npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02566-w

Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms

Jun-Seok Sohn[1], Byeong-Gwan Ha[2], SoHyun Park[3], Jae-Jin Kim[4,5], Eojin Lee[4], Hyangkyeong Oh[4], San Lee[6] ✉ & Eunjoo Kim[4,5] ✉

Mental health chatbots have proliferated rapidly, yet their effectiveness remains unclear. This systematic review and meta-analysis included randomized controlled trials comparing chatbots with any control condition for depressive and/or anxiety outcomes. PubMed, Embase, PsycINFO, Scopus and Web of Science were searched from January 2017 to October 2025. Risk of bias was assessed using the revised Cochrane tool. Pooled effect sizes (Hedges' *g*) were calculated using random-effects models. Of the 39 eligible studies, 38 (n = 7,401) were analyzed for depression and 34 (n = 7,621) for anxiety. Chatbots produced statistically significant reductions in depressive (g = 0.31, 95% CI [0.17, 0.46]) and anxiety symptoms (g = 0.28, 95% CI [0.05, 0.51]) compared with controls. Subgroup analyses for depressive symptoms showed larger effects in clinical and subclinical than in nonclinical samples (p = 0.001). Contemporary chatbots thus appear to alleviate depressive and anxiety symptoms, especially in individuals with greater depressive severity. (PROSPERO registration: CRD42024598761).

Mental health disorders represent a significant global health burden, affecting an estimated 970 million people (one in eight individuals) in 2019, with depressive and anxiety disorders being the most prevalent[1]. The COVID-19 pandemic further exacerbated this issue, showing about a 25% increase in cases of major depressive disorder (MDD) and anxiety disorders (AD) during the first year of the pandemic[2]. Despite the growing need, access to mental health services remains limited globally due to various barriers, including resource shortages, geographical limitations, cost, and the stigma associated with mental health[3].

In response, digital mental healthcare, encompassing technologies such as teletherapy, mobile applications, and wearable devices, has become a viable approach to overcome these challenges[4]. Within this domain, conversational agent interventions (CAIs), particularly chatbots, are promising. Chatbots are computer programs that simulate human conversations through text, image, or voice interfaces, and they offer several notable advantages: continuous (24/7) availability, anonymity, and cost-effectiveness[5–7]. Together, these features can make mental health support more accessible, especially for people in remote areas or for those who encounter barriers to traditional care.

The evolution of chatbots has been driven by advances in Natural Language Processing (NLP) technologies. Early chatbots utilized rule-based NLP, operating on predefined linguistic rules and keywords. While these systems offered high explainability, they demonstrated limited flexibility in handling complex language nuances, constraining their use to basic tasks such as appointment scheduling and information sharing[8].

The emergence of AI-based NLP marked a transformation in chatbot capabilities. These systems, incorporating machine learning and deep learning techniques, enabled more dynamic and personalized interactions, showing remarkable success in managing linguistic ambiguities and complex language tasks[7]. Notable examples include Woebot[9] and Wysa[10], which deliver cognitive behavioral therapy for depression and anxiety through natural conversations. These AI-powered chatbots demonstrate the potential for sophisticated therapeutic interventions, though they still face challenges in achieving truly natural conversations[11].

The latest breakthrough came with Large Language Models (LLMs), such as OpenAI's Chat GPT[12], and Google's Gemini[13]. These models demonstrate unprecedented capabilities in generating human-like text and understanding context, with particular significance in psychiatry given the central role of language in mental health assessment and treatment[14]. This

[1]Department of Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea. [2]Department of Psychiatry, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea. [3]NAVER Cloud, Seongnam, Republic of Korea. [4]Institute of Behavioral Sciences in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea. [5]Department of Psychiatry, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea. [6]Working Mind Institute, Seongnam, Republic of Korea. ✉e-mail: sanlee@womi.kr; ejkim96@yuhs.ac

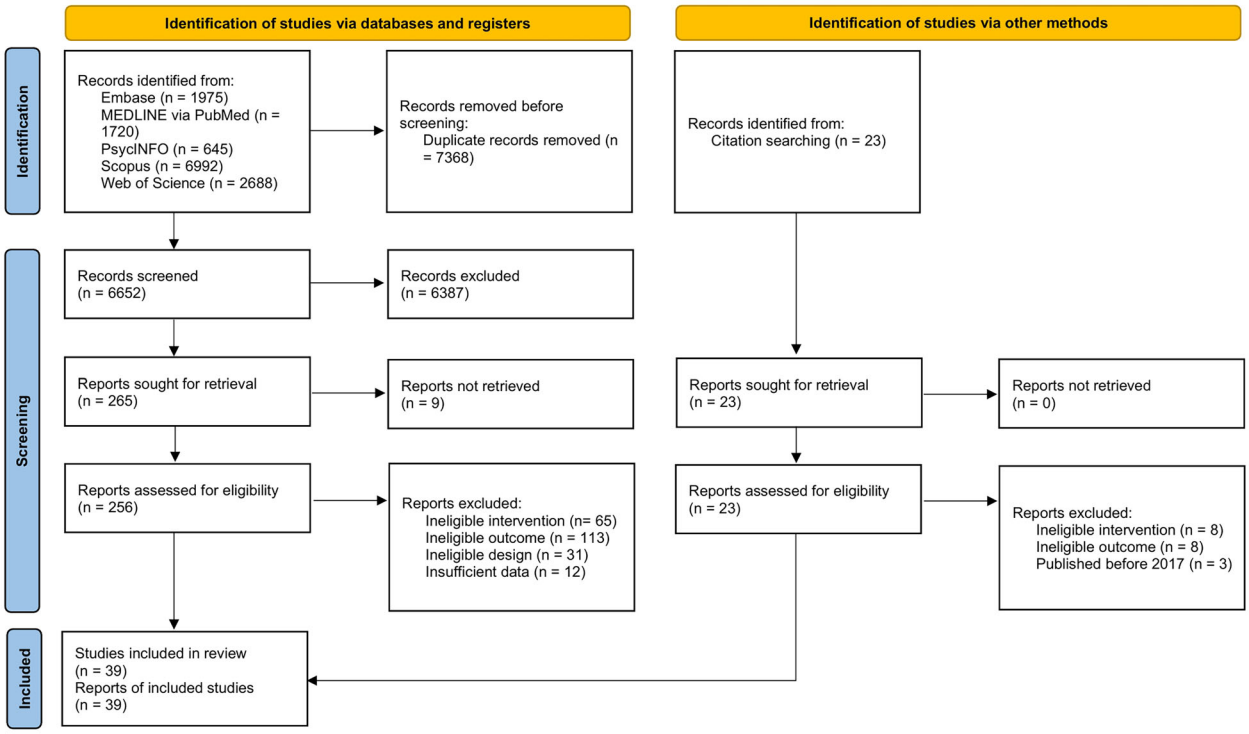

Fig. 1 | PRISMA flow diagram of study selection process. Records were identified through database and citation searching. After duplicate removal, records were screened and excluded based on title/abstract and full-text review. A total of 39 studies were included in the final analyses.

advancement has sparked a rapid proliferation of mental health chatbots with diverse purposes, such as psychological counseling, cognitive behavioral therapy, emotion identification, and personalized mental health journaling[15–17].

Despite the growing prevalence of mental health chatbots, several critical challenges remain. First, there is a notable scarcity of standardized evaluation guidelines for mental health chatbots, hindering systematic assessment of their clinical validity and long-term efficacy[18]. This lack of standardization is further complicated by the inherent complexity of machine learning algorithms, which makes it difficult to trace how these chatbots arrive at specific recommendations or responses—a crucial concern in healthcare applications where understanding the rationale behind clinical suggestions is essential for patient safety and accountability[19]. Most critically, in real-world medical settings, the potential for immediate harm exists when chatbots generate inappropriate or incorrect responses, particularly given the absence of real-time clinical verification mechanisms[20]. Addressing these challenges requires developing robust evaluation frameworks, improving algorithm transparency, and especially establishing empirical evidence through rigorous reviews.

Previous systematic reviews have evaluated the impact of conversational agents on mental health[21,22]. However, these reviews generally addressed a broad range of psychological outcomes, including overall well-being, stress, and general distress, rather than focusing on specific clinical symptom domains. In contrast, the present review provides a symptom-specific synthesis by focusing exclusively on depressive and anxiety symptoms, thereby yielding more homogeneous effect estimates that can more directly inform the efficacy of chatbot-based interventions for these two prevalent and diagnostically defined conditions.

Additionally, one recent review examined NLP-based self-administered interventions for reducing depressive and anxiety symptoms; however, its scope included interactive voice response systems, sentiment-analysis tools, and virtual assistants[23]. In contrast, the current review is deliberately confined to chatbot interventions that engage users through turn-taking, text-based conversational exchanges, thereby improving construct homogeneity and enabling mechanism-aligned comparisons across trials.

Moreover, given the rapid advancement of LLMs since late 2022 and the ensuing proliferation of chatbots, updating the evidence base with recent randomized controlled trials (RCTs) is both timely and essential.

Taken together, these developments and gaps in the current evidence highlight the need for a comprehensive evaluation of chatbot interventions in managing depressive and anxiety symptoms. Such evaluation is crucial for understanding the potential role of chatbots in addressing the growing global mental health burden while maintaining high standards of clinical care.

This systematic review and meta-analysis aimed to evaluate the effectiveness of chatbot interventions in managing depressive and anxiety symptoms. We synthesized evidence from RCTs to outline both clinical and implementation characteristics of chatbot interventions, examined their overall effectiveness compared to control conditions, and analyzed outcomes for depressive and anxiety symptoms separately. We also investigated potential moderating factors, including intervention characteristics and participant profiles.

Results

Study selection

The flow of study selection is presented in a flow diagram adapted from the template proposed in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (see Fig. 1)[24]. The initial database search yielded 14,020 records. Following the removal of duplicates, 6,652 records were screened, of which 265 records warranted full-text evaluations. After this process, 35 studies met the inclusion criteria. Subsequent citation tracking and reference list screening of these initially included studies identified 4 additional eligible studies. Thus, the final analysis comprised 39 studies[9,25–62].

Study characteristics

Table 1 presents the major characteristics of studies included in the review. A total of 39 studies were included, and the majority of studies were conducted in the United States ($n = 10$), followed by China ($n = 7$), Japan ($n = 4$), and Hong Kong ($n = 3$). Several studies were also carried out in

https://doi.org/10.1038/s41746-026-02566-w

Table 1 | Study characteristics of the included studies

Study and sample characteristics					Intervention characteristics			Outcomes		
Author (Year)	**Region**	**Duration**	**Sample size (Female, %)**	**Sample mean age (SD)**	**Population type**	**Primary purpose**	**Chatbot name**	**Response generation approach**	**Depression scale**	**Anxiety scale**
Bird et al. (2018)[42]	UK	immediate	171 (N/A)	21.3 (6.1)	Nonclinical [college students]	Resolving mildly stressful long-standing problems	MYLO	Retrieval-based	DASS-21	DASS-21
Chan et al. (2024)[50]	Hong Kong	6 weeks	129 (76%)	34.1 (12.1)	Clinical [adults with clinically significant insomnia]	Coaching digital CBT for insomnia (dCBTi)	Sleep Sensei	Retrieval-based	PHQ-9	GAD-7
Chen et al. (2025)[53]	Hong Kong	5 months	124 (N/A)	N/A	Nonclinical [paretns recruited through two school principal's networks]	Reducing anxiety and depression levels in the general population	N/A	Generative	PHQ-9	GAD-7
Chua et al. (2024)[51]	Singapore	5 months	118 (50%)	32.6 (4.0)	Nonclinical [heterosexual couples recruite from antenatal clinics]	Supporting parents across the perinatal period	Parentbot	Retrieval-based	EPDS	STAI
Danieli et al. (2022)[29]	Italy	8 weeks	45 (35, 78%)	55.6 (5.1)	Subclinical [workers with stress symptoms and mild-to moderate anxiety]	Promoting mental health and well-being	TEO	Retrieval-based	PHQ-8	GAD-7
de Graaff et al. (2025)[54]	Jordan	8 weeks	60 (82%)	19.7 (1.2)	Subclinical [young people with elevated psychological distress]	Adressing symptoms of depression and anxiety in young people	STARS	Retrieval-based	HSCL-25	HSCL-25
Fitzpatrick et al. (2017)[9]	U.S.	2 weeks	70 (47, 67%)	22.2 (2.3)	Subclinical [young adults with symptoms of depression and anxiety]	Delivering CBT	Woebot	Retrieval-based	PHQ-9	GAD-7
Fitzsimmons-Craft et al. (2022)[32]	U.S.	6 months	700 (N/A)	21.1 (3.1)	Nonclinical [women at high risk for an eating disorder]	Reducing eating disorder risk factors	Tessa	Retrieval-based	PHQ-8	GAD-7
Gong et al. (2020)[45]	Australia	12 months	187 (75, 42%)	57.0 (10.0)	Clinical [adults with type 2 diabetes]	Supporting diabetes self-management	Laura	Retrieval-based	HADS-D	HADS-A
Greer et al. (2019)[47]	U.S.	4 weeks	45 (36, 80%)	25.0 (2.9)	Subclinical [young people after cancer treatment]	Delivering positive psychology skills and promoting well-being	Vivibot	Retrieval-based	PROMIS	PROMIS
He et al. (2022)[21]	China	1 week	148 (55, 37%)	18.8 (0.9)	Subclinical [young adults with depressive symptoms]	Reducing depressive symptoms	XiaoE	Retrieval-based	PHQ-9	-
Heinz et al. (2025)[55]	U.S.	4 weeks	210 (60%)	33.9 (11.0)	Clinical [people with significant symptoms of MDD, GAD, or CHR-FED]	Treating clinical-level mental health symptoms	Therabot	Generative	PHQ-9	GAD-Q-IV
Hunt et al. (2021)[38]	U.S.	8 weeks	121 (91, 75%)	32.0 (10.2)	Clinical [adults diagnosed with IBS]	Applying CBT to IBS	Zemedy	Retrieval-based	PHQ-9; DASS-D	DASS-A
Jang et al. (2021)[44]	South Korea	4 weeks	46 (26, 56%)	25.1 (7.5)	Subclinical [adults with significant attention problems]	Improving attention deficit and its associated psychiatric symptoms	Todaki	Retrieval-based	QIDS-SR	SAS
Karkosz et al. (2021)[35]	Poland	2 weeks	81 (58, 72%)	25.7 (4.6)	Subclinical [young adults with depressive or anxiety symptoms]	Providing mental health support	Fido	Retrieval-based	CERD-R; PHQ-9	STAI
Kleinau et al. (2024)[33]	Malawi	8 weeks	836 (553, 66%)	N/A	Nonclinical [healthcare workers]	Improving mental well-being	Vitalk	Retrieval-based	PHQ-9	GAD-7
Liu et al. (2024)[40]	China	4 days	107 (N/A)	N/A	Nonclinical [general population]	Enhancing psychological resilience	Philobot	Generative	PHQ-9	GAD-7
Liu et al. (2022)[48]	China	16 weeks	83 (46, 55%)	23.1 (1.8)	Nonclinical [college students]	Allevating depression	XiaoNan	Generative	PHQ-9	GAD-7

