included studies in five domains: (1) bias arising from the randomization process, (2) bias due to deviations from intended interventions, (3) bias due to missing outcome data, (4) bias in measurement of the outcome, and (5) bias in selection of the reported result. Domain 1 was assessed at the study level, whereas the other domains were assessed at the result level. Each domain was rated as low risk of bias, some concerns, or high risk of bias, and an overall judgment for each study was derived based on these domain-level assessments. The results showed that 64.29% (9/14) of studies were rated as having some concerns and 35.71% (5/14) as high risk, with no study judged to be at low risk of bias.

Table 5. Risk of bias using the Cochrane RoB2 for randomized control trials (RCTs).

Study	Domain 1 Randomization	Domain 2 Deviation from intended interventions	Domain 3 Missing outcome data	Domain 4 Measurement of the outcomes	Domain 5 Selection of the reported results	Overall risk of bias
Ali et al, 2024 [60]	L[a]	L	L	SC[b]	L	SC
Chen et al, 2025 [50]	SC	H[c]	SC	SC	SC	H
Drouin et al, 2022 [59]	H	SC	L	H	L	H
Gan et al, 2025 [53]	L	SC	L	L	L	SC
He et al, 2022 [54]	L	L	L	SC	L	SC
Heinz et al, 2024 [55]	L	SC	L	SC	L	SC
Liu Ivan et al, 2022 [58]	SC	SC	H	SC	L	H
McFadyen et al, 2024 [61]	SC	SC	L	SC	L	SC
Romanovskyi et al, 2021 [63]	L	H	SC	SC	SC	H
Sabour, 2023 [64]	L	L	SC	SC	SC	SC
Wang & Farb, 2024 [65]	L	SC	L	L	L	SC
Wang et al, 2024 [51]	L	SC	SC	SC	SC	SC
Yahagi et al, 2024 [66]	L	H	SC	SC	L	H
Zheng et al, 2025 [67]	L	SC	L	SC	SC	SC

[a]L: low risk.
[b]SC: some concerns.
[c]H: high risk.

Discussion

Principal Findings

This review provides the first systematic synthesis and meta-analysis focused on GenAI chatbots for mental health outcomes, including 26 articles in the systematic review and 14 RCTs in meta-analysis. Overall, our results indicate a small-to-moderate but statistically significant average effect, suggesting that GenAI mental health chatbot interventions may be effective in reducing mental health issues. However, wide prediction intervals and substantial between-study heterogeneity indicate that these benefits are not consistent across studies or populations. This was similar to previous meta-analyses on the effectiveness of rule-based, retrieval-based, and GenAI chatbots [13,15,16,18]. Yet, it is crucial to note that effectiveness varies depending on chatbots' design and target outcome.

Social-Oriented Chatbots Are More Effective Than Task-Oriented Chatbots

Social function emerged as the most consistent moderator across different models. We found that social-oriented chatbots are more effective than task-oriented chatbots, although this result should be interpreted cautiously given the limited number of included studies and the high heterogeneity of effects. This pattern aligns with previous literature on social chatbots leading to better consumer satisfaction [69] and social outcomes for older adults [70]. Decades of research demonstrate that perceived social support is a protective factor against stress, depression, and anxiety [71, 72]. Social chatbots can simulate supportive relationships, offering emotional validation, empathy, and companionship, even if users cognitively recognize the artificiality of the interaction [73]. This aligns with the Computers Are Social Actors (CASA) paradigm [74], which shows that humans often respond to machines using the same social heuristics they apply to human partners. In contrast, task-oriented chatbots, lacking this socio-emotional dimension, primarily provide informational rather than emotional support, limiting their impact on distress.

The effect may also be because social interactions with AI chatbots facilitate therapeutic alliance, one of the most effective factors in psychotherapy [75,76]. From psychotherapy research, the common factors model emphasizes that therapeutic alliance, empathy, and relational depth are among the strongest predictors of positive clinical outcomes [77, 78]. Social chatbots, by offering empathic, personalized, and emotionally attuned exchanges, can foster trust and disclosure, which may help reduce negative mental health issues [69]. Task-oriented chatbots, by contrast, often lack the flexibility to respond empathetically, limiting their capacity to generate the relational bonds essential for emotional relief.

An important implication of this finding is that developers should consider integrating relational design principles, including empathy, warmth, and social support, into conversational systems. Designing AI interactions that communicate acceptance and genuine care may enhance users' emotional engagement and psychological well-being, aligning chatbot interactions more closely with the therapeutic mechanisms that underpin effective human support.

