Review

Large Language Models (LLMs) in Mental Health: Scoping Review of Clinical Promise, Human Factors, and Implementation Challenges

Reza Babaei[1], MS; Samuel Cheng[1], Prof Dr, PhD; Dan Li[2], Prof Dr, PhD; Jie Cao[3], Prof Dr, PhD; Shangqing Zhao[3], Prof Dr, PhD

[1]School of Electrical and Computer Engineering, The University of Oklahoma, Norman, OK, United States
[2]Hudson College of Public Health, University of Oklahoma Harold Hamm Health Campus, Oklahoma City, OK, United States
[3]School of Computer Science, The University of Oklahoma, Norman, OK, United States

Corresponding Author:

Samuel Cheng, Prof Dr, PhD
School of Electrical and Computer Engineering
The University of Oklahoma
4502 E41st #4W404 Schusterman Center
Norman, OK 74135
United States
Phone: 1 9184010409
Email: samuel.cheng@ou.edu

Abstract

Background: The rapid integration of large language models (LLMs) into mental health care presents clinical opportunities alongside important ethical challenges. Despite a growing literature proposing LLMs for diagnostic, therapeutic, and supportive roles, a comprehensive synthesis of their empirical efficacy, technical frameworks, and clinical safety profile remains limited.

Objective: This scoping review mapped evidence on the clinical outcomes and reported efficacy of LLM-based interventions, characterized user psychological impacts, examined commercial deployment patterns, and identified long-term behavioral implications and governance needs.

Methods: Following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework, we searched 5 sources: PubMed, arXiv, Semantic Scholar, OpenAlex (queried via their public APIs), and PsycINFO (searched manually through an institutional ProQuest library interface). Record identification used conventional keyword and Boolean queries; AI was not used to discover or retrieve records. A semiautomated pipeline then applied LLM-assisted title and abstract screening (Google Gemini models) against the review eligibility criteria after deduplication (n=792 unique records screened). Full-text assessment yielded 241 included sources for qualitative synthesis; studies reporting quantitative clinical or psychological outcomes were summarized as a structured narrative organized by study design, publication status, evidentiary maturity, psychological construct, and AI-authorship disclosure conditions. Agreement between the model's screening decisions and independent, blinded human reviewers was subsequently quantified on stratified samples in a post hoc validation.

Results: Of 803 records identified (792 after deduplication), 317 advanced to full-text review, and 241 sources were included. Thematic synthesis identified 5 core themes: (1) clinical applications, (2) technical frameworks, (3) evaluation and metrics, (4) challenges and limitations, and (5) governance and future directions. Across the quantitative evidence base, randomized controlled trials tended to report modest or null between-group effects, whereas uncontrolled designs reported larger contrasts that cannot be causally attributed to the LLM because of the limitations of their study designs. Where AI authorship was experimentally disclosed, studies reported lower empathy-related ratings relative to comparators.

Conclusions: The evidence base for LLMs in mental health is early stage, heterogeneous, and methodologically variable. LLM-based tools show promise in circumscribed settings; however, effects are often modest and tightly conditioned by implementation context, disclosure practices, and human factors including user expectations, trust, and engagement. Future research should prioritize well-controlled longitudinal trials alongside robust governance frameworks and human-in-the-loop safeguards to protect safety, equity, and privacy.

JMIR Ment Health 2026;13:e101785; doi: 10.2196/101785

Keywords: large language models; generative artificial intelligence; ChatGPT; mental health; psychotherapy; digital mental health; chatbots; human factors; scoping review

Introduction

The advent of large language models (LLMs)—exemplified by OpenAI's ChatGPT, Meta's LLaMA, and Google's Gemini—has led to major changes across the medical sciences [1-3]. In mental health care, LLMs are being deployed as diagnostic adjuncts, conversational therapeutic agents, and scalable emotional support platforms [4,5], capable of simulating empathetic dialogue, synthesizing complex psychiatric literature, and delivering structured psychological interventions at scale [6]. Commercial enthusiasm has substantially outpaced clinical evaluation [2,7], yet early evidence suggests meaningful potential in accessibility and structured support—particularly in settings with limited mental health resources [4,6]. As deployment reaches vulnerable populations, the field must balance these opportunities with careful attention to patient safety, clinical efficacy, and the ethical implications of delegating psychiatric care to nonsentient systems [1,2,4,5].

Several evidence syntheses have begun to map this space, but each has a bounded scope. A recent scoping review has cataloged human-assessed generative-task applications of LLMs in mental health care [1,8], and a systematic review has charted the evolution of mental health chatbots from rule-based systems to LLMs [9]; another systematic review has surveyed their capabilities and limitations across clinical contexts [10], and empirically oriented reviews have examined ChatGPT's clinical use [11] and weighed the benefits and harms of LLMs in digital mental health [6,12]. These syntheses have tended to focus on a single model family (eg, ChatGPT), a single task class (eg, generative tasks or capability benchmarking), or a capabilities-versus-risks framing. Less well integrated is a synthesis that spans the full translational pipeline—clinical applications, technical adaptation frameworks, evaluation methodology, safety limitations, and governance—while foregrounding the human factors (perceived empathy, trust, disclosure and transparency effects, engagement, and over-reliance) that increasingly shape real-world outcomes. Such a synthesis also interprets the quantitative evidence through the lens of study-design strength and disclosure conditions. This review addresses this gap.

This review synthesizes the current landscape of LLM applications in mental health, guided by 4 core research questions (RQs):

1. What are the clinical outcomes and therapeutic efficacy of using LLM-based technologies compared to traditional human-to-human psychiatric and psychological interactions?
2. What does the synthesized literature indicate about relational and process-level dynamics—including therapeutic alliance, perceived empathy, trust, and engagement trajectories—when LLM-based interactions are substituted for human connection, and where do the principal empirical gaps lie?
3. How are LLM technologies being implemented and scaled within the broader context of the commercial self-help and digital wellness industries?
4. What does current evidence suggest about downstream longitudinal outcomes of sustained reliance on LLM-based mental health support—including agency, coping behavