Review

# Effectiveness of Generative AI–Based Interventions in Behavioral Medicine: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Yexuan Xiao[1,2,3,4], MSc; Mengya Zhu[3,4], BSc; Tiansu Liu[4], BSc; Yilin He[4], MSc; Yuchen Ye[2,3,4], BSc; Yuhe Zhang[5], MSc; Yingchao Zhu[4], MSc; Dong Roman Xu[6], PhD; Vivian W Q Lou[7,8], PhD; Wan Fen Yip[9], PhD; Bin Sheng[10], PhD; John Torous[11], PhD; Tien Yin Wong[1,12,13], MD, PhD; Yih Chung Tham[13,14,15,16], MD, PhD; Nan Jiang[2,3,4], PhD

[1]School of Biomedical Engineering, Tsinghua Medicine, Tsinghua University, Beijing, China
[2]School of Healthcare Management, Tsinghua University, Beijing, China
[3]Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China
[4]Institute for Hospital Management, Tsinghua University, Shenzhen, China
[5]Peking Union Medical College Hospital, Beijing, China
[6]School of Health Management, Southern Medical University, Guangzhou, China
[7]Department of Social Work and Social Administration, University of Hong Kong, Hong Kong, China (Hong Kong)
[8]Sau Po Centre on Ageing, University of Hong Kong, Hong Kong, China (Hong Kong)
[9]Health Services and Outcomes Research, National Healthcare Group, Singapore, Singapore
[10]School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
[11]Beth Israel Deaconess Medical Center, Harvard medical school, Harvard University, Boston, MA, United States
[12]Beijing Visual Science and Translational Eye Research Institute, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China
[13]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore
[14]Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[15]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[16]Ophthalmology and Visual Science Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore

**Corresponding Author:**

Nan Jiang, PhD
School of Healthcare Management, Tsinghua University
Beijing 100084
China
Phone: 86 13717588712
Email: jiangnan1221@tsinghua.edu.cn

## Abstract

**Background:** Generative AI (GenAI), particularly large language models, is increasingly being used in behavioral medicine to provide scalable and personalized support. However, evidence from randomized controlled trials (RCTs) remains heterogeneous, and the contribution of generative capabilities beyond therapeutic content and intervention design remains unclear.

**Objective:** This systematic review and meta-analysis aimed to evaluate the effectiveness of GenAI-based interventions across behavioral medicine and to characterize the intervention features and trial designs relevant to their clinical evaluation and implementation.

**Methods:** We searched PubMed, Embase, Web of Science, and Scopus for RCTs published from January 1, 2020, to March 31, 2026. Eligible studies compared GenAI-based interventions with inactive controls, usual care, nongenerative interventions, or active comparators. The primary outcomes were depressive and anxiety symptoms. Secondary outcomes included affect, engagement, empathy, and therapeutic alliance. Risk of bias was assessed using the revised Cochrane risk-of-bias tool for randomized trials (RoB 2), and certainty of evidence was assessed using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach. Random-effects models were used to calculate pooled Hedges *g* values and 95% CIs.

**Results:** Of the 19,475 records identified, 20 RCTs met the eligibility criteria. Of these, 14 (70%) were at overall high risk of bias, 4 (20%) raised some concerns, and 2 (10%) were at low risk. GenAI-based interventions were associated with a small reduction in depressive symptoms in the primary random-effects analysis (Hedges $g$=–0.21, 95% CI –0.41 to –0.02; $I^2$=63%), although the Hartung-Knapp-Sidik-Jonkman (HKSJ) sensitivity analysis yielded a CI that included 0 (95% CI –0.46 to 0.04). Anxiety symptoms did not significantly improve ($g$=–0.14, 95% CI –0.31 to 0.02; $I^2$=63%). Negative affect improved, whereas positive affect did not significantly improve. Point estimates favored larger benefits against waitlist controls, but formal between-subgroup tests did not support differences by comparator type for depression ($P$=.20) or anxiety ($P$=.28). Engagement generally favored GenAI-based interventions. Narrative findings suggested greater empathy compared with rule-based systems, mixed empathy findings compared with human-delivered care, and no clear differences in therapeutic alliance compared with active conversational comparators. Follow-up beyond 3 months was limited, and safety reporting was sparse. The certainty of evidence ranged from low to very low.

**Conclusions:** GenAI-based interventions may provide small, outcome-specific benefits in behavioral medicine, but the available evidence does not establish whether these benefits are attributable specifically to generative capabilities. Effects appear to depend on intervention design, human involvement, comparator selection, and implementation context. Current evidence, therefore, supports viewing GenAI as a design-dependent augmentation tool rather than a standalone therapeutic innovation. Future trials should prioritize active comparators, longer follow-up, and transparent reporting of model characteristics and safety outcomes.

**Trial Registration:** PROSPERO CRD420251180679; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251180679

***J Med Internet Res 2026;28:e108789;*** doi: 10.2196/108789

**Keywords:** generative AI; behavioral medicine; large language models; digital health; randomized controlled trials; systematic review; meta-analysis

## Introduction

Behavioral medicine is an interdisciplinary field that applies behavioral, psychological, social, and biomedical knowledge to the prevention, diagnosis, and management of disease, with a particular focus on health-related behaviors and long-term self-management [1]. It is central to chronic disease prevention and management because many health outcomes are shaped by modifiable behavioral processes that influence adherence and sustained self-management [2]. Around half of the global chronic disease burden has been attributed to such behaviors [3], indicating that behavioral interventions are important for improving population health and reducing long-term health care expenditure [4]. By targeting these processes and supporting sustained engagement, behavioral medicine can generate benefits that extend beyond symptom control alone [5].

However, behavioral medicine interventions are often difficult to implement consistently in real-world settings [6]. Unlike interventions that can be delivered through relatively fixed protocols, behavioral care typically requires ongoing adjustment to individual motivation, psychological state, social circumstances, and environmental triggers [7]. This makes strategies such as coaching, nudging, cognitive behavioral techniques, and adherence support highly dependent on human input, contextual judgment, and sustained follow-up. In practice, these demands create major barriers to scale-up, particularly in settings affected by limited workforce capacity, long waits for care, and difficulty sustaining engagement over time [5]. Earlier digital approaches, including rule-based chatbots, sought to extend access, but their scripted logic and limited responsiveness often reduced their ability to provide meaningful support as users' needs evolved over time [8].

There is therefore growing interest in tools that can deliver more adaptive and scalable behavioral support. Generative AI (GenAI), particularly large language and multimodal models, enables dynamic content generation and interactive communication. These systems produce context-sensitive, conversational outputs in response to user input, enabling repeated, adaptive interactions that resemble human-delivered behavioral support [9]. They further enable dynamic personalization and deliver intervention components through conversational formats that integrate coaching, psychoeducation, and behavioral guidance, while providing empathic and naturalistic responses that might enhance engagement and sustain participation [10]. However, these generative and interactive capabilities also introduce new and potentially amplified risks. Technical fluency does not ensure clinical reliability. GenAI systems remain prone to factual errors, bias, unsafe responses, and limited interpretability. Parallel concerns have been raised in oncology, including model instability, cybersecurity, privacy, limited transparency, and the need for rigorous human oversight [11]. Such risks are especially important in behavioral care, where interventions often involve repeated engagement and sensitive health needs [12].

Although interest in GenAI for health care has expanded rapidly, its effectiveness in behavioral medicine has not been clearly established. Previous reviews have classified mental-health chatbots as rule-based, machine learning–based, or large language model (LLM)–based systems and summarized their intervention effects [13,14]. Other reviews have focused on specific outcomes, such as depression and anxiety [14], or limited their scope to selected health behaviors or patient populations [15,16]. However, relatively few have systematically quantified the effects of GenAI-based interventions [13,16]. For example, Zhang et

al [17] included 26 quantitative studies and 14 randomized controlled trials (RCTs) in a review of generative and hybrid mental-health chatbots. However, their meta-analysis was limited to negative mental-health outcomes and excluded studies in which the control condition also involved a chatbot.

Consequently, current evidence remains insufficient to determine whether GenAI provides incremental benefits over active or human-delivered interventions [14,17]. Further synthesis is needed to establish how confidently the reported benefits can be interpreted, considering risk of bias, inconsistency, and imprecision [16,18]. It is also necessary to examine whether trials adequately report model versions and configurations, adverse-event monitoring, and clinicians' roles in intervention delivery and safety oversight. These details are essential for judging the reliability, reproducibility, and clinical applicability of GenAI-based interventions [13,18,19]. To address these gaps, we conducted a systematic review and meta-analysis restricted to peer-reviewed RCTs of GenAI-based interventions in behavioral medicine, synthesizing their effects on behavioral, psychological, and health-related outcomes and examining whether observed benefits were associated with therapeutic structure, comparator intensity, human involvement, and other intervention or trial characteristics rather than generative capability itself.