https://doi.org/10.1038/s41746-026-02566-w

Table 1 (continued) | Study characteristics of the included studies

Study and sample characteristics					Intervention characteristics			Outcomes		
Author (Year)	**Region**	**Duration**	**Sample size (Female, %)**	**Sample mean age (SD)**	**Population type**	**Primary purpose**	**Chatbot name**	**Response generation approach**	**Depression scale**	**Anxiety scale**
MacNeill et al. (2024)[34]	Canada	4 weeks	68 (47, 69%)	42.9 (11.3)	Clinical [people with arthritis or diabetes]	Improving mental health	Wysa	Retrieval-based	PHQ-9	GAD-7
Maeda et al. (2020)[46]	Japan	10 days	927 (927, 100%)	28.8 (3.6)	Nonclinical [women aged 20 to 34 years]	Providing fertility and preconception health education	N/A	Retrieval-based	-	STAI
Nicol et al. (2022)[31]	U.S.	12 weeks	17 (15, 88%)	14.7 (1.7)	Clinical [adolescents diagnosed with depression and anxiety]	Delivering CBT	Woebot-GenZ	Retrieval-based	PHQ-9	GAD-7
Ogawa et al. (2022)[30]	Japan	5 months	20 (9, 45%)	66.0 (8.8)	Clinical [adults diagnosed with Parkinson's disease]	Improving smile and speech in PD	N/A	Retrieval-based	BDI-II	-
Oh et al. (2020)[37]	South Korea	4 weeks	41 (20, 49%)	41.0 (11.9)	Clinical [patients with panic disorder]	Relieving panic symptoms	Todaki	Retrieval-based	HADS-D	HADS-A
Prochaska et al. (2021)[28]	U.S.	8 weeks	180 (117, 65%)	40.8 (12.1)	Subclinical [adults who screened positive for substance misuse]	Reducing substance misuse	Woebot-SUDs	Retrieval-based	PHQ-8	GAD-7
Reilly et al. (2024)[52]	U.S.	7 weeks	42 (19%)	53.7 (15.2)	Clinical [veterans with noncancer chronic pain]	Delievering acceptance and commitment therapy for chronic pain	VACT-CP	Retrieval-based	PHQ-9	-
Romanovskyi et al. (2021)[39]	Ukraine	4 weeks	82 (39, 48%)	20.9 (1.2)	Subclinical [students who identified their tendency to depression, anxiety, and low mood]	Reducing the tendency to anxiety, depression and experiencing negative emotional states	Elomia	Generative	PHQ-9	GAD-7
Sabour et al. (2023)[25]	China	3 weeks	247 (190, 77%)	30.9 (7.5)	Nonclinical [general population]	Reducing mental distress	Emohaa	Retrieval-based; Generative	PHQ-9	GAD-7
Sharp et al. (2025)[56]	Australia	Immediate	60 (83%)	30.9 (11.7)	Clinical [people on waitlists for eading disorder treatment]	Delievering single-session interventions	ED ESSI	Retrieval-based	DASS-21	DASS-21
Six et al. (2025)[57]	U.S.	2 weeks	209 (19%)	20.0 (2.2)	Nonclinical [college student with and without depressive symptoms]	Delievering brief CBT	AirHeart	Retrieval-based	PHQ-8	-
Suharwardy et al. (2023)[41]	U.S.	6 weeks	192 (192, 100%)	34.0 (N/A)	Nonclinical [general postpartum population]	Mood management	Woebot	Retrieval-based	PHQ-9; EPDS	GAD-7
Tong et al. (2025)[58]	Hong Kong	10 days	285 (76%)	26.5 (8.4)	Nonclinical [people aged over 18 years]	Promoting mental health self-care and mental well-being	N/A	Retrieval-based	PHQ-9	GAD-7
Ulrich et al. (2024)[27]	Swiss	24–54 days	140 (103, 74%)	26.7 (6.3)	Subclinical [university students experiencing stress]	Coaching stress management	MISHA	Retrieval-based	PHQ-9	GAD-7
Ulrich et al. (2024)[49]	Swiss	24–60 days	198 (172, 87%)	38.7 (12.1)	Subclinical [adults with frequent headaches]	Improving mental well-being	BalanceUP	Retrieval-based	PHQ-9	GAD-7
Vereschagin et al. (2024)[36]	Canada	30 days	1489 (1045, 70%)	20.0 (N/A)	Nonclinical [college students]	Improving mental health and substance use outcomes	Minder	Retrieval-based	PHQ-9	GAD-7
Xu et al. (2025)[59]	China	16 weeks	84 (49%)	23.3 (1.1)	Subclinical [college students with depressive symptoms]	Delievering CBT	Neil	Retrieval-based	PHQ-9	GAD-7
Yasukawa et al. (2024)[26]	Japan	8 weeks	143 (52, 36%)	41.4 (11.1)	Subclinical [employees with subthreshold depression]	Improving adherence to iCBT	EPO	Retrieval-based	PHQ-9	GAD-7

Table 1 (continued) | Study characteristics of the included studies

Study and sample characteristics					Intervention characteristics			Outcomes		
Author (Year)	Region	Duration	Sample size (Female, %)	Sample mean age (SD)	Population type	Primary purpose	Chatbot name	Response generation approach	Depression scale	Anxiety scale
Ye et al. (2025)[60]	China	2 weeks	40 (60%)	12.6 (N/A)	Nonclinical [secondary school students aged between 12 and 14 years]	Mitigating depression and anxiety among students	WarmGPT	Generative	CDI	SCARED
Yokotani et al. (2025)[61]	Japan	8 weeks	310 (55%)	19.0 (1.3)	Nonclinical [university students]	Reducing symptoms of anxiety by encouraging negative emotional expressions	UP chatbot	Retrieval-based	ODSIS	OASIS
Zhao et al. (2025)[62]	China	28 days	657 (62%)	20.6 (2.0)	Subclinical [young adults with mild depressive or anxiety symptoms]	alleviating negative emotions	Douyin compainon bot	Generative	PHQ-9	GAD-7

Among the scales, the underlined ones were selected for the meta-analyses.
BDI-II Beck Depression Inventory-II, *CBT* Cognitive Behavioral Therapy, *CDI* Children's Depression Inventory, *CESD-R* Centerfor Epidemiologic Studies Depression Scale-Revised, *CHR-FED* Clinically High Risk for Feeding and Eating Disorders, *DASS* Depression Anxiety and Stress Scale, *EPDS* Edinburgh Postnatal Depression Scale, *GAD* Generalized Anxiety Disorder, *HADS* Hospital Anxiety and Depression Scale, *HSCL* Hopkins Symptom CheckList, *MDD* Major Depressive Disorder, *OASIS* Overall Anxiety Severity and Impairment Scale, *ODSIS* Overall Depression Severity and Impairment Scale, *PHQ* Patient Health Questionnaire, *PROMIS* Patient-Reported Outcomes Measurement Information System, *QIDS-SR* Quick Inventory of Depressive Symptomatology-Self Report, *SAS* Zung Self-Rating Anxiety Scale, *SCARED* Screen for Child Anxiety Related Disorders, *STAI* State-Trait Anxiety Inventory.

Australia, South Korea, Canada, and Switzerland ($n = 2$ each). Single studies were conducted in the United Kingdom, Singapore, Italy, Jordan, Poland, Malawi, and Ukraine ($n = 1$ each).

Regarding population characteristics, most studies targeted non-clinical samples ($n = 15$), followed by sub-clinical ($n = 14$) and clinical populations ($n = 10$).

A variety of self-report instruments were employed to assess depressive symptoms, with the Patient Health Questionnaire-9 (PHQ-9) being the most frequently used measure ($n = 23$), followed by the Patient Health Questionnaire-8 (PHQ-8) ($n = 4$). Other tools included the Depression subscale of the Hospital Anxiety and Depression Scale (HADS-D), Depression Anxiety and Stress Scale-21 (DASS-21), Edinburgh Postnatal Depression Scale (EPDS), Beck Depression Inventory-II (BDI-II), Quick Inventory of Depressive Symptomatology–Self Report (QIDS-SR), Patient-Reported Outcomes Measurement Information System (PROMIS), Hopkins Symptom Checklist (HSCL-25), Children's Depression Inventory (CDI), Center for Epidemiologic Studies Depression Scale-Revised (CESD-R), and Overall Depression Severity and Impairment Scale (ODSIS).

For anxiety outcomes, the Generalized Anxiety Disorder-7 (GAD-7) was the most commonly administered instrument ($n = 21$). Other measures included the Anxiety subscale of the Hospital Anxiety and Depression Scale (HADS-A), State-Trait Anxiety Inventory (STAI), Depression Anxiety and Stress Scale-21 (DASS-21), Zung Self-Rating Anxiety Scale (SAS), Patient-Reported Outcomes Measurement Information System (PROMIS), Hopkins Symptom Checklist (HSCL-25), Screen for Child Anxiety Related Emotional Disorders (SCARED), Generalized Anxiety Disorder Questionnaire-IV (GAD-Q-IV), and Overall Anxiety Severity and Impairment Scale (OASIS), indicating substantial methodological diversity in outcome assessment across studies.

As shown in Fig. 2, the number of RCTs examining chatbot interventions has increased steadily over time. Among the studies included, 31 employed retrieval-based chatbots, 8 utilized generative AI-based chatbots, and one study incorporated both approaches. Notably, half of the studies using generative AI-based chatbots were published in 2025.

Risk of Bias

Risk of bias assessment was conducted using the revised Cochrane Risk of Bias tool for randomized trials (RoB 2), and the results are summarized in Supplementary Fig. S2. Among the 39 included studies, 35 were judged to have a high overall risk of bias. The most frequently flagged domain was Domain 4 (bias in measurement of the outcome). In many cases, outcome measurements were based on participant self-reports.

Result of Synthesis

Meta-analysis of studies revealed therapeutic effects of chatbots, with varying levels of statistical significance across conditions. For depressive outcomes, 38 RCTs ($N = 7{,}401$) demonstrated that participants interacting with chatbots showed statistically significant symptom reductions compared with various control conditions ($g = 0.31$, 95% CI [0.17, 0.46]), accompanied by significant heterogeneity ($Q = 153.10$, $p < 0.001$, I^2 = 85.34%) (see Fig. 3)[9,25–45,47–62]. Analysis of anxiety outcomes from 34 RCTs ($N = 7{,}621$) also indicated a statistically significant effect of chatbots compared with control conditions ($g = 0.28$, 95% CI [0.05, 0.51]), with significant heterogeneity ($Q = 207.91$, $p < 0.001$, I^2 = 94.32%) (see Fig. 4)[9,25,27–29,31–42,44–51,53–56,58–62].

Publication bias

For depressive outcomes, statistically significant publication bias was detected through both Egger's test ($t = 3.31$, df = 36, $p = 0.002$) and modified Egger's test ($t = 2.28$, df = 36, $p = 0.007$). For anxiety outcomes, publication bias analysis showed no significant asymmetry in either Egger's test ($t = 1.71$, df = 32, $p = 0.098$) or modified Egger's test ($t = 1.11$, df = 32, $p = 0.276$).

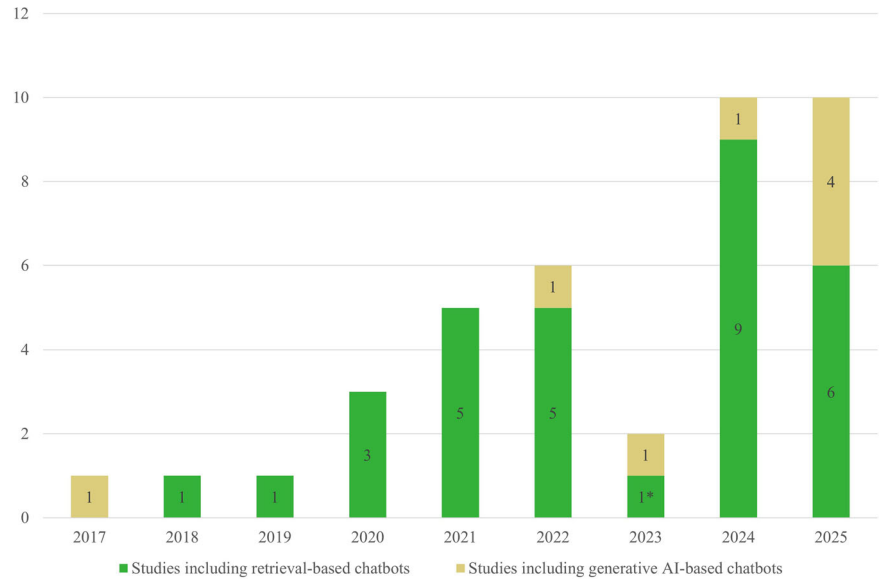

Fig. 2 | Growth of chatbot intervention studies over time, categorized by chatbot type. * One study published in 2023 included both retrieval-based and generative AI-based chatbot components and was categorized under generative AI-based chatbots.

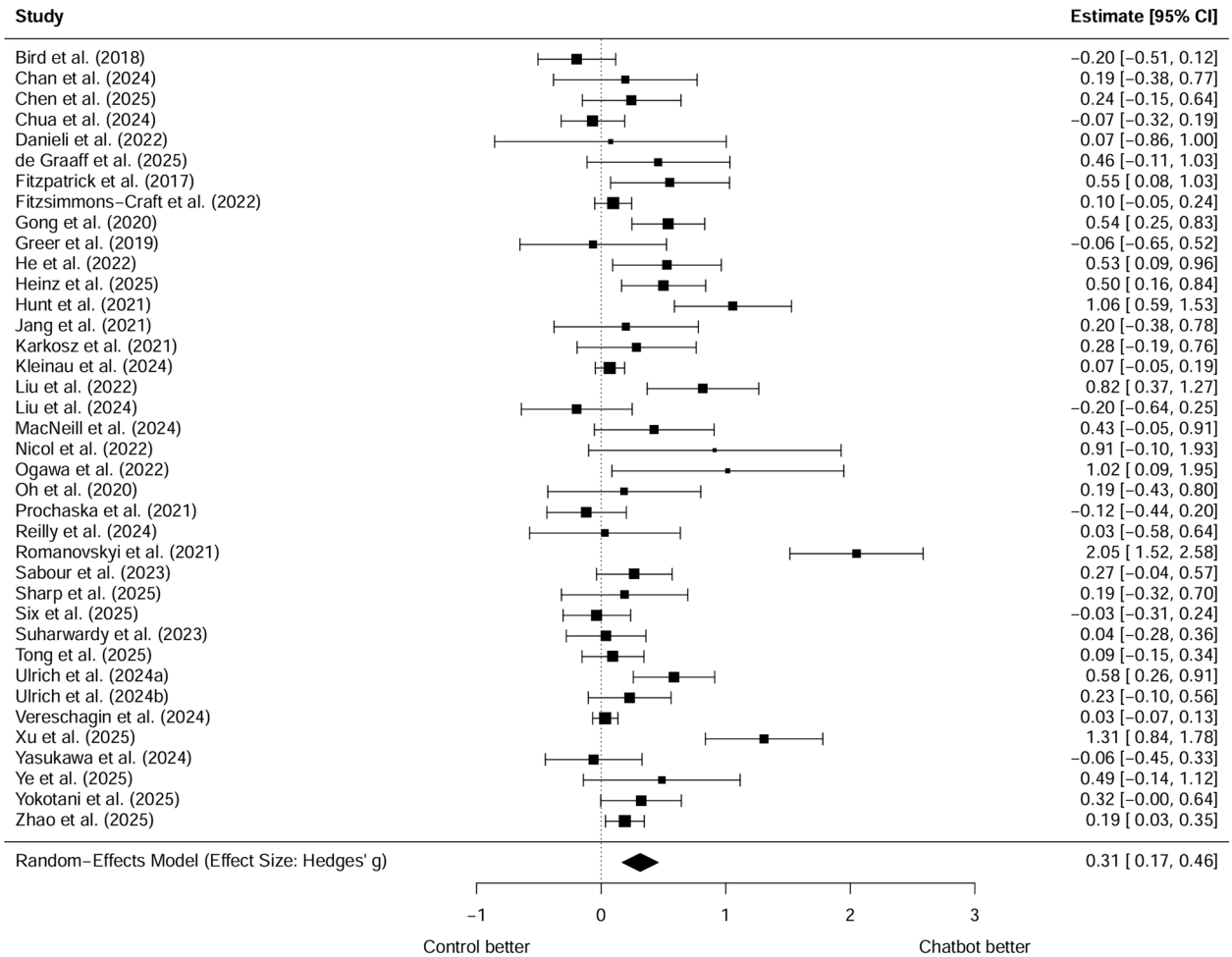

Fig. 3 | The effect of chatbot on depressive symptoms. Forest plot of studies reporting the effect of chatbot on depressive symptoms.

Subgroup analyses

A statistically significant subgroup difference for depressive symptom outcomes was observed based on participants' clinical status ($p = 0.001$), with larger effect sizes found in clinical ($g = 0.64$, 95% CI [0.20, 1.09]) and sub-clinical populations ($g = 0.34$, 95% CI [0.13, 0.55]) than in non-clinical populations ($g = 0.07$, 95% CI [−0.01, 0.15]) (see Fig. 5). No other moderators (including intervention primary purpose, chatbot response-generation method, type of control group, study duration, or assessment scales) showed significant difference (see Supplementary material).