Outcome Subgroup: GenAI Chatbots Are Most Effective in Treating Depression

Among the outcome subgroups (depression, anxiety, stress, negative moods), effect sizes were positive across all groups, but only the depression subgroup demonstrated a statistically significant effect (ES=.49, *P*=.041). Depression and anxiety are the most studied outcomes in existing studies of GenAI mental health chatbots. This finding is unsurprising, given that both disorders are not only the most prevalent [79] but also highly comorbid [80]. While our results suggest GenAI chatbots' promise in addressing depression, these technologies should be positioned as supplementary instead of replacement treatments. Depression management typically requires long-term care, and approximately half of patients relapse after an initial episode [81]. Effective treatment requires careful examination of patients' medical history, symptom trajectory, and sustained therapeutic alliances, which cannot be fully replicated by current GenAI systems. Therefore, GenAI chatbots might act as complementary supports to enhance counselors' efficiency and extend access to care. Indeed, in our systematic review, 69.23% of GenAI interventions incorporated some forms of human assistance instead of relying solely on fully autonomous GenAI chatbot experience. Future studies could explore how to deliver more targeted, personalized, and sustainable treatment through optimal combination of human expertise with GenAI technology.

In contrast, despite plenty of studies focusing on depression and anxiety, there was a severe lack of studies focusing on negative mood and stress. This imbalance reflects a broader gap in the literature, particularly regarding the role of GenAI chatbots in managing more severe or complex mental health conditions. The increasing severity and complexity of global mental health challenges highlight the limited application of chatbots to severe mental health conditions such as suicidality, schizophrenia, or substance use disorders. Taken together, these gaps indicate that while GenAI chatbots may serve as valuable adjuncts to care, they should not be viewed as standalone solutions for addressing the full spectrum of mental health needs. Rather, their role lies in complementing human-delivered services and expanding access to support, especially in contexts where resources are scarce.

Most Interventions Took Place in WEIRD Countries

Most of the studies (58%, 15/26) took place in non-WEIRD countries, such as China. While comparing across continents, there is a severe lack of GenAI chatbot studies from Europe. One explanation could be the stringent and comprehensive AI regulation in European countries introduced by the European Union [82]. The self-regulatory AI market in other countries, such as China, United States, and United Kingdom, might help the local AI development in mental health areas. However, these cross-national observations are descriptive rather than inferential; our study did not test the effects of cultural or regulatory differences statistically. Large language models are frequently trained on datasets predominantly sourced from WEIRD contexts [83]. Results suggested that there are some systematic differences between WEIRD and non-WEIRD countries in terms of age and recruitment type. Consequently, when these models are deployed in non-WEIRD contexts, they may not fully grasp or appropriately respond to culturally specific nuances or local dialects.

With the global shortage of mental health resources, especially in non-WEIRD countries, it is essential to examine how cultural differences shape the adoption and effectiveness of GenAI chatbots for mental health [84]. Cultural beliefs and stigma influence willingness to seek digital support, while differences in language and communication styles affect the perceived appropriateness of chatbot responses [85,86]. Training AI systems with culturally representative data and considering local ethical and regulatory contexts may improve trust, relevance, and uptake [87]. Attention is needed regarding the generalizability of the intervention results across diverse cultural and socioeconomic contexts. Further research is needed to adapt these interventions to non-WEIRD contexts, taking into account local cultural nuances and resource availability [88].

Lack of Studies on Adolescents, Older Adults, and Applications in Diverse Settings

Eighty-one percent of studies from the systematic review focus on early- and middle-aged adults (18–50 y old), with only one study investigating adolescents (<18 y old) and three studies focusing on the older adults (>50 y old). This might be attributed to a cautious attitude towards GenAI's impact on youth and the lack of research focus on older adults because of the potential concern regarding their technological skills. In the increasingly aging society, GenAI chatbots have great potential to provide companions for the older adults to reduce their sense of loneliness [89]. For future researchers, the impact of GenAI chatbots on these two age groups is worth more investigation to ensure future more targeted usage.

As for settings, although a few studies implement AI chatbots in therapy procedures or educational settings, most studies have not yet streamlined GenAI chatbots in usual care procedures. Future research could investigate ways to integrate GenAI chatbots in existing treatment processes or programs to ensure sustainability of benefiting from AI chatbots. Apart from clinical settings, the application of GenAI chatbots in reducing anxiety and depression in educational settings is equally important, but only four studies investigated this area. Future studies could explore more diverse settings, including medical, educational, and therapy settings.

Ethical Considerations

The growing use of social chatbots in mental health contexts raises significant ethical concerns that cannot be overlooked. News reports of suicide cases linked to interaction with AI companion chatbots [90,91] highlight the urgent risks, reinforcing findings from prior studies on the dark side of AI companionship, including emotional dependency, manipulation, privacy violations, and social isolation [92,93]. These dangers are particularly acute in mental health settings, where users may be especially vulnerable and the generative nature of AI systems can produce responses that are unpredictable, inappropriate, or even harmful.

