Vol. 7, No. 1, 2026, pp. 1555-1571
DOI: https://doi.org/10.29210/993770

Contents lists available at Elora Center

Lentera Negeri

Journal homepage: https://lentera.eloracenter.org/lentera

Conversational AI for mental health: a systematic review of effectiveness and design

Anisa Sholiha Mia*), Rudyanto Rudyanto
Universitas Negeri Padang, Indonesia

Article Info

Article history:
Received Sept 04th, 2026
Revised Sept 06th, 2026
Accepted Sept 07th, 2026

Keyword:
Large language models,
Conversational agents,
Mental health,
Digital interventions,
Psychological well-being

ABSTRACT

This systematic review synthesised evidence on the applications, effectiveness, engagement, acceptability, and responsible design of large language model (LLM)- and AI-based conversational agents for mental health and psychological well-being. Following the PRISMA 2020 framework, Scopus, PubMed, and ScienceDirect were searched for English-language publications from 2023 to 2026. The search identified 388 records; 16 duplicates were removed, 372 records were screened, 56 full texts were assessed, and 21 eligible publications were retained. Because the corpus combined randomized and quasi-experimental studies with single-arm, qualitative, developmental, protocol, conceptual, and review publications, findings were synthesised thematically rather than pooled in a meta-analysis. Four themes emerged: effectiveness and outcomes; design and development; assessment and evaluation; and user perception, acceptability, and ethics. Controlled studies reported promising short-term improvements in depression, anxiety, stress, and well-being, while feasibility and qualitative studies indicated generally favourable acceptability but also concerns about engagement decline, privacy, emotional dependence, bias, transparency, and crisis management. The evidence supports conversational AI primarily as a supervised adjunct to mental-health care rather than an autonomous clinical decision-maker. Major limitations were heterogeneous designs and outcomes, limited long-term follow-up, geographic concentration, and incomplete reporting of screening reliability and search dates in the original review record. Future research should use adequately powered, multi-site trials, standardised outcomes, explicit safety benchmarks, and transparent human-oversight procedures.

Corresponding Author:
Anisa Sholiha Mia,
anisasholihamia@unp.ac.id

Introduction

Mental ill-health is a major public health concern and remains undertreated in many settings because of structural barriers including workforce shortages, uneven service distribution, cost, and stigma (Sinha et al., 2026; Ganesh & Venkateswaramurthy, 2025). These constraints have increased interest in psychologically supportive interventions that are accessible, scalable, and potentially less costly than conventional services. Digital mental-health interventions are particularly relevant to this challenge because they can extend support beyond conventional service hours and locations.

Within digital mental health, conversational technologies have received sustained attention because they can provide interactive psychoeducation, coaching, assessment support, and behavioural guidance. Earlier systems were predominantly rule-based, whereas more recent platforms increasingly use generative AI and LLMs to produce open-ended, context-sensitive responses (Kenny & Parsons, 2024; Nadeem et al., 2026).

This technological shift creates new opportunities for personalised support but also raises questions about reliability, safety, transparency, and clinical integration.

Recent studies have examined conversational agents for stress management, psychological distress, lifestyle support, motivational interviewing, and clinical-reasoning education (Chakraborty et al., 2026; Comulada et al., 2026; Sharma et al., 2026; Shenoi et al., 2026). Intervention studies suggest potential benefits, but the evidence remains heterogeneous in design, target population, agent architecture, outcome measures, and follow-up duration.

The technology is also becoming more multimodal and adaptive. Current systems may incorporate voice, facial, or behavioural signals, while prompt engineering, fine-tuning, explainability, and user-centred design are increasingly used to align agent behaviour with therapeutic principles and safety requirements (Md Zuki et al., 2026; Terry, 2026; Akazua et al., 2026). These developments make it important to distinguish technological capability from clinical effectiveness.

Despite rapid growth, the evidence base is still difficult to interpret because many publications are pilot, feasibility, developmental, conceptual, or single-arm studies. Differences in populations, intervention intensity, comparator conditions, and outcomes limit direct quantitative comparison and make it difficult to identify which design features reliably improve mental-health outcomes.

Methodologically, evaluation standards have not developed at the same pace as the technology. Trust, transparency, bias, privacy, crisis response, and implementation are discussed across the literature, but these domains are assessed using different frameworks and metrics (Maher et al., 2026; De Gagne & Gris, 2026). Consequently, a systematic synthesis is needed that considers effectiveness together with design, user experience, and responsible deployment.

These gaps motivate a systematic synthesis that maps where conversational AI is being used, what outcomes have been reported, and which design and governance factors are repeate-ly identified as important for responsible deployment. The review therefore treats effectiveness, acceptability, and safety as related but distinct dimensions of evidence.

The first objective is to map the contexts, populations, and delivery formats in which LLM- and AI-based conversational agents have been applied to mental health and psychological well-being.

The second objective is to synthesise evidence on effectiveness, engagement, and acceptability, distinguishing controlled intervention evidence from single-arm, qualitative, developmental, and review evidence.

The third objective is to identify design, ethical, and implementation factors-including trust, transparency, customisation, explainability, crisis management, equity, and human oversight-that shape responsible deployment.

RQ1: In what mental-health and psychological well-being contexts, populations, and delivery formats are LLM- and AI-based conversational agents being applied?

RQ2: What effectiveness, engagement, and acceptability do conversational-agent interventions demonstrate for mental-health and psychological well-being outcomes?

RQ3: Which design, ethical, and implementation factors shape the responsible deployment of conversational AI in mental health, and what research gaps remain?

Method

Research Design and Framework

We conducted a PRISMA 2020-guided systematic review with thematic narrative synthesis. The design was selected because the literature combines heterogeneous intervention, observational, qualitative, developmental, protocol, conceptual, and review publications. The unit of synthesis was therefore the eligible publication rather than a homogeneous set of randomized trials. PRISMA 2020 guided identification, screening, eligibility assessment, and reporting (Page et al., 2021). No quantitative meta-analysis was planned because the corpus differed substantially in populations, technologies, comparators, outcomes, and reporting formats.

Search Strategy

We composed a three-block Boolean strategy covering conversational-AI technologies, mental-health outcomes, and intervention/evaluation concepts. Searches were restricted to title, abstract, and keyword fields where supported. The master conceptual query was:

> *("large language model*" OR "LLM" OR "generative AI" OR "conversational agent*" OR "chatbot*" OR "conversational AI" OR "ChatGPT") AND ("mental health" OR "psychological well-being" OR "depression" OR "anxiety" OR "stress" OR "psychotherap*" OR "counsel*") AND ("intervention" OR "support" OR "coaching" OR "assessment" OR "effectiveness")*

Truncation (*) captured plural and derivational forms. The three conceptual blocks were adapted to database-specific syntax while preserving the same retrieval logic. Publication year was limited to 2023–2026 and language to English. The exact calendar day(s) of the 2026 search were not retained in the original search record; this is now stated explicitly as a reporting limitation rather than reconstructed.

Database and Information Sources

Three databases were searched: Scopus, PubMed, and ScienceDirect. Scopus was selected for multidisciplinary coverage, PubMed for biomedical and clinical literature, and ScienceDirect for additional health and behavioural-science coverage. The searches were run during 2026 for publications dated 2023–2026. The database yields were Scopus (n = 139), PubMed (n = 149), and ScienceDirect (n = 100), giving 388 retrieved records before deduplication.

Eligibility Criteria

Eligibility was predefined using the PICOS framework and operationalised through the inclusion and exclusion criteria in Table 2. Because this review intentionally mapped an emerging field, eligible publication types included empirical intervention studies, feasibility and development studies, qualitative studies, protocols, conceptual/framework papers, and relevant reviews. A sample-size calculation was not applicable because inclusion was determined by all records meeting the predefined criteria rather than by prospective sampling or saturation.