# Methods

## Study Design

This systematic review and meta-analysis was conducted in accordance with Cochrane methodological guidance and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement [20]. The completed PRISMA 2020 checklist is provided in Checklist 1. The review was prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD420251180679).

## Search Strategy and Selection Criteria

Eligible studies were RCTs in which GenAI was used as part of a behavioral medicine intervention. Behavioral medicine was defined as interventions targeting health-related behaviors, psychological processes, or self-management in physical or mental health contexts. Studies were included irrespective of clinical condition, encompassing both clinical populations (eg, individuals with diagnosed mental or physical disorders) and nonclinical populations (eg, community or student samples), provided that the intervention aimed to modify behavioral or psychological outcomes. Eligible interventions included systems using generative models (eg, LLMs or comparable architectures) to deliver interactive, adaptive, or personalized support, including conversational coaching, cognitive-behavioral approaches, psychoeducation, and adherence support. Interventions could be delivered via mobile applications, web-based platforms, or other digital interfaces, and could involve varying levels of human support (eg, clinician oversight or hybrid delivery models). Interventions were included if a generative component played a substantive role in user interaction or content generation. Comparators included waitlist, usual care, nongenerative digital interventions, or active controls such as rule-based chatbots or human-delivered interventions.

We searched 4 databases (PubMed, Embase, Web of Science, and Scopus) from January 1, 2020, to March 31, 2026, for English-language articles. We excluded conference abstracts without full texts, study protocols, nonrandomized studies, and reports lacking sufficient detail on participants, interventions, comparators, or outcomes after attempts to contact the authors. The search strategy was developed iteratively and is presented in Table S1 in Multimedia Appendix 1. We also manually screened the reference lists of relevant studies and reviews. We also searched clinical trial registries (including ClinicalTrials.gov and WHO ICTRP [World Health Organization International Clinical Trials Registry Platform]) and screened preprint servers to identify ongoing or unpublished studies; however, only peer-reviewed RCTs were included in the quantitative synthesis. Records were imported into EndNote (version 21; Clarivate), and duplicates were removed. Two reviewers (YX and MZ) independently screened titles and abstracts and assessed full texts for eligibility. Disagreements were resolved through discussion, with a third reviewer (NJ) consulted when necessary.

## Data Analysis

Two reviewers (YX and TL) independently extracted data using a standardized data collection form in Microsoft Excel. Extracted information included study design, participant characteristics, intervention and comparator details, outcome measures, and follow-up duration. Disagreements were resolved through discussion. When data were incomplete or unclear, corresponding authors were contacted for additional information. Where required, missing SDs were calculated from reported statistics (eg, CIs).

Two reviewers (YX and TL) independently assessed the risk of bias using the revised Cochrane risk-of-bias tool for randomized trials (RoB 2), implemented using the official Cochrane Excel decision-support tool (version ROB2_IRPG_beta_v9; RoB 2 Development Group). Detailed risk-of-bias judgments are presented in Table S3 in Multimedia Appendix 1. The following domains were evaluated: bias arising from the randomization process, bias due to deviations from intended interventions, bias due to missing outcome data, bias in measurement of the outcome, and bias in selection of the reported result. Each domain was judged as low risk of bias, some concerns, or high risk of bias. Disagreements were resolved through discussion, with a third reviewer (NJ) consulted when necessary.

The primary outcomes of interest were symptom-based measures of mental health, including anxiety and depressive symptoms (eg, Generalized Anxiety Disorder-7 [GAD-7] and Patient Health Questionnaire-9 [PHQ-9]). Secondary outcomes included related psychological and behavioral constructs, such as affect (eg, Positive and Negative Affect Schedule [PANAS]) and therapeutic alliance (eg, Working Alliance Inventory [WAI]). When studies were not sufficiently comparable for quantitative pooling, findings were

synthesized narratively. Random-effects meta-analyses were conducted with 95% CIs when studies were considered sufficiently comparable in terms of outcome constructs, despite variation in intervention formats and comparator types. A random-effects model was used to account for expected clinical and methodological heterogeneity across studies, including differences in intervention design (eg, GenAI chatbot and hybrid digital support), comparator conditions (eg, waitlist, usual care, and active control), and study populations. Studies were pooled when they measured conceptually similar outcomes, even if different instruments were used, with effect sizes standardized where appropriate.

Meta-analyses were conducted separately for each outcome construct when at least 3 studies provided sufficiently comparable between-group data. Outcomes reported by fewer than 3 studies or without comparable between-group data were synthesized narratively. Statistical heterogeneity was considered alongside clinical and methodological differences and the magnitude and direction of study-specific effects; when heterogeneity was considerable and individual effects were substantially inconsistent in magnitude or direction, narrative synthesis was preferred over a pooled estimate. Continuous outcomes were summarized as standardized mean differences (SMDs; Hedges *g*), with absolute values of 0.2, 0.5, and 0.8 interpreted as small, moderate, and large effects, respectively. For comparative outcomes, change scores were preferentially used when sufficient summary statistics were available; when change-score data were unavailable, comparable postintervention means and SDs were used. Data sources, derivation methods, assumptions, and detailed numerical calculations are presented in Table S7 in Multimedia Appendix 2 (data sources and effect size derivation). Comparator type was examined through subgroup analyses and narrative exploration of heterogeneity, where feasible. Between-subgroup differences were assessed using a mixed-effects meta-regression with comparator type as a moderator, restricted maximum likelihood (REML) estimation, and Hartung-Knapp-Sidik-Jonkman (HKSJ)–adjusted inference. Funnel plots and the Egger test were planned to assess small-study effects but were not formally undertaken for outcomes with fewer than 10 contributing studies. When available, effect estimates based on intention-to-treat analyses were preferentially extracted.

For the 3-arm Emohaa trial, we extracted outcomes for the full intervention, comprising cognitive behavioral therapy (CBT)–Bot (Conversational AI Group at Tsinghua University) and the generative emotional support–Bot (Conversational AI Group at Tsinghua University) vs the waitlist control; the CBT-Bot-only arm was not included. The shared control group therefore contributed only once to each outcome-specific meta-analysis. For Liu et al [21], only substudy 3 contributed to the quantitative synthesis. The 2 independent patient subtrials in Liu et al [22] were described separately and synthesized narratively. Multiarm trials by Sun et al [23] and Zhang et al [24] were also synthesized narratively, so their shared controls did not contribute repeated observations to the pooled analyses.

All analyses were conducted in R (version 4.3.2; R Foundation for Statistical Computing) using the metagen() function in the *meta* package. Random-effects models were fitted with between-study variance ($\tau^2$) estimated using the REML method, and CIs for the pooled estimates were calculated using the classic normal-approximation (Wald/z-based) method. We also performed a sensitivity analysis using the HKSJ adjustment and calculated a prediction interval to assess the robustness and expected range of the pooled effect. When multiple time points were reported, the primary end point or the closest time point to the end of the intervention was used for pooling.

Two reviewers (YX and TL) independently assessed the certainty of evidence for each outcome using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework and GRADEpro GDT (McMaster University and Evidence Prime) software. Detailed assessments are presented in Table S4 in Multimedia Appendix 1. Evidence from randomized trials was initially rated as high certainty and downgraded across 5 domains: risk of bias, inconsistency, indirectness, imprecision, and publication bias. The overall certainty of the evidence was classified as high, moderate, low, or very low. AI-specific reporting was examined with reference to the CONSORT-AI (Consolidated Standards of Reporting Trials–AI) extension.

### Ethical Considerations

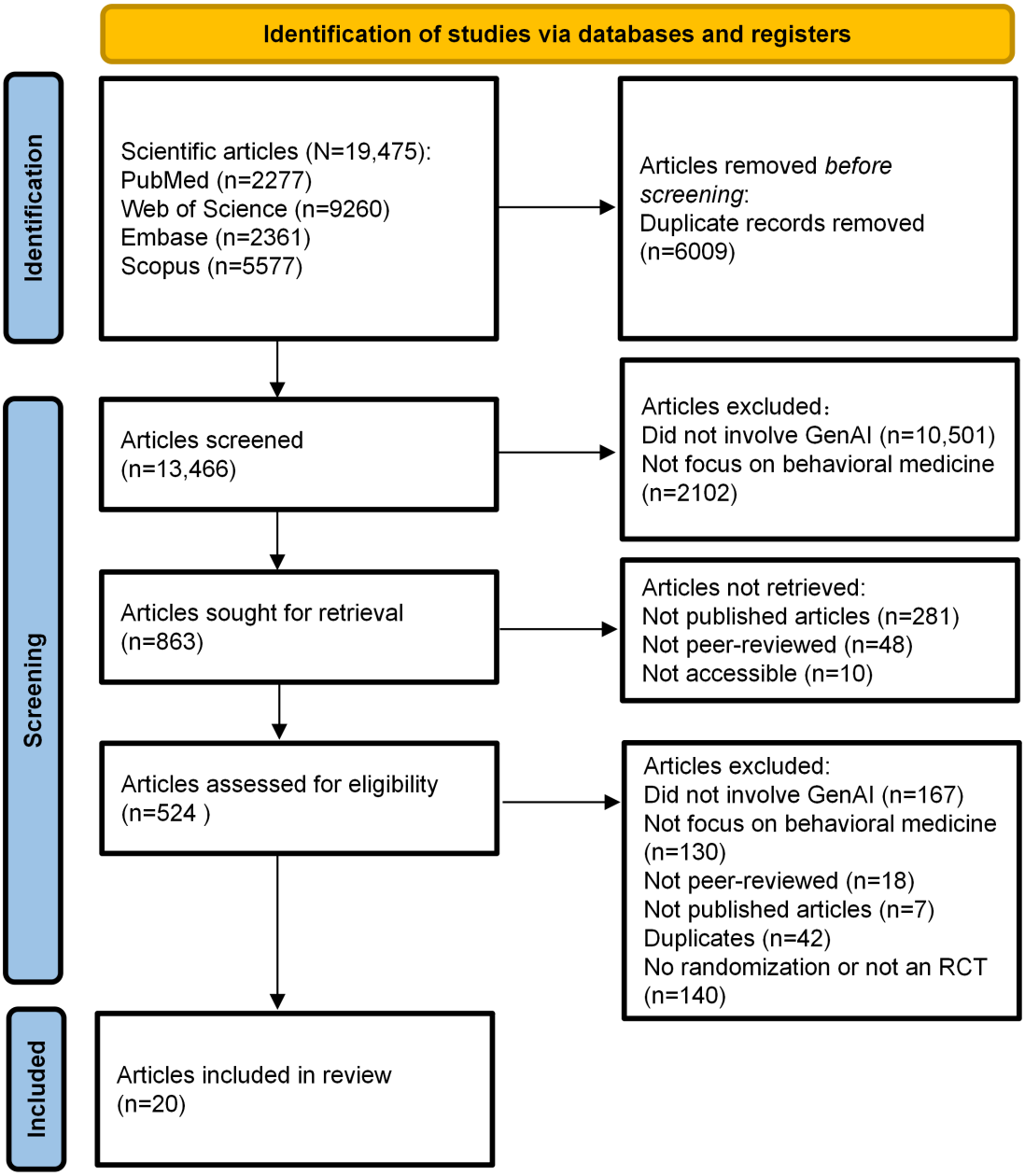
This study is a systematic review and meta-analysis of previously published studies. No new data involving human participants were collected; therefore, ethical approval and informed consent were not required.

## Results

### Study Characteristics

Of the 19,475 records identified, 20 RCTs met the eligibility criteria. The included trials were published between 2023 and 2026 and were conducted mainly in China and the United States. The main characteristics of the included studies are presented in Table S2 in Multimedia Appendix 1. Sample sizes ranged from 40 [25] to 2113 [26]. Across all trials, 8252 participants were included. Most studies enrolled adults, including general adult populations, university students, and participants with depression or anxiety, although several focused on behavioral or clinical contexts such as procrastination and perioperative care [23,27-31]. The study selection process is presented in the PRISMA flow diagram in Figure 1.

**Figure 1.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for the search process. GenAI: generative AI; RCT: randomized controlled trial.




Intervention duration ranged from single-session or one-off consultations to structured programs lasting 6 to 8 weeks, with some perioperative studies extending follow-up to 1 to 6 months [23,24,28,30,32,33]. Interventions were delivered mainly as conversational agents for mental health support, behavioral coaching, or clinical communication and education, often with CBT elements or personalized feedback [21,34,35]. Human involvement ranged from fully automated delivery to hybrid models with clinician oversight, particularly in perioperative settings [23,27,33].

## Risk of Bias

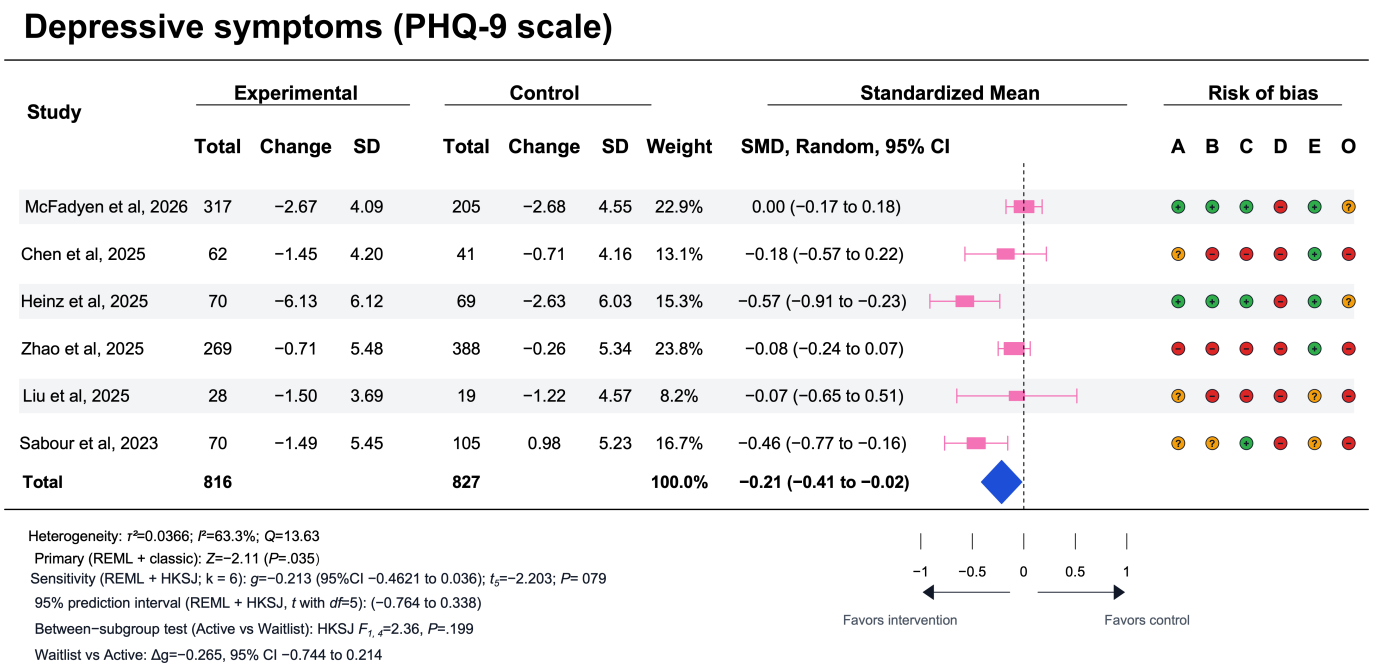
Overall, the included trials were characterized by substantial methodological limitations (Table S3 in Multimedia Appendix 1). Among 20 trials, 14 (70%) were judged to be at high overall risk of bias, 4 (20%) raised some concerns, and 2 (10%) were rated at low risk of bias. The most common sources of bias were measurement of the outcome, which was rated as high risk in 17 (85%) trials, and deviations from intended interventions, which were rated as high risk in 10 (50%) trials. These concerns were reflected in the subsequent GRADE assessments, with certainty of the evidence rated as very low across the pooled outcomes.

## Depressive Symptoms

Six trials reported PHQ-9 outcomes (Figure 2). The pooled random-effects meta-analysis showed a small reduction in depressive symptoms favoring GenAI-based interventions, which was statistically significant under conventional inference (Hedges $g$=–0.21, 95% CI –0.41 to –0.02; $P$=.04; $I^2$=63.3%). In the HKSJ sensitivity analysis, the pooled estimate was unchanged, but the CI crossed 0 (Hedges $g$=–0.21, 95% CI –0.46 to 0.04; $P$=.08). The clearest effect was observed in the Therabot trial (AI and Mental

Health Laboratory at Dartmouth College's Geisel School of Medicine), in which the mean PHQ-9 change at 4 weeks was –6.13 (SD 6.12) in the intervention group and –2.63 (SD 6.03) in the waitlist group [28]. By contrast, studies using more active comparators, such as nurse hotline support or chatbot-based controls, generally showed smaller or nonsignificant between-group differences (Hedges $g$=–0.03, 95% CI –0.18 to 0.12; $I^2$=0%; Figure S1 in Multimedia Appendix 1) [21,32]. However, the quality of evidence was very low for PHQ-9 due to risk of bias, inconsistency, and imprecision.

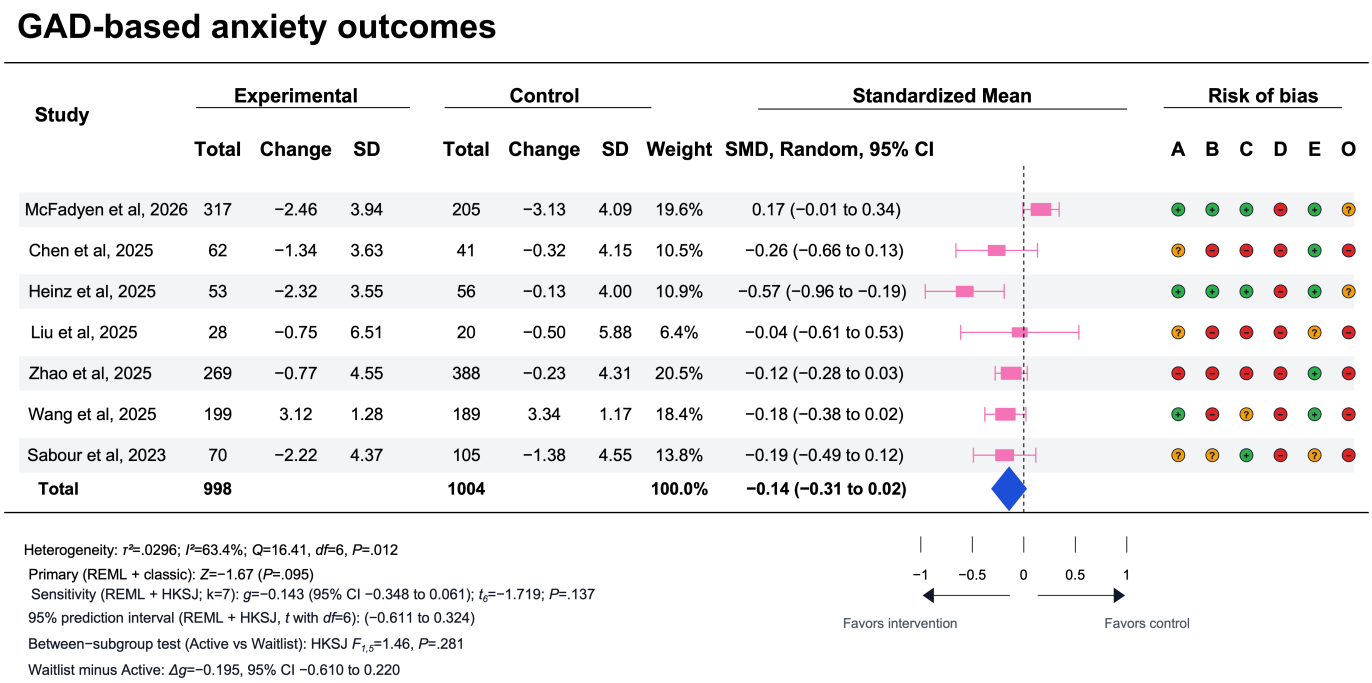
**Figure 2.** Forest plot of pooled standardized mean differences (SMDs; Hedges $g$) in depressive symptoms measured using the Patient Health Questionnaire-9 (PHQ-9) across randomized controlled trials comparing generative AI (GenAI)–based interventions with control conditions. Squares represent study-specific effect estimates, and horizontal lines indicate 95% CIs. Diamonds represent pooled estimates from random-effects models. Risk-of-bias judgments were assessed for each study using the revised Cochrane risk-of-bias tool for randomized trials (RoB 2) tool. The risk-of-bias domains included A: randomization process; B: deviations from intended interventions; C: missing outcome data; D: measurement of the outcome; E: selection of the reported result; and O: overall risk of bias. HKSJ: Hartung-Knapp-Sidik-Jonkman; REML: restricted maximum likelihood [15,21,28,30,35,36].




## Anxiety Symptoms

Twelve trials reported anxiety outcomes using generalized anxiety disorder (GAD)–based or State-Trait Anxiety Inventory (STAI)–based measures. Of the 9 trials reporting GAD-based outcomes, Wang et al [37] lacked sufficient summary statistics to calculate a between-group effect estimate, and Li et al [33] reported GAD-7 values that were inconsistent with the stated scoring range; neither was included in the quantitative synthesis. The remaining 7 trials contributed to the meta-analysis (Figure 3), which showed a small, nonsignificant reduction in anxiety favoring GenAI-based interventions (Hedges $g$=–0.14, 95% CI –0.31 to 0.02; $I^2$=63.4%). Heterogeneity was substantial. For example, the mean Generalized Anxiety Disorder Questionnaire-IV (GAD-Q-IV) change at 4 weeks in the Therabot trial was –2.32 (SD 3.55) in the intervention group and –0.13 (4.00) in the waitlist group [28], whereas the mean GAD-7 change in McFadyen et al [30] was –2.46 (3.94) in the GenAI group and –3.13 (4.09) in the active control group. Similar patterns were observed for GAD-based outcomes, with smaller effects vs active comparators (Hedges $g$=–0.06, 95% CI –0.28 to 0.17; $I^2$=63.6%) and larger effects vs waitlist controls (Hedges $g$=–0.25, 95% CI –0.49 to –0.01; $I^2$=56.3%), as shown in Figure S2 in Multimedia Appendix 1.

**Figure 3.** Forest plot of pooled standardized mean differences (SMDs; Hedges *g*) in anxiety symptoms measured using generalized anxiety disorder (GAD)–based measures (Generalized Anxiety Disorder-7 [GAD-7] and Generalized Anxiety Disorder Questionnaire-IV [GAD-Q-IV]) across randomized controlled trials comparing generative AI (GenAI)–based interventions with control conditions. Squares represent study-specific effect estimates, and horizontal lines indicate 95% CIs. The diamond represents the pooled estimate from the random-effects model. Risk-of-bias judgments were assessed for each study using the revised Cochrane risk-of-bias tool for randomized trials (RoB 2) tool. Risk-of-bias domains included A: randomization process; B: deviations from intended interventions; C: missing outcome data; D: measurement of the outcome; E: selection of the reported result; and O: overall risk of bias. HKSJ: Hartung-Knapp-Sidik-Jonkman; REML: restricted maximum likelihood [15,21,28,30,31,35,36].

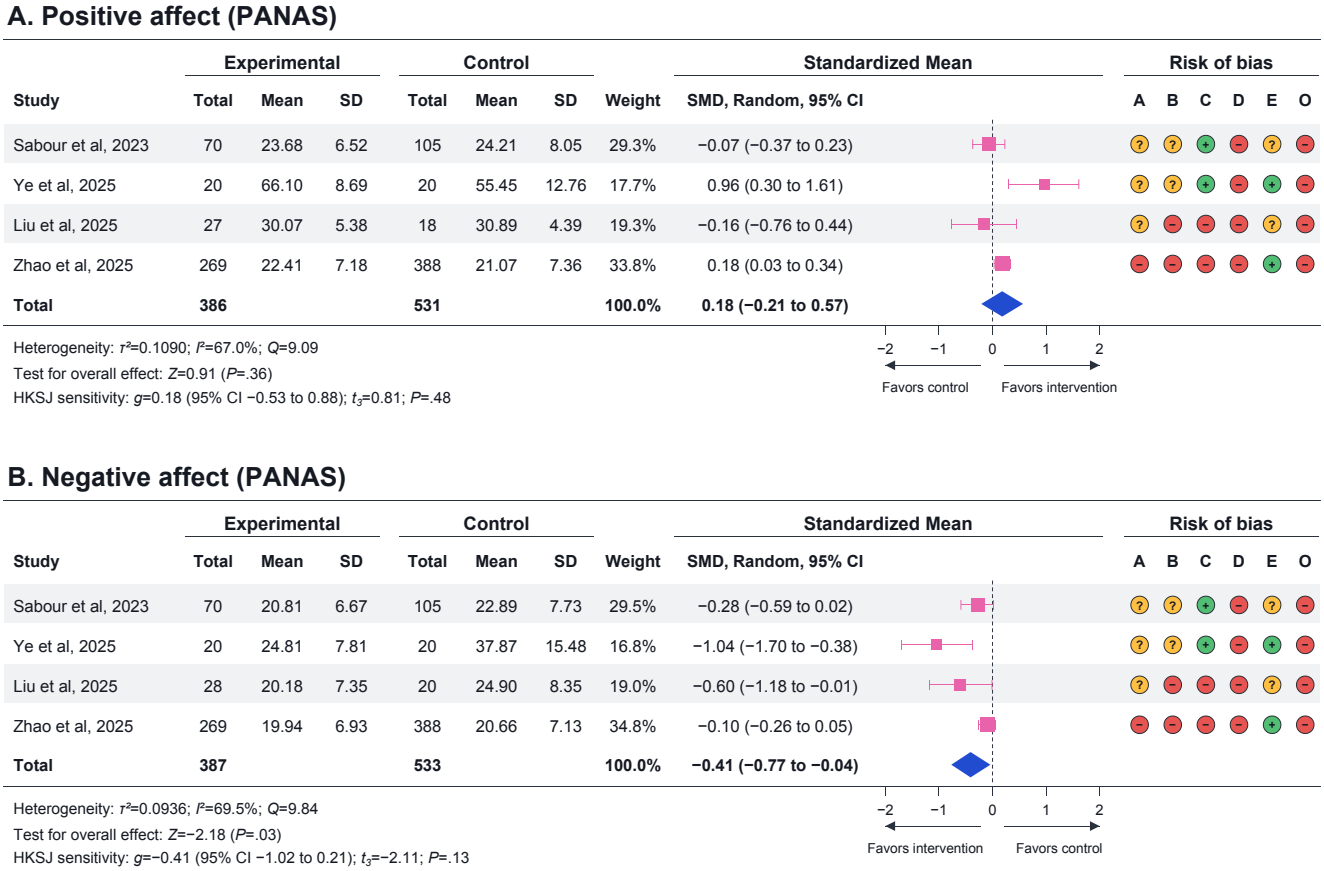



Three trials assessed anxiety using STAI-based measures and were synthesized narratively because of substantial clinical and methodological heterogeneity. Song et al [38] reported a larger calculated reduction in state anxiety with GenAI-supported postdischarge care than with control care (–6.18 vs –0.13). Yahagi et al [39] found a significant group-by-time interaction favoring ChatGPT (Open AI)-assisted education (*P*=.02), although the overall group effect was not significant. In Zhang et al [24], the attending-urologist group had the lowest immediate postintervention anxiety score, followed by the ChatGPT and Ernie Bot (Baidu) groups (median 9, IQR 7-9; median 10, IQR 9-11; and median 12, IQR 10-13, respectively). Overall, the STAI findings did not demonstrate a consistent advantage of GenAI. Certainty was very low for the GAD-based meta-analysis and very low for the narrative STAI evidence.

## Affect-Related Outcomes

Four trials reported PANAS-family outcomes (Figure 4). In random-effects meta-analysis, GenAI-based interventions were associated with a small, nonsignificant improvement in positive affect (Hedges *g*=0.18, 95% CI –0.21 to 0.57; $I^2$=67.0%) and a modest reduction in negative affect (Hedges *g*=–0.41, 95% CI –0.77 to –0.04; $I^2$=69.5%). In HKSJ sensitivity analyses, the pooled estimates were unchanged, but the CIs included 0 for both positive affect (95% CI –0.53 to 0.88; *P*=.48) and negative affect (95% CI –1.02 to 0.21; *P*=.13). In Ye et al [25], the mean postintervention negative affect was 24.81 (SD 7.81) in the intervention group and 37.87 (SD 15.48) in the control group, whereas in Liu et al [21] positive affect was similar between groups, but negative affect remained lower in the intervention group. The quality of evidence was ranked as very low for PANAS due to risk of bias, inconsistency, and imprecision.

**Figure 4.** Forest plots of pooled standardized mean differences (SMDs; Hedges *g*) in affective outcomes measured using the Positive and Negative Affect Schedule (PANAS) across randomized controlled trials comparing generative AI (GenAI)–based interventions with control conditions. Squares represent study-specific effect estimates, and horizontal lines indicate 95% CIs. Diamonds represent pooled estimates from random-effects models. Risk-of-bias judgments were assessed for each study using the revised Cochrane risk-of-bias tool for randomized trials (RoB 2) tool. Risk-of-bias domains included A: randomization process; B: deviations from intended interventions; C: missing outcome data; D: measurement of the outcome; E: selection of the reported result; and O: overall risk of bias. HKSJ: Hartung-Knapp-Sidik-Jonkman [21,25,35,36].




## *Hospital Anxiety and Depression Scale–Based Outcomes*

Three trials reported Hospital Anxiety and Depression Scale (HADS)–based outcomes and were synthesized narratively because of differences in clinical setting, assessment timing, and reporting format. In Akdogan et al [40], the ChatGPT-assisted counseling group had lower median Hospital Anxiety and Depression Scale–Anxiety subscale (HADS-A) and Hospital Anxiety and Depression Scale–Depression subscale (HADS-D) scores before the second chemotherapy cycle than the standard-education group (HADS-A: median 6, range 0-17 vs median 8, range 1-17; *P*=.002; HADS-D: median 6, range 0-20 vs median 9, range 0-19; *P*<.001). In Gan et al [27], mean HADS-A scores were lower with ChatGPT-assisted than surgeon-led informed consent immediately after consent (mean 10.48, SD 3.84 vs mean 12.75, SD 4.12; *P*=.04) and on postoperative day 5 (mean 8.33, SD 3.20 vs mean 10.71, SD 3.83; *P*=.01), whereas HADS-D scores showed no clear between-group differences. In Sun et al [23], mean HADS-A and HADS-D scores at 1 month after discharge were lower in the AI-assisted physician-voice and patient self-voice education groups than in the traditional-education group (HADS-A: mean 5.2, SD 2.5 and mean 4.3, SD 2.2 vs mean 6.5, SD 2.9; HADS-D: mean 4.6, SD 2.3 and mean 3.7, SD 2.0 vs mean 5.8, SD 2.6; both overall 3-group comparisons *P*<.001).

## *Therapeutic Alliance*

Two trials reported therapeutic alliance using WAI-based measures and were summarized narratively. In the comparison with a rule-based conversational system, mean Working Alliance Inventory–Short Revised (WAI-SR) bond scores at the end-of-trial assessment were identical in both groups (mean 3.9, SD 1.1) [34]. In the Therabot trial, the mean overall WAI-SR score at 4 weeks was 3.59 (SD 1.27) among intervention participants, with no corresponding comparator-arm assessment [28]. These findings suggest that participants perceived a therapeutic alliance with GenAI, but they do not establish superiority over active conversational comparators.

## *Medication Adherence*

Two trials reported adherence-related outcomes. In Sun et al [23], mean adherence scores were higher in both GenAI groups than in the control group, at 24.7 (SD 3.5) and 26.6 (SD 2.8) vs 21.6 (SD 4.2). In Wang et al

[31], GenAI-supported communication was associated with more regular medication use, lower opioid use, and higher analgesic satisfaction, although overall analgesic use did not differ.

### Other Nonpooled Clinical Outcomes

Several additional trials assessed outcomes using less frequently used or study-specific measures, including insomnia severity, perioperative anxiety, treatment-related distress, health knowledge, and patient-reported communication or satisfaction [23,24,26,39,40]. Improvements were generally observed in these proximal, intervention-specific outcomes, such as reduced perioperative anxiety and improved treatment-related understanding or communication.

### Engagement

Five studies reported engagement-related outcomes. Compared with less-interactive digital controls, GenAI was generally associated with higher frequency and longer duration of use. McFadyen et al [30] reported open rates 2.4 times higher and engagement duration 3.8 times longer in the GenAI group than in the digital workbook group. When comparators were interactive, differences were smaller; Campellone et al [34] found no clear advantage for GenAI over a rule-based system in total active days, session counts, or conversational exchanges. In the Therabot trial, participants sent a mean of 260 messages over a mean of 24 days and spent a mean total of 6.18 hours with the system, although the waitlist design did not permit between-group comparison [28].

### Empathy

Three studies reported empathy-related outcomes. Compared with a rule-based system, GenAI showed higher empathic listening and reflection accuracy, with 98% (286/292) vs 69% (178/258) [34]. When compared with human-delivered care, findings were mixed. In Liu et al [22], empathy scores were similar between groups in appointment triage and preexamination preparation, but lower in the GenAI-supported group in radiology clinic services. In Zhang et al [24], ChatGPT outperformed Ernie Bot on empathy, whereas physician-led education remained superior for anxiety reduction.

### Follow-Up

Nine studies included longitudinal follow-up, mostly within 1 month. Short-term improvements in depression, anxiety, and related outcomes were observed at 2 to 8 weeks or at 30-day follow-up in several trials [23,28,31,36]. Evidence beyond 3 months was limited and less consistent. In Li et al [33], early improvements attenuated over time, with no clear between-group differences at 6 months.

### Transparency and Governance

Across 20 RCTs, reporting on governance was consistently less developed than reporting on effectiveness (Table S5 in Multimedia Appendix 1). All studies named the model or system used, and 17 of 20 described prompt design or system architecture in substantive detail. Using the CONSORT-AI extension as a benchmark (Table S6 in Multimedia Appendix 1) [41,42], we found that only 25% (n=5) of trials reported inclusion and exclusion criteria for input data, 30% (n=6) described procedures for assessing and handling poor-quality or unavailable input data, and 25% (n=5) stated whether and how the AI intervention and/or its code could be accessed. Safety monitoring or adverse-event reporting was documented in 8 trials, factuality or hallucinations assessment in 10, and privacy or data-governance reporting in 13. Human oversight or review was more frequent (n=14) but usually took the form of clinician review or retrospective auditing rather than embedded real-time governance. Only a minority of interventions reported clinician involvement in model development, typically through expert-curated content (board-certified psychiatrist and clinical psychologist), multidisciplinary clinical input (clinical psychologists and mental health experts), or real-world clinical data (development based on real-world clinical dialogue data and expert knowledge) [22,28,37].