Addressing these challenges requires concerted, multi-stakeholder efforts involving policymakers, technology developers, clinicians, and end-users. Robust regulatory frameworks, ethical guidelines, and oversight mechanisms are essential to ensure that generative AI chatbots are designed, deployed, and monitored in ways that safeguard user well-being [94,95]. This involves co-designing systems with input from mental health professionals and users; conducting systematic auditing and debiasing of training datasets; establishing safeguards to clearly delineate the boundaries of chatbot outputs; and ensuring that systems are regularly evaluated against therapeutic objectives [96,97]. Only through such comprehensive efforts can the potential benefits of GenAI chatbots be realized while minimizing risks to vulnerable populations.

Limitation

Readers should be aware of a few limitations when interpreting the results. Among the 12 studies included in the meta-analysis, readers should be aware that some RCTs might have bias due to large baseline differences and differential attrition. First, some studies' baseline differences exceed 0.25 SD, a threshold proposed by What Works Clearinghouse [98]. For example, Jeong [99] reported 0.30 SD for depression, McFadyen et al [61] reported 0.26 SD for anxiety, Sabour [64] reported 0.33 SD for depression. Second, some studies exceed 15% differential attrition between treatment and control groups, a threshold proposed by What Works Clearinghouse [98]. For instance, Chen et al [50] reported a differential attrition rate of 34% and He et al [54] reported 24 %. Third, the meta-analysis only analyzed a small sample of 12 studies. Although our number of effect sizes (n=37) is relatively large in terms of outcomes analyzed, it is small in terms of statistical analyses that combine empirical studies [100]. Readers should note that a small sample reduces statistical power for performing moderator analyses, and consequently, the capacity to obtain more precise estimates of the effect size via moderators [101]. Lastly, the risk of bias showed a mix of studies' qualities between some concerns and high risk, which means that the GenAI mental health chatbot RCTs still need methodological improvement, and the results should be interpreted with caution.

Conclusion

In conclusion, this systematic review has highlighted the promising yet inconsistent potential of GenAI chatbots in addressing mental health issues. Meta-regression findings indicate that social-oriented chatbots, as opposed to task-oriented ones, demonstrate greater effectiveness, though with wide variability and uncertainty. While these interventions are promising, their benefits come with risks that cannot be ignored. This review also identifies several research gaps, emphasizing the need for further investigation into adolescent and older adult populations, better serving users in non-WEIRD countries, analysis with mental health disorders other than anxiety and depression, integration of chatbots into existing therapy frameworks, and exploration within diverse settings. Given the substantial heterogeneity, moderate risk of bias, and small number of available RCTs, conclusions should be drawn with caution, viewing current findings as a foundation for more rigorous future studies rather than as definitive evidence of efficacy.

Acknowledgments

We thank the JMIR editor and reviewers for their constructive feedback during the revision. The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision:
- Code generation
- Visualization
- Text generation
The GAI tool used was: ChatGPT-4.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by: Collective responsibility

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets and R code generated or analyzed during this study are available in the GitHub Generative-AI-Mental-Health-Chatbot repository [42].

Authors’ Contributions

QZ: Conceptualization, Formal analysis, Investigation, Visualization, Interpretation, Validation, Writing – original draft, Writing – review & editing, Supervision
RZ: Conceptualization, Interpretation, Writing – original draft, Writing – review & editing
YX: Interpretation, Writing – original draft, Writing – review & editing
YS: Data curation, Writing – original draft, Writing – review & editing
CT: Data curation, Writing – review & editing, Validation
F-HL: Data curation, Writing – review & editing, Visualization

Conflicts of Interest

None declared.

Multimedia Appendix 1

GenAI mental health chatbots’ comparison in features and target outcomes (organized by year of launch).
[DOCX File (Microsoft Word File), 2220 KB-Multimedia Appendix 1]

Multimedia Appendix 2

A comparison of the past nine systematic reviews on GenAI mental health chatbots.
[DOCX File (Microsoft Word File), 2707 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Search keywords’ details for database searching.
[DOCX File (Microsoft Word File), 2190 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Search string details recorded for database searching (Round 1: November 1, 2024; Round 2: March 5, 2025).
[DOCX File (Microsoft Word File), 2191 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Univariate HKSJ-SJ meta-regression model results for models with and without moderators.
[DOCX File (Microsoft Word File), 3053 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Statistical details of the univariate HKSJ model.
[DOCX File (Microsoft Word File), 2189 KB-Multimedia Appendix 6]

Checklist 1

PRISMA 2020 checklist.
[DOCX File (Microsoft Word File), 2198 KB-Checklist 1]

References

1. Freeman M. The World Mental Health Report: transforming mental health for all. World Psychiatry. Oct 2022;21(3):391-392. [doi: 10.1002/wps.21018] [Medline: 36073688]
2. Prince M, Patel V, Saxena S, et al. No health without mental health. Lancet. Sep 8, 2007;370(9590):859-877. [doi: 10.1016/S0140-6736(07)61238-0] [Medline: 17804063]
3. About mental health. Centers for Disease Control and Prevention; 2025. URL: https://www.cdc.gov/mental-health/about/index.html [Accessed 2016-03-16]
4. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. eClinicalMedicine. Dec 2022;54:101675. [doi: 10.1016/j.eclinm.2022.101675]
5. Schwartz E. The global mental health crisis. Project Hope. 2025. URL: https://www.projecthope.org/news-stories/story/the-global-mental-health-crisis-10-numbers-to-note/ [Accessed 2025-03-16]
6. Henderson C, Evans-Lacko S, Thornicroft G. Mental illness stigma, help seeking, and public health programs. Am J Public Health. May 2013;103(5):777-780. [doi: 10.2105/AJPH.2012.301056] [Medline: 23488489]
7. Yusefi AR, Sharifi M, Nasabi NS, Rezabeigi Davarani E, Bastani P. Health human resources challenges during COVID-19 pandemic; evidence of a qualitative study in a developing country. PLOS ONE. 2022;17(1):e0262887. [doi: 10.1371/journal.pone.0262887] [Medline: 35073374]
8. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health. Oct 18, 2019;6(10):e14166. [doi: 10.2196/14166] [Medline: 31628789]
9. Boucher EM, Harake NR, Ward HE, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices. Dec 2021;18(sup1):37-49. [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
10. Oh H, Pickering TA, Schiffman J, LaBrie JW, Soffer-Dudek N, Pedersen ER. Web-based personalized normative feedback to decrease stigma and increase intentions to seek help. Stigma Health. May 2, 2024;10(3):420-427. [doi: 10.1037/sah0000518]
11. Bond RR, Mulvenna MD, Potts C, O'Neill S, Ennis E, Torous J. Digital transformation of mental health services. Npj Ment Health Res. Aug 22, 2023;2(1):13. [doi: 10.1038/s44184-023-00033-y] [Medline: 38609479]
12. Bhatt S. Digital mental health: role of artificial intelligence in psychotherapy. Ann Neurosci. Apr 2025;32(2):117-127. [doi: 10.1177/09727531231221612] [Medline: 39544658]
13. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med. Dec 19, 2023;6(1):236. [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
14. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry. Jul 2019;64(7):456-464. [doi: 10.1177/0706743719828977] [Medline: 30897957]
15. Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. J Affect Disord. Jul 1, 2024;356:459-469. [doi: 10.1016/j.jad.2024.04.057] [Medline: 38631422]
16. He Y, Yang L, Qian C, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res. Apr 28, 2023;25:e43862. [doi: 10.2196/43862] [Medline: 37115595]
17. Al-Amin M, Ali MS, Salam A, et al. History of generative artificial intelligence (AI) chatbots: past, present, and future development. arXiv. Preprint posted online on Feb 4, 2024. [doi: 10.48550/arXiv.2402.05122]
18. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform. Dec 2019;132:103978. [doi: 10.1016/j.ijmedinf.2019.103978] [Medline: 31622850]
19. Dsouza R, Sahu S, Patil R, Kalbande DR. Chat with bots intelligently: a critical review & analysis. Presented at: 2019 International Conference on Advances in Computing, Communication and Control (ICAC3); Dec 20-21, 2019; Mumbai, India. 2019.[doi: 10.1109/ICAC347590.2019.9036844]
20. Kumar A, Shankar A, Behl A, Chakraborty D, Gundala RR. Anthropomorphic generative AI chatbots for enhancing customer engagement, experience and recommendation. JCM. Jun 11, 2025;42(4):472-483. [doi: 10.1108/JCM-06-2024-6922]
21. Siddals S, Torous J, Coxon A. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. npj Mental Health Res. Oct 27, 2024;3(1):48. [doi: 10.1038/s44184-024-00097-4]
22. Yin Y, Jia N, Wakslak CJ. AI can help people feel heard, but an AI label diminishes this impact. Proc Natl Acad Sci USA. Apr 2, 2024;121(14):e2319112121. [doi: 10.1073/pnas.2319112121]