Table 1. PICOS Framework Guiding Study Eligibility

Element	Specification
Population	Individuals engaging with digital mental-health support, including general adults, university and school students, athletes, perinatal women, and workplace or clinical populations.
Intervention	LLM- or AI-based conversational agents, chatbots, virtual coaches, or generative-AI dialogue systems delivering mental-health or psychological well-being support, assessment, or coaching.
Comparator	Waitlist or usual-care controls, alternative digital tools, pre–post comparison, or no comparator (single-arm, qualitative, developmental, and review designs).
Outcomes	Mental-health symptoms (depression, anxiety, stress), psychological well-being, engagement and acceptability, and diagnostic or assessment performance.
Study design	Peer-reviewed empirical studies (RCT, quasi-experimental, cohort, mixed-methods, qualitative), development and feasibility studies, protocols, and reviews.

Note. PICOS = Population, Intervention, Comparator, Outcomes, Study design.

Table 2. Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Language	English only	Non-English publications
Document type	Peer-reviewed article or review	Conference paper, conference review, editorial, book chapter, note, erratum
Publication period	2023–2026	Published before 2023

Criterion	Inclusion	Exclusion
Subject focus	Conversational agent central to a mental-health or psychological well-being application	AI applied outside a conversational-agent modality (e.g., wearable-only, sentiment analysis only, social gaming)
Population/outcome	Mental health or psychological well-being as a primary target or outcome	Primary target a physical disease (e.g., chronic pain, oncology, exercise physiology) without a psychological outcome
Accessibility	Full text available	Full text unavailable or abstract only
Relevance	Directly addresses application, effectiveness, or design of conversational AI	Tangential mention only; information-seeking behaviour or branding scope

Note. Criteria were applied sequentially during title/abstract screening and full-text assessment.

Study Selection Process

Study selection proceeded in three stages: deduplication; title/abstract screening; and full-text eligibility assessment. The original screening record did not preserve paired reviewer decisions or a formally calculated Cohen's kappa statistic. Accordingly, no post-hoc coefficient is reported, because doing so without the underlying independent decisions would be methodologically invalid. This limitation is acknowledged explicitly, and future updates should use at least two independent screeners with a documented agreement statistic and adjudication procedure.

Quality Assessment - FICO Framework

Methodological adequacy was appraised using the structured FICO framework used in the original review. FICO evaluates Focus (clarity of aims/questions), Information (methods, sample, and procedures), Context (setting, population, and technology), and Outcome (measures and reported findings). Because the corpus included protocols, conceptual papers, reviews, and heterogeneous empirical designs, FICO was used as an evidence-interpretability appraisal framework rather than as a universal validated risk-of-bias instrument. Appraisal informed the confidence and emphasis given to findings; no publication was excluded solely because of a lower appraisal profile.

Data Extraction Procedure

A standardised extraction template captured authorship, year, country/setting, publication type and design, population and sample, conversational-agent technology, intervention characteristics, outcome measures, and principal findings. Extraction was performed directly from the primary publications and checked against the source records used for synthesis.

Network and Bibliometric Analysis Methodology

Descriptive bibliometric mapping complemented the qualitative synthesis. The temporal distribution of publications, the geographic distribution by first-author country, and the distribution of study designs and themes were tabulated and visualised. These descriptive analyses characterised the maturity, reach, and methodological composition of the field rather than serving as a formal co-citation or co-authorship network analysis.

Data Analysis and Synthesis

Given the heterogeneity of designs, populations, technologies, comparators, and outcomes, a thematic narrative synthesis was conducted rather than a meta-analysis. Findings were coded inductively and grouped into higher-order themes aligned with the three research questions. Quantitative effect sizes reported by individual studies were retained descriptively when available, but no pooled effect, confidence interval, or heterogeneity statistic (I^2) was calculated because the studies did not provide sufficiently comparable estimates and measurement structures for a defensible common effect model.

Reporting and Documentation

The review was reported with reference to the PRISMA 2020 checklist, and record flow is presented in Figure 1 (Page et al., 2021). Numerical counts were reconciled across the abstract, methods, results, and PRISMA diagram. The review was not prospectively registered in PROSPERO, and the exact search-day metadata and paired screening decisions were not retained; both issues are reported as limitations.

Results and Discussions

Study Selection Results

The database searches returned 388 records: Scopus (n = 139), PubMed (n = 149), and ScienceDirect (n = 100). After removal of 16 duplicates, 372 records entered title/abstract screening and 316 were excluded as clearly outside scope. Fifty-six full texts were assessed; 35 were excluded because they were not conversational-agent studies (n = 14), did not focus on mental health or psychological well-being (n = 12), or were ineligible publication types/inaccessible (n = 9). Thus, 21 eligible publications were included in the thematic synthesis. The PRISMA flow in Figure 1 now reflects these reconciled counts.