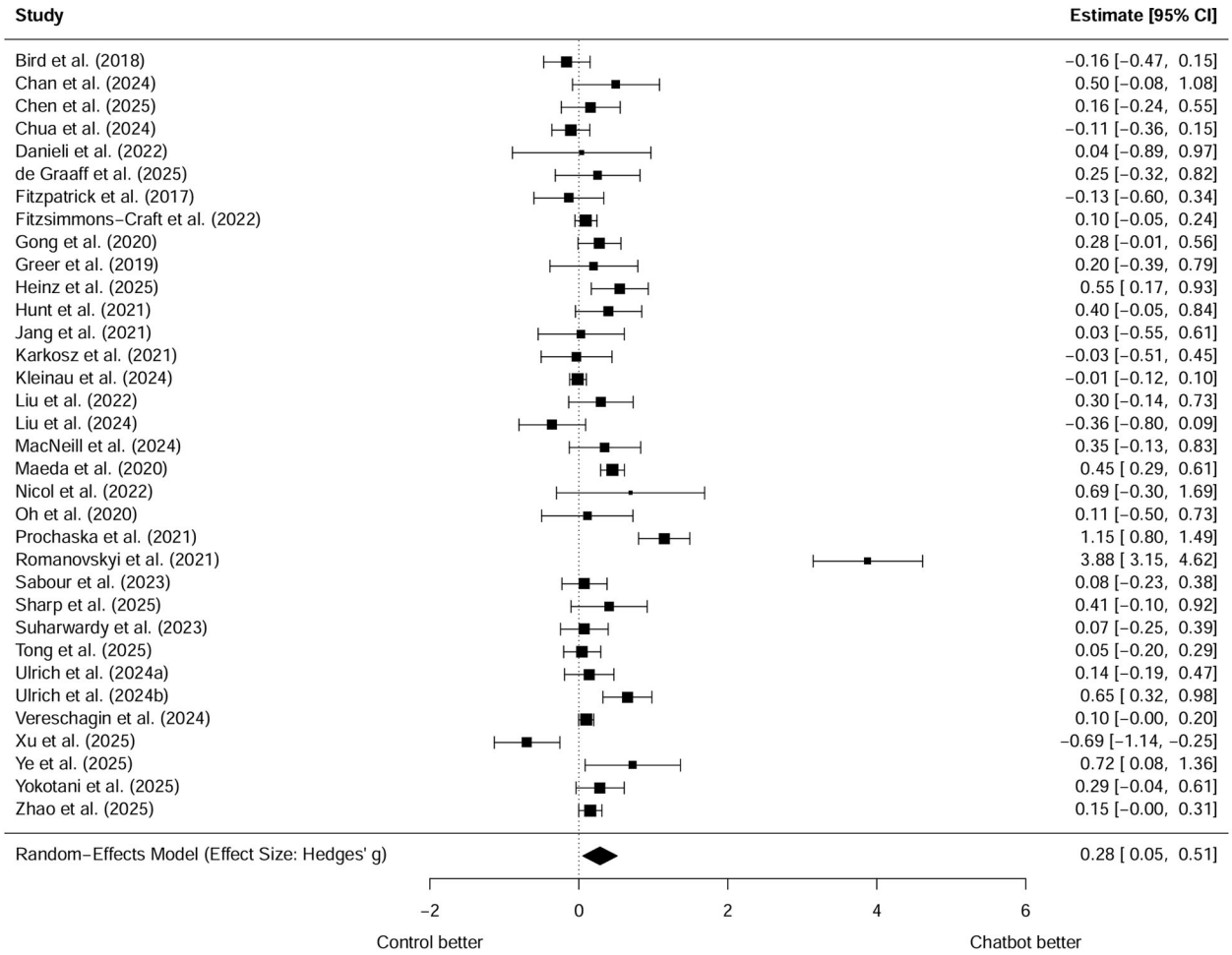

Fig. 4 | The effect of chatbot on anxiety symptoms. Forest plot of studies reporting the effect of chatbot on anxiety.

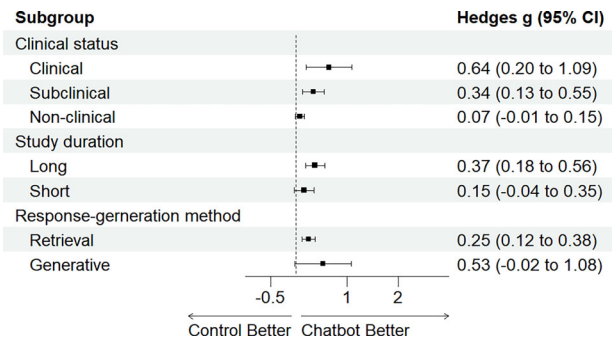

Fig. 5 | Subgroup analyses of the effect of chatbot on depressive symptoms. Selected subgroup analyses are presented according to clinical status, study duration, and response-generation method.

Discussion

In this systematic review and meta-analysis, we synthesized evidence on the effectiveness of chatbots in mental health care. Overall, chatbot interventions showed statistically significant benefits in alleviating both depressive ($g = 0.31$) and anxiety ($g = 0.28$) symptoms compared with various control conditions. Subgroup analyses further revealed that these beneficial effects on depressive symptoms were more pronounced among individuals in clinical or subclinical populations than among those in nonclinical populations.

Our findings extend prior evidence on the benefits of chatbots (or conversational agents) in mental health care by incorporating a larger and more recent pool of RCTs. He et al. (2023) reported statistically significant short-term effects of conversational agent interventions on depressive symptoms ($g = 0.29$), generalized anxiety symptoms ($g = 0.27$), and specific anxiety symptoms ($g = 0.47$)[21]. In contrast, Li et al. (2023) found a statistically significant effect of AI-based conversational agents on depressive symptoms ($g = 0.64$) but no significant improvement in anxiety outcomes[22]. Recently, Villarreal-Zegarra et al. (2024) observed statistically significant effects for depressive symptoms ($g = 0.82$) and smaller but significant effects for anxiety symptoms ($g = 0.27$) in their synthesis of self-administered NLP-based interventions[23].

Compared with these earlier reviews, the present study included an additional 20 RCTs, thereby expanding the evidence base with more recent and methodologically diverse trials. Furthermore, whereas other meta-analyses included at most three generative AI-based chatbots, our review included eight such studies, capturing the most recent technological advances in large language model-driven conversational agents.

The subgroup analyses revealed a statistically significant moderation of treatment effects by participants' clinical status. Specifically, the chatbot produced moderate effects among individuals in clinical population and small effects among those in the subclinical range, whereas its impact was negligible in nonclinical populations.

These findings help clarify previous research, which has shown inconsistent results regarding whether treatment effects differ by baseline symptom severity. For example, one study reported statistically significantly greater reductions in psychological distress among clinical and subclinical groups relative to nonclinical samples[22], whereas another, using a three-level classification (clinical, symptomatic, general), detected no between-group differences in anxiety, quality of life, or stress outcomes[21]. By employing a

comparable three-tier framework, the present study extends prior work in two ways. First, it provides statistically significant evidence that chatbots can alleviate depressive symptoms among users with clinical or near-clinical levels of distress, thereby supporting the role of chatbots as "bridge interventions" that expand access to mental health care and complement more intensive treatments. Second, the markedly attenuated effects in nonclinical participants underscore the limitations of a one-size-fits-all approach. They highlight the need for modular and personalized intervention strategies that flexibly adapt therapeutic content to an individual's symptom severity and treatment goals, rather than delivering an identical protocol to all users.

Nevertheless, the interpretation of these subgroup findings warrants caution. Subgroup analyses in meta-analyses are inherently observational and examine between-study rather than within-study differences, making them vulnerable to confounding by study-level characteristics[63]. Moreover, the greater apparent improvement among clinical samples may partly reflect regression to the mean (RTM), a statistical phenomenon in which individuals with extreme baseline scores tend to move toward the average upon retesting, independent of any true intervention effect[64,65]. Participants with higher initial symptom levels thus have greater potential for natural reduction over time, whereas those with low baseline scores (i.e., nonclinical samples) show limited scope for further improvement. Accordingly, observed moderation by clinical status should be interpreted with caution.

Although a recent review reported significantly larger effect sizes for generative AI-based conversational agents than for retrieval-based systems in alleviating psychological distress[22], our subgroup analysis detected no statistically significant differences between the two chatbot types for either depressive or anxiety outcomes. Notably, the subgroup meta-analysis of generative-AI-based chatbots did not reach statistical significance. However, this finding should be interpreted with caution, as it may reflect insufficient statistical power due to the small number of included studies rather than a lack of efficacy. Further RCTs using generative AI-based chatbots are required to ensure sufficient statistical power and support more robust conclusions.

Although the point estimate for the long-duration studies was approximately two to three times larger than that for short-duration studies, the formal statistical test for subgroup differences was not significant. However, the absence of a statistically significant subgroup difference does not imply that short- and long-duration chatbot interventions are equivalent. Several methodological factors may have obscured a true duration effect: (i) limited statistical power due to the relatively small number of studies (nine or ten in the short-duration subgroup); (ii) substantial between-study heterogeneity; (iii) a coarse binary classification of duration; and (iv) potential confounding by intervention dose and fidelity. Given these limitations, the findings should be interpreted with caution and not be used to justify abbreviated interventions. The absence of a statistically significant difference indicates insufficient evidence to confirm a duration effect rather than evidence that no such effect exists.

There are well-recognized concerns regarding the safety of digital interventions in psychiatry, including the potential risks of generative AI models with unrestricted conversational outputs[66,67]. However, unlike conventional RCTs such as drug studies, these issues are rarely reported in RCTs evaluating digital interventions. Among the studies included in our review, only a few explicitly addressed safety considerations. For instance, Nicol et al. (2022) implemented safety protocols to detect suicide risk among users[31]; Prochaska et al. (2021) reported no serious adverse events related to study participation[28]; and Heinz et al. (2025) documented 15 instances of staff intervention for participant safety concerns (e.g., expressions of suicidal ideation) and 13 interventions to correct inappropriate chatbot responses, such as the provision of medical advice[55]. In contrast, 23 of the 39 included studies did not report any systematic safety monitoring or adverse-event data. This pattern underscores an important issue in mental health chatbot research: while efficacy outcomes are increasingly well characterized, safety monitoring and transparent adverse-event reporting remain underdeveloped.

This review has several limitations. First, Egger-type tests indicated possible publication bias for depressive outcomes but not for anxiety outcomes. However, these tests are acknowledged to have reduced accuracy when between-study heterogeneity is high, so the results should be interpreted with caution[68]. Second, most of the studies relied on participant-completed self-report scales rather than clinician-rated instruments. The exclusive use of self-reported measures can magnify treatment effects through shared-method variance and response biases, restricting the generalizability of our findings. Third, most trials were assessed to be at high risk of bias, chiefly because participant blinding could not be maintained throughout the intervention period. While initial group allocation might have been blinded, participants likely became aware of their group assignment during the intervention period, potentially influencing their responses on post-intervention self-reported outcome measures. Fourth, the evidence base for both retrieval-based agents and, more recently, generative AI-based chatbots is still emerging. This early stage is reflected in marked variability in trial design, intervention protocols, control conditions, and outcome definitions. Such methodological diversity translated into high statistical heterogeneity across both outcomes, limiting the precision and generalizability of our pooled estimates even under random-effects modelling.

In sum, our findings indicate that contemporary chatbot-based mental health interventions yield small yet statistically significant benefits in alleviating both depressive and anxiety symptoms. Moreover, as chatbots showed greater effectiveness among users with clinical or sub-clinical symptom severity but limited impact in non-clinical populations, future interventions would ideally be tailored to individual symptom profiles rather than employing a uniform approach for all users. In this context, the recent rise of generative AI-based chatbots may offer a promising avenue for delivering more personalized and adaptive mental health interventions. However, the transition to generative AI also introduces important considerations regarding safety, reliability, and ethical use, particularly in relation to sensitive mental health data and the potential for hallucinated or inappropriate outputs[69].

To consolidate and extend this evidence base, future trials are encouraged to (a) adopt standardized clinician-rated outcome measures; (b) implement prespecified safety protocols and systematically report adverse events associated with chatbot use; (c) rigorously compare generative AI-based and retrieval-based chatbots using standardized designs that equate key parameters such as session length, frequency, and conversational objectives; (d) report long-term follow-up; (e) recruit socio-culturally and linguistically diverse samples to enhance generalizability. Addressing these points could enhance the precision of effect estimates and provide stronger evidence for the effective and safe deployment of chatbots in mental health care.

Methods

The findings of this study were reported in accordance with the PRISMA 2020 statement[24]. The review protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) (CRD42024598761).

Eligibility criteria

The eligibility criteria were established utilizing the PICOS framework, which consists of the following components:

(1) Population: participants from all demographic characteristics and groups; (2) Intervention: chatbots as the primary intervention tool; (3) Comparator: any type of comparison group; (4) Outcome: depressive and anxiety symptoms, measured through validated scales (e.g., Patient Health Questionnaire-9, Generalized Anxiety Disorder-7, or similar questionnaires); and (5) Study design: RCTs only.

The following exclusion criteria were applied: (1) Studies not available as full-text articles or written in languages other than English; (2) Commentaries, editorials, case reports, reviews, conference abstracts, or non-

peer-reviewed literature; and (3) primary studies employing quasi-experimental, non-randomized, or uncontrolled designs.

Search strategy

To identify potentially relevant articles, the following databases were searched: MEDLINE (via PubMed), Embase, PsycINFO, Scopus, and Web of Science. The search initially covered publications from 1 January 2017 to 10 December 2024. This period was selected based on preliminary results indicating that the majority of relevant articles were published during this time[70]. Furthermore, this approach helps maintain the validity of the findings by focusing on recent developments and excluding outdated or less pertinent research[71]. To ensure inclusion of the most recent evidence, the same search strategy was repeated on 13 October 2025 to capture newly published RCTs, which identified 13 additional eligible trials.

The search strategy was developed based on three core concepts: (1) intervention (chatbot, conversational agent, or similar digital conversational systems), (2) outcome (depressive, anxiety or related emotional symptoms), and (3) underlying computational techniques (natural language processing, machine learning, or large language models, or related text-based AI methods).

Search terms were initially drafted by a single reviewer [SL] and further refined through team discussions. The final search results were exported into Microsoft Excel, where duplicates were removed manually using built-in sorting and filtering functions. Manual searches of the reference lists of relevant reviews were also conducted to identify additional eligible studies. The detailed search strings for each database are provided in Supplementary Table S1.

Study selection and screening

Title and abstract screening were performed manually by two independent reviewers [JSS and BGH]. Full-text screening was subsequently conducted. Any discrepancies between reviewers were resolved through discussion, with arbitration by a third reviewer [EK] when necessary.

Data extraction

The extracted information encompassed basic study characteristics (title, authors, publication date, country, study design), participant demographics (target population, sample size, mean age, and gender distribution), intervention and control specifications, and outcome measures. Study results that assessed depressive and anxiety symptoms were extracted, including means and standard deviations at baseline, post-intervention, and follow-up periods where available, or alternative statistical metrics such as effect sizes, standard errors, and confidence intervals (CIs). For studies reporting both intention-to-treat (ITT) and per protocol (PP) analyses, ITT data were prioritized for extraction.

Risk of bias assessment

Risk of bias in the included studies was evaluated using the RoB 2[72]. The assessment encompassed five specific 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 outcome measurement; and (5) bias in selection of reported results.

Two reviewers [JSS and BGH] independently conducted the assessment for each included study. For each domain, reviewers assigned one of three possible judgements: low risk, high risk, or some concerns. Discrepancies in assessments were resolved through consensus discussions between the reviewers. In cases where consensus could not be reached, a third reviewer [EK] served as an arbiter. Following the standardized protocol, an overall risk of bias judgement was synthesized for each study.

Meta-analysis

The between-group differences were calculated using post-intervention outcomes from the intervention and control groups. Effect sizes were calculated as standardized mean differences (SMDs), specifically Hedges' g, with corresponding 95% CIs for both depressive and anxiety symptoms.

Meta-analyses were performed in R software (version 4.4.3) with the metafor package (version 4.6-0) using a random-effects model. The between-study variance was estimated using restricted maximum likelihood (REML), followed by the calculation of the overall effect size. CI for the summary effect was computed using the Hartung-Knapp-Sidik-Jonkman adjustment, which provides more conservative estimates by accounting for uncertainty in the variance estimation[11].

