Review

# Effectiveness of Conversational Agents for Depression and Anxiety: Systematic Review and Network Dose-Response Meta-Analysis of Agent Types, Dose, and Human Support

Pengyu Yang[1], PhD; Ping Lei Chui[1], PhD; Mingxin He[2], MSN; Yao Wu[3], MSN; Pengfei Cheng[4], PhD; Minxiang Li[5,6], MSN; Haojie Sun[7], BSN; Lurong Cai[7], BSN; Hua Zhang[6,8], PhD

[1]Faculty of Nursing, Universiti Malaya, Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia

[2]Department of Intensive Care Unit, Peking University Shenzhen Hospital, Shenzhen, Guangdong, China

[3]Department of Nursing, Leshan Vocational and Technieal College, Leshan, Sichuan, China

[4]Faculty of Medicine, Macau University of Science and Technology, Macau, Macau, China

[5]Department of Nursing, The Second Affiliated Hospital of Hainan Medical University, Haikou, Hainan, China

[6]School of Nursing, Hainan Medical University, Haikou, Hainan, China

[7]Department of Psychiatry, Ziyang Psychosis Hospital, Ziyang, Sichuan, China

[8]School of Nursing, Hainan Health Vocational College, Haikou, Hainan, China

**Corresponding Author:**
Ping Lei Chui, PhD
Faculty of Nursing
Universiti Malaya
Lembah Pantai
Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, 50603
Malaysia
Phone: 60 12 7128893
Email: chuipinglei@um.edu.my

## Abstract

**Background:** Conversational agents (CAs) are increasingly used to deliver scalable interventions for depression and anxiety, but existing evidence has often treated these tools as a single category. The comparative effectiveness of different agent types, intervention doses, and levels of human support remains unclear.

**Objective:** This review systematically evaluated the comparative effectiveness of various CA modalities and levels of human support, as well as potential dose-response relationships across different CA types.

**Methods:** A comprehensive search of PubMed, Embase, Web of Science, Scopus, Ovid, CINAHL, and Psychology and Behavioral Sciences Collection (PBSC) databases was conducted from inception to December 2025. Eligible studies were randomized controlled trials (RCTs) evaluating CA-based interventions. We conducted a random-effects network meta-analysis (NMA) using postintervention outcomes and assessed inconsistency and risk of bias (RoB). Evidence certainty was assessed using the CINeMA (confidence in network meta-analysis) framework. Where data permitted, model-based network dose-response meta-analysis (MBNMA) was performed to examine dose-response associations.

**Results:** A total of 41 studies (n=7577) were included. NMA showed that AI-enhanced intensive human support (AEIHS) appeared to produce greater symptom improvement for depression and anxiety compared with pure human digital guidance (PHDG; Hedges *g*=–4.28, 95% CI –5.56 to –2.99) and open-ended multimodal-enhanced chatbot (OEMM; Hedges *g*=–4.49, 95% CI –6.31 to –2.66), whereas most other comparisons were not statistically significant. MBNMA indicated that goal-oriented conversational agent (GOCA) and highly structured modular chatbot (HSMC) appeared to exhibit nonlinear effects, with preliminary evidence suggesting potentially enhanced response at low-to-moderate doses, though these findings require further investigation. However, publication bias was detected for depression (Egger *P*=.02), and the overall certainty of evidence was very low, primarily due to RoB, indirectness, and imprecision.

**Conclusions:** While hybrid CA models with enhanced human involvement appeared to show potential advantages over fully automated systems, the evidence base remains limited. The very low certainty of evidence constrains confidence in comparative



XSL•FO
RenderX

rankings and dose recommendations. Future high-quality trials are needed to establish optimal CA architectures and intervention doses for scalable mental health care.

**Trial Registration:** PROSPERO CRD420251145207; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251145207

***(J Med Internet Res 2026;28:e104186)*** doi: 10.2196/104186

**KEYWORDS**

conversational agent; chatbot; dose-response relationship; depression; anxiety; systematic review

## Introduction

In recent years, the global burden of mental health disorders has continued to rise, with depression and anxiety among the most prevalent and disabling conditions worldwide [1]. It is estimated that more than 280 million people are affected by depression and over 300 million by anxiety disorders, contributing to annual economic losses exceeding US $1 trillion [2]. Beyond their economic impact, these conditions significantly impair quality of life and impose long-term burdens on individuals, families, and health care systems. Consequently, the need for effective and scalable mental health interventions has become an urgent global priority.

Conventional treatments, including psychotherapy (eg, cognitive behavioral therapy [CBT]) and pharmacotherapy, have demonstrated efficacy in reducing symptoms of depression and anxiety [3,4]. However, their real-world implementation remains constrained by challenges, such as limited accessibility, high dropout rates, and persistent stigma associated with seeking mental health care [5,6]. These barriers are particularly pronounced in low- and middle-income countries, where treatment coverage may be as low as 20%, compared with approximately 50% in high-income settings [7]. Even in well-resourced systems, gaps in service provision persist [7]. Collectively, these limitations underscore the need for alternative approaches that are accessible, scalable, and acceptable to diverse populations.

Against this backdrop, technology-based interventions have emerged as promising solutions. Among them, AI-driven conversational agents (CAs), commonly referred to as chatbots, have gained increasing attention in mental health care [8]. These systems are designed to simulate human-like interactions and deliver psychological support in real time, offering a potentially scalable and cost-effective complement to traditional services [9]. Emerging evidence suggests that chatbot-based interventions may reduce symptoms of depression and anxiety, with some studies reporting effects comparable to established therapeutic approaches [10]. Platforms such as Woebot (Woebot Health) and Wysa exemplify this development by integrating CBT-based techniques, psychoeducation, and emotional support within accessible digital interfaces [10,11]. Their ability to provide 24/7 support, overcome geographical barriers, and deliver personalized content further enhances their appeal, particularly in resource-limited settings [12,13].

Despite these advantages, the effectiveness of a fully automated chatbot intervention remains a topic of ongoing debate. While such systems can provide psychoeducation and low-intensity support, they may be insufficient to address the complexity and heterogeneity of mental health disorders [10]. This has contributed to growing interest in hybrid models that integrate AI-driven interventions with varying levels of human support, including input from therapists, coaches, or trained facilitators. Conceptually, these approaches seek to leverage the scalability of AI with the clinical insight and adaptability of human providers. However, the extent to which human involvement improves outcomes and the conditions under which it is most beneficial remain unclear. In addition, chatbot interventions differ widely in their theoretical frameworks (eg, CBT and mindfulness), delivery formats, and intervention intensity, further complicating direct comparisons across studies [14].

Although the evidence base for AI-driven and hybrid mental health interventions is rapidly expanding, several critical methodological and conceptual gaps remain unaddressed in the current literature. Recently published systematic reviews and meta-analyses have underscored the overall therapeutic potential of CAs in alleviating depressive and anxiety symptoms. For instance, a meta-analysis of text-based turn-taking CAs (39 trials, 7401 participants) found modest but statistically significant depression reduction (Hedges $g$=0.31, 95% CI 0.17-0.46) [15], while CBT-oriented CA reviews (29 trials) reported larger postintervention depression effects (Hedges $g$=–0.55, 95% CI –0.70 to –0.40) [16]. However, they predominantly rely on conventional pairwise comparisons or broad aggregate groupings that cannot integrate direct and indirect evidence, rank interventions comprehensively, or compare untested combinations. Critically, prior reviews acknowledged varying intensities of human involvement in CA systems [15-17]. Emerging evidence highlights that interaction design features (such as emotional responsiveness, structured feedback, and dialogue depth) and adherence patterns significantly influence clinical utility [18], yet existing pairwise frameworks fail to simultaneously contrast distinct agent architectures alongside varying intensities of human support. Additionally, none has positioned human support level as a formally comparable intervention arm within a unified analytical network. Most importantly, despite growing recognition that intervention “dose” (eg, duration and frequency of interaction) dictates engagement and therapeutic response [15,19], no study has systematically modeled nonlinear dose-response relationships across heterogeneous CA modalities. This is a notable omission because dose-response curves often exhibit nonmonotonic patterns (eg, thresholds or diminishing returns), and identifying optimal “prescriptive doses” is essential for implementation.

To address these distinct gaps, the present study advances the literature by adopting a more sophisticated analytical architecture with 3 core elements of novelty. First, we use a



XSL•FO
RenderX

network meta-analysis (NMA) framework to simultaneously synthesize direct and indirect evidence across 8 distinct intervention classifications, encompassing granular variations in agent capabilities (goal-oriented, highly structured modular, open-ended multimodal) and systematically ranked tiers of human support (minimal AI-assisted guidance, routine AI-augmented therapist involvement, and intensive human-led frameworks with AI integration). This is the first synthesis to position human support as a structural network node, enabling formal comparative ranking. Second, we conduct model-based network dose-response meta-analysis (MBNMA) using nonlinear Emax modeling, moving beyond binary efficacy estimates to capture dose-response trajectories and identify whether optimal effects exist at specific exposure thresholds. This approach directly parallels dose-response research in psychotherapy and will illuminate whether "more" CA interaction uniformly improves outcomes or whether diminishing returns or plateaus occur. By reconciling previous conflicting findings regarding agent types, human oversight, and intervention dose within a unified Bayesian framework, this review provides a rigorous, implementation-focused evidentiary roadmap designed to guide clinicians, developers, and policymakers in optimizing the architectural and clinical design of scalable digital mental health interventions.

## Methods

### Study Design

This study was conducted in accordance with the PRISMA-NMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Network Meta-Analyses) guidelines [20] (Multimedia Appendix 1) and the Cochrane Handbook for Systematic Reviews of Interventions [21]. The review protocol was prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD420251145207).

### Search Strategy

A comprehensive literature search was performed from database inception to December 2025 across PubMed, Embase, Web of Science, Scopus, Ovid, CINAHL, and the PBSC. The search strategy combined MeSH and free-text terms related to CAs (eg, chatbot, CA, large language model, AI) and mental health outcomes (eg, depression, anxiety), using Boolean operators ("AND", "OR"). The full search strategy is provided in Multimedia Appendix 2. No language or publication status restrictions were applied. Two reviewers (PY and MH) independently screened titles and abstracts and assessed the full-text eligibility of identified studies. Any discrepancies between reviewers were resolved through consensus discussion; when consensus could not be achieved, a third reviewer (PLC) provided adjudication.

### Eligibility Criteria

Eligibility criteria were defined according to the PICOS (Participants, Interventions, Comparators, Outcomes, and Study Design) framework. Randomized controlled trials (RCTs) evaluating AI-based CA interventions targeting depression and/or anxiety were included. No restrictions were applied to participant characteristics to enhance generalizability. Detailed inclusion and exclusion criteria are provided in S1.1 in Multimedia Appendix 3.

### Data Extraction

Two reviewers (PY and MH) independently extracted data from eligible studies using a standardized data extraction form. Any discrepancies were resolved through discussion with a third reviewer (PLC). Extracted data included study characteristics (first author and publication year), participant characteristics (age and sample size), study design, intervention and comparator details (type, duration, and interaction intensity), and relevant outcome measures. When necessary, corresponding authors were contacted to obtain missing data. Detailed data coding and management procedures are described in S1.2.1 and S1.2.2 in Multimedia Appendix 3.

### Risk of Bias and Certainty of Evidence

Risk of bias (RoB) was independently assessed by 2 reviewers (PY and MH) using the Cochrane RoB 2 tool [22], with any disagreements resolved through discussion with a third reviewer (PLC). The certainty of evidence for network estimates was evaluated using the CINeMA (confidence in network meta-analysis) framework [23,24]. This approach assesses within-study bias, reporting bias, indirectness, imprecision, heterogeneity, and incoherence for each comparison [23].