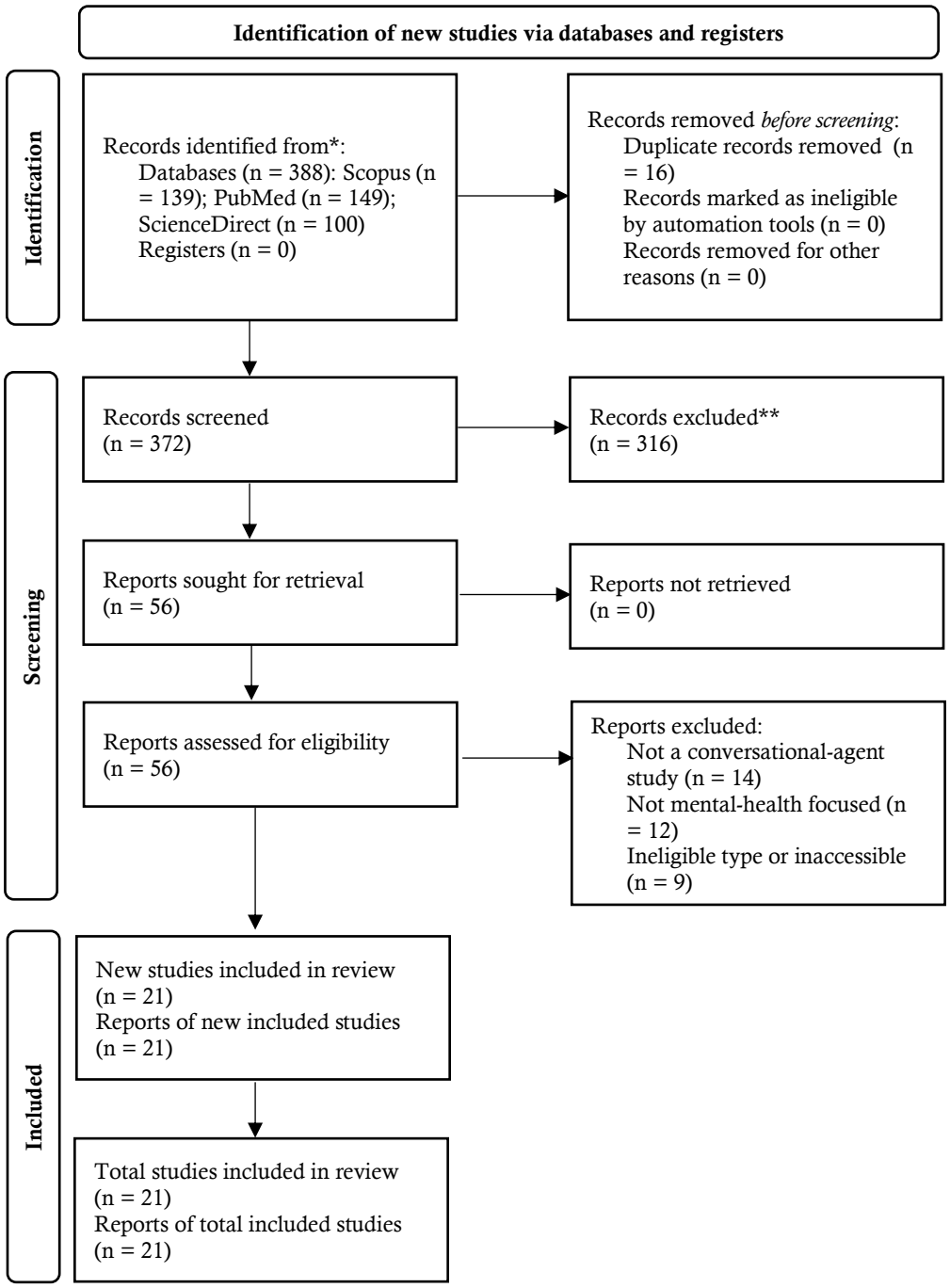

Figure 1. PRISMA 2020 Flow Diagram of the Study Selection Process

Descriptive Characteristics

The 21 eligible publications cover 2023–2026 and represent 11 countries in the study table. Switzerland and the United States contributed four publications each, Australia three, Singapore and Italy two each, and Canada, China, Türkiye, Brazil, the United Kingdom, and India one each. The corpus therefore has a marked concentration in high-income settings. Publication types included randomized and quasi-experimental studies, single-arm and observational evaluations, feasibility/development studies, qualitative work, protocols, conceptual/framework papers, and reviews.

Table 3. Summary of Included Studies

Title	Author(s)	Year	Country	Design
Development of "LvL UP 1.0": a smartphone-based, conversational agent-delivered holistic lifestyle intervention for the prevention of non-communicable diseases and common mental disorders	Castro et al.	2023	Singapore	Intervention development (MOST/MRC framework)
Providing Self-Led Mental Health Support Through an Artificial Intelligence–Powered Chat Bot (Leora) to Meet the Demand of Mental Health Care	van der Schyff et al.	2023	Australia	Conceptual / viewpoint
A Digital Lifestyle Program for Psychological Distress, Wellbeing and Return-to-Work: A Proof-of-Concept Study	Brinsley et al.	2023	Australia	Retrospective cohort (pre–post)
Can digital health researchers make a difference during the pandemic? Results of the single-arm, chatbot-led Elena+: Care for COVID-19 interventional study	Ollier et al.	2023	Switzerland	Single-arm interventional study
A Chatbot-Delivered Stress Management Coaching for Students (MISHA App): Pilot Randomized Controlled Trial	Ulrich et al.	2024a	Switzerland	Pilot randomized controlled trial
Virtual Standardized LLM-AI Patients for Clinical Practice	Kenny & Parsons	2024	United States	Conceptual / technical review
A Brief Review of the Efficacy in Artificial Intelligence and Chatbot-Generated Personalized Fitness Regimens	Bays et al.	2024	United States	Narrative review
Effectiveness of the Minder Mobile Mental Health and Substance Use Intervention for University Students: Randomized Controlled Trial	Vereschagin et al.	2024	Canada	Randomized controlled trial
Development and Evaluation of a Smartphone-Based Chatbot Coach to Facilitate a Balanced Lifestyle in Individuals With Headaches (BalanceUP App): Randomized Controlled Trial	Ulrich et al.	2024b	Switzerland	Randomized controlled trial
Describing the Framework for AI Tool Assessment in Mental Health and Applying It to a Generative AI Obsessive-Compulsive Disorder Platform: Tutorial	Golden & Aboujaoude	2024	United States	Framework tutorial / review
VITA: A Multi-Modal LLM-Based System for Longitudinal, Autonomous and Adaptive Robotic Mental Well-Being Coaching	Spitale et al.	2025	United Kingdom	System development + pilot + real-world study
Artificial Intelligence (AI) Generated Health Counseling For Mental Illness Patients	Ganesh & Venkateswaramurthy	2025	India	Narrative review
Prevention of postpartum depression via a digital acceptance and commitment therapy-based intervention: protocol for a pilot usability study	Rizzi et al.	2025	Italy	Pilot usability protocol
Pilot Quasi-Experimental Research on the Effectiveness of the Woebot AI Chatbot for	Qi	2025	China	Quasi-experimental (pilot)

Title	Author(s)	Year	Country	Design
Reducing Mild Depression Symptoms among Athletes				
"Shaping ChatGPT into my Digital Therapist": A thematic analysis of social media discourse on using generative artificial intelligence for mental health	Luo et al.	2025	United States	Qualitative thematic analysis
Maintained effectiveness of a smartphone-based chatbot coach (BalanceUP) for mental well-being in headache sufferers: preliminary findings from a single-arm 6-month follow-up study	Ulrich et al.	2025	Switzerland	Single-arm 6-month follow-up
The Potential of AI Chatbots as Diagnostic Tools in Mental Health: Evaluating Exercise Dependence Symptoms	Saraç et al.	2025	Türkiye	Expert-rated evaluation (case vignette)
World Health Organization Evidence-Based Self-Help Plus Intervention for Stress Management via Chatbot: Protocol for Adaptation to a Tech-Enabled Model	Fietta et al.	2025	Italy	Adaptation protocol
Cadê o Kauê? Co-design and acceptability testing of a chat-story aimed at enhancing youth participation in the promotion of mental health in Brazil	Pavarini et al.	2025	Brazil	Co-design & acceptability (mixed methods)
Feasibility of the LvL UP digital lifestyle coaching intervention designed to prevent non-communicable diseases and common mental disorders	Mair et al.	2026	Singapore	Feasibility (mixed methods, in-the-wild)
Explainable Conversational Agents for Mobile Health Coaching Systems: Trust Factors, Progress and Opportunities	Akazua et al.	2026	Australia	Scoping review (PRISMA-ScR)

Note. Full titles are reported without abbreviation. Country denotes the first-author country or study setting.

Table 4. Classification of Included Studies by Theme, Technology, and Key Findings

Author(s), Year	Theme/Focus	Technology/Intervention	Key Findings
Castro et al. (2023)	Design & development	Conversational agent (chatbot); CBT-based holistic lifestyle	Developed a scalable chatbot-delivered intervention structured around three pillars (Move More, Eat Well, Stress Less) targeting physical and mental well-being for prevention of NCDs and common mental disorders.
van der Schyff et al. (2023)	Perception, acceptability & ethics	AI-powered chatbot (Leora); self-led support	Argued that AI chatbots can deliver accessible, discreet, round-the-clock self-led support for minimal-to-mild anxiety and depression, while foregrounding trust, transparency, bias and equity challenges.
Brinsley et al. (2023)	Effectiveness & outcomes	Chatbot-led virtual health coach + telehealth	A 6-week chatbot-led lifestyle program produced significant improvements in psychological distress, depression, anxiety, well-being and return-to-work confidence (medium-to-large effects) among workers' compensation claimants.
Ollier et al. (2023)	Effectiveness & outcomes	Chatbot (Elena+); pandemic lifestyle coaching	A chatbot delivering coaching across seven health areas showed strong public demand and significant reductions in depression over 42 days, with behavioral activation mediating outcomes.

Author(s), Year	Theme/Focus	Technology/Intervention	Key Findings
Ulrich et al. (2024a)	Effectiveness & outcomes	Conversational agent (MISHA); stress-management coaching	In a pilot RCT (N=140), the MISHA chatbot improved perceived stress, depression and somatic symptoms with medium effect sizes relative to a waitlist control among students.
Kenny & Parsons (2024)	Assessment, diagnosis & frameworks	LLM-based virtual standardized patients (NLP/NLU/NLG)	Positioned LLMs as an advance for building believable virtual standardized patients to train therapists, while noting persistent challenges in embodiment, realism and safe integration.
Bays et al. (2024)	Perception, acceptability & ethics	AI chatbots for exercise/fitness prescription	Reviewed evidence that AI chatbots can generate personalized, low-cost fitness guidance supporting physical and mental well-being, but concluded that human–AI collaboration and further validation are needed.
Vereschagin et al. (2024)	Effectiveness & outcomes	Chatbot-delivered content + peer coaching (Minder)	In an RCT (N=1489), the Minder app-delivering evidence-based content via an automated chatbot plus peer coaching-significantly improved anxiety, depression and alcohol-use outcomes in university students.
Ulrich et al. (2024b)	Effectiveness & outcomes	Conversational agent (BalanceUP); behavior-change coaching	The BalanceUP conversational-agent coach improved mental well-being (anxiety/depression composite) versus waitlist control in adults with frequent headaches, demonstrating feasibility of CA-delivered behavior change.
Golden & Aboujaoude (2024)	Assessment, diagnosis & frameworks	FAITA-Mental Health evaluation framework; generative-AI OCD tool	Introduced and applied the FAITA-Mental Health framework (credibility, user experience, crisis management, agency, equity, transparency) to a ChatGPT-store OCD tool, revealing key strengths and safety gaps.
Spitale et al. (2025)	Design & development	Multimodal LLM-based robotic coach (VITA); reinforcement learning	VITA, an LLM-based multimodal robotic coach, adapted autonomously to users' facial valence and speech; adaptive configurations were perceived more positively than pre-scripted ones across lab and 4-week workplace studies.
Ganesh & Venkateswaramurthy (2025)	Perception, acceptability & ethics	AI chatbots, virtual assistants, VR therapy	Synthesized how AI-generated counseling (chatbots, virtual coaches, VR therapy) can widen access, provide 24/7 support and reduce stigma, while cautioning on reliability and the need for clinical oversight.
Rizzi et al. (2025)	Design & development	Virtual coach (REA); ACT-based; wearable triangulation	Detailed a protocol for REA, an ACT-based virtual coach, to prevent postpartum depression, combining chatbot-delivered psychoeducation with smartwatch data and validated usability/engagement measures.
Qi (2025)	Effectiveness & outcomes	Woebot AI chatbot; CBT-based	Among 84 college athletes, daily interaction with the Woebot chatbot progressively reduced mild depression from weeks 5–8 versus no significant change in controls, supporting CAs as an adjunct for athlete mental health.

Author(s), Year	Theme/Focus	Technology/Intervention	Key Findings
Luo et al. (2025)	Perception, acceptability & ethics	Generative AI (ChatGPT) used for self-help	Reddit users appropriated ChatGPT as a "digital therapist" for symptom management, self-discovery and companionship, valuing availability and perceived objectivity but raising privacy and emotional-dependence concerns.
Ulrich et al. (2025)	Effectiveness & outcomes	Conversational agent (BalanceUP); follow-up	A 6-month follow-up indicated that mental well-being gains from the BalanceUP conversational-agent coach were largely sustained after intervention completion in headache sufferers.
Saraç et al. (2025)	Assessment, diagnosis & frameworks	LLM chatbots (Claude-3.5, GPT-4o, Gemini-1.5)	Sport psychologists judged that leading LLM chatbots identified major exercise-dependence symptoms (Claude-3.5 strongest) but struggled with severity and primary/secondary differentiation, supporting an assistive not autonomous role.
Fietta et al. (2025)	Design & development	Chatbot (ALBA); WHO Self-Help Plus	Described adapting WHO's evidence-based Self-Help Plus stress-management program into a chatbot app (ALBA) via user-centered and service design for pregnant women and women with breast cancer.
Pavarini et al. (2025)	Design & development	Chat-story / chatbot narrative; co-design	Co-designed a chat-story with adolescents that was acceptable across multiple testing phases and enhanced peer-support and collective-action skills for youth mental-health promotion.
Mair et al. (2026)	Design & development	Conversational agent (LvL UP); 'talk and tools'	An in-the-wild study in Singapore supported LvL UP's feasibility and acceptability; engagement peaked in the first eight days and users rated the CA-based app enjoyable and easy to use.
Akazua et al. (2026)	Perception, acceptability & ethics	Explainable AI (XAI) conversational agents; mHealth coaching	Mapped evidence on conversational agents and explainable AI in mobile health coaching, identifying trust-related design criteria and opportunities for transparent, interpretable agents in underserved settings.

Note. Key findings are summarised from each study's reported results.

The time distribution of the corpus (Figure 2) shows increased publication activity after 2024, coinciding with the broader emergence of generative-AI and LLM applications. However, temporal growth should not be interpreted as evidence of increasing effectiveness; it more directly reflects growth in research activity and diversification of study designs.

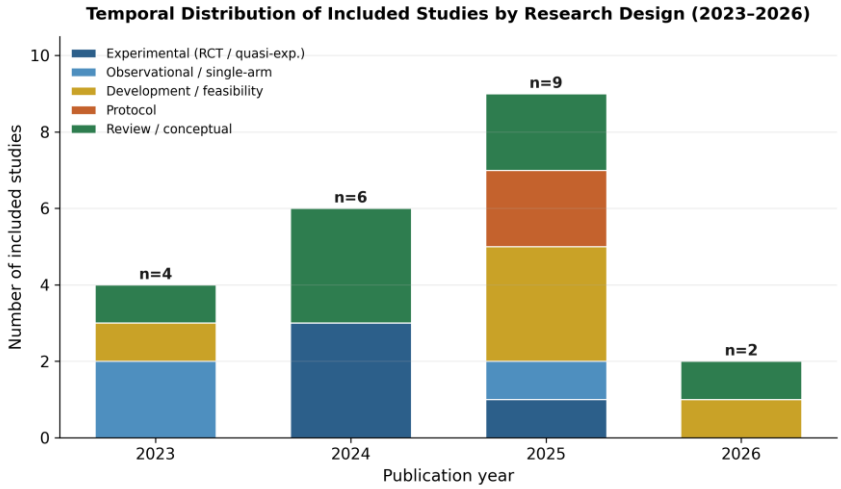

Figure 2. Temporal Distribution of Included Studies by Research Design (2023–2026)