Heterogeneity analysis

Heterogeneity between studies was assessed using multiple approaches: the Cochran Q statistics to examine the presence of heterogeneity, the I^2 and H^2 statistics to quantify the proportion of variance attributable to heterogeneity, and the between-study variance (τ^2) to assess the magnitude of variance in effect sizes across studies.

Publication bias analysis

To assess potential publication bias, we implemented multiple analytical approaches. First, funnel plots were generated for visual inspection of asymmetry. Second, Egger's regression test was conducted to examine funnel plot asymmetry[73]. Given the concerns about potential false-positive results when applying Egger's test to SMDs, we additionally performed the Pustejovsky-Rodgers' approach[74].

Subgroup analysis

Subgroup analyses were conducted to explore whether the pooled effects differed according to multiple categorical moderators: (1) participants' clinical status (clinical, sub-clinical, non-clinical), (2) the primary therapeutic aim of the intervention (direct vs indirect targeting of depressive/anxiety symptoms), (3) the chatbot's response-generation method (retrieval-based vs generative), (4) type of control condition (wait-list/assessment-only, informational/attentional, other active interventions), (5) study duration (long ≥ 4 weeks vs short < 4 weeks), and (6) assessment scales employed. All subgroup analyses were carried out separately for depressive and anxiety outcomes.

For each moderator, we fitted an independent random-effect model identical to the main analysis. This produced the subgroup's pooled SMD, its 95% CI, between-study variance, and the number of contributing comparisons. The set of subgroup estimates was then analyzed in a fixed-effect meta-regression. The resulting omnibus Q test, with degrees of freedom equal to the number of subgroups minus one, evaluated the null hypothesis that all subgroup effects are equal. A two-sided $p < 0.05$ was interpreted as evidence of a statistically significant difference between subgroups.

Data availability

The data generated or analyzed during this study are included in this published article and its supplementary information files. Further inquiries can be directed to the corresponding author.

Received: 8 August 2025; Accepted: 9 March 2026;
Published online: 25 March 2026

References

1. Institute of Health Metrics and Evaluation. Global Health Data Exchange (GHDx). https://vizhub.healthdata.org/gbd-results. (2025).
2. Santomauro, D. F. et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *Lancet*. **398**, 1700–712 (2021).
3. Mongelli, F., Georgakopoulos, P. & Pato, M. T. Challenges and Opportunities to Meet the Mental Health Needs of Underserved and Disenfranchised Populations in the United States. *Focus (Am. Psychiatr. Publ.)* **18**, 16–24 (2020).
4. Tal, A. & Torous, J. The digital mental health revolution: Opportunities and risks. *Psychiatr. Rehabil. J.* **40**, 263–265 (2017).
5. Jabir, A. I. et al. Attrition in Conversational Agent-Delivered Mental Health Interventions: Systematic Review and Meta-Analysis. *J. Med Internet Res* **26**, e48168 (2024).

6. Tudor Car, L. et al. Conversational Agents in Health Care: Scoping Review and Conceptual Analysis. *J. Med Internet Res.* **22**, e17158 (2020).
7. Laymouna, M. et al. Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review. *J. Med Internet Res.* **26**, e56930 (2024).
8. Mnasri, M. Recent advances in conversational NLP: Towards the standardization of Chatbot building. *arXiv preprint arXiv:1903.09025* (2019).
9. Fitzpatrick, K. K., Darcy, A. & Vierhile, M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment. Health.* **4**, e19 (2017).
10. Inkster, B., Sarda, S. & Subramanian, V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR Mhealth Uhealth.* **6**, e12106 (2018).
11. Veroniki, A. A. et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth. Methods* **7**, 55–79 (2016).
12. Achiam, J. et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023)..
13. Team, G. et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
14. Demszky, D. et al. Using large language models in psychology. *Nat. Rev. Psychol.* **2**, 688–701 (2023).
15. Peretz, G., Taylor, C. B., Ruzek, J. I., Jefroykin, S. & Sadeh-Sharvit, S. Machine Learning Model to Predict Assignment of Therapy Homework in Behavioral Treatments: Algorithm Development and Validation. *JMIR Form. Res* **7**, e45156 (2023).
16. Tanana, M. J. et al. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behav. Res Methods* **53**, 2069–2082 (2021).
17. Chen, Y. et al. Structured Dialogue System for Mental Health: An LLM Chatbot Leveraging the PM+ Guidelines. *Springer*, *Singapore*. **15170**, 262–271 (2025).
18. Koh, J., Tng, G. Y. Q. & Hartanto, A. Potential and Pitfalls of Mobile Mental Health Apps in Traditional Treatment: An Umbrella Review. *J. Pers. Med.* **12**, https://doi.org/10.3390/jpm12091376.(2022).
19. Abd-Alrazaq, A. A. et al. An overview of the features of chatbots in mental health: A scoping review. *Int J. Med Inf.* **132**, 103978 (2019).
20. Shiferaw, M. W., Zheng, T., Winter, A., Mike, L. A. & Chan, L.-N. Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. *BMC Med. Inform. Decis. Mak.* **24**, 404 (2024).
21. He, Y. et al. Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. *J. Med Internet Res* **25**, e43862 (2023).
22. Li, H., Zhang, R., Lee, Y. C., Kraut, R. E. & Mohr, D. C. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digit Med* **6**, 236 (2023).
23. Villarreal-Zegarra, D. et al. Self-Administered Interventions Based on Natural Language Processing Models for Reducing Depressive and Anxiety Symptoms: Systematic Review and Meta-Analysis. *JMIR Ment. Health.* **11**, e59560 (2024).
24. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
25. Sabour, S. et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. *Front Digit Health* **5**, 1133987 (2023).
26. Yasukawa, S. et al. A chatbot to improve adherence to internet-based cognitive-behavioural therapy among workers with subthreshold depression: a randomised controlled trial. *BMJ Ment. Health* **27**, https://doi.org/10.1136/bmjment-2023-300881.(2024).
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* **12**, e54945 (2024).
28. Prochaska, J. J. et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug Alcohol Depend.* **227**, 108986 (2021).
29. Danieli, M. et al. Assessing the Impact of Conversational Artificial Intelligence in the Treatment of Stress and Anxiety in Aging Adults: Randomized Controlled Trial. *JMIR Ment. Health* **9**, e38067 (2022).
30. Ogawa, M. et al. Can AI make people happy? The effect of AI-based chatbot on smile and speech in Parkinson's disease. *Parkinsonism Relat. Disord.* **99**, 43–46 (2022).
31. Nicol, G., Wang, R., Graham, S., Dodd, S. & Garbutt, J. Chatbot-Delivered Cognitive Behavioral Therapy in Adolescents With Depression and Anxiety During the COVID-19 Pandemic: Feasibility and Acceptability Study. *JMIR Form. Res* **6**, e40242 (2022).
32. Fitzsimmons-Craft, E. E. et al. Effectiveness of a chatbot for eating disorders prevention: A randomized clinical trial. *Int J. Eat. Disord.* **55**, 343–353 (2022).
33. Kleinau, E. et al. Effectiveness of a chatbot in improving the mental wellbeing of health workers in Malawi during the COVID-19 pandemic: A randomized, controlled trial. *PLoS One* **19**, e0303370 (2024).
34. MacNeill, A. L., Doucet, S. & Luke, A. Effectiveness of a Mental Health Chatbot for People With Chronic Diseases: Randomized Controlled Trial. *JMIR Form. Res* **8**, e50025 (2024).
35. Karkosz, S., Szymański, R., Sanna, K. & Michałowski, J. Effectiveness of a Web-based and Mobile Therapy Chatbot on Anxiety and Depressive Symptoms in Subclinical Young Adults: Randomized Controlled Trial. *JMIR Form. Res* **8**, e47960 (2024).
36. Vereschagin, M. et al. Effectiveness of the Minder Mobile Mental Health and Substance Use Intervention for University Students: Randomized Controlled Trial. *J. Med Internet Res* **26**, e54287 (2024).
37. Oh, J., Jang, S., Kim, H. & Kim, J. J. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. *Int J. Med Inf.* **140**, 104171 (2020).
38. Hunt, M., Miguez, S., Dukas, B., Onwude, O. & White, S. Efficacy of Zemedy, a Mobile Digital Therapeutic for the Self-management of Irritable Bowel Syndrome: Crossover Randomized Controlled Trial. *JMIR Mhealth Uhealth* **9**, e26152 (2021).
39. Romanovskyi, O., Pidbutska, N. & Knysh, A. in *International Conference on Computational Linguistics and Intelligent Systems*.
40. Liu, I., Chen, W., Ge, Q., Song, D. & Ni, S. in *Proceedings of the Tenth International Symposium of Chinese CHI* 216–221 (Association for Computing Machinery, Guangzhou, China and Online, China, 2024).
41. Suharwardy, S. et al. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. *AJOG Glob. Rep.* **3**, 100165 (2023).
42. Bird, T., Mansell, W., Wright, J., Gaffney, H. & Tai, S. Manage Your Life Online: A Web-Based Randomized Controlled Trial Evaluating the Effectiveness of a Problem-Solving Intervention in a Student Sample. *Behav. Cogn. Psychother.* **46**, 570–582 (2018).
43. He, Y. 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* **24**, e40719 (2022).
44. Jang, S. et al. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. *Int J. Med Inf.* **150**, 104440 (2021).
45. Gong, E. et al. My Diabetes Coach, a Mobile App-Based Interactive Conversational Agent to Support Type 2 Diabetes Self-Management: Randomized Effectiveness-Implementation Trial. *J. Med Internet Res* **22**, e20322 (2020).
46. Maeda, E. et al. Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial. *Reprod. Biomed. Online* **41**, 1133–1143 (2020).