### Measures of Treatment Effect

For continuous outcomes in the NMA, treatment effects were expressed as standardized mean differences (SMDs). Mean differences (MDs) with corresponding SDs and sample sizes were converted to SMD values. When studies reported multiple postintervention time points, data at the end of the intervention period were extracted. Where multiple outcome measures were available, selection prioritized the most clinically relevant validated scales: for depression, Patient Health Questionnaire-9 (PHQ-9)>Patient Health Questionnaire-8 (PHQ-8)>Beck Depression Inventory II (BDI-II)>Center for Epidemiologic Studies Depression Scale (CES-D)>Depression Anxiety Stress Scale (DASS-21) depression>Patient-Reported Outcomes Measurement Information System (PROMIS) depression; for anxiety, Generalized Anxiety Disorder 7-item scale (GAD-7)>Generalized Anxiety Disorder Questionnaire (GAD-Q-V)>Hospital Anxiety and Depression Scale (HADS) anxiety>Screen for Child Anxiety Related Emotional Disorders (SCARED)>Self-Rating Anxiety Scale (SAS)>Behavioral Dysexecutive Syndrome Inventory/Beck Anxiety Inventory (BDSI/BAI)>PROMIS anxiety>DASS-21 anxiety. The pooled effect estimate was calculated using Hedges *g* as SMD with small-study correction. Following the convention of included studies, negative effect sizes indicate greater symptom improvement, with larger negative values representing stronger treatment effects on depressive and anxiety symptoms. When SD was not reported, it was derived from SEs, CIs, *P* values, or t-statistics using established computational methods [21]. For studies reporting change scores without corresponding SDs, a correlation coefficient of 0.5 was assumed to estimate variance [25]. To examine the transitivity assumption underlying NMA, we assessed the distribution of key potential effect modifiers



XSL•FO
RenderX

across treatment nodes and direct comparisons, including participant age, clinical population status and percentage, baseline symptom measurement values where reported, intervention duration, weekly effective interaction time in minutes, and primary outcome scale type. The transitivity principle posits that populations from similar target groups could theoretically be randomly assigned to any available intervention, implying that effect modifiers should be sufficiently comparable across network nodes. Given observed imbalances in the dataset, subgroup analyses were conducted restricting comparisons by intervention duration, primary outcome scale, participant age, and population type. Furthermore, we conducted sensitivity analyses by excluding high-risk studies and those with small sample sizes (ie, papers with fewer than 30 participants) [26] to assess the stability of the estimates.

For dose-response modeling, MBNMA was conducted using binary outcomes, with data comprising sample size, weekly intervention dose, and intervention type (detailed dose categorization provided in S1.2.3 in Multimedia Appendix 3). Binary outcomes were derived using a 2-tiered extraction strategy (S1.2.4 in Multimedia Appendix 3). The complete statistical derivation formulas are presented in S1.2.5 in Multimedia Appendix 3. The dose-response relationship was parameterized as $f(x, \theta)$, where x represents dose and $\theta$ denotes dose-response parameters. A Bayesian random-effects model was implemented with binomial likelihood and a logit link function. Treatment effects were estimated on the log odds ratio (OR) scale and subsequently exponentiated to obtain ORs, with depression and anxiety symptom improvement or remission as the binary outcome. Given limited data density, control arms (NMA categories 7: pure human digital guidance [PHDG] and 8: waitlist, usual care, health information and education, and general health app [UC]) were pooled into a single reference group (pure human digital guidance, waitlist, usual care, health information and education, and general health app [Placebo_0]). Because the model estimates relative effects, this reference category was assigned a baseline effect of 0. To generate model predictions across the dose range, the baseline response E0 was set to 0.2 based on literature [27]. Clinically interpretable outcome scales (PHQ-9 for depression, GAD-7 for anxiety) were assessed as effect modifiers using stratified subgroup analyses to evaluate transitivity and to determine whether differing scale definitions influenced the dose-response findings. Sensitivity analyses were further performed using the derived binary outcome indicators. The direction of the dose-response effects, posterior mean point estimates, and corresponding credible intervals (CrIs) were compared with those from the main analysis to assess whether the overall conclusions were influenced by the choice of outcome definition.

### Statistical Analysis

NMA was conducted to synthesize direct and indirect evidence across multiple interventions using Stata (StataCorp LLC) software. Pairwise meta-analysis was performed first to estimate g and 95% CIs for depressive and anxiety outcomes across 8 intervention categories: goal-oriented conversational agent (GOCA), highly structured modular chatbot (HSMC), open-ended multimodal-enhanced chatbot (OEMM), AI-assisted minimal human support (AIMHS), AI-assisted routine human support (AIRHS), AI-enhanced intensive human support (AEIHS), PHDG, and UC. The frequentist NMA framework generated relative treatment effects and ranking probabilities for all interventions, with consistency between direct and indirect evidence assessed via node-splitting analysis. Further details of the statistical analyses are provided in S1.3.1 in Multimedia Appendix 3.

MBNMA was conducted for selected intervention types (GOCA, HSMC, and OEMM) to examine the association between intervention dose (defined as weekly effective interaction time) and treatment outcomes. A Bayesian random-effects model was implemented in R (R Foundation) to model dose-response relationships, and model fit was compared across different functional forms (eg, the Emax model and the unrelated mean effects [UME] model) [28]. Further details of the statistical analyses are provided in S1.3.2 in Multimedia Appendix 3.

## Results

### Study Selection and Characteristics

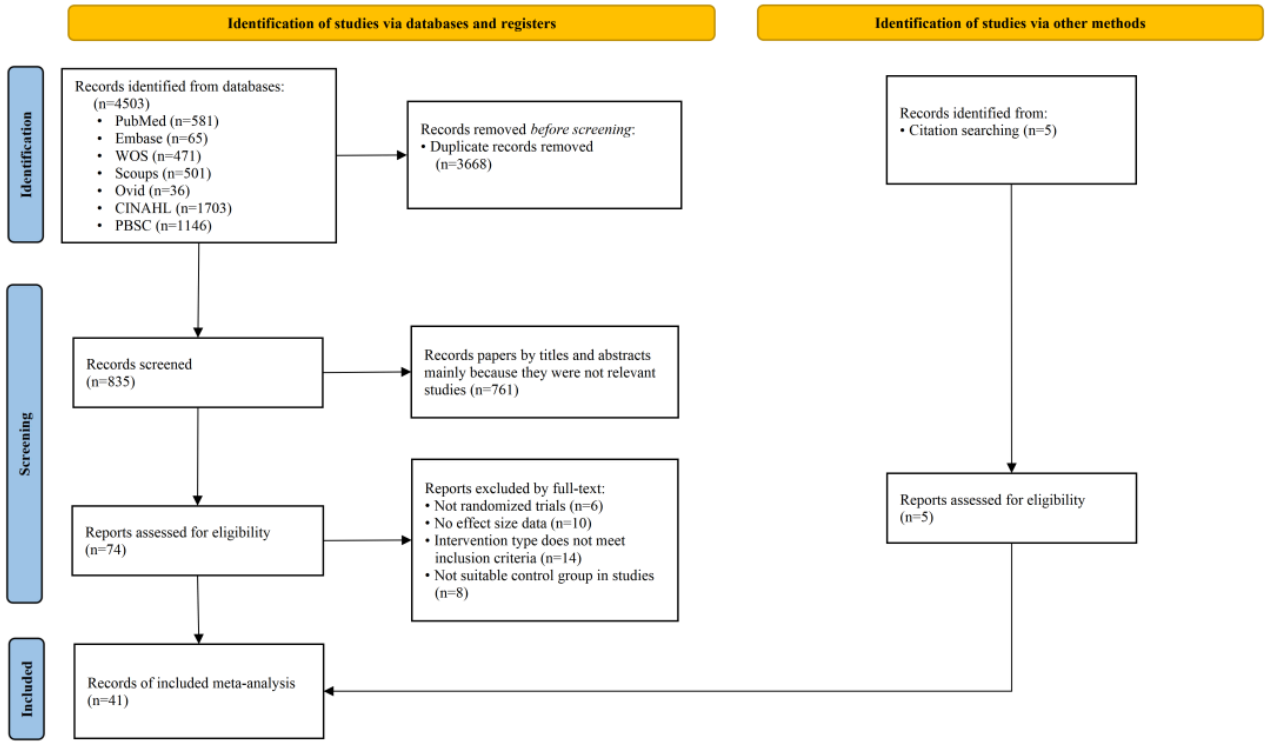
The study selection process is presented in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1). Of 4503 records identified, 74 full-text papers were assessed for eligibility, and 41 papers were ultimately included in the review.

A total of 7577 participants were included in the analysis, with an overall mean age of 30.3 (SD 6.8) years. Among the included studies, 38 reported depression outcomes (n=6985), while 32 reported anxiety outcomes (n=6617). Only 23 studies provided sufficient information on weekly interaction time, with intervention durations ranging from 1 to 16 weeks and effective weekly engagement ranging from 7.5 to 105 minutes. Only 2 studies reported remission rates [29,30], whereas no study directly reported response counts. The remaining studies derived response estimates from continuous outcomes. Detailed study characteristics are presented in Multimedia Appendix 4 (Table S1) [10,29-68].



XSL•FO
RenderX

**Figure 1.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection.



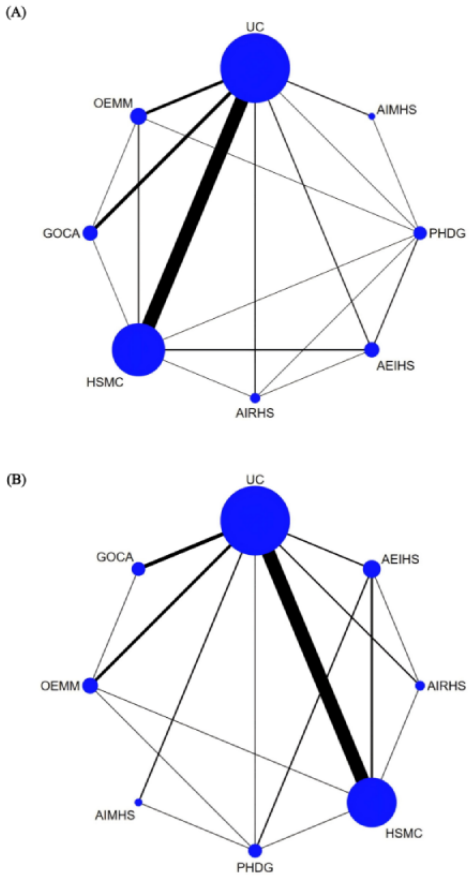

## NMA

Eight intervention categories included the following: GOCA, HSMC, OEMM, AIMHS, AIRHS, AEIHS, PHDG, and UC. The network structures for depression and anxiety outcomes are presented in Figure 2A and B. Node size represents the number of participants, while edge thickness reflects the number of direct comparisons between interventions. Due to the limited reporting of follow-up results, this analysis includes only postintervention data. Subsequent consistency checks revealed significant violations within the networks of evidence for both depression and anxiety, likely reflecting unmeasured effect modifiers not adequately captured by the existing data. Consequently, caution is warranted when interpreting network ranking results; similarly, the reliability of individual effect estimates, particularly those based solely on indirect evidence, should be assessed conservatively. The full league table is provided in Table 1, and ranking probabilities are shown in Figure 3. Table S2 in Multimedia Appendix 4 provides a breakdown of the studies and participant characteristics in the direct comparison.



XSL•FO
RenderX

**Figure 2.** Network map of comparisons for depression (A) and anxiety (B). AEIHS: AI-enhanced intensive human support; AIMHS: AI-assisted minimal human support; AIRHS: AI-assisted routine human support; GOCA: goal-oriented conversational agent; HSMC: highly structured modular chatbot; OEMM: open-ended multimodal-enhanced chatbot; PHDG: pure human digital guidance; UC: waitlist, usual care, health information and education, and general health app.






XSL•FO
RenderX

**Table 1.** Network league table showing results of network meta-analysis using the Hedges g and 95% CI for depressive and anxiety symptoms.

| | UC[a,b] | PHDG[b,c] | OEMM[b,d] | HSMC[b,e] | GOCA[b,f] | AIRHS[b,g] | AIMHS[b,h] | AEIHS[b,i] |
|---|---|---|---|---|---|---|---|---|
| Depression | | | | | | | | |
| UC | —[j] | 0.78 (–0.40 to 1.96) | 0.62 (–0.66 to 1.89) | 0.35 (–0.43 to 1.13) | 0.51 (–0.61 to 1.63) | 0.32 (–0.46 to 1.10) | 0.75 (–0.59 to 2.09) | –0.02 (–1.24 to 1.21) |
| PHDG | –0.78 (–1.96 to 0.40) | — | –0.16 (–1.71 to 1.38) | –0.43 (–1.60 to 0.74) | –0.27 (–1.69 to 1.14) | –0.46 (–1.63 to 0.70) | –0.03 (–0.66 to 0.60) | –4.28 (–5.56 to –2.99)[k] |
| OEMM | –0.62 (–1.89 to 0.66) | 0.16 (–1.38 to 1.71) | — | –0.27 (–1.27 to 0.74) | –0.11 (–0.72 to 0.50) | –0.30 (–1.56 to 0.96) | 0.14 (–1.53 to 1.80) | –0.45 (–1.55 to 0.66) |
| HSMC | –0.35 (–1.13 to 0.43) | 0.43 (–0.74 to 1.60) | 0.27 (–0.74 to 1.27) | — | 0.16 (–0.64 to 0.96) | –0.03 (–0.79 to 0.72) | 0.40 (–0.93 to 1.73) | 0.82 (–0.35 to 1.98) |
| GOCA | –0.51 (–1.63 to 0.61) | 0.27 (–1.14 to 1.69) | 0.11 (–0.50 to 0.72) | –0.16 (–0.96 to 0.64) | — | –0.19 (–1.29 to 0.91) | 0.24 (–1.31 to 1.79) | –0.26 (–1.02 to 0.50) |
| AIRHS | –0.32 (–1.10 to 0.46) | 0.46 (–0.70 to 1.63) | 0.30 (–0.96 to 1.56) | 0.03 (–0.72 to 0.79) | 0.19 (–0.91 to 1.29) | — | 0.43 (–0.89 to 1.76) | –0.69 (–1.78 to 0.39) |
| AIMHS | –0.75 (–2.09 to 0.59) | 0.03 (–0.60 to 0.66) | –0.14 (–1.80 to 1.53) | –0.40 (–1.73 to 0.93) | –0.24 (–1.79 to 1.31) | –0.43 (–1.76 to 0.89) | — | 0.86 (–0.55 to 2.28) |
| AEIHS | 0.02 (–1.21 to 1.24) | 4.28 (2.99 to 5.56)[k] | 0.45 (–0.66 to 1.55) | –0.82 (–1.98 to 0.35) | 0.26 (–0.50 to 1.02) | 0.69 (–0.39 to 1.78) | –0.86 (–2.28 to 0.55) | — |
| Anxiety | | | | | | | | |
| UC | — | –0.01 (–1.75 to 1.74) | 0.01 (–1.62 to 1.64) | 0.06 (–1.15 to 1.26) | –0.14 (–0.67 to 0.38) | 0.21 (–0.99 to 1.41) | –0.01 (–2.08 to 2.05) | –0.11 (–1.29 to 1.08) |
| PHDG | 0.01 (–1.74 to 1.75) | — | 0.02 (–2.04 to 2.07) | 0.06 (–1.67 to 1.79) | –0.14 (–1.95 to 1.68) | 0.21 (–1.52 to 1.95) | –0.01 (–1.12 to 1.10) | –0.29 (–2.06 to 1.49) |
| OEMM | –0.01 (–1.64 to 1.62) | –0.02 (–2.07 to 2.04) | — | 0.04 (–1.06 to 1.15) | –0.15 (–1.87 to 1.56) | 0.19 (–1.43 to 1.81) | –0.03 (–2.36 to 2.31) | –4.49 (–6.31 to –2.66)[k] |
| HSMC | –0.06 (–1.26 to 1.15) | –0.06 (–1.79 to 1.67) | –0.04 (–1.15 to 1.06) | — | –0.20 (–1.51 to 1.11) | 0.15 (–1.03 to 1.34) | –0.07 (–2.13 to 1.99) | 0.09 (–1.22 to 1.40) |
| GOCA | 0.14 (–0.38 to 0.67) | 0.14 (–1.68 to 1.95) | 0.15 (–1.56 to 1.87) | 0.20 (–1.11 to 1.51) | — | 0.35 (–0.96 to 1.66) | 0.13 (–2.00 to 2.26) | –0.26 (–1.44 to 0.93) |
| AIRHS | –0.21 (–1.41 to 0.99) | –0.21 (–1.95 to 1.52) | –0.19 (–1.81 to 1.43) | –0.15 (–1.34 to 1.03) | –0.35 (–1.66 to 0.96) | — | –0.22 (–2.28 to 1.84) | –0.04 (–1.72 to 1.64) |
| AIMHS | 0.01 (–2.05 to 2.08) | 0.01 (–1.10 to 1.12) | 0.03 (–2.31 to 2.36) | 0.07 (–1.99 to 2.13) | –0.13 (–2.26 to 2.00) | 0.22 (–1.84 to 2.28) | — | 0.12 (–2.09 to 2.34) |
| AEIHS | 0.11 (–1.08 to 1.29) | 0.29 (–1.49 to 2.06) | 4.49 (2.66 to 6.31)[k] | –0.09 (–1.40 to 1.22) | 0.26 (–0.93 to 1.44) | 0.04 (–1.64 to 1.72) | –0.12 (–2.34 to 2.09) | — |

[a]UC: waitlist, usual care, health information and education, and general health app.

[b]95% CI.

[c]PHDG: pure human digital guidance.

[d]OEMM: open-ended multimodal-enhanced chatbot.

[e]HSMC: highly structured modular chatbot.

[f]GOCA: goal-oriented conversational agent.

[g]AIRHS: AI-assisted routine human support.

[h]AIMHS: AI-assisted minimal human support.

[i]AEIHS: AI-enhanced intensive human support.

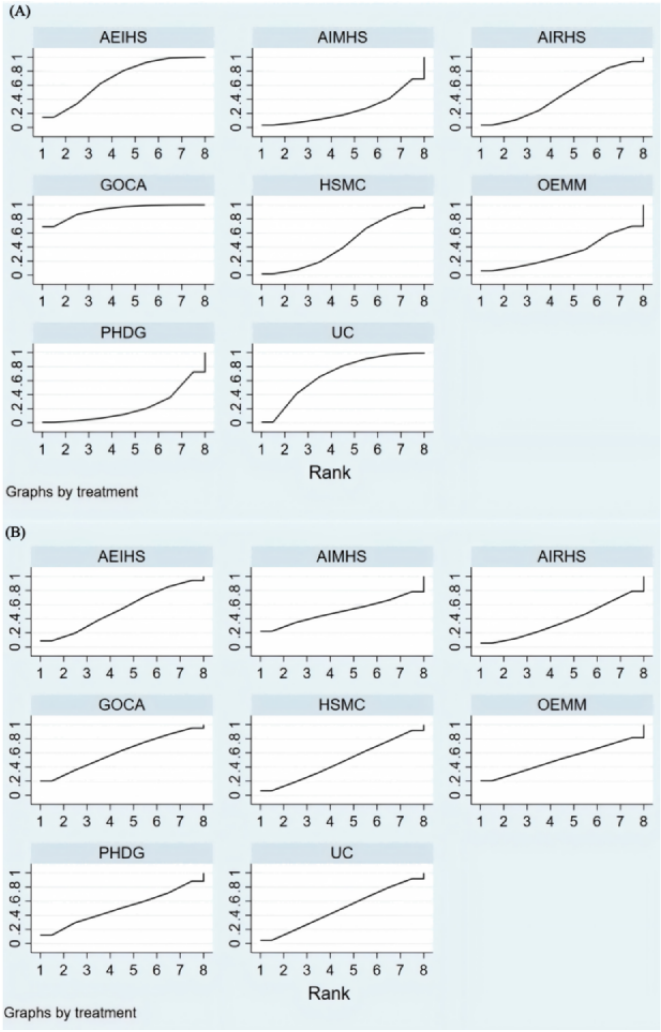
[j]Not available.

[k]meaningful result.



XSL•FO
RenderX

**Figure 3.** Surface under the cumulative ranking curve (SUCRA) showing the relative ranking probabilities of different interventions for (A) depression and (B) anxiety. AEIHS: AI-enhanced intensive human support; AIMHS: AI-assisted minimal human support; AIRHS: AI-assisted routine human support; GOCA: goal-oriented conversational agent; HSMC: highly structured modular chatbot; OEMM: open-ended multimodal-enhanced chatbot; PHDG: pure human digital guidance; UC: waitlist, usual care, health information and education, and general health app.




## Depression

The NMA of depression symptoms revealed substantial global inconsistency ($P$<.05, Figure S1 in Multimedia Appendix 5), necessitating adoption of an inconsistency model for analysis. Local inconsistency parameter testing identified 2 of 14 parameters reaching statistical significance ($P$<.05, Table S1 in Multimedia Appendix 5). Node-splitting analysis further indicated significant discrepancies between direct and indirect evidence for 2 specific comparisons: AEIHS vs HSMC ($P$=.01) and AEIHS vs UC ($P$=.001; Table S2 in Multimedia Appendix 5). The surface under the cumulative ranking curve (SUCRA) scores positioned GOCA highest (Figure 3A, Table S3 in Multimedia Appendix 5). The majority of pairwise comparisons yielded no statistically significant treatment differences, with 95% CI encompassing 0. A single statistically significant contrast emerged between AEIHS and PHDG, demonstrating superior symptom improvement under AEIHS (k=2, N=75, Hedges $g$=–4.28, 95% CI –5.56 to –2.99; Table 1 and Table S2 in Multimedia Appendix 4) [31,32]. This comparison was based on sparse direct evidence, and although sensitivity analyses (excluding high RoB trials and small sample studies) supported the robustness of this unusually large effect (AEIHS vs PHDG: Hedges $g$=–4.28, 95% CI –5.57 to –2.99; Tables S4 and S5 in Multimedia Appendix 5), the broader network inconsistencies and very low overall certainty of evidence warrant considerable caution. The ranking results and relative treatment merits based on this estimate should not be regarded as definitive.

## Moderation by Participant Characteristics in Depression

In the adult (18-59 years) and nonclinical population subgroups, AEIHS produced significantly greater improvements in depressive symptoms than OEMM. The corresponding effect estimates (Hedges $g$) of –4.28 (95% CI –5.55 to –3.00) for the 18-59 years subgroup and Hedges $g$=–4.28 (95% CI –5.51 to –3.05) for the nonclinical subgroup. In both subgroups, GOCA ranked first based on SUCRA. Nevertheless, global inconsistency was statistically significant ($P$<.05). Local node-splitting analyses also suggested inconsistency for some comparisons ($P$<.05), indicating that direct and indirect evidence were not fully aligned. In addition, the Egger test suggested potential small-study effects in the adults subgroup ($P$=.02). After adjustment using the trim-and-fill method, the pooled effect remained Hedges $g$=–0.17 (95% CI –0.23 to –0.10), indicating that the findings should still be interpreted with



XSL•FO
RenderX

caution. Detailed results are provided in Table S6 in Multimedia Appendix 5.

### Moderation by Scale and Intervention Design Characteristics in Depression

For the PHQ-9 scale, AEIHS demonstrated significantly superior symptom improvement relative to PHDG (MD –8.73, 95% CI –13.41 to –4.04), with GOCA occupying the top position in the SUCRA rankings. However, this network exhibited substantial global inconsistency (*P*=.001), and local comparisons revealed inconsistency between HSMC and OEMM, as well as between HSMC and UC (*P*<.05). The Egger test indicated potential publication bias within this network (*P*=.04); following adjustment via the trim-and-fill method with inclusion of 4 imputed studies, the pooled effect size (Hedges *g*) was –1.58 (95% CI –2.14 to –1.02), necessitating cautious interpretation of these findings. Among studies with intervention durations of 4 to 8 weeks, AEIHS showed statistically significant superiority over UC in reducing depressive symptoms (Hedges *g*=–1.16, 95% CI –2.21 to –0.11). OEMM demonstrated significant symptom improvement advantages compared to GOCA (Hedges *g*=–0.91, 95% CI –1.63 to –0.19), HSMC (Hedges *g*=–1.45, 95% CI –2.50 to –0.41), PHDG (Hedges *g*=–1.88, 95% CI –3.31 to –0.45), and UC (Hedges *g*=–1.10, 95% CI –1.66 to –0.55), while showing inferiority relative to both AIMHS (Hedges *g*=1.85, 95% CI 0.24-3.47) and AIRHS (Hedges *g*=1.42, 95% CI 0.38-2.46). This network demonstrated both excellent global consistency (*P*=.75) and local consistency (*P*>.05), with no evidence of publication bias on the Egger test (*P*=.07). The cumulative distribution of ranking probabilities indicated that OEMM achieved the highest SUCRA value. Detailed results are presented in Table S6 in Multimedia Appendix 5.