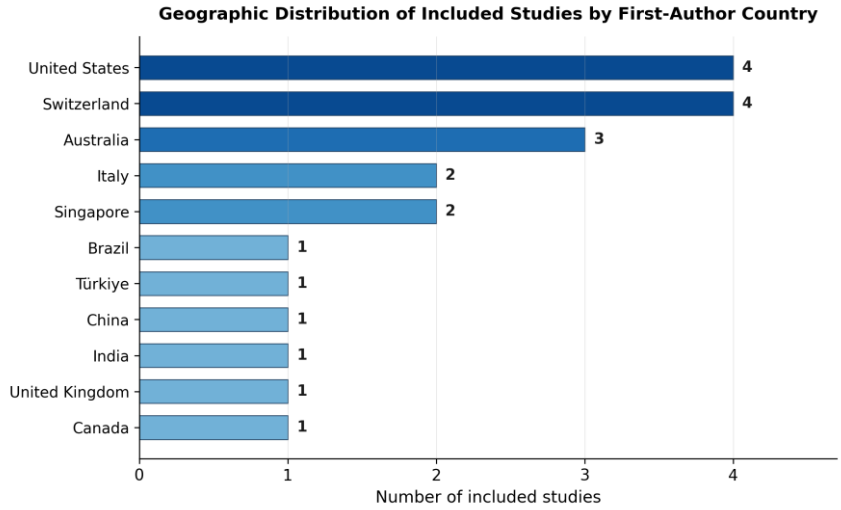

Figure 3. Geographic Distribution of Included Studies by First-Author Country

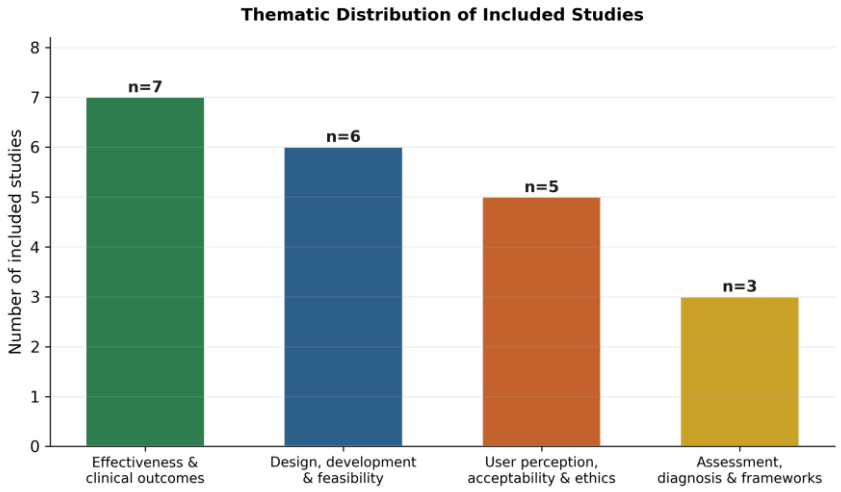

Figure 4. Thematic Distribution of Included Studies

Thematic Synthesis

Findings for RQ1 - Applications, Populations, and Delivery Formats

The included publications show that conversational agents are being applied across student mental-health support, psychological distress, lifestyle coaching, headache-related well-being, postpartum prevention, athlete mental health, clinical training, and population mental-health promotion (Ulrich et al., 2024; Vereschagin et al., 2024; Brinsley et al., 2023; Rizzi et al., 2025; Qi, 2025). Delivery formats include mobile applications, chat-based self-help, peer-supported interventions, multimodal conversational systems, and embodied robotic agents. The dominant pattern is movement toward context-specific content and adaptive interaction rather than a single generic chatbot model.

Chatbot agents could be delivered via mobile phones or as multimodal systems through voice and even embodied robots. The majority of the interventions were mobile phone-based and self-directed, nevertheless, the collection of papers includes a multimodal and LLM-based robotic coach that could learn from facial expressions and vocal cues of the users over repeated interaction (Spitale et al., 2025), and a narrative style or "chat-story" co-created with the youth to foster peer support (Pavarini et al., 2025). Due to the pandemic the use of chatbots increased and chatbot-led programs were offered as lifestyle coaches across various health fields during the lockdown periods (Ollier et al., 2023). Chatbots have also been used in sport and exercise contexts to help with athletes, including reducing depressive symptoms and assessing exercise addictions (Qi, 2025; Saraç et al., 2025).

In sum, all these works tell us that conversational AI has gone so far beyond chatbots for general mental health but has found applications at the grassroots level and particular setting levels. The number of populations covered, students, athletes, perinatal women, headache sufferers and workers' compensation groups, and the general population, leads one to conclude that the technology is appreciated primarily for its scalability and ease of access, and that developers are more and more adjusting the content as well as the interface to match the specific needs of the users rather than just making one-size-fits-all tool.

Findings for RQ2 - Effectiveness, Engagement, and Acceptability

The controlled and quasi-experimental publications provide preliminary evidence of benefit, but the magnitude and durability of effects should be interpreted cautiously. The MISHA pilot RCT reported medium-sized improvements in perceived stress, depression, and somatic symptoms, the Minder RCT reported improvements in anxiety and depression outcomes, BalanceUP reported improved mental well-being relative to a waitlist, and the Woebot pilot reported reductions in mild depressive symptoms among athletes (Ulrich et al., 2024a; Vereschagin et al., 2024; Ulrich et al., 2024b; Qi, 2025). Single-arm and proof-of-concept studies add supportive but weaker evidence. These findings cannot be treated as pooled evidence because intervention content, comparator, outcome instruments, follow-up periods, and statistical reporting differ across publications.

Furthermore, a considerable number of studies focused on the sustainability of the effects, a topic that is frequently avoided in digital health research. A follow-up study lasting six months showed that the improvements gained through using a conversational agent as a coach were retained largely after the completion of the intervention (Ulrich et al., 2025). Also, feasibility and developmental studies constantly pointed out high acceptability and technological acceptance, though engagement appeared to spike in the initial stages and then decrease thereafter, a real-life study indicating that engagement was highest after only eight days usage (Mair et al., 2026; Castro et al., 2023). Qualitative feedback from the users of digital therapeutics revealed reasons why individuals choose such tools: people who use artificial intelligence language models in their capacity as therapists liked the fact that it is always available, the perceived objectivity of the machine and its non-judging nature, while voicing their concerns around privacy and being emotionally reliant on AI.

The evidence therefore supports a cautious interpretation: short-term symptom improvement and favourable user experience recur across several studies, but the corpus does not justify a generalised clinical-effectiveness claim. In particular, the available publications are heterogeneous in study design and outcome measurement, and the longitudinal evidence remains limited. The literature also mixes empirical evidence with protocols, conceptual papers, and reviews, which are informative for design and implementation but should not be treated as intervention-effect estimates.

Findings for RQ3 -Design, Ethics, and Responsible Deployment

The corpus agrees most prominently that conversational agents should operate as assistance only under human supervision instead of being full decision-makers for clinical decisions. A thorough review of the leading LLM chatbot by experts concluded that, although they accurately pinpointed the main signs of

exercise dependence, they lacked the severity calibration and they weren't able to make accurate differential diagnoses so that they were, at best, clinicians' first step of a medical examination aid (Saraç et al., 2025). In the same way, the article on virtual standardised patients highlighted that LLMs might support and enhance the learning of clinical doctors without taking away their judgment. (Kenny & Parsons, 2024).

Concerns about design and regulation played a major role. A special evaluation tool for AI mental-health tools identified such areas for mental health AI as credibility, crisis management, user authority, fairness, and openness and the fact that applying it revealed several issues of a publicly available AI chatbot, specifically crisis-related capabilities. (Golden & Aboujaoude, 2024) The aspect of how an AI is built up or the explainability becomes very important as far as design is concerned and the scoping review has uncovered trust-related elements and has pointed out that transparent, easy-going-to-understand agents will be necessary for mobile health coaching especially among disadvantaged segments of population (Akazua et al., 2026). At the theoretical level, van der Schyff and the team have once again confirmed the importance of trust, bias, and fairness as the starting points of responsible AI use (van der Schyff et al., 2023).

Last but not least, in different studies, responsible design has actually been put into practice through collaboration and research-based development. Therapeutic methods such as acceptance and commitment therapy and the Self-Help Plus programme of the World Health Organization were not only adapted into conversational scripts but also the content of the interventions was, to some extent, decided on or created by users (Rizzi et al., 2025; Fietta et al., 2025; Pavarini et al., 2025). What these methods show is that the paradigm of clinical development is changing where research evidence, active usership, openness, and safety are seen as core aspects and not as add-ons. The major problems still left include the need for unified safety measures, crisis-resolution plans, and equal availability of solutions for communities that are hardly represented in current datasets.