23. Habicht J, Dina LM, Stylianou M, Harper R, Hauser TU, Rollwage M. Generative AI-enabled therapy support tool improves clinical outcomes and patient engagement in NHS talking therapies. PsyArXiv. Preprint posted online on Apr 10, 2024. [doi: 10.31234/osf.io/mj46k]
24. Liu I, Liu F, Xiao Y, Huang Y, Wu S, Ni S. Investigating the key success factors of chatbot-based positive psychology intervention with retrieval- and generative pre-trained transformer (GPT)-based chatbots. International Journal of Human–Computer Interaction. Jan 2, 2025;41(1):341-352. [doi: 10.1080/10447318.2023.2300015]
25. Liu AR, Pataranutaporn P, Maes P. Chatbot companionship: a mixed-methods study of companion chatbot usage patterns and their relationship to loneliness in active users. arXiv. Preprint posted online on Oct 28, 2024. [doi: 10.48550/arXiv.2410.21596]
26. Wang Y, Li S. Tech vs. tradition: ChatGPT and mindfulness in enhancing older adults' emotional health. Behav Sci (Basel). Oct 10, 2024;14(10):923. [doi: 10.3390/bs14100923]
27. Ulrich S, Lienhard N, Künzli H, Kowatsch T. A chatbot-delivered stress management coaching for students (MISHA App): pilot randomized controlled trial. JMIR Mhealth Uhealth. Jun 26, 2024;12(1):e54945. [doi: 10.2196/54945] [Medline: 38922677]
28. Vinyals O, Le Q. A neural conversational model. arXiv. Preprint posted online on Jul 22, 2015. [doi: 10.48550/arXiv.1506.05869]
29. Vowels LM, Vowels MJ, Sweeney S, Hatch SG, Darwiche J. The evaluation of efficacy, feasibility, and technical outcomes of a GPT-4O-based chatbot Amanda for relationship support: a randomized controlled trial. PsyArXiv. Preprint posted online on Sep 20, 2024. [doi: 10.31234/osf.io/q4yh7]
30. Zheng S. The effects of chatbot use on foreign language reading anxiety and reading performance among Chinese secondary school students. Computers and Education: Artificial Intelligence. Dec 2024;7:100271. [doi: 10.1016/j.caeai.2024.100271]
31. Maples B, Cerit M, Vishwanath A, Pea R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. Npj Ment Health Res. Jan 22, 2024;3(1):4. [doi: 10.1038/s44184-023-00047-6] [Medline: 38609517]
32. Zhang Q, Neitzel A. Choosing the right tool for the job: screening tools for systematic reviews in education. J Res Educ Eff. Jul 2, 2024;17(3):513-539. [doi: 10.1080/19345747.2023.2209079]
33. Beyebach M, Neipp MC, Solanes-Puchol Á, Martín-Del-Río B. Bibliometric differences between WEIRD and non-WEIRD countries in the outcome research on solution-focused brief therapy. Front Psychol. 2021;12:754885. [doi: 10.3389/fpsyg.2021.754885] [Medline: 34867649]
34. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews a product from the ESRC methods programme. Semantic Scholar. 2006. URL: https://www.semanticscholar.org/paper/Guidance-on-the-Conduct-of-Narrative-Synthesis-in-A-Popay-Roberts/ed8b23836338f6fdea0cc55e161b0fc5805f9e27 [Accessed 2025-12-12]
35. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36(3). [doi: 10.18637/jss.v036.i03]
36. Lipsey MW, Wilson DB. Practical Meta-Analysis. Thousand Oaks (CA): Sage Publications; 2001.
37. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. Apr 2010;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
38. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol. Nov 14, 2015;15:99. [doi: 10.1186/s12874-015-0091-1] [Medline: 26573817]
39. van Aert RCM, Jackson D. A new justification of the Hartung-Knapp method for random-effects meta-analysis based on weighted least squares regression. Res Synth Methods. Dec 2019;10(4):515-527. [doi: 10.1002/jrsm.1356] [Medline: 31111673]
40. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. Feb 18, 2014;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
41. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open. Jul 12, 2016;6(7):e010247. [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
42. Generative-AI-mental-health-chatbot. GitHub. 2025. URL: https://github.com/qiyangzh/Generative-AI-Mental-Health-Chatbot [Accessed 2025-11-08]
43. Zhang Q. Generative AI chatbots as therapeutic tools: a systematic review and meta-analysis of their role in mitigating mental health issues. OSF. 2025. URL: https://osf.io/9daj7 [Accessed 2025-11-08]
44. Vevea JL, Woods CM. Publication bias in research synthesis: sensitivity analysis using a priori weight functions. Psychol Methods. 2005;10(4):428-443. [doi: 10.1037/1082-989X.10.4.428]

45. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. Oct 12, 2016;355:i4919. [doi: 10.1136/bmj.i4919]
46. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898. [doi: 10.1136/bmj.l4898]
47. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. Psychol Methods. Sep 2015;20(3):375-393. [doi: 10.1037/met0000011] [Medline: 24773356]
48. Al Mazroui K, Alzyoudi M. The role of ChatGPT in mitigating loneliness among older adults: an exploratory study. ONLINE J COMMUN MEDIA TECHNOL. Oct 1, 2024;14(4):e202444. [doi: 10.30935/ojcmt/14777]
49. Carl N, Nguyen L, Haggenmüller S, et al. Comparing patient’s confidence in clinical capabilities in urology: large language models versus urologists. Eur Urol Open Sci. Dec 2024;70:91-98. [doi: 10.1016/j.euros.2024.10.009] [Medline: 39507511]
50. Chen C, Lam KT, Yip KM, et al. Comparison of an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: pilot randomized controlled trial. JMIR Hum Factors. Mar 6, 2025;12:e65785. [doi: 10.2196/65785] [Medline: 40048637]
51. Wang C, Zou B, Du Y, Wang Z. The impact of different conversational generative AI chatbots on EFL learners: an analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. System. Dec 2024;127:103533. [doi: 10.1016/j.system.2024.103533]
52. Çakmak F. Chatbot-human interaction and its effects on EFL students’ L2 speaking performance and anxiety. Novitas-ROYAL (Research on Youth and Language). 2022:113-131. URL: https://eric.ed.gov/?id=EJ1365002 [Accessed 2025-12-12]
53. Gan W, Ouyang J, She G, et al. CHATGPT’s role in alleviating anxiety in total knee arthroplasty consent process: a randomized controlled trial pilot study. Int J Surg. Mar 1, 2025;111(3):2546-2557. [doi: 10.1097/JS9.0000000000002223] [Medline: 39903546]
54. He Y, Yang L, Zhu X, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res. Nov 21, 2022;24(11):e40719. [doi: 10.2196/40719]
55. Heinz MV, Mackin D, Trudeau B, et al. Evaluating Therabot: a randomized control trial investigating the feasibility and effectiveness of a generative AI therapy chatbot for depression, anxiety, and eating disorder symptom treatment. PsyArXiv. Preprint posted online on Jun 14, 2024. [doi: 10.31234/osf.io/pjqmr]
56. Habicht J, Viswanathan S, Carrington B, Hauser TU, Harper R, Rollwage M. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. Nat Med. Feb 2024;30(2):595-602. [doi: 10.1038/s41591-023-02766-x] [Medline: 38317020]
57. Kimani E, Bickmore T, Trinh H, Pedrelli P. You’ll be great: virtual agent-based cognitive restructuring to reduce public speaking anxiety. Presented at: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII); Sep 3-6, 2019; Cambridge, United Kingdom. [doi: 10.1109/ACII.2019.8925438]
58. Liu I, Chen W, Ge Q, Song D, Ni S. Enhancing psychological resilience with chatbot-based cognitive behavior therapy: a randomized control pilot study. Presented at: Chinese CHI ’22: Proceedings of the Tenth International Symposium of Chinese CHI; Oct 22-23, 2022; Guangzhou, China and Online China. [doi: 10.1145/3565698.3565787]
59. Drouin M, Sprecher S, Nicola R, Perkins T. Is chatting with a sophisticated chatbot as good as chatting online or FTF with a stranger? Comput Human Behav. Mar 2022;128:107100. [doi: 10.1016/j.chb.2021.107100]
60. Ali M, Rehman S, Cheema E. Impact of artificial intelligence on the academic performance and test anxiety of pharmacy students in objective structured clinical examination: a randomized controlled trial. Int J Clin Pharm. Aug 2025;47(4):1034-1041. [doi: 10.1007/s11096-025-01876-5] [Medline: 39903358]
61. McFadyen J, Habicht J, Dina LM, Harper R, Hauser TU, Rollwage M. AI-enabled conversational agent increases engagement with cognitive-behavioral therapy: a randomized controlled trial. medRxiv. Psychiatry and Clinical Psychology. Preprint posted online on Nov 2, 2024. [doi: 10.1101/2024.11.01.24316565]
62. Hu M, Chua XCW, Diong SF, Kasturiratna K, Majeed NM, Hartanto A. AI as your ally: The effects of AI‐assisted venting on negative affect and perceived social support. Applied Psych Health & Well. Feb 25, 2025;17(1). [doi: 10.1111/aphw.12621]
63. Romanovskyi O, Pidbutska N, Knysh A. Elomia chatbot: the effectiveness of artificial intelligence in the fight for mental health. Presented at: COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems; Apr 22-23, 2021; Lviv, Ukraine. URL: https://ceur-ws.org/Vol-2870/paper89.pdf [Accessed 2025-10-08]
64. Sabour S, Zhang W, Xiao X, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health. May 4, 2023;5. [doi: 10.3389/fdgth.2023.1133987]
65. Wang Y, Farb NAS. Chatbot-based interventions for mental health support. PsyArXiv. Preprint posted online on Aug 29, 2024. [doi: 10.31234/osf.io/xj7cz]

66. Yahagi M, Hiruta R, Miyauchi C, Tanaka S, Taguchi A, Yaguchi Y. Comparison of conventional anesthesia nurse education and an artificial intelligence chatbot (CHATGPT) intervention on preoperative anxiety: a randomized controlled trial. J Perianesth Nurs. Oct 2024;39(5):767-771. [doi: 10.1016/j.jopan.2023.12.005] [Medline: 38520470]
67. Zheng YB, Zhou YX, Chen XD, Ye XD. The influence of large language models as collaborative dialogue partners on EFL English oral proficiency and foreign language anxiety. Computer Assisted Language Learning. Jan 22, 2025:1-27. [doi: 10.1080/09588221.2025.2453191]
68. Habicht J, Dina LM, McFadyen J, et al. Generative AI-enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: real-world observational study. J Med Internet Res. Mar 10, 2025;27:e60435. [doi: 10.2196/60435] [Medline: 40063074]
69. Cai N, Gao S, Yan J. How the communication style of chatbots influences consumers' satisfaction, trust, and engagement in the context of service failure. Humanit Soc Sci Commun. May 28, 2024;11(1). [doi: 10.1057/s41599-024-03212-0]
70. Chattaraman V, Kwon WS, Gilbert JE, Ross K. Should AI-Based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. Comput Human Behav. Jan 2019;90:315-330. [doi: 10.1016/j.chb.2018.08.048]
71. Roohafza HR, Afshar H, Keshteli AH, et al. What's the role of perceived social support and coping styles in depression and anxiety? J Res Med Sci. Oct 2014;19(10):944-949. [Medline: 25538777]
72. Huang Y, Su X, Si M, et al. The impacts of coping style and perceived social support on the mental health of undergraduate students during the early phases of the COVID-19 pandemic in China: a multicenter survey. BMC Psychiatry. Dec 2021;21(1). [doi: 10.1186/s12888-021-03546-y]
73. Li H, Zhang R. Finding love in algorithms: deciphering the emotional contexts of close encounters with AI chatbots. OSF Preprints. Preprint posted online on Jul 22, 2024. [doi: 10.31219/osf.io/xd4k7]
74. Nass C, Moon Y. Machines and mindlessness: social responses to computers. Journal of Social Issues. Jan 2000;56(1):81-103. [doi: 10.1111/0022-4537.00153]
75. Flückiger C, Del Re AC, Wampold BE, Symonds D, Horvath AO. How central is the alliance in psychotherapy? A multilevel longitudinal meta-analysis. J Couns Psychol. 2012;59(1):10-17. [doi: 10.1037/a0025749]
76. Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. J Consult Clin Psychol. 2000;68(3):438-450. [doi: 10.1037/0022-006X.68.3.438]
77. Elvins R, Green J. The conceptualization and measurement of therapeutic alliance: an empirical review. Clin Psychol Rev. Oct 2008;28(7):1167-1187. [doi: 10.1016/j.cpr.2008.04.002] [Medline: 18538907]
78. Ferrario A, Sedlakova J, Trachsel M. The role of humanization and robustness of large language models in conversational artificial intelligence for individuals with depression: a critical analysis. JMIR Ment Health. Jul 2, 2024;11:e56569. [doi: 10.2196/56569] [Medline: 38958218]
79. Mental disorders. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [Accessed 2025-10-06]
80. Mental health conditions: depression and anxiety. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/tobacco/campaign/tips/diseases/depression-anxiety.html#three [Accessed 2025-10-06]
81. Kupfer DJ. Long-term treatment of depression. J Clin Psychiatry. May 1991;52 Suppl:28-34. [Medline: 1903134]
82. Siegmann C, Anderljung M. The Brussels effect and artificial intelligence: how EU regulation will impact the global AI market. APSA Preprints. Preprint posted online on Aug 23, 2022. [doi: 10.33774/apsa-2022-vxtsl]
83. Naous T, Ryan MJ, Ritter A, Xu W. Having beer after prayer? Measuring cultural bias in large language models. Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1); Aug 11-16, 2024; Bangkok, Thailand. 2024.[doi: 10.18653/v1/2024.acl-long.862]
84. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci. Jun 2010;33(2-3):61-83; [doi: 10.1017/S0140525X0999152X] [Medline: 20550733]
85. Chentsova-Dutton YE, Ryder AG. Cultural-clinical psychology: from cultural scripts to contextualized treatments. In: Handbook of Cultural Psychology. 2nd ed. Guilford Press; 2020:732-759.
86. Xiong Y, Yang L. Asian international students' help-seeking intentions and behavior in American postsecondary institutions. Int J Intercult Relat. Jan 2021;80:170-185. [doi: 10.1016/j.ijintrel.2020.11.007]
87. Farhud DD, Zokaei S. Ethical issues of artificial intelligence in medicine and healthcare. Iran J Public Health. Oct 27, 2021. [doi: 10.18502/ijph.v50i11.7600] [Medline: 35223619]
88. Bernal G, Domenech Rodríguez M, editors. Cultural Adaptations: Tools for Evidence-Based Practice with Diverse Populations. American Psychological Association; 2012. [doi: 10.1037/13752-000]