### Anxiety

In the NMA of anxiety outcomes, global inconsistency was statistically significant (*P*<.05, Figure S2 in Multimedia Appendix 5), necessitating adoption of an inconsistency model. Local inconsistency parameter testing revealed that only 1 of 11 parameters achieved statistical significance (*P*<.05, Table S1 in Multimedia Appendix 5). Node-splitting analysis further demonstrated that direct and indirect evidence showed no statistically meaningful differences for any pairwise comparison (*P*>.05, Table S7 in Multimedia Appendix 5). Overall, only AEIHS demonstrated a statistically significant advantage over OEMM (Hedges *g*=–4.49, 95% CI –6.31 to –2.66, Table 1), indicating greater reductions in anxiety symptoms; most other pairwise comparisons failed to reach statistical significance. Notably, no direct head-to-head trials contributed to the AEIHS vs OEMM comparison in the available evidence base (k=0), meaning this estimate rests entirely on indirect evidence through the network. SUCRA rankings (Figure 3B and Table S3 in Multimedia Appendix 5) identified GOCA as relatively higher-ranked. Sensitivity analyses excluding high RoB trials and small sample size studies yielded similar effect magnitudes (Hedges *g*=–4.49, 95% CI –6.56 to –2.41; Hedges *g*=–4.49, 95% CI –6.34 to –2.63, respectively, Tables S4 and S5 in Multimedia Appendix 5). However, given the observed network inconsistency and the very low certainty of the evidence overall, these findings should be interpreted with caution and viewed as exploratory rather than definitive conclusions.

### Moderation by Participant Characteristics in Anxiety

In the adult (18-59 years) and nonclinical population subgroups, AEIHS demonstrated statistically significant superiority over HSMC for anxiety symptom improvement, with comparable effect estimates (Hedges *g*) of –4.48 (95% CI –5.53 to –3.44) in adults and Hedges g=–4.48 (95% CI –6.76 to –2.21) in nonclinical populations. Additionally, within the adult subgroup, AIMHS showed a significant advantage over UC in anxiety reduction (Hedges *g*=–0.21, 95% CI –0.31 to –0.11). Regarding ranking patterns, SUCRA analysis revealed divergent hierarchies: adults demonstrated the highest SUCRA for AIMHS, while nonclinical populations showed the highest SUCRA for GOCA. Although the Egger test did not indicate significant publication bias across subgroups (adults: *P*=.61; nonclinical: *P*=.28), global inconsistency testing revealed statistically significant heterogeneity in both populations (*P*<.05). Subsequent node-splitting analysis identified multiple instances of direct-indirect evidence conflict: in the adult population, inconsistency was detected between AEIHS and UC (*P*=.01) as well as between HSMC and UC (*P*=.02); in the nonclinical population, AEIHS vs UC similarly demonstrated significant inconsistency (*P*=.03). These findings underscore the need for cautious interpretation of effect estimates and ranking results, with detailed comparisons presented in Table S8 in Multimedia Appendix 5.

### Moderation by Scale and Intervention Design Characteristics in Anxiety

For the GAD-7 scale, pairwise network comparisons revealed that OEMM demonstrated significantly greater symptom improvement compared to GOCA (MD=–5.23, 95% CI –8.76 to –1.70), HSMC (MD=–4.26, 95% CI –6.88 to –1.63), and UC (MD=–5.25, 95% CI –7.69 to –2.80). AEIHS (MD=–3.38, 95% CI –6.11 to –0.65) and PHDG (MD=–3.29, 95% CI –6.48 to –0.10) likewise demonstrated significant symptom reduction relative to UC. SUCRA ranking indicated OEMM as the highest-ranked intervention. Although the Egger test revealed no significant small sample publication bias (*P*=.10) and global inconsistency testing approached but did not reach statistical significance (*P*=.057), node-splitting analysis detected a statistically significant direct-indirect evidence inconsistency between HSMC and OEMM (*P*=.02), necessitating cautious interpretation of these results. For the subgroup of studies with intervention duration of 4 to 8 weeks, OEMM exhibited superior symptom improvement over UC (Hedges *g*=–1.52, 95% CI –2.25 to –0.79), GOCA (Hedges *g*=–1.40, 95% CI –2.42 to –0.38), and HSMC (Hedges *g*=–1.30, 95% CI –2.10 to –0.50), yet demonstrated relative disadvantage when compared to AEIHS (Hedges *g*=1.15, 95% CI 0.11-2.20), AIMHS (Hedges *g*=1.20, 95% CI 0.16-2.24), and AIRHS (Hedges *g*=1.34, 95% CI 0.25-2.43). SUCRA ranking identified OEMM as the highest-ranked intervention. Notably, global inconsistency testing showed no statistical significance (*P*=.86), and all node-splitting comparisons revealed no evidence of inconsistencies (*P*>.05). The Egger test similarly detected no significant bias (*P*=.08). However, given the sparsity of the



XSL•FO
RenderX

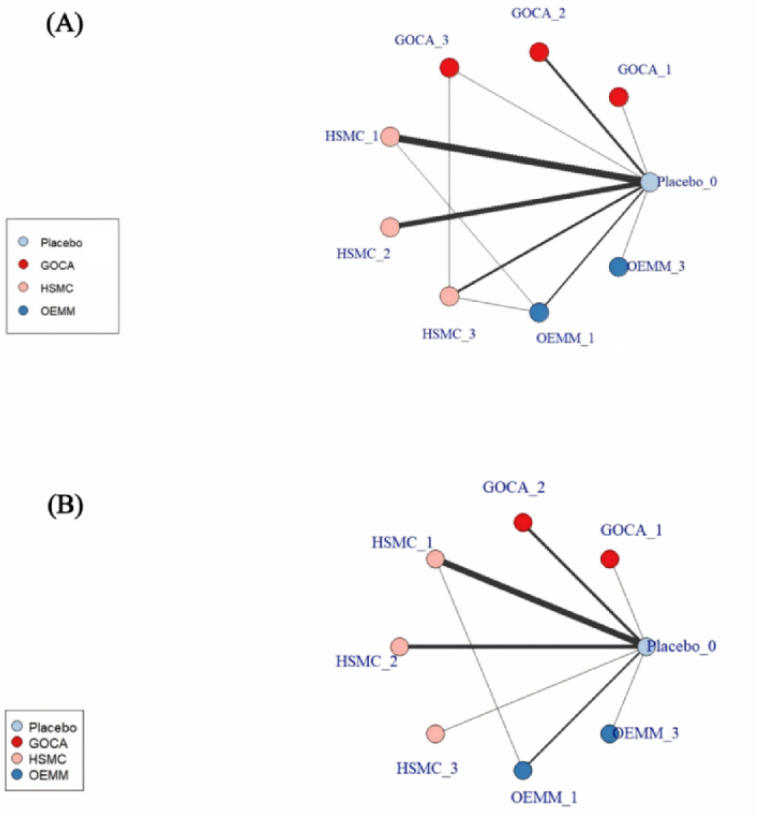
network, effect estimates and ranking results should still be interpreted with caution. Detailed results are presented in Table S8 in Multimedia Appendix 5.

## Dose-Response Relationships

### *Depression*

The network geometry comprised 9 intervention nodes spanning a control (Placebo_0) and 3 agent types (GOCA, HSMC, and OEMM). However, the OEMM_2 node was absent due to insufficient data from included papers (Figure 4A). Preliminary dose-effect assessment on the link scale exhibited a non-linear relationship across doses (Figure 5A). The Emax and UME models yielded comparable deviance information criterion (DIC) values (256.2 vs 256.7), with between-study heterogeneity SD 0.788 (95% CrI 0.529-1.23). Markov Chain Monte Carlo (MCMC) sampling diagnostics demonstrated stable trace plots with adequate mixing; in conjunction with R-hat=1.001 and effective sample size (ESS)=2994.76, these findings confirmed sufficient model convergence (Figure S1A in Multimedia Appendix 6). Deviance plots, fitted plots, and calibration plots (Figures S2A, S3A, and S4A in Multimedia Appendix 6) showed good agreement between fitted and observed values overall, with minor departures only at high doses. Node-splitting and direct-indirect posterior density overlap analyses (Figure S5A in Multimedia Appendix 6) indicated that consistency assumptions held for OEMM across doses (OEMM_1: *P*=.61; OEMM_3: *P*=.53) and for HSMC_2 (*P*=.26) and GOCA_2 (*P*=.07); conversely, remaining interventions exhibited separated posterior density peaks (*P*<.05), indicating local inconsistency. The MBNMA prediction curve displayed a non-monotonic pattern (Figure 6A). Relative treatment effects (Table S1 in Multimedia Appendix 6) showed that, compared with placebo, low-dose GOCA_1 demonstrated a stronger therapeutic advantage (OR 4.77, 95% CrI 3.34-6.80), and HSMC_1 also yielded significant benefit (OR 1.43, 95% CrI 1.13-1.81). All OEMM dose groups had 95% CrIs covering 1, indicating no statistically significant difference vs placebo. Notably, high-dose HSMC_3 (OR 0.54, 95% CrI 0.37-0.78) and GOCA_3 (OR 0.25, 95% CrI 0.07-0.82) appeared to have diminishing effects compared to placebo. In pairwise comparisons with GOCA_1 positioned on the left, direct contrasts with higher-dose GOCA variants or other intervention types yielded OR <1 with 95% CrIs excluding 1, indicating diminishing marginal benefits at higher-dose GOCA variants and other agent types. SUCRA analysis revealed that GOCA achieved relatively leading ranking in terms of Emax (0.259), while the 3 agent types showed highly comparable ED50 rankings (Figure 7A), collectively suggesting GOCA possesses the most favorable dose-response potential. Detailed results are presented in Table S2 in Multimedia Appendix 6. Sensitivity analyses showed that the advantage of low-dose GOCA_1 vs placebo and the reduced effect of high-dose GOCA_3 remained consistent, with limited changes in point estimates and substantial overlap in CrIs (DIC: Emax 243.3 vs UME 240.7; SD 0.80, 95% CrI 0.51-1.22; R-hat=1.01, ESS=2999), indicating robustness of the overall conclusions (Table S3 in Multimedia Appendix 6).

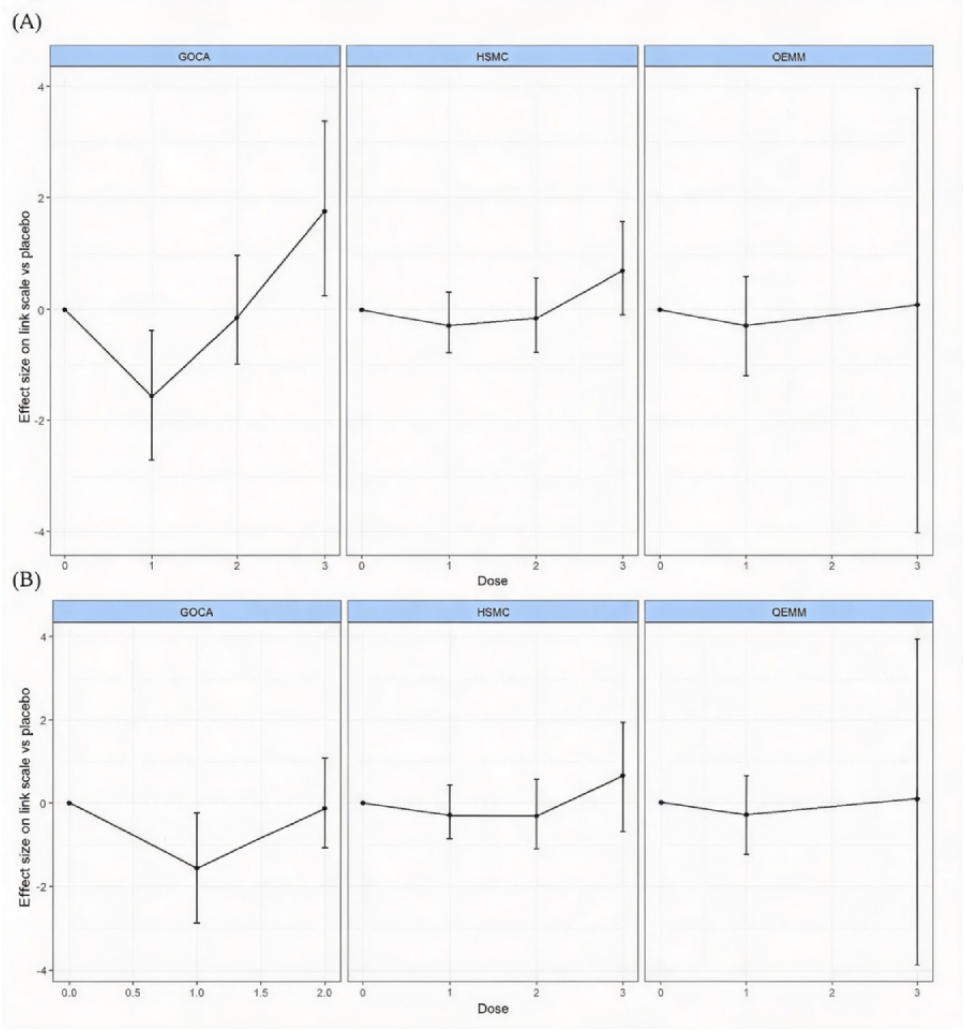
**Figure 4.** Network geometry indicating the number of participants in each arm (size of points) and the number of comparisons between arms (thickness of lines) for (A) depression and (B) anxiety. GOCA: goal-oriented conversational agent; HSMC: highly structured modular chatbot; OEMM: open-ended multimodal-enhanced chatbot.






XSL•FO
RenderX

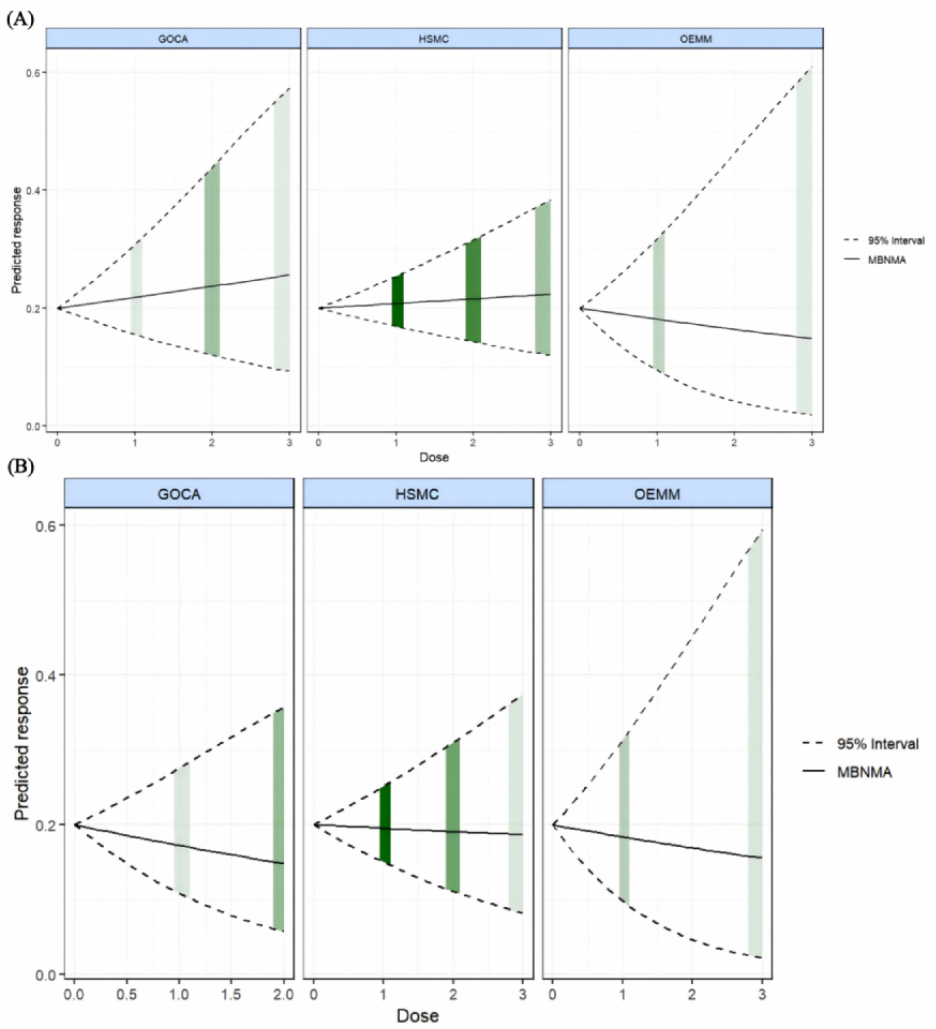
**Figure 5.** Dose-response curves comparing different types of conversational agent interventions with placebo for (A) depression and (B) anxiety. GOCA: goal-oriented conversational agent; HSMC: highly structured modular chatbot; OEMM: open-ended multimodal-enhanced chatbot.






XSL•FO
RenderX

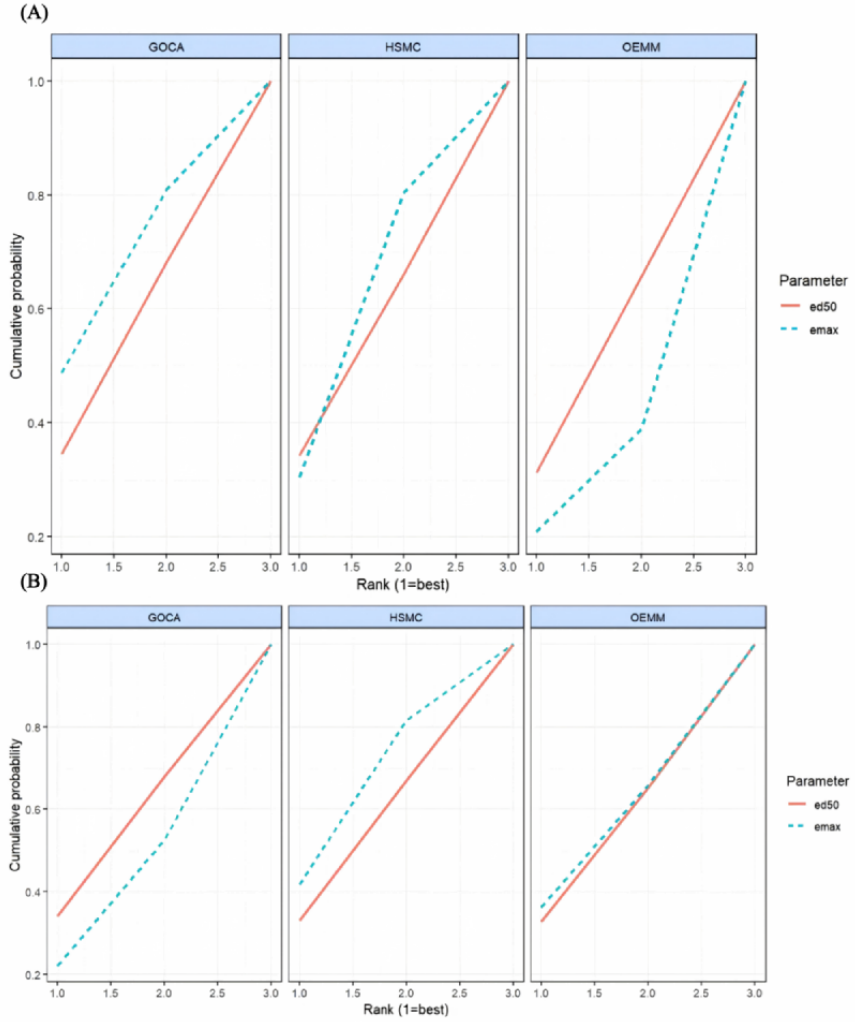
**Figure 6.** Dose-response prediction curves for different types of conversational agents in (A) depression and (B) anxiety. GOCA: goal-oriented conversational agent; HSMC: highly structured modular chatbot; MBNMA: model-based network dose-response meta-analysis; OEMM: open-ended multimodal-enhanced chatbot.






XSL•FO
RenderX

**Figure 7.** Cumulative ranking plots of Emax and ED50 parameters for different types of conversational agents in (A) depression and (B) anxiety. GOCA: goal-oriented conversational agent; HSMC: highly structured modular chatbot; OEMM: open-ended multimodal-enhanced chatbot.




### *Anxiety*

The network geometry for anxiety outcomes encompassed 9 intervention nodes across a control (Placebo_0) and 3 chatbot types (GOCA, HSMC, and OEMM). However, 2 nodes were absent from the network due to limited data availability: GOCA_3 and OEMM_2 (Figure 4B). Preliminary dose-response relationship assessment confirmed nonlinear patterns (Figure 5B). MCMC sampling demonstrated stable trajectories and adequate mixing of total residual deviance (Figure S1B in Multimedia Appendix 6), with convergence diagnostics confirming R-hat=1.001 and ESS reaching 3000. The Emax model yielded a DIC of 214, closely comparable to the UME model (213.6), with between-study heterogeneity SD 0.742 (95% CrI 0.457-1.24). Deviance plots, fitted vs observed value plots, and calibration plots (Figures S2B, S3B, and S4B in Multimedia Appendix 6) revealed minor deviations at high doses. Node-splitting and posterior density overlap analysis (Figure S5B in Multimedia Appendix 6) indicated that consistency assumptions were satisfied for OEMM at all dose levels (OEMM_1 and OEMM_3: $P$=.58) and HSMC_2 ($P$=.12) vs placebo; however, posterior density separation for HSMC_1, HSMC_3, GOCA_1, and GOCA_2 ($P$<.05) indicated substantial local inconsistency between direct and indirect evidence. The dose-response prediction curves revealed marginally declining response rates with increasing dose for GOCA and OEMM, while HSMC remained relatively stable with notably widened CIs at higher doses (Figure 6B). Compared with placebo, low-dose GOCA_1 (OR 4.77, 95% CrI 3.34-6.80) and HSMC_1 (OR 1.44, 95% CrI 1.13-1.82) demonstrated statistically significant efficacy. Although high-dose HSMC_3 also differed from placebo (OR 0.52, 95% CrI 0.34-0.78), the estimate was below 1, suggesting diminishing benefit. All OEMM dose groups had 95% CrIs that included 1, indicating no statistically significant differences vs placebo. In pairwise comparisons, GOCA_1 and HSMC_2 were placed on the left (GOCA_1 vs GOCA_2: OR 0.29, 95% CrI 0.15-0.58; HSMC_2 vs HSMC_3: OR 0.37, 95% CrI 0.20-0.67), indicating diminishing returns for middose GOCA_2 relative to GOCA_1 and for HSMC_3 relative to HSMC_2 (Table S1 in Multimedia Appendix 6). Cumulative ranking probability plots (Figure 7B) indicated that HSMC ranked relatively higher for Emax (0.254), while the 3 CA types showed similar ED50 rankings, suggesting that HSMC possesses superior maximum efficacy potential, whereas GOCA demonstrates more pronounced dose-response sensitivity at lower dose ranges (Table S2 in Multimedia Appendix 6). Sensitivity analyses showed that low-dose GOCA_1 outperformed GOCA_2 and that HSMC_2 outperformed



XSL•FO
RenderX

HSMC_3 and remained stable. Posterior means and CrIs closely overlapped with those from the main analysis (DIC: Emax 208.2 vs UME 209.5; SD 0.75, 95% CrI 0.45-1.25; R-hat=1.00; ESS=2882), indicating robustness of the overall conclusions (Table S3 in Multimedia Appendix 6).

### Moderation by Scale Characteristics

In the subgroup analyses with the assessment scale as a moderator, the network geometry differed between the PHQ-9 and GAD-7 subgroups: compared with the PHQ-9 network, the GAD-7 network lacked the GOCA_3 and OEMM_3 nodes (Figure S6 in Multimedia Appendix 6). In the preliminary exploration, both the depression and anxiety evidence networks showed nonlinear patterns (Figure S7 in Multimedia Appendix 6). For both subgroups, MCMC sampling converged well, with stable total residual deviance trajectories and adequate mixing (Figure S8 in Multimedia Appendix 6), where R-hat ranged from 1.00 to 1.01 and ESS were close to or above 1000. For model fit, Emax models outperformed UME models in both scales, with DIC values of 176 vs 178.6 for PHQ-9 and 188.8 vs 190.2 for GAD-7. The between-study heterogeneity SDs were 0.924 and 0.802, respectively. Diagnostic plots including bias, fit, and calibration indicated that the observed proportions and the model-based posterior predictive probabilities across treatment arms were closely aligned, demonstrating good calibration and predictive performance (Figures S9-S11 in Multimedia Appendix 6). Node-splitting and direct-vs-indirect posterior density checks indicated that, in the PHQ-9 subgroup, all dose levels of OEMM ($P$>.05) and HSMC_2 ($P$=.20) met the consistency assumption, whereas HSMC_1, HSMC_3, GOCA_1, and GOCA_3 showed significant local inconsistency ($P$<.05). In the GAD-7 subgroup, aside from HSMC_2 ($P$=.12), the posterior density distributions for the remaining direct-vs-indirect comparisons were significantly separated ($P$<.05), suggesting local statistical inconsistency (Figure S12 in Multimedia Appendix 6).

Relative efficacy analyses (Table S4 in Multimedia Appendix 6) showed that low-dose GOCA_1 (PHQ-9: OR 4.76, 95% CrI 3.34-6.80; GAD-7: OR 4.65, 95% CrI 3.27-6.61) and HSMC_1 (PHQ-9: OR 1.76, 95% CrI 1.13-2.74; GAD-7: OR 1.50, 95% CrI 1.17-1.91) demonstrated significantly greater efficacy than placebo across both depression and anxiety scales. High-dose HSMC_3 showed diminishing benefit relative to placebo on both scales (PHQ-9: OR 0.54, 95% CrI 0.37-0.78; GAD-7: OR 0.51, 95% CrI 0.34-0.77). Pairwise comparisons of GOCA_1 vs GOCA_2 across both PHQ-9 and GAD-7 scales revealed OR <1 with 95% CrIs excluding 1, demonstrating diminished returns for GOCA_2 relative to GOCA_1. Cumulative ranking (Figure S13 in Multimedia Appendix 6) indicated that for the Emax, GOCA was optimal in the depression (PHQ-9) subgroup (0.261), whereas HSMC was optimal in the anxiety (GAD-7) subgroup (0.249). The ED50 rankings were highly similar between the 2 subgroups. The predicted curves suggested that in the PHQ-9 subgroup, the response rate under GOCA increased slightly with dose, while OEMM showed a modest decline; in the GAD-7 subgroup, all 3 intervention types showed relatively stable or slightly decreasing stability. Moreover, high-dose intervals exhibited substantially widened predictive variability (Figure S14 in Multimedia Appendix 6), indicating earlier sensitivity to clinically meaningful effects at lower doses in both subgroups, as summarized in Table S5 in Multimedia Appendix 6.

Collectively, the moderation analysis by assessment scale substantiated the robustness of the primary MBNMA findings, excluding potential confounding from measurement heterogeneity. However, residual statistical inconsistency persisted among certain local network nodes, warranting cautious interpretation of the results. The current evidence base supports the following exploratory characterization: at low doses (0-30 min/wk), GOCA may harbor greater therapeutic potential for depression outcomes; at moderate doses (30-60 min/wk), HSMC appeared to show superior treatment durability and consolidated its top-ranking SUCRA position for anxiety outcomes, whereas marginal gains for GOCA became attenuated; at high doses (>60 min/wk), all interventions showed diminishing returns, with evidence of an efficacy plateau or declining trajectory.

### Publication Bias

For depression, the comparison-adjusted funnel plot inspection suggested mild asymmetry, indicating potential small-study effects or publication bias in Multimedia Appendix 7 (Figure S1). However, given the sparse nature of the underlying evidence network and the limited number of trials contributing to specific comparisons, the diagnostic ability of funnel-plot methods was substantially reduced. Consequently, supplementary approaches were used to assess small-study effects and publication bias. The results of the Egger regression intercept approach ($P$=.02) indicated significant evidence of publication bias in the pairwise meta-analysis. The Duval and Tweedie trim-and-fill analysis suggested that 7 studies were missing, and the adjusted Hedges g using the trim-and-fill approach was –0.17 (95% CI –0.22 to –0.11). The result of the classic fail-safe N suggested that 510 nil or null reports would be needed to raise the $P$ value associated with the average effect above an $\alpha$ level of 0.05.

For anxiety, the comparison-adjusted funnel plots did not reveal any evidence of asymmetry in Multimedia Appendix 7 (Figure S2). No significant publication bias was observed. The results of the Egger regression intercept approach ($P$=.14) indicated no significant evidence of publication bias in the pairwise meta-analysis. The Duval and Tweedie trim-and-fill analysis suggested that 4 studies were missing, and the adjusted g using the trim-and-fill approach was –0.19 (95% CI –0.27 to –0.11). The result of classic fail-safe N suggested that 496 nil or null reports would be needed to raise the $P$ value associated with the average effect above an $\alpha$ level of 0.05.

### RoB and Certainty of Evidence

RoB assessment (Figure 8) indicated that 12 studies were at low risk, 20 had some concerns, and 9 were at high RoB. Study-level assessments are provided in Multimedia Appendix 8 (Figure S1) [10,29-68].

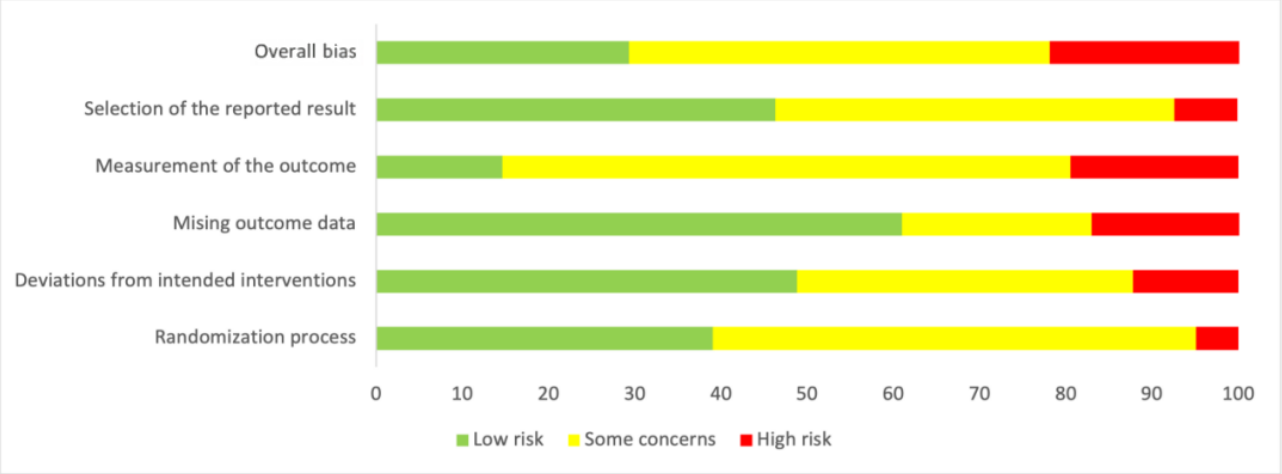
According to the CINeMA framework, the overall certainty of evidence was very low. This was primarily driven by concerns regarding RoB, indirectness, and imprecision. Many studies relied on self-reported outcomes, and a substantial proportion



XSL•FO
RenderX

were small-scale or pilot trials, contributing to wide CIs and limited robustness of estimates. Detailed assessments are presented in Multimedia Appendix 8 (Figures S2 and S3).

**Figure 8.** Risk of bias (RoB) summary (Cochrane RoB 2) across included studies.




## Discussion

### Principal Findings

This systematic review and NMA, encompassing 38 studies on depression (n=6985) and 32 studies on anxiety (n=6617), integrates standardized pairwise and network comparisons with MBNMA to explore whether CA architecture, levels of human support, and weekly usage duration influence short-term symptom reduction. However, given the predominantly very low to low certainty of evidence (CINeMA assessment), widespread statistical nonsignificance across pairwise comparisons, and considerable uncertainty in nonlinear dose-response parameters, our findings warrant cautious interpretation.

A prominent feature of our NMA was widespread lack of statistical significance across most pairwise comparisons. Rather than indicating therapeutic equivalence, this likely reflects inherent homogeneity of digital mental health interventions [69]. Most CAs share common therapeutic "active ingredients" (eg, cognitive restructuring and behavioral activation), resulting in smaller effect size differentials than those typically observed in pharmacological trials [70]. This issue is further compounded by a star-shaped network topology with limited head-to-head comparisons, modest sample sizes per comparison, and detection of a small-study effect in depression analysis (Egger $P$=.02). These factors collectively reduce statistical power and increase uncertainty in pairwise effect estimates [71]. SUCRA rankings are susceptible to considerable uncertainty arising from model estimation variance and random error, and their interpretation must be integrated with direct pairwise comparisons and CI width [72].

First, AEIHS appeared to show a statistically significant advantage over PHDG in depression. However, this comparison derives predominantly from sparse network evidence and cannot confirm synergistic enhancement mechanisms. While one might speculate that AI provides temporal coverage through immediate 24/7 tracking and human support offers therapeutic alliance and crisis management [10,73], such mechanistic interpretations remain tentative pending direct comparative evidence. Second, AEIHS appeared to show a significant advantage over OEMM in anxiety. Mechanistically, evidence-based anxiety treatment relies fundamentally on structured exposure hierarchies and graduated habituation protocols [74]. The highly structured, human-supported nature of AEIHS may better ensure fidelity in the delivery of these essential components, whereas the unstructured, user-led dialogue of OEMM lacks such safeguards. But this observation should be interpreted cautiously. Because the apparent association cannot establish causality, particularly given that this comparison lacked direct head-to-head evidence and the network exhibited notable inconsistency.

Our MBNMA identified suggestive, but highly uncertain, dose-response patterns across chatbot subtypes, which should be interpreted as preliminary observations rather than definitive evidence for dosing guidance. The analyses suggest that GOCA may have exhibited greater symptom improvements, particularly regarding depression, at lower doses (0-30 min/wk) compared to higher doses, a pattern that could align with skill acquisition theory if therapeutic components are effectively delivered early in exposure. In other words, we view the sequence of "goal orientation, repetitive practice, skill internalization, and enhanced self-efficacy" as an explanatory hypothesis: GOCA may facilitate perceptible therapeutic progress within a short period of exposure through more structured goals and feedback, thereby yielding greater symptom relief during the early stages [75]. However, this finding derives from limited data density, and the mechanisms underlying this apparent pattern remain speculative. Conversely, HSMC appeared to be associated with sustained benefits at moderate doses (30-60 minutes per week), particularly for anxiety outcomes. Because therapeutic content is compartmentalized into discrete, skill-focused modules, core skills may be acquired relatively quickly [76]; prolonged usage may primarily serve as review, yielding diminishing marginal returns [77]. This dose range may warrant consideration in future research, though the evidence remains exploratory.

Finally, OEMM yielded highly dispersed and unstable dose-response estimates. Its open-ended design means that therapeutic exposure may be entirely dependent on individual



XSL•FO
RenderX

user interaction patterns, resulting in unpredictable group-level fidelity [78]. Consequently, the optimal dosing strategy for open dialogue remains unclear and is unlikely to follow a simple "more is better" relationship. Given the sparse network geometry, limited sample sizes contributing to specific dose-response comparisons, the reliance of rankings on indirect evidence, and predominantly low to very low certainty of the underlying evidence base, these dose-response patterns should be regarded as exploratory observations generating plausible hypotheses for future hypothesis-driven research, rather than as evidence supporting specific dosing recommendations.

### Implications for Practice and Design

The overall evidence base remains limited. For example, AEIHS-related evidence derives from only 4 trials, with the depression AEIHS-PHDG comparison informed by 2 small studies from the same research group (combined n=75). Evidence regarding OEMM use-time patterns is sparse, and high-dose data are particularly thin. Diagnostic checks indicate deviation in high-dose predictions, and most GOCA and HSMC nodes exhibit local inconsistency. Given these limitations, these exploratory findings do not support definitively recommending any single CA strategy as first-line treatment. Rather, AI-based agents should continue to be positioned as complementary, adjunctive tools. The combination of CA with human support appears promising but requires confirmation in sufficiently powered head-to-head trials. While emerging patterns suggest a dose-response relationship, practitioners should not assume that simply excluding OEMM, or increasing frequency and duration, will improve outcomes. The emerging suggestive pattern that 0-60 minutes/week may be associated with symptom improvements is based on sparse dose data and should be considered hypothesis-generating only, not a dosing recommendation; it requires re-examination in adequately powered studies. Engagement metrics should be evaluated in relation to symptom trajectories rather than treated as direct clinical improvement surrogates. In depression, a small-study effect was detected: trim-and-fill analysis identified 7 missing studies, yielding corrected effect size (Hedges *g*) of –0.17 (95% CI –0.22 to –0.11), suggesting that average treatment effects may be inflated.

### Strengths and Limitations

The strengths of this review lie in its 2D classification of CA interventions, the joint use of NMA and MBNMA, comprehensive evaluation of inconsistency through global, local, and node-splitting approaches, and formal assessment of small-study effects alongside evidence certainty. Several limitations, however, constrain the strength of the inferences that can be drawn. First, main networks violated consistency assumptions, with inconsistency concentrated in highest-effect comparisons, raising the possibility that observed heterogeneity reflects transitivity violations and not only genuine treatment differences. Second, only 23 of 41 trials reported interaction time with sufficient detail. Dose was operationalized as time alone, overlooking interaction quality, module completion, and semantic depth. The estimates may conflate "higher engagement" with "more effective dose" rather than capturing active component potency. Third, the synthesis includes only postintervention outcomes, precluding assessment of durability, symptom recurrence, or long-term trajectories. RoB was low in only 12 of 41 trials; most outcomes were self-reported; and many studies were small pilot trials, collectively contributing to very low CINeMA ratings. Future research should prioritize multiarm trials with direct comparisons across CA subtypes and support intensities, conduct individual-participant-data meta-analysis, and include ≥6-month follow-up. Exposure should use multidimensional, standardized definitions (module completion, interaction turns, session quality, and behavior change indicators).

### Conclusions

This systematic review suggests that differences across most CA categories are small and statistically uncertain. Networks exhibited notable inconsistency, and depression evidence showed publication bias indicators. Amid this heterogeneity, AEIHS appeared to show greater symptom reduction relative to PHDG and OEMM, though with substantial uncertainty. In dose-response analyses, GOCA appeared to show greater gains at low doses (0-30 min/wk) and HSMC at moderate doses (30-60 min/wk), while OEMM yielded unstable response trajectories. Given extremely low certainty of evidence, these findings should be interpreted as exploratory observations generating hypotheses for future research, not as established evidence supporting clinical practice recommendations. Confirmation requires adequately powered multiarm head-to-head trials with standardized dose reporting.

## Funding

This study was funded by grants from the 2026 foreign expert program of Hainan Province (number H20260407025E). The funder had no role in the study design, data collection, analysis, interpretation of the results, or the writing of this paper.

## Data Availability

All data supporting the findings of this systematic review are included in the main text and its supplementary information. For further details or additional datasets, please contact the corresponding author.



XSL•FO
RenderX

## Generative AI Use

The authors declare that Claude Opus 5.5 (Anthropic PBC) was used to assist with parts of the English proofreading. After using this tool or service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

## Authors' Contributions

Conceptualization: PLC (lead), HZ (equal)
Investigation: PY (lead), ML (equal), MH, YW, PC, HS, LC
Data curation: PY (lead), ML (equal), MH, YW, PC, HS, LC
Formal analysis: PY (lead), ML (equal), MH, YW, PC, HS, LC
Methodology: PY (lead), PLC (supporting), HZ (supporting)
Validation: PY (lead), ML (equal), MH, YW, PC, HS, LC
Supervision: PLC (lead), HZ (equal)
Writing – original draft: PY (lead), ML (equal)
Writing – review & editing: PY (lead), ML (supporting), PLC (supporting), HZ (supporting)

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

PRISMA NMA checklist.
[DOCX File , 25 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

Search strategies and results of this study.
[DOCX File , 19 KB-Multimedia Appendix 2]

## Multimedia Appendix 3

Methods involved in the study.
[DOCX File , 49 KB-Multimedia Appendix 3]

## Multimedia Appendix 4

Characteristics of included studies.
[DOCX File , 55 KB-Multimedia Appendix 4]

## Multimedia Appendix 5

Network meta-analysis.
[DOCX File , 396 KB-Multimedia Appendix 5]

## Multimedia Appendix 6

Dose network meta- analysis.
[DOCX File , 4008 KB-Multimedia Appendix 6]

## Multimedia Appendix 7

Publication bias.
[DOCX File , 235 KB-Multimedia Appendix 7]

## Multimedia Appendix 8

Risk of bias and certainty of evidence.
[DOCX File , 1406 KB-Multimedia Appendix 8]

## References



XSL•FO
RenderX

1. Barker MM, Beresford B, Bland M, Fraser LK. Prevalence and incidence of anxiety and depression among children, adolescents, and young adults with life-limiting conditions: a systematic review and meta-analysis. JAMA Pediatr. 2019;173(9):835-844. [FREE Full text] [doi: 10.1001/jamapediatrics.2019.1712] [Medline: 31282938]
2. Mental health at work. World Health Organization. Sep 15, 2026. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work [accessed 2026-10-29]
3. Zhang W, Yan Y, Wu Y, Yang H, Zhu P, Yan F, et al. Medicinal herbs for the treatment of anxiety: a systematic review and network meta-analysis. Pharmacol Res. 2022;179:106204. [doi: 10.1016/j.phrs.2022.106204] [Medline: 35378276]
4. Butler AC, Chapman JE, Forman EM, Beck AT. The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clin Psychol Rev. 2006;26(1):17-31. [doi: 10.1016/j.cpr.2005.07.003] [Medline: 16199119]
5. Cuijpers P, Karyotaki E, Weitz E, Andersson G, Hollon SD, van Straten A. The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. J Affect Disord. 2014;159:118-126. [doi: 10.1016/j.jad.2014.02.026] [Medline: 24679399]
6. Corrigan P. How stigma interferes with mental health care. Am Psychol. 2004;59(7):614-625. [doi: 10.1037/0003-066X.59.7.614] [Medline: 15491256]
7. Moitra M, Santomauro D, Collins PY, Vos T, Whiteford H, Saxena S, et al. The global gap in treatment coverage for major depressive disorder in 84 countries from 2000-2019: a systematic review and Bayesian meta-regression analysis. PLoS Med. 2022;19(2):e1003901. [FREE Full text] [doi: 10.1371/journal.pmed.1003901] [Medline: 35167593]
8. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT, Li YJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. J Biomed Sci. 2025;32(1):45. [FREE Full text] [doi: 10.1186/s12929-025-01131-z] [Medline: 40335969]
9. Raptopoulou A, Komnidis A, Bamidis PD, Astaras A. Human-robot interaction for social skill development in children with ASD: a literature review. Healthcare Technol Lett. 2021;8(4):90-96. [FREE Full text] [doi: 10.1049/htl2.12013] [Medline: 34295506]
10. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health. 2017;4(2):e19. [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
11. 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. 2018;6(11):e12106. [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
12. Rousmaniere T, Goldberg SB, Torous J. Large language models as mental health providers. Lancet Psychiatry. 2026;13(1):7-9. [doi: 10.1016/S2215-0366(25)00269-X] [Medline: 40939602]
13. Kola L, Kohrt BA, Hanlon C, Naslund JA, Sikander S, Balaji M, et al. COVID-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health. Lancet Psychiatry. 2021;8(6):535-550. [FREE Full text] [doi: 10.1016/S2215-0366(21)00025-0] [Medline: 33639109]
14. Hofmann SG, Sawyer AT, Witt AA, Oh D. The effect of mindfulness-based therapy on anxiety and depression: a meta-analytic review. J Consult Clin Psychol. 2010;78(2):169-183. [FREE Full text] [doi: 10.1037/a0018555] [Medline: 20350028]
15. Sohn JS, Ha BG, Park S, Kim JJ, Lee E, Oh H, et al. Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms. npj Digit Med. 2026;9(1):377. [FREE Full text] [doi: 10.1038/s41746-026-02566-w] [Medline: 41882250]
16. Gong B, Yao N, Xie H, Huang C, Kishimoto T, Berenbaum H, et al. Efficacy, user engagement, and acceptability of cognitive behavioral therapy-oriented psychological chatbots for adults with depressive and/or anxiety symptoms: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res. 2026;28:e82677. [FREE Full text] [doi: 10.2196/82677] [Medline: 42101333]
17. Satake Y, Costello H, Naran N, Ishimaru D, Ikeda M, Howard R. Autonomous conversational agents for loneliness, social isolation, depression, and anxiety in older people without cognitive impairment: systematic review and meta-analysis. Psychol Med. 2026;56:e27. [doi: 10.1017/S0033291725103073] [Medline: 41556104]
18. Huang T, Li S, Wang Y, Liu W. Therapeutic interaction features of AI chatbots in depression interventions: systematic review and meta-analysis. J Med Internet Res. 2026;28:e88697. [doi: 10.2196/88697] [Medline: 42379144]
19. Leung WKC, Lam SC, Chan BCL, Chow JNL, Wong YYY, Ng F, et al. Chatbot interventions for improving mental health among people in Asia: a systematic review and meta-analysis of randomised controlled trials. BMJ Health Care Inform. 2026;33(1):e101479. [FREE Full text] [doi: 10.1136/bmjhci-2025-101479] [Medline: 41534952]
20. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med. 2015;162(11):777-784. [FREE Full text] [doi: 10.7326/M14-2385] [Medline: 26030634]
21. Tian S, Liang Z, Tian M, Qiu F, Yu Y, Mou H, et al. Comparative efficacy of various exercise types and doses for depression in older adults: a systematic review of paired, network and dose-response meta-analyses. Age Ageing. 2024;53(10):afae211. [FREE Full text] [doi: 10.1093/ageing/afae211] [Medline: 39348911]
22. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898. [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]




XSL•FO
RenderX

23. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, Chaimani A, Del Giovane C, Egger M, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. PLoS Med. 2020;17(4):e1003082. [FREE Full text] [doi: 10.1371/journal.pmed.1003082] [Medline: 32243458]
24. Papakonstantinou T, Nikolakopoulou A, Higgins JPT, Egger M, Salanti G. CINeMA: software for semiautomated assessment of the confidence in the results of network meta-analysis. Campbell Syst Rev. 2020;16(1):e1080. [FREE Full text] [doi: 10.1002/cl2.1080] [Medline: 37131978]
25. Lopez P, Radaelli R, Taaffe D, Newton R, Galvão DA, Trajano G, et al. Resistance training load effects on muscle hypertrophy and strength gain: systematic review and network meta-analysis. Med Sci Sports Exerc. 2021;53(6):1206-1216. [FREE Full text] [doi: 10.1249/MSS.0000000000002585] [Medline: 33433148]
26. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365-376. [doi: 10.1038/nrn3475] [Medline: 23571845]
27. Pedder H, Dias S, Boucher M, Bennetts M, Mawdsley D, Welton NJ. Methods to assess evidence consistency in dose-response model based network meta-analysis. Stat Med. 2022;41(4):625-644. [doi: 10.1002/sim.9270] [Medline: 34866221]
28. Mawdsley D, Bennetts M, Dias S, Boucher M, Welton N. Model-based network meta-analysis: a framework for evidence synthesis of clinical trial data. CPT Pharmacometrics Syst Pharmacol. 2016;5(8):393-401. [FREE Full text] [doi: 10.1002/psp4.12091] [Medline: 27479782]
29. 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. 2022;6(11):e40242. [FREE Full text] [doi: 10.2196/40242] [Medline: 36413390]
30. Sandoval LR, Buckey JC, Ainslie R, Tombari M, Stone W, Hegel MT. Randomized controlled trial of a computerized interactive media-based problem solving treatment for depression. Behav Ther. 2017;48(3):413-425. [FREE Full text] [doi: 10.1016/j.beth.2016.04.001] [Medline: 28390503]
31. Danieli M, Ciulli T, Mousavi SM, Silvestri G, Barbato S, Di Natale L, 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. 2022;9(9):e38067. [FREE Full text] [doi: 10.2196/38067] [Medline: 36149730]
32. Danieli M, Ciulli T, Mousavi SM, Riccardi G. A conversational artificial intelligence agent for a mental health care app: evaluation study of its participatory design. JMIR Form Res. 2021;5(12):e30053. [FREE Full text] [doi: 10.2196/30053] [Medline: 34855607]
33. Chan WS, Cheng WY, Lok SHC, Cheah AKM, Lee AKW, Ng ASY, 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. Aug 07, 2024;11:e51716. [FREE Full text] [doi: 10.2196/51716] [Medline: 39110971]
34. Sabour S, Zhang W, Xiao X, Zhang Y, Zheng Y, Wen J, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health. 2023;5:1133987. [FREE Full text] [doi: 10.3389/fdgth.2023.1133987] [Medline: 37214342]
35. Ulrich S, Gantenbein AR, Zuber V, Von Wyl A, Kowatsch T, Künzli H. 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. Jan 24, 2024;26:e50132. [FREE Full text] [doi: 10.2196/50132] [Medline: 38265863]
36. Zhao Y, Qian W, Chen Y, Wu D, Luo Y, Gao C, 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. Oct 2025;17(5):e70067. [doi: 10.1111/aphw.70067] [Medline: 40910958]
37. 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. May 20, 2021;9(5):e26152. [FREE Full text] [doi: 10.2196/26152] [Medline: 33872182]
38. de Graaff AM, Habashneh R, Fanatseh S, Keyan D, Akhtar A, Abualhaija A, et al. Evaluation of a guided chatbot intervention for young people in Jordan: feasibility randomized controlled trial. JMIR Ment Health. Feb 05, 2025;12:e63515. [FREE Full text] [doi: 10.2196/63515] [Medline: 39908538]
39. 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. Sep 2018;46(5):570-582. [FREE Full text] [doi: 10.1017/S1352465817000820] [Medline: 29366432]
40. Jang S, Kim JJ, Kim SJ, Hong J, Kim S, Kim E. 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 Inform. Jun 2021;150:104440. [doi: 10.1016/j.ijmedinf.2021.104440] [Medline: 33799055]
41. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. 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. Oct 31, 2019;7(10):e15018. [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
42. 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. Sep 01, 2025;43(9). [doi: 10.1097/CIN.0000000000001349] [Medline: 40704906]