Comparative and Critical Analysis

Comparing the methodological approaches across the corpus shows a field in transition. Randomized trials are present but remain a minority relative to feasibility, development, single-arm, qualitative, protocol, and review publications. The strongest quantitative evidence comes from a small subset of controlled studies, whereas much of the remaining literature addresses feasibility, acceptability, design, or governance. This explains why a meta-analysis was not performed: there is no sufficiently common combination of intervention, comparator, population, outcome construct, timing, and statistical representation to justify a pooled estimate without substantial and potentially misleading transformation.

Still, a rather clear progression with the changing of time is noticeable. Earlier studies focused on dialogue-based or rule-based chatbots and mainly on feasibility, while the newer works brought into play multimodal sensing, reinforcement learning, and LLM-driven adaptation together with a focus on explanation and evaluation methods (Spitale et al., 2025; Akazua et al., 2026). The same pattern was followed in a number of areas: generative models gradually started being used in coaching, education, and clinical-reasoning aids (Chang et al., 2026; Sharma et al., 2026; Comulada et al., 2026). So the main lesson drawn is that the issue that prevents the field from moving forward mainly is not the technological aspect but the methodology one.

Discussion

The synthesis suggests that conversational AI has progressed beyond isolated proof-of-concept demonstrations, but it remains an emergent rather than established clinical modality. Controlled studies indicate promising short-term improvements in several outcomes, while qualitative and feasibility studies indicate generally favourable acceptability. However, user satisfaction and perceived empathy should not be equated with clinical efficacy or diagnostic validity.

The findings suggest that behaviour-change principles, therapeutic communication strategies, explainability, and user-centred design can be operationalised in human–AI interaction. At the same time, the distinction between perceived relational quality and clinical competence remains essential: an agent may be engaging without being sufficiently reliable for diagnosis, crisis management, or autonomous treatment decisions.

For practitioners, developers, and policymakers, the evidence supports positioning conversational agents as supervised adjuncts within stepped-care pathways. Development should use validated therapeutic content where appropriate, transparent disclosure of AI involvement, explicit escalation and crisis pathways, routine safety testing, privacy-by-design practices, and human oversight for high-risk or diagnostically ambiguous cases.

The present findings are consistent with earlier systematic and scoping syntheses showing promising but heterogeneous effects of conversational agents. A 2023 systematic review and meta-analysis found significant pooled reductions in depression and distress but not overall psychological well-being, while a 2024 systematic review and meta-analysis reported reductions in depression and anxiety but questioned the persistence of gains at follow-up (Li et al., 2023; [VERIFY EXACT 2024 REVIEW CITATION]). A 2025 scoping review of reviews identified 17 relevant reviews and likewise highlighted variation in methods, outcomes, and evidence quality. Recent 2026 work on generative-AI mental-health chatbots further emphasises intervention design, user experience, and safety as important emerging dimensions. The present review extends this literature by restricting its synthesis to conversational-agent applications in 2023–2026 and integrating effectiveness with design and governance rather than treating these as separate domains.

Contradictions in the literature. Tensions are apparent between studies emphasising strong user enthusiasm and perceived therapeutic value (Luo et al., 2025; van der Schyff et al., 2023) and appraisals highlighting reliability limitations and safety gaps (Saraç et al., 2025; Maher, 2026). These divergences likely reflect differences between user-reported experience and expert-assessed performance, underscoring that acceptability and clinical adequacy are distinct constructs that must be evaluated separately.

Three gaps are prominent. First, adequately powered randomized trials with long-term follow-up and harmonised outcomes remain scarce. Second, safety and crisis-management performance require direct benchmarking rather than indirect claims based on user satisfaction. Third, evidence is geographically concentrated in high-income settings, leaving transferability to lower-resource and linguistically diverse contexts uncertain.

Limitations of this review. First, the search was restricted to English-language publications from three databases and to 2023–2026, so relevant evidence indexed elsewhere or published in other languages may have been missed. Second, the corpus is methodologically heterogeneous and includes empirical studies, development/feasibility publications, protocols, conceptual/framework papers, and reviews; therefore, quantitative effect estimates were not pooled. Third, the original search record did not retain the exact calendar date(s) of the database search or paired screening decisions, so a valid post-hoc Cohen's kappa coefficient cannot be reconstructed. Fourth, the evidence is vulnerable to internal-validity threats such as self-selection, lack of active controls, attrition, regression to the mean, and measurement reactivity in single-arm studies. External validity is constrained by the geographic concentration of the corpus and under-representation of populations with different languages, health systems, digital access, and cultural expectations. Finally, AI-delivered outcomes may be influenced by novelty effects, platform updates, model-version changes, and differences in implementation fidelity that are not consistently reported.

Future research should prioritise multi-site randomized trials with prospective power calculations based on a prespecified minimally important difference, anticipated variance, alpha level, statistical power, clustering where applicable, and expected attrition. Trials should pre-register primary outcomes and analysis plans, use standardized outcome batteries, report confidence intervals and adverse events, and include follow-up of at least six months where clinically appropriate. In parallel, researchers should establish reproducible safety and crisis-response benchmarks, report model/version and prompt changes, document human-oversight procedures, and conduct equity-focused evaluations in under-represented regions and populations.

In direct response to the research questions: conversational AI is being applied across diverse mental-health contexts, populations, and delivery formats, with increasing tailoring to user needs (RQ1); controlled studies indicate promising short-term improvements and generally favourable acceptability, but the heterogeneous and partly non-empirical evidence base does not support strong pooled clinical claims (RQ2); and responsible deployment depends on transparency, explainability, evidence-anchored and participatory design, privacy and safety governance, crisis escalation, equity, and sustained human oversight (RQ3).

Conclusions

This systematic review synthesised 21 eligible publications published between 2023 and 2026 to map the application, effectiveness, acceptability, and responsible design of LLM- and AI-based conversational agents for mental health and psychological well-being. The evidence indicates promising short-term benefits across depression, anxiety, stress, and well-being, but the corpus is too heterogeneous in design, intervention content, outcomes, and reporting to support pooled quantitative estimates. The review therefore treats controlled intervention findings separately from feasibility, qualitative, conceptual, protocol, and review

evidence. The main practical implication is that conversational agents should currently be deployed as supervised adjuncts rather than autonomous clinical decision-makers. Future studies should use prospectively powered, multi-site designs, standardized outcome and safety metrics, explicit human-oversight procedures, and equity-focused recruitment to determine whether these tools can be integrated safely, effectively, and fairly into routine mental-health care.

Acknowledgments

The Authors would like to thank Faculty of Sport Sciences, Universitas Negeri Padang for providing excellent facilities, as well as for being very supportive of this research from the beginning. Authors also thank the team members who had generously shared their brilliant ideas, engaging discussion, and academic supports during the study.