47. Greer, S. et al. Use of the Chatbot "Vivibot" to Deliver Positive Psychology Skills and Promote Well-Being Among Young People After Cancer Treatment: Randomized Controlled Feasibility Trial. *JMIR Mhealth Uhealth.* **7**, e15018 (2019).
48. Liu, H., Peng, H., Song, X., Xu, C. & Zhang, M. Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Inter.* **27**, 100495 (2022).
49. Ulrich, S. et al. Development and Evaluation of a Smartphone-Based Chatbot Coach to Facilitate a Balanced Lifestyle in Individuals With Headaches (BalanceUP App): Randomized Controlled Trial. *J. Med Internet Res.* **26**, e50132 (2024).
50. Chan, W. S. et al. Assessing the Short-Term Efficacy of Digital Cognitive Behavioral Therapy for Insomnia With Different Types of Coaching: Randomized Controlled Comparative Trial. *JMIR Ment. Health* **11**, e51716 (2024).
51. Chua, J. Y. X. et al. The effectiveness of Parentbot - a digital healthcare assistant - on parenting outcomes: A randomized controlled trial. *Int J. Nurs. Stud.* **160**, 104906 (2024).
52. Reilly, E. D. et al. Virtual Coach-Guided Online Acceptance and Commitment Therapy for Chronic Pain: Pilot Feasibility Randomized Controlled Trial. *JMIR Form. Res* **8**, e56437 (2024).
53. Chen, C. 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* **12**, e65785 (2025).
54. de Graaff, A. M. et al. Evaluation of a Guided Chatbot Intervention for Young People in Jordan: Feasibility Randomized Controlled Trial. *JMIR Ment. Health* **12**, e63515 (2025).
55. Heinz, M. V. et al. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI* **2**, AIoa2400802 (2025).
56. Sharp, G., Dwyer, B., Randhawa, A., McGrath, I. & Hu, H. The Effectiveness of a Chatbot Single-Session Intervention for People on Waitlists for Eating Disorder Treatment: Randomized Controlled Trial. *J. Med Internet Res* **27**, e70874 (2025).
57. Six, S., Schlesener, E., Hill, V., Babu, S. V. & Byrne, K. Impact of Conversational and Animation Features of a Mental Health App Virtual Agent on Depressive Symptoms and User Experience Among College Students: Randomized Controlled Trial. *JMIR Ment. Health* **12**, e67381 (2025).
58. Tong, A. C. Y., Wong, K. T. Y., Chung, W. W. T. & Mak, W. W. S. Effectiveness of Topic-Based Chatbots on Mental Health Self-Care and Mental Well-Being: Randomized Controlled Trial. *J. Med Internet Res* **27**, e70436 (2025).
59. Xu, S. & Ma, T. Depression intervention using AI chatbots with social cues: a randomized trial of effectiveness. *J. Affect Disord.* **389**, 119760 (2025).
60. Ye, X., Shan, X., Tu, Y. & Zhang, Y. Examining the Efficacy of Large Language Models for Mitigating Depression and Anxiety Among Chinese Students: A Randomized Controlled Trial. *Comput Inform. Nurs.* **43**, https://doi.org/10.1097/cin.0000000000001349. (2025).
61. Yokotani, K., Ito, M., Ihara, N. & Shigeeda, Y. A unified protocol chatbot reduces anxiety by encouraging university students' negative emotional expressions: A randomized controlled trial. *Computers Hum. Behav. Rep.* **19**, 100770 (2025).
62. Zhao, Y. et al. Effect of an AI agent trained on a large language model (LLM) as an intervention for depression and anxiety symptoms in young adults: A 28-day randomized controlled trial. *Appl Psychol. Health Well Being* **17**, e70067 (2025).
63. Higgins JPT et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (Cochrane, 2024). www.cochrane.org/handbook.
64. Barnett, A. G., van der Pols, J. C. & Dobson, A. J. Regression to the mean: what it is and how to deal with it. *Int J. Epidemiol.* **34**, 215–220 (2005).
65. Morton, V. & Torgerson, D. J. Regression to the mean: treatment effect without the intervention. *J. Eval. Clin. Pr.* **11**, 59–65 (2005).
66. Blease, C. & Rodman, A. Generative Artificial Intelligence in Mental Healthcare: An Ethical Evaluation. *Curr. Treat. Options Psychiatry* **12**, 5 (2024).
67. Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M. & Househ, M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *J. Med Internet Res* **22**, e16021 (2020).
68. van Aert, R. C., Wicherts, J. M. & van Assen, M. A. Conducting Meta-Analyses Based on p Values: Reservations and Recommendations for Applying p-Uniform and p-Curve. *Perspect. Psychol. Sci.* **11**, 713–729 (2016).
69. Sohn, J.-S., Lee, E., Kim, J.-J., Oh, H.-K. & Kim, E. Implementation of generative AI for the assessment and treatment of autism spectrum disorders: a scoping review. *Front. Psychiatr.* **16**, 1628216 (2025).
70. Denecke, K. & May, R. Developing a Technical-Oriented Taxonomy to Define Archetypes of Conversational Agents in Health Care: Literature Review and Cluster Analysis. *J. Med Internet Res* **25**, e41583 (2023).
71. Chiu, Y. H., Lee, Y. F., Lin, H. L. & Cheng, L. C. Exploring the Role of Mobile Apps for Insomnia in Depression: Systematic Review. *J. Med Internet Res* **26**, e51110 (2024).
72. Sterne, J. A. C. et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *Bmj* **366**, l4898 (2019).
73. Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *Bmj* **315**, 629–634 (1997).
74. Pustejovsky, J. E. & Rodgers, M. A. Testing for funnel plot asymmetry of standardized mean differences. *Res Synth. Methods* **10**, 57–71 (2019).

Acknowledgements

This research was supported by a grant of the Research and Development (R&D) project funded by the National Center for Mental Health (grant number: MHER25C04). The funder had no role in the study design; data collection, analysis, or interpretation; manuscript preparation; or decision to submit the manuscript for publication. We thank Dr. Vincent Kipkorir for feedback on the original draft and methodological insights regarding systematic reviews.

Author contributions

Conceptualization: J.-S.S., H.-K.O., S.L., and E.K.; Methodology: J.-S.S., S.L., and S.P.; Data curation and investigation: J.-S.S., B.-G.H., and E.K.; Writing-original draft: J.-S.S.; Writing-review and editing: J.-S.S., B.-G.H., S.P., J.-J.K., E.L., H.-K.O., S.L., and E.K.; Supervision: J.-J.K., S.L., and E.K.; All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-02566-w.

Correspondence and requests for materials should be addressed to San Lee or Eunjoo Kim.