XSL•FO
RenderX

43. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, 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. [FREE Full text] [doi: 10.2196/40719] [Medline: 36355633]
44. 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 Interv. Mar 2022;27:100495. [FREE Full text] [doi: 10.1016/j.invent.2022.100495] [Medline: 35059305]
45. 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. Mar 20, 2024;8:e47960. [FREE Full text] [doi: 10.2196/47960] [Medline: 38506892]
46. 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:e54945. [FREE Full text] [doi: 10.2196/54945] [Medline: 38922677]
47. Fitzsimmons-Craft EE, Chan WW, Smith AC, Firebaugh ML, Fowler LA, Topooco N, et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. Int J Eat Disord. Mar 2022;55(3):343-353. [doi: 10.1002/eat.23662] [Medline: 35274362]
48. Reilly ED, Kelly MM, Grigorian HL, Waring ME, Quigley KS, Hogan TP, et al. Virtual coach-guided online acceptance and commitment therapy for chronic pain: pilot feasibility randomized controlled trial. JMIR Form Res. Nov 08, 2024;8:e56437. [FREE Full text] [doi: 10.2196/56437] [Medline: 39514264]
49. Kang B, Hong M. Digital interventions for reducing loneliness and depression in Korean college students: mixed methods evaluation. JMIR Form Res. Sep 12, 2024;8:e58791. [FREE Full text] [doi: 10.2196/58791] [Medline: 39264705]
50. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. JMIR Form Res. Aug 12, 2021;5(8):e20678. [FREE Full text] [doi: 10.2196/20678] [Medline: 34092548]
51. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. J Med Internet Res. Apr 30, 2025;27:e70436. [FREE Full text] [doi: 10.2196/70436] [Medline: 40306635]
52. 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. May 21, 2025;27:e70874. [FREE Full text] [doi: 10.2196/70874] [Medline: 40397933]
53. Kleinau E, Lamba T, Jaskiewicz W, Gorentz K, Hungerbuehler I, Rahimi D, 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. 2024;19(5):e0303370. [FREE Full text] [doi: 10.1371/journal.pone.0303370] [Medline: 38805444]
54. Oh J, Jang S, Kim H, Kim JJ. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inform. Aug 2020;140:104171. [doi: 10.1016/j.ijmedinf.2020.104171] [Medline: 32446158]
55. Burton C, Tatar AS, McKinstry B, Matheson C, Matu S, Moldovan R, et al. Help4Mood Consortium. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. J Telemed Telecare. Sep 2016;22(6):348-355. [doi: 10.1177/1357633X15609793] [Medline: 26453910]
56. Romanovskyi O, Pidbutska N, Knysh A. Elomia chatbot: the effectiveness of artificial intelligence in the fight for mental health. 2021. Presented at: Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021); April 22–23, 2021:1215-1224; Lviv, Ukraine. URL: https://ceur-ws.org/Vol-2870/paper89.pdf
57. Suharwardy S, Ramachandran M, Leonard SA, Gunaseelan A, Lyell DJ, Darcy A, et al. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. AJOG Glob Rep. Aug 2023;3(3):100165. [FREE Full text] [doi: 10.1016/j.xagr.2023.100165] [Medline: 37560011]
58. Nguyen QC, Aparicio EM, Jasczynski M, Doig AC, Yue X, Mane H, et al. Rosie, a health education question-and-answer chatbot for new mothers: randomized pilot study. JMIR Form Res. Jan 12, 2024;8:e51361. [FREE Full text] [doi: 10.2196/51361] [Medline: 38214963]
59. Six S, Schlesener E, Hill V, Babu SV, 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. Apr 11, 2025;12:e67381. [FREE Full text] [doi: 10.2196/67381] [Medline: 40215483]
60. Johnson C, Egan SJ, Carlbring P, Shafran R, Wade TD. Artificial intelligence as a virtual coach in a cognitive behavioural intervention for perfectionism in young people: a randomised feasibility trial. Internet Interv. Dec 2024;38:100795. [FREE Full text] [doi: 10.1016/j.invent.2024.100795] [Medline: 39717697]
61. Vereschagin M, Wang AY, Richardson CG, Xie H, Munthali RJ, Hudec KL, et al. Effectiveness of the Minder mobile mental health and substance use intervention for university students: randomized controlled trial. J Med Internet Res. Mar 27, 2024;26:e54287. [FREE Full text] [doi: 10.2196/54287] [Medline: 38536225]
62. MacNeill AL, Doucet S, Luke A. Effectiveness of a mental health chatbot for people with chronic diseases: randomized controlled trial. JMIR Form Res. May 30, 2024;8:e50025. [FREE Full text] [doi: 10.2196/50025] [Medline: 38814681]




XSL•FO
RenderX

63. Prochaska JJ, Vogel EA, Chieng A, Baiocchi M, Maglalang DD, Pajarito S, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. Drug Alcohol Depend. Oct 01, 2021;227:108986. [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108986] [Medline: 34507061]
64. Chen C, Lam KT, Yip KM, So HK, Lum TYS, Wong ICK, 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 06, 2025;12:e65785. [FREE Full text] [doi: 10.2196/65785] [Medline: 40048637]
65. Gleason MM, Flom M, Rapoport S, Williams A, Birch A, Wells NK, et al. A relational agent intervention for adolescents seeking mental health treatment: outcomes from a randomized controlled trial within a children's outpatient hospital. JAACAP Open. Dec 2025;3(4):1033-1045. [FREE Full text] [doi: 10.1016/j.jaacop.2025.02.002] [Medline: 41367963]
66. Yasukawa S, Tanaka T, Yamane K, Kano R, Sakata M, Noma H, 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. Jan 10, 2024;27(1). [FREE Full text] [doi: 10.1136/bmjment-2023-300881] [Medline: 38199786]
67. Weiner LS, Crowley RN, Sheeber LB, Koegler FH, Davis JF, Wells M, et al. Engagement, acceptability, and effectiveness of the self-care and coach-supported versions of the vira digital behavior change platform among young adults at risk for depression and obesity: pilot randomized controlled trial. JMIR Ment Health. Sep 19, 2024;11:e51366. [FREE Full text] [doi: 10.2196/51366] [Medline: 39298763]
68. Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI. Apr 2025;2(4):AIoa2400802. [doi: 10.1056/AIoa2400802] [Medline: 42932588]
69. Linardon J. Can acceptance, mindfulness, and self-compassion be learned by smartphone apps? A systematic and meta-analytic review of randomized controlled trials. Behav Ther. 2020;51(4):646-658. [doi: 10.1016/j.beth.2019.10.002] [Medline: 32586436]
70. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry. 2019;18(3):325-336. [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
71. Cuijpers P, Noma H, Karyotaki E, Cipriani A, Furukawa TA. Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression: a network meta-analysis. JAMA Psychiatry. 2019;76(7):700-707. [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.0268] [Medline: 30994877]
72. Singh S, Murad MH, Fumery M, Dulai PS, Sandborn WJ. First- and second-line pharmacotherapies for patients with moderate to severely active ulcerative colitis: an updated network meta-analysis. Clin Gastroenterol Hepatol. 2020;18(10):2179-2191.e6. [FREE Full text] [doi: 10.1016/j.cgh.2020.01.008] [Medline: 31945470]
73. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health. 2018;5(4):e64. [FREE Full text] [doi: 10.2196/mental.9782] [Medline: 30545815]
74. Craske MG, Treanor M, Conway CC, Zbozinek T, Vervliet B. Maximizing exposure therapy: an inhibitory learning approach. Behav Res Ther. 2014;58:10-23. [FREE Full text] [doi: 10.1016/j.brat.2014.04.006] [Medline: 24864005]
75. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med. 2023;53(2):567-579. [FREE Full text] [doi: 10.1017/S0033291721001999] [Medline: 34047264]
76. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res. 2009;11(2):e13. [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
77. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry. 2017;74(4):351-359. [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
78. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical metrics used to evaluate health care chatbots: scoping review. J Med Internet Res. 2020;22(6):e18301. [FREE Full text] [doi: 10.2196/18301] [Medline: 32442157]


## Abbreviations

**AEIHS:** AI-enhanced intensive human support
**AIMHS:** AI-assisted minimal human support
**AIRHS:** AI-assisted routine human support
**BDI-II:** Beck Depression Inventory II
**BDSI/BAI:** Behavioral Dysexecutive Syndrome Inventory/Beck Anxiety Inventory
**CA:** conversational agent
**CBT:** cognitive behavioral therapy
**CES-D:** Center for Epidemiologic Studies Depression Scale
**CINeMA:** confidence in network meta-analysis



XSL•FO
RenderX

**CrI:** credible interval
**DASS-21:** Depression Anxiety Stress Scale
**DIC:** deviance information criterion
**ESS:** effective sample size
**GAD-7:** Generalized Anxiety Disorder 7-item scale
**GAD-Q-V:** Generalized Anxiety Disorder Questionnaire
**GOCA:** goal-oriented conversational agent
**HADS:** Hospital Anxiety and Depression Scale
**HSMC:** highly structured modular chatbot
**MBNMA:** model-based network dose-response meta-analysis
**MCMC:** Markov Chain Monte Carlo
**MD:** mean difference
**NMA:** network meta-analysis
**OEMM:** open-ended multimodal-enhanced chatbot
**OR:** odds ratio
**PHDG:** pure human digital guidance
**PHQ-8:** Patient Health Questionnaire-8
**PHQ-9:** Patient Health Questionnaire-9
**PICOS:** Participants, Interventions, Comparators, Outcomes, and Study Design
**Placebo_0:** pure human digital guidance, waitlist, usual care, health information and education, and general health app
**PRISMA:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses
**PRISMA-NMA:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Network Meta-Analyses
**PROMIS:** Patient-Reported Outcomes Measurement Information System
**PROSPERO:** International Prospective Register of Systematic Reviews
**RCT:** randomized controlled trial
**RoB:** risk of bias
**SAS:** Self-Rating Anxiety Scale
**SCARED:** Screen for Child Anxiety Related Emotional Disorders
**SMD:** standardized mean difference
**SUCRA:** surface under the cumulative ranking curve
**UC:** waitlist, usual care, health information and education, and general health app
**UME:** unrelated mean effect

*Edited by M Balcarras; submitted 09.Jun.2026; peer-reviewed by Z Quan, M Chatzimina; comments to author 16.Sep.2026; revised version received 26.Oct.2026; accepted 26.Oct.2026; published 13.Nov.2026*

*Please cite as:*
*Yang P, Chui PL, He M, Wu Y, Cheng P, Li M, Sun H, Cai L, Zhang H*
*Effectiveness of Conversational Agents for Depression and Anxiety: Systematic Review and Network Dose-Response Meta-Analysis of Agent Types, Dose, and Human Support*
*J Med Internet Res 2026;28:e104186*
*URL: https://www.jmir.org/2026/1/e104186*
*doi: 10.2196/104186*
*PMID:*

©Pengyu Yang, Ping Lei Chui, Mingxin He, Yao Wu, Pengfei Cheng, Minxiang Li, Haojie Sun, Lurong Cai, Hua Zhang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 13.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.



XSL•FO
RenderX