89. Pani B, Crawford J, Allen KA. Can generative artificial intelligence foster belongingness, social support, and reduce loneliness? A conceptual analysis. In: Applications of Generative AI. Springer; 2024:261-276. [doi: 10.1007/978-3-031-46238-2_13]
90. Bhuiyan J. ChatGPT encouraged Adam Raine's suicidal thoughts. His family's lawyer says OpenAI knew it was broken. Guardian News and Media. 2025. URL: https://www.theguardian.com/us-news/2025/aug/29/chatgpt-suicide-openai-sam-altman-adam-raine [Accessed 2025-10-06]
91. Duffy C. 'There are no guardrails.' this mom believes an AI chatbot is responsible for her son's suicide. Cable News Network. 2024. URL: https://edition.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit [Accessed 2025-10-06]
92. Zhang R, Li H, Meng H, Zhan J, Gan H, Lee YC. The dark side of AI companionship: a taxonomy of harmful algorithmic behaviors in human-AI relationships. CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025:1-17. [doi: 10.1145/3706598.3713429]
93. Laestadius L, Bishop A, Gonzalez M, Illenčík D, Campos-Castillo C. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New Media & Society. Oct 2024;26(10):5923-5941. [doi: 10.1177/14614448221142007]
94. Goktas P, Grzybowski A. Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy AI. J Clin Med. Feb 27, 2025;14(5):1605. [doi: 10.3390/jcm14051605] [Medline: 40095575]
95. Thieme A, Hanratty M, Lyons M, et al. Designing human-centered AI for mental health: developing clinically relevant applications for online CBT treatment. ACM Trans Comput-Hum Interact. Apr 30, 2023;30(2):1-50. [doi: 10.1145/3564752]
96. Lee KS, Yeung J, Kurniawati A, Chou DT. Designing human-centric AI mental health chatbots: a case study of two apps. In: Information Systems for Intelligent Systems: Proceedings of ISBM 2024. Springer; 2025:432-452. [doi: 10.1007/978-981-96-1747-0_36]
97. Grabb D, Lamparth M, Vasan N. Risks from language models for automated mental healthcare: ethics and structure for implementation (Extended Abstract). AIES '24: Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society. 2024;7(1). [doi: 10.1609/aies.v7i1.31654]
98. What Works Clearinghouse Procedures Handbook (Version 41). Institute of Education Sciences, US Department of Education; 2020.
99. Jeong H, Yoo JH, Goh M. Virtual agents in internet-based cognitive behavioral therapy: enhancing engagement and alleviating depression. Presented at: 2023 IEEE International Conference on Agents (ICA); Dec 4-6, 2023; Kyoto, Japan. [doi: 10.1109/ICA58824.2023.00019]
100. Turner RM, Bird SM, Higgins JPT. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PLoS ONE. 2013;8(3):e59202. [doi: 10.1371/journal.pone.0059202] [Medline: 23544056]
101. Hedges LV, Pigott TD. The power of statistical tests for moderators in meta-analysis. Psychol Methods. Dec 2004;9(4):426-445. [doi: 10.1037/1082-989X.9.4.426] [Medline: 15598097]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
CBT: cognitive behavioral therapy
df: degrees of freedom
GenAI: generative artificial intelligence
GPT: generative pre-trained transformer
LLM: large language model
LSTM: long short-term memory
NA: not available
NLG: natural language generation
NLP: natural language processing
PI: prediction intervals
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
REML: restricted maximum likelihood
RoB2: revised Cochrane risk-of-bias tool for randomized trials
ROBINS-I: Risk Of Bias In Non-randomised Studies – of Interventions
SC: some concerns
SE: standard error
SMD: standardized mean difference
WEIRD: western, educated, industrialized, rich, and democratic

Edited by Gunther Eysenbach; peer-reviewed by Bruna Oewel, Fujian Song; submitted 29.May.2025; final revised version received 13.Nov.2025; accepted 18.Nov.2025; published 16.Dec.2025

Please cite as:
Zhang Q, Zhang R, Xiong Y, Sui Y, Tong C, Lin FH
Generative AI Mental Health Chatbots as Therapeutic Tools: Systematic Review and Meta-Analysis of Their Role in Reducing Mental Health Issues
J Med Internet Res 2025;27:e78238
URL: https://www.jmir.org/2025/1/e78238
doi: 10.2196/78238