## Discussion

### Principal Findings

To our knowledge, this is the first systematic review and meta-analysis to synthesize randomized evidence on behavioral, psychological, and health-related outcomes of GenAI-based interventions across behavioral medicine and to examine whether observed benefits track therapeutic structure, comparator intensity, and human involvement rather than generative capability alone. The pattern of findings suggests that intervention design, comparator context, and implementation features may shape observed effects more strongly than generative capability alone. In the primary random-effects analysis, GenAI-based interventions were associated with a small reduction in depressive symptoms (PHQ-9, Hedges $g$=–0.21, 95% CI –0.41 to –0.02). However, the HKSJ sensitivity analysis produced a wider CI that included 0 (95% CI –0.46 to 0.04; $P$=.08). Evidence for anxiety outcomes remained inconclusive (GAD-based outcomes, $g$=–0.14, 95% CI –0.31 to 0.02; $I^2$=63%). Negative affect favored GenAI-based interventions ($g$=–0.41, 95% CI –0.77 to –0.04), whereas positive affect did not improve significantly ($g$=0.18, 95% CI –0.21 to 0.57). Point estimates suggested larger effects compared with waitlist controls, but formal between-subgroup tests did not support comparator-related differences for depression ($P$=.20) or anxiety ($P$=.28). Interpretation remains constrained by the small number of studies, substantial heterogeneity, and high overall risk of bias in 14 of 20 (70%) trials. The certainty of evidence ranged from low to very low.

### Patterns Differ by Outcome and Comparator

A consistent pattern across the included studies was that benefits were more apparent for proximal, intervention-specific outcomes than for broader transdiagnostic outcomes. Improvements were most consistently observed in domains such as treatment understanding, communication, perioperative anxiety, health knowledge, and medication adherence,

which align closely with the immediate functions of many GenAI-based systems. By contrast, broader psychological outcomes were less consistently improved. This pattern suggests that current GenAI-based interventions might be better understood as tools that more reliably support proximal processes and short-term coping than broad psychological enhancement. Similar asymmetry has been reported in digital behavioral health interventions, where effects are often clearer for distress reduction and task-specific processes than for improvements in well-being or flourishing [43].

Pooled effect estimates were numerically more favorable for GenAI-based interventions compared with waitlist controls than with active comparators, although formal subgroup tests did not establish differences between comparator types. Similar patterns have been reported in digital mental health research, with larger effects against minimal controls and smaller effects against active comparators [44]. These findings highlight the importance of evaluating the incremental benefit of GenAI over existing digital and clinician-delivered care. Comparisons with credible active controls are particularly informative for determining the clinical value of adding generative capabilities to behavioral interventions.

## Generative Capability and Therapeutic Design

The clearest effects were observed in therapeutically structured systems rather than in lower-intensity conversational or educational applications. In particular, the Therabot trial combined expert fine-tuning, explicit therapeutic logic, repeated interaction, and human safety oversight, and produced the clearest signals for both depression and anxiety [28]. Similar patterns were observed in other interventions, including semigenerative CBT-based chatbots using predefined interaction pathways [29] and conversational agents with graded therapeutic dialogue structures [36]. Generative systems generally showed stronger effects on engagement, empathic listening, and reflection than less interactive controls, but these advantages did not clearly translate into stronger alliance when comparators were already conversational. This distinction is important because engagement, empathy, and alliance are related but not interchangeable constructs. In digital mental health interventions, alliance may arise not only from generative capability, but also from responsiveness, continuity, perceived attentiveness, and the overall interaction framework. These findings suggest that benefits of GenAI are driven less by generative capability alone than by structured therapeutic design and sustained interaction, with effectiveness further shaped by expert tuning and clear clinical targets.

## Scope and Maturity of Current Evidence

Limited reporting of model and system characteristics might have contributed to between-study heterogeneity. Most trials identified the generative model or platform used, but reporting of prompt or system design, safety monitoring, hallucination or factuality assessment, privacy safeguards, data governance, and domain-expert involvement was often incomplete. Future RCTs of GenAI-based behavioral medicine interventions should prospectively follow CONSORT-AI and emerging LLM-specific reporting guidance, with clear documentation of model version, access mode, configuration and prompt details, input-data eligibility and handling, human-AI interaction, safety and error-monitoring procedures, governance safeguards, and access to intervention materials or code. Such reporting would improve comparability and help disentangle model-specific effects from other intervention components.

Heterogeneity was likely influenced by differences in outcome measurement. Most RCTs relied on self-reported symptom scales rather than clinician-confirmed diagnoses or objective behavioral and physiological measures, which might be more sensitive to short-term changes but less reflective of sustained clinical and functional improvement. Broader outcome frameworks that integrate self-report with multimodal, real-time, and wearable-derived measures, such as digital phenotyping, might help to better capture clinically meaningful change.

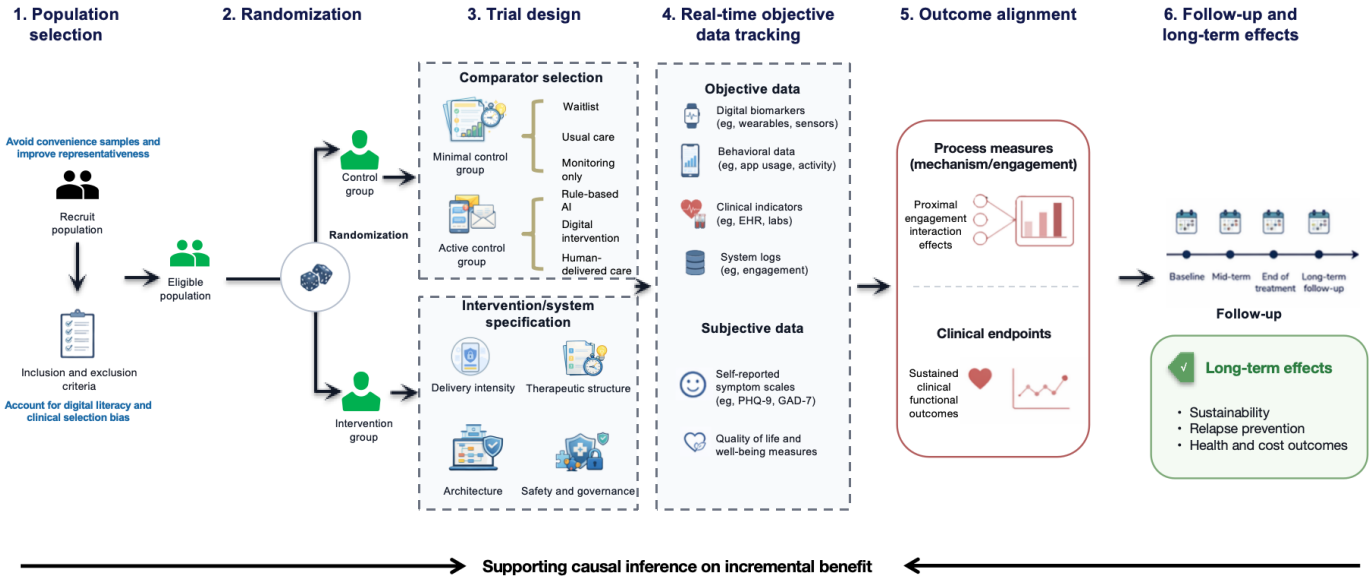
Interpretation of these findings should also take into account the early stage of the evidence base. Most trials assessed outcomes within 1 month, and relatively few extended beyond 3 months. The current literature is therefore better suited to demonstrating short-term feasibility and early signals of benefit rather than durable clinical effectiveness. The limited longer-term evidence suggests that early gains might attenuate over time, but this remains insufficiently studied. It remains unclear whether short-term improvements in proximal outcomes, such as understanding, communication, adherence, or distress reduction, translate into durable clinical benefit.

## Implications for Future Trial Design and Clinical Implementation

A key implication of our findings is that GenAI-based interventions may provide small, outcome-specific benefits and acceptable user engagement but should not be conceptualized as a standalone therapeutic modality, as their impact depends on the structure and intensity of the intervention in which they are embedded. On this basis, we outline a framework to guide future trial design and interpretation (Figure 5). In the current evidence base, generative capability is often introduced alongside differences in therapeutic structure, delivery intensity, and human support, which limits causal interpretation of observed effects. Future trials would be more informative if they specify whether the contrast of interest concerns generative interaction, therapeutic content, personalization, or implementation intensity, and that comparator conditions are selected to match the underlying causal contrast. More detailed reporting of system architecture, prompting strategies, safety mechanisms, and human oversight will also be needed to distinguish model-related effects from broader features of intervention design. Outcome frameworks should similarly be aligned with this logic, so that process measures and clinical end points can be interpreted as part of the same causal pathway rather than as separate classes of findings. Such specification is

also necessary to support the interpretation of incremental benefit. Without such design and reporting refinement, it will remain difficult to determine when GenAI offers meaningful advantages over existing digital interventions.