References

Akazua, L.O., Zhou, J., Chen, F., Shafiabady, N., Tian, G., Holzinger, A., & Müller, H. (2026). Explainable Conversational Agents for Mobile Health Coaching Systems: Trust Factors, Progress and Opportunities. Machine Learning and Knowledge Extraction, 8(6). https://doi.org/10.3390/make8060144

Alsanoosy, T., & Qarah, F. (2026). An Arabic language benchmark and hybrid BERT-BiGRU-CapsNet model for optimism and pessimism detection. Scientific reports. https://doi.org/10.1038/s41598-026-50842-2

Bays, D.K., Verble, C., & Powers Verble, K.M. (2024). A Brief Review of the Efficacy in Artificial Intelligence and Chatbot-Generated Personalized Fitness Regimens. Strength and Conditioning Journal, 46(4), 485–492. https://doi.org/10.1519/SSC.0000000000000831

Brinsley, J., Singh, B., & Maher, C.A. (2023). A Digital Lifestyle Program for Psychological Distress, Wellbeing and Return-to-Work: A Proof-of-Concept Study. Archives of Physical Medicine and Rehabilitation, 104(11), 1903–1912. https://doi.org/10.1016/j.apmr.2023.04.023

Castro, O., Mair, J.L., Salamanca-Sanabria, A., Alattas, A., Keller, R., Zheng, S., Jabir, A., Lin, X., Frese, B.F., Lim, C.S., Santhanam, P., van Dam, R.M., Car, J., Lee, J., Tai, E.S., Fleisch, E., von Wangenheim, F., Tudor Car, L., Müller-Riemenschneider, F., & Kowatsch, T. (2023). Development of "LvL UP 1.0": a smartphone-based, conversational agent-delivered holistic lifestyle intervention for the prevention of non-communicable diseases and common mental disorders. Frontiers in Digital Health, 5. https://doi.org/10.3389/fdgth.2023.1039171

Chakraborty, S., Mendro, N., & Li, L. (2026). Leveraging Student-Athlete Mental Health Through an AI-Augmented Mobile Platform: The ThriveNudge Study Protocol. Behavioral sciences (Basel, Switzerland). https://doi.org/10.3390/bs16020268

Chang, E., Che, X., & Zhang, J. (2026). Designing and evaluating LLM-driven coaching agents for VR public speaking practice. Frontiers in psychology. https://doi.org/10.3389/fpsyg.2026.1886780

Cheng, Y., Gong, T., & Zhao, H. (2026). Psychological-physical synergy of athletes based on artificial intelligence and deep learning. Scientific reports. https://doi.org/10.1038/s41598-026-45920-4

Comulada, W.S., Swendeman, D., Ho, Y.X., Streck, J.M., Lewis, D., Rezai, R., Warren, D., Pachas, G., Chandler, M., Jackson, J.L., & Gelberg, L. (2026). Development, Feasibility, Acceptability, and Usability of an Artificial Intelligence-Powered Chatbot (Suzy) to Support Patients in Substance Use Disorder Recovery: Multiphase Study. JMIR formative research. https://doi.org/10.2196/84683

De Gagne, J.C., & Gris, I. (2026). AI disclosure uncertainty in nursing education: A pedagogical scaffold for professional learning. Nurse education today. https://doi.org/10.1016/j.nedt.2026.107097

Dergaa, I., Dergaa, M.A., Razzak, M., Ceylan, H.İ., Stefanica, V., Muntean, R.I., & Guelmami, N. (2026). Generative artificial intelligence and large language models in sports medicine: a scoping review of applications, accuracy, and ethical implications. Frontiers in public health. https://doi.org/10.3389/fpubh.2026.1843535

El Kinany, K., Abdelhamid, A., Benedict, A., Aly, M., Abdelkarim, O., & El Rhazi, K. (2026). Digital communication tools used to promote healthy lifestyles: an overview of reviews from the BRIDGE project. Frontiers in digital health. https://doi.org/10.3389/fdgth.2026.1805124

Fietta, V., Rizzi, S., Gios, L., Pavesi, M.C., De Luca, C., Gabrielli, S., Monaro, M., Navarin, N., Gadotti, E., Mayora-Ibarra, O., Purgato, M., Barbui, C., & Forti, S. (2025). World Health Organization

Evidence-Based Self-Help Plus Intervention for Stress Management via Chatbot: Protocol for Adaptation to a Tech-Enabled Model. JMIR Research Protocols, 14. https://doi.org/10.2196/69644
Fu, L., Burns, R., Xie, Y., Shen, J., Zhe, S., Estabrooks, P., & Bai, Y. (2026). The Development and Use of AI Chatbots for Health Behavior Change: Scoping Review. Journal of medical Internet research. https://doi.org/10.2196/79677
Ganesh M, S., & Venkateswaramurthy, N. (2025). Artificial Intelligence (AI) Generated Health Counseling For Mental Illness Patients. Current Psychiatry Research and Reviews, 21(3), 269–283. https://doi.org/10.2174/0126660822277500240109050359
Ghimire, A., & Qiu, Y. (2026). Recalibrating the value of presence: A qualitative study on generative AI, attendance, and professional formation in undergraduate nursing. Nurse education today. https://doi.org/10.1016/j.nedt.2026.107111
Golden, A., & Aboujaoude, E. (2024). Describing the Framework for AI Tool Assessment in Mental Health and Applying It to a Generative AI Obsessive-Compulsive Disorder Platform: Tutorial. JMIR Formative Research, 8. https://doi.org/10.2196/62963
Hasan, M.M., Rahat, M.F.R., Anjum, N., Sakib, N., & Moore, B.A. (2026). DigiMindReady: Enhancing Military Readiness Through Edge AI-Driven Wellness, Education, and Digital Discipline Through Privacy-First mHealth Innovation. Military medicine. https://doi.org/10.1093/milmed/usag101
Kenny, P.G., & Parsons, T.D. (2024). Virtual Standardized LLM-AI Patients for Clinical Practice. Annual Review of CyberTherapy and Telemedicine, 22, 177–182.
Li, T., Wang, L., Zhang, C., Liu, Y., Qiu, Y., & Liang, G. (2026). Associations between 24-hour movement behavior profiles and mental health among Chinese school students: A latent profile analysis. Journal of Exercise Science & Fitness. https://doi.org/10.1016/j.jesf.2026.200515
Zhong, W., Luo, J., & Zhang, H. (2024). The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. Journal of Affective Disorders, 356, 459–469. https://doi.org/10.1016/j.jad.2024.04.057
Olisaeloka, L., Richardson, C. G., Wang, A. Y., Munthali, R. J., & Vigo, D. V. (2026). Generative AI mental health chatbots: A scoping review of intervention design and user experience. npj Digital Medicine. https://doi.org/10.1038/s41746-026-02972-0
Mayor, E. (2025). Chatbots and mental health: A scoping review of reviews. Current Psychology, 44, 13619–13640. https://doi.org/10.1007/s12144-025-08094-2
Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. npj Digital Medicine, 6, 236. https://doi.org/10.1038/s41746-023-00979-5
Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. PLoS Medicine, 6(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100
Luo, X., Ghosh, S., Tilley, J.L., Besada, P., Wang, J., & Xiang, Y. (2025). "Shaping ChatGPT into my Digital Therapist": A thematic analysis of social media discourse on using generative artificial intelligence for mental health. Digital Health, 11. https://doi.org/10.1177/20552076251351088
Maher, J., Byszewski, A., & Lochnan, H. (2026). Whose Voice is it Anyway? Artificial Intelligence and the New Crisis of Authenticity in Medical Education. Perspectives on medical education. https://doi.org/10.5334/pme.2265
Mair, J.L., Jabir, A.I., Salamanca-Sanabria, A., Castro, O., Zheng, S., Keller, R., Frese, B.F., Lim, C.S., Alattas, A., Negi, S., Shenoi, A., van Dam, R.M., Tai, E., Fleisch, E., von Wangenheim, F., Car, L.T., Müller-Riemenschneider, F., & Kowatsch, T. (2026). Feasibility of the LvL UP digital lifestyle coaching intervention designed to prevent non-communicable diseases and common mental disorders. Scientific Reports, 16(1). https://doi.org/10.1038/s41598-025-30960-z
Md Zuki, M.A., Mohamad Ali, N., & Chaw, J.K. (2026). Multimodal Sentiment and Emotion Analysis Framework for Personalized Health Coaching Messages: Proof-of-Concept Study. JMIR formative research. https://doi.org/10.2196/79558
Mwogosi, A., Marungu, J., Wasanda, N., Kabage, F., Kingako, D., & Nestory, V. (2026). Perceptions for integrating IoT in electronic health records for mental health follow-up: an exploratory study in Tanzania. Mental Health and Digital Technologies. https://doi.org/10.1108/MHDT-07-2025-0043
Nadeem, F., Berta, C., Dayal, D., Rahaman, C., Smitherman, A., Harrell, M., Powell, E., Minor, B., Evely, T., Brabston, E., & Momaya, A. (2026). Assessing the Accuracy and Reliability of ChatGPT-4 to