**Figure 5.** Framework for a randomized controlled trial design to evaluate generative AI (GenAI)–based behavioral interventions. EHR: electronic health record. GAD-7: Generalized Anxiety Disorder-7; PHQ-9: Patient Health Questionnaire-9.




## Strengths and Limitations

This review has some limitations. Variations in study design and comparator choices limited the ability to distinguish intervention effects from nonspecific effects. The number of RCTs for some conditions was limited, resulting in low statistical power and requiring cautious interpretation. Pooling across heterogeneous populations, outcomes, and intervention characteristics might have introduced additional variability and reduced the specificity of conclusions regarding which populations benefit most. Only English-language studies were included, which might have excluded relevant evidence from other regions. Subgroup analyses were limited by the small number of studies within key categories, including comparator type, intervention structure, and outcome domains.

A key strength of this review is its focus on RCTs and its structured analytic approach to an emerging and heterogeneous evidence base, allowing clearer insight into the conditions under which GenAI interventions might be effective. These findings also identify several priorities for the design, evaluation, and reporting of future GenAI trials in behavioral medicine.

## Conclusions

In conclusion, GenAI-based interventions may provide small, outcome-specific benefits and acceptable user engagement in behavioral medicine, but their impact is neither uniform nor intrinsic to the generative capability itself. Rather, effectiveness appears to arise from how these systems are embedded within structured therapeutic designs, sustained interaction frameworks, and appropriate levels of human support, suggesting that GenAI functions primarily as an augmentative component within broader intervention architectures rather than as a standalone therapeutic modality. Future research should prioritize rigorous comparative designs, transparent reporting of model and system characteristics, and evaluation of long-term outcomes and safety within real-world implementation contexts to determine when, how, and for whom these interventions can deliver clinically meaningful benefit.

### Generative AI Use

During the preparation of this work, the authors used ChatGPT-5 (OpenAI) to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

### Funding

This work was supported by the Ministry of Education of the People's Republic of China (PRC; 171-BHZX). The funder had no role in the study design, data collection and analysis, preparation of the manuscript, or the decision to submit for publication.

### Data Availability

All data analyzed in this systematic review and meta-analysis were extracted from published randomized controlled trials and are included in this article and its Multimedia Appendices. The R code used to conduct the meta-analyses is available from the corresponding author upon reasonable request.

## Authors' Contributions

NJ and YX conceived the idea and need for the systematic review and contributed to the study conception and design. YX registered the protocol on PROSPERO (International Prospective Register of Systematic Reviews). YX created the search strategy and performed the literature search. MZ, TL, YH, YY, Y Zhang, and Y Zhu conducted the literature screening and data extraction. YX and TL independently performed quality assessment of the included studies under the supervision of NJ. YX performed the meta-analyses. NJ and YX drafted the manuscript, figures, and tables. All authors critically reviewed the manuscript for important intellectual content. NJ, YCT, and TYW provided the overall supervision and project administration. All authors read and approved the final version of the manuscript. NJ and YCT are joint guarantors and take full responsibility for the integrity and accuracy of the work.

## Conflicts of Interest

JT is the editor in chief of *JMIR Mental Health* at the time of this publication. JT had no involvement in the editorial review and processing of this manuscript.

## Multimedia Appendix 1

Framework for designing and evaluating generative AI–based randomized controlled trials, covering population selection, trial design, outcome assessment, follow-up, and long-term effects.
[DOCX File (Microsoft Word File), 804 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

Data sources and effect size derivation.
[DOCX File (Microsoft Word File), 93 KB-Multimedia Appendix 2]

## Checklist 1

PRISMA checklist.
[DOCX File (Microsoft Word File), 273 KB-Checklist 1]

## References

1. Wilson DK. Behavior matters: the relevance, impact, and reach of behavioral medicine. Ann Behav Med. Feb 2015;49(1):40-48. [doi: 10.1007/s12160-014-9672-1] [Medline: 25559044]
2. Arigo D, Jake-Schoffman DE, Wolin K, Beckjord E, Hekler EB, Pagoto SL. The history and future of digital health in the field of behavioral medicine. J Behav Med. Feb 2019;42(1):67-83. [doi: 10.1007/s10865-018-9966-z] [Medline: 30825090]
3. GBD 2023 Disease and Injury and Risk Factor Collaborators. Burden of 375 diseases and injuries, risk-attributable burden of 88 risk factors, and healthy life expectancy in 204 countries and territories, including 660 subnational locations, 1990-2023: a systematic analysis for the Global Burden of Disease Study 2023. Lancet. Oct 18, 2025;406(10513):1873-1922. [doi: 10.1016/S0140-6736(25)01637-X] [Medline: 41092926]
4. Hesse BW. Riding the wave of digital transformation in behavioral medicine. Ann Behav Med. Dec 1, 2020;54(12):960-967. [doi: 10.1093/abm/kaaa093] [Medline: 33416839]
5. Sezgin E, McKay I. Behavioral health and generative AI: a perspective on future of therapies and patient care. Npj Ment Health Res. Jun 7, 2024;3(1):25. [doi: 10.1038/s44184-024-00067-w] [Medline: 38849499]
6. Romay Barja M, Bilos IB, Bratu EC, et al. The relevance of WHO behavioural health insights. Lancet. Jun 7, 2025;405(10494):2045-2046. [doi: 10.1016/S0140-6736(25)00835-9] [Medline: 40483104]
7. Prestwich A, Kenworthy J, Conner M. Health Behavior Change: Theories, Methods and Interventions. Routledge; 2023. ISBN: 1003302416
8. Torous J, Linardon J, Goldberg SB, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. World Psychiatry. Jun 2025;24(2):156-174. [doi: 10.1002/wps.21299] [Medline: 40371757]
9. Ning Y, Teixayavong S, Shang Y, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. Lancet Digit Health. Nov 2024;6(11):e848-e856. [doi: 10.1016/S2589-7500(24)00143-2] [Medline: 39294061]
10. Torous J, Topol EJ. Assessing generative artificial intelligence for mental health. Lancet. Jun 11, 2025;406(10504). [doi: 10.1016/S0140-6736(25)01237-1] [Medline: 40516569]
11. Mundinger A. Artificial intelligence in breast oncology. Holist Integ Oncol. 2025;4(1):33. [doi: 10.1007/s44178-025-00174-4]

12. Malgaroli M, Schultebraucks K, Myrick KJ, et al. Large language models for the mental health community: framework for translating code to care. Lancet Digit Health. Apr 2025;7(4):e282-e285. [doi: 10.1016/S2589-7500(24)00255-3] [Medline: 39779452]
13. Hua Y, Siddals S, Ma Z, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. World Psychiatry. Oct 2025;24(3):383-394. [doi: 10.1002/wps.21352] [Medline: 40948070]
14. Sohn JS, Ha BG, Park S, et al. Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms. NPJ Digit Med. Mar 25, 2026;9(1):377. [doi: 10.1038/s41746-026-02566-w] [Medline: 41882250]
15. Chen J, Hu RZ, Zhuang YX, et al. Natural language processing chatbot-based interventions for improvement of diet, physical activity, and tobacco smoking behaviors: systematic review. JMIR mHealth uHealth. Jun 11, 2025;13:e66403. [doi: 10.2196/66403] [Medline: 40503914]
16. Feng X, Tian L, Ho GWK, Yorke J, Hui V. The effectiveness of AI chatbots in alleviating mental distress and promoting health behaviors among adolescents and young adults: systematic review and meta-analysis. J Med Internet Res. Nov 26, 2025;27:e79850. [doi: 10.2196/79850] [Medline: 41313175]
17. Zhang Q, Zhang R, Xiong Y, Sui Y, Tong C, Lin FH. Generative AI mental health chatbots as therapeutic tools: systematic review and meta-analysis of their role in reducing mental health issues. J Med Internet Res. Dec 16, 2025;27:e78238. [doi: 10.2196/78238] [Medline: 41401240]
18. Yang H, Chang F, Muroi F, Liu Z, Zhang W, Cai J. Commercial AI-based mental health chatbots as low-intensity adjuncts to psychotherapy: effectiveness, adherence, and safety - a systematic review and meta-analysis. Psychother Psychosom. May 11, 2026:1-24. [doi: 10.1159/000552072] [Medline: 42113705]
19. Hua Y, Na H, Li Z, et al. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med. Apr 30, 2025;8(1):230. [doi: 10.1038/s41746-025-01611-4] [Medline: 40307331]
20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
21. Liu I, Liu F, Xiao Y, Huang Y, Wu S, Ni S. Investigating the key success factors of chatbot-based positive psychology intervention with retrieval- and generative pre-trained transformer (GPT)-based chatbots. Int J Hum-Comput Interact. Jan 2, 2025;41(1):341-352. [doi: 10.1080/10447318.2023.2300015]
22. Liu Z, Orouji F, Yang F, et al. A fine-tuned large language model chatbot for multi-scenario radiology cancer care: randomized controlled trial on interaction optimization, emotional support, and provider burnout reduction. J Transl Med. Feb 14, 2026;24(1):314. [doi: 10.1186/s12967-026-07738-6] [Medline: 41689094]
23. Sun Y, Xu S, Jin H, et al. Effectiveness of AI-assisted patient health education using voice cloning and ChatGPT: prospective randomized controlled trial. J Med Internet Res. Mar 19, 2026;28:e81387. [doi: 10.2196/81387] [Medline: 41855493]
24. Zhang H, Wang X, Luo H, et al. Comparison of preoperative education by artificial intelligence versus traditional physicians in perioperative management of urolithiasis surgery: a prospective single-blind randomized controlled trial conducted in China. Front Med (Lausanne). Jun 25, 2025;12:1543630. [doi: 10.3389/fmed.2025.1543630]
25. 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 1, 2025;43(9):e01349. [doi: 10.1097/CIN.0000000000001349] [Medline: 40704906]
26. Li S, Li Y, Zhou S, et al. A community-codesigned LLM-powered chatbot for primary care: a randomized controlled trial. Nat Health. 2026;1(2):238-250. [doi: 10.1038/s44360-025-00021-w] [Medline: 41659358]
27. Gan W, Ouyang J, She G, et al. ChatGPT's role in alleviating anxiety in total knee arthroplasty consent process: a randomized controlled trial pilot study. Int J Surg. Mar 1, 2025;111(3):2546-2557. [doi: 10.1097/JS9.0000000000002223] [Medline: 39903546]
28. Heinz MV, Mackin DM, Trudeau BM, 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]
29. Lee S, Jeong J, Kim M, Lee S, Kim SP, Jung D. Development of a mobile intervention for procrastination augmented with a semigenerative chatbot for university students: pilot randomized controlled trial. JMIR mHealth uHealth. Apr 10, 2025;13:e53133. [doi: 10.2196/53133] [Medline: 40208664]
30. McFadyen J, Habicht J, Dina LM, Harper R, Hauser TU, Rollwage M. Increasing engagement with cognitive-behavioral therapy (CBT) using generative AI: a randomized controlled trial (RCT). Commun Med (Lond). Jan 15, 2026;6(1):129. [doi: 10.1038/s43856-025-01321-8] [Medline: 41540194]
31. Wang L, Peng Y, Zhang G, et al. Application of narrative and AI-assisted follow-up after voluntary medical male circumcision: multicenter, double-blind, prospective, randomized controlled trial. J Med Internet Res. Nov 17, 2025;27:e68573. [doi: 10.2196/68573] [Medline: 41248462]

32. Chen C, Lam KT, Yip KM, et al. Comparison of an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: pilot randomized controlled trial. JMIR Hum Factors. Mar 6, 2025;12:e65785. [doi: 10.2196/65785] [Medline: 40048637]
33. Li J, Zhang Y, Zhang Z, et al. A randomized controlled trial of a WeChat-based artificial intelligence agent for postoperative care in orthopedic patients. NPJ Digit Med. Jan 17, 2026;9(1):105. [doi: 10.1038/s41746-025-02269-8]
34. Campellone TR, Flom M, Montgomery RM, et al. Safety and user experience of a generative artificial intelligence digital mental health intervention: exploratory randomized controlled trial. J Med Internet Res. May 23, 2025;27:e67365. [doi: 10.2196/67365] [Medline: 40408143]
35. Sabour S, Zhang W, Xiao X, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health. 2023;5:1133987. [doi: 10.3389/fdgth.2023.1133987] [Medline: 37214342]
36. Zhao Y, Qian W, Chen 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. Oct 2025;17(5):e70067. [doi: 10.1111/aphw.70067] [Medline: 40910958]
37. Wang Y, Li X, Zhang Q, Yeung D, Wu Y. Effect of a cognitive behavioral therapy-based AI chatbot on depression and loneliness in Chinese university students: randomized controlled trial with financial stress moderation. JMIR mHealth uHealth. Aug 29, 2025;13:e63806. [doi: 10.2196/63806] [Medline: 40882177]
38. Song Y, Yan T, Jia F, Chen L, Li H. Developing generative AI for value co-creation: an intervention-based randomized field experiment in a healthcare context. J Oper Manag. Mar 2026;72(2):197-232. [doi: 10.1002/joom.70012]
39. Yahagi M, Hiruta R, Miyauchi C, Tanaka S, Taguchi A, Yaguchi Y. Comparison of conventional anesthesia nurse education and an artificial intelligence chatbot (ChatGPT) intervention on preoperative anxiety: a randomized controlled trial. J Perianesth Nurs. Oct 2024;39(5):767-771. [doi: 10.1016/j.jopan.2023.12.005] [Medline: 38520470]
40. Akdogan O, Uyar GC, Yesilbas E, et al. Effect of a ChatGPT-based digital counseling intervention on anxiety and depression in patients with cancer: a prospective, randomized trial. Eur J Cancer. May 15, 2025;221:115408. [doi: 10.1016/j.ejca.2025.115408] [Medline: 40215593]
41. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. Sep 2020;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
42. Feuerriegel S, Barrie C, Crockett MJ, et al. A reporting checklist for large language models in behavioural science. Nat Hum Behav. Jul 2026;10(7):1182-1186. [doi: 10.1038/s41562-026-02492-7] [Medline: 42265331]
43. Andersson G, Titov N, Dear BF, Rozental A, Carlbring P. Internet-delivered psychological treatments: from innovation to implementation. World Psychiatry. Feb 2019;18(1):20-28. [doi: 10.1002/wps.20610] [Medline: 30600624]
44. Karyotaki E, Efthimiou O, Miguel C, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry. Apr 1, 2021;78(4):361-371. [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]


## Abbreviations

**CBT:** cognitive behavioral therapy
**CONSORT-AI:** Consolidated Standards of Reporting Trials–AI
**GAD:** generalized anxiety disorder
**GAD-7:** Generalized Anxiety Disorder-7
**GAD-Q-IV:** Generalized Anxiety Disorder Questionnaire-IV
**GDT:** Guideline Development Tool
**GenAI:** generative AI
**GRADE:** Grading of Recommendations Assessment, Development, and Evaluation
**HADS:** Hospital Anxiety and Depression Scale
**HADS-A:** Hospital Anxiety and Depression Scale–Anxiety subscale
**HADS-D:** Hospital Anxiety and Depression Scale–Depression subscale
**HKSJ:** Hartung-Knapp-Sidik-Jonkman
**LLM:** large language model
**PANAS:** Positive and Negative Affect Schedule
**PHQ-9:** Patient Health Questionnaire-9
**PRISMA:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses
**PROSPERO:** International Prospective Register of Systematic Reviews
**RCT:** randomized controlled trial
**REML:** restricted maximum likelihood
**RoB 2:** revised Cochrane risk-of-bias tool for randomized trials
**SMD:** standardized mean difference

**STAI:** State-Trait Anxiety Inventory
**WAI:** Working Alliance Inventory
**WAI-SR:** Working Alliance Inventory–Short Revised
**WHO ICTRP:** World Health Organization International Clinical Trials Registry Platform

*Edited by Matthew Balcarras; peer-reviewed by Jia Xue, Xiang Qi; submitted 05.Aug.2026; initial comments to author 28.Aug.2026; final revised version received 14.Sep.2026; accepted 02.Oct.2026; published 27.Nov.2026*

*Please cite as:*
*Xiao Y, Zhu M, Liu T, He Y, Ye Y, Zhang Y, Zhu Y, Xu DR, Lou VWQ, Yip WF, Sheng B, Torous J, Wong TY, Tham YC, Jiang N*
*Effectiveness of Generative AI–Based Interventions in Behavioral Medicine: Systematic Review and Meta-Analysis of Randomized Controlled Trials*
*J Med Internet Res 2026;28:e108789*
*URL: https://www.jmir.org/2026/1/e108789*
*doi: 10.2196/108789*

© Yexuan Xiao, Mengya Zhu, Tiansu Liu, Yilin He, Yuchen Ye, Yuhe Zhang, Yingchao Zhu, Dong Roman Xu, Vivian W Q Lou, Wan Fen Yip, Bin Sheng, John Torous, Tien Yin Wong, Yih Chung Tham, Nan Jiang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 27.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.