Answer Clinical EHR Messages in Sports Medicine. Southern medical journal. https://doi.org/10.14423/SMJ.0000000000001977

Nan, J., Purpura, S., Jaiswal, S., Afshar, H., Maric, V., Manchanda, J.K., Taylor, C.T., Ramanathan, D., & Mishra, J. (2026). Personalized machine learning guided intervention for optimizing lifestyle behaviors in depression: a pilot study. NPP - digital psychiatry and neuroscience. https://doi.org/10.1038/s44277-026-00062-3

Ollier, J., Suryapalli, P., Fleisch, E., Wangenheim, F.V., Mair, J.L., Salamanca-Sanabria, A., & Kowatsch, T. (2023). Can digital health researchers make a difference during the pandemic? Results of the single-arm, chatbot-led Elena+: Care for COVID-19 interventional study. Frontiers in Public Health, 11. https://doi.org/10.3389/fpubh.2023.1185702

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Pankow, K., & Vella, S.A. (2026). MindMoves: Exploring the acceptability and feasibility of a digital mental health literacy program for collegiate athletes. Performance Enhancement & Health. https://doi.org/10.1016/j.peh.2026.100450

Pavarini, G., Murta, S.G., Mendes, J.A.D.A., Siston, F.R., de Souza, R.R.A., de Oliveira da Cunha, R., Alves Ferreira, J., de Santos, V.H.D.L., Seabra, B.T.R., Ferrario, N., Battan, L.S., Matera, B., Coni, I., Pereira, L.R., Iglesias, T., Savi, A.L., Foscarini, D., Calil, K., Corvello, F., & de Oliveira, T.M.G. (2025). Cadê o Kauê? Co-design and acceptability testing of a chat-story aimed at enhancing youth participation in the promotion of mental health in Brazil. Journal of Child Psychology and Psychiatry and Allied Disciplines, 66(5), 697–715. https://doi.org/10.1111/jcpp.14078

Qi, Y. (2025). Pilot Quasi-Experimental Research on the Effectiveness of the Woebot AI Chatbot for Reducing Mild Depression Symptoms among Athletes. International Journal of Human-Computer Interaction, 41(1), 452–459. https://doi.org/10.1080/10447318.2023.2301256

Rizzi, S., Nicoletti, A.E., De Luca, C., Poggianella, S., Dalmonego, C., Paoli, C., Marroni, D., Burlon, B., Chiodega, V., Dallafior, M., Pertile, R., Rizzi, S., Valentini, L., Alberi, I., Forti, S., & Taddei, F. (2025). Prevention of postpartum depression via a digital acceptance and commitment therapy-based intervention: protocol for a pilot usability study. Frontiers in Psychiatry, 16. https://doi.org/10.3389/fpsyt.2025.1633229

Saraç, H., Yüzakı, E., & Aşçı, F.H. (2025). The Potential of AI Chatbots as Diagnostic Tools in Mental Health: Evaluating Exercise Dependence Symptoms. Journal of Technology in Behavioral Science. https://doi.org/10.1007/s41347-025-00567-2

Sartor, F., Rosito, S.A., Raimo, M., Beato, M., & Bertollo, M. (2026). Artificial intelligence in sport psychology: Implications for the identification and development of talent. Psychology of sport and exercise. https://doi.org/10.1016/j.psychsport.2025.102977

Semeraro, F., Montomoli, J., & Gamberini, L. (2026). The AI dispatcher copilot: beyond cardiac arrest recognition to dynamic large language model-assisted Tele-CPR. Resuscitation. https://doi.org/10.1016/j.resuscitation.2026.111180

Sharma, A., Dreicer, J.J., & Parsons, A.S. (2026). Using generative AI to support clinical reasoning coaching: a theory-informed approach. Diagnosis (Berlin, Germany). https://doi.org/10.1515/dx-2026-0013

Shenoi, A., Li, T., Jabir, A.I., Pitkethly, A., Fleisch, E., Kowatsch, T., & Mair, J.L. (2026). Structured Large Language Model Workflows for Motivational Interviewing in Health Behavior Change: Proof-of-Concept Study. JMIR formative research. https://doi.org/10.2196/94036

Sinha, G.R., Power, S.R., Singh, R.S., & Sinha, A.K. (2026). Understanding mental health language, help-seeking, and barriers to care among emerging adults in India. Mental Health & Prevention. https://doi.org/10.1016/j.mhp.2026.200538

Spitale, M., Axelsson, M., & Gunes, H. (2025). VITA: A Multi-Modal LLM-Based System for Longitudinal, Autonomous and Adaptive Robotic Mental Well-Being Coaching. ACM Transactions on Human-Robot Interaction, 14(2). https://doi.org/10.1145/3712265

Terry, P.E. (2026). How Should We Prompt Engineer Our Health Coach Bots? Stoicism, Epicureanism or Something Else?. American journal of health promotion : AJHP. https://doi.org/10.1177/08901171261416224

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, 45. https://doi.org/10.1186/1471-2288-8-45

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management, 14(3), 207–222. https://doi.org/10.1111/1467-8551.00375

Trani, L., Fontana, A., Greco, A.G., Egidi, F., Cacioppo, M., & Sideli, L. (2026). Doping prevention in adolescence: findings from a chat-bot assisted educational program for high school students. The Journal of sports medicine and physical fitness. https://doi.org/10.23736/S0022-4707.26.17642-7

Ulrich, S., Gantenbein, A.R., Zuber, V., Von Wyl, A., Kowatsch, T., & Künzli, H. (2024). Development and Evaluation of a Smartphone-Based Chatbot Coach to Facilitate a Balanced Lifestyle in Individuals With Headaches (BalanceUP App): Randomized Controlled Trial. Journal of Medical Internet Research, 26(1). https://doi.org/10.2196/50132

Ulrich, S., Künzli, H., Zuber, V., Kowatsch, T., & Gantenbein, A.R. (2025). Maintained effectiveness of a smartphone-based chatbot coach (BalanceUP) for mental well-being in headache sufferers: preliminary findings from a single-arm 6-month follow-up study. mHealth, 11. https://doi.org/10.21037/mhealth-25-9

Ulrich, S., Lienhard, N., Künzli, H., & Kowatsch, T. (2024). A Chatbot-Delivered Stress Management Coaching for Students (MISHA App): Pilot Randomized Controlled Trial. JMIR mHealth and uHealth, 12. https://doi.org/10.2196/54945

van der Schyff, E.L., Ridout, B., Dip, G., Amon, K.L., Forsyth, R., Cert, G., & Campbell, A.J. (2023). Providing Self-Led Mental Health Support Through an Artificial Intelligence–Powered Chat Bot (Leora) to Meet the Demand of Mental Health Care. Journal of Medical Internet Research, 25. https://doi.org/10.2196/46448

Vereschagin, M., Wang, A.Y., Richardson, C.G., Xie, H., Munthali, R.J., Hudec, K.L., Leung, C., Wojcik, K.D., Munro, L., Halli, P., Kessler, R.C., & Vigo, D.V. (2024). Effectiveness of the Minder Mobile Mental Health and Substance Use Intervention for University Students: Randomized Controlled Trial. Journal of Medical Internet Research, 26(1). https://doi.org/10.2196/54287

Wei, Y., Farizan, N.H., & Ji, C. (2026). A psychological education model integrating artificial intelligence-based BERT in physical education. Scientific reports. https://doi.org/10.1038/s41598-026-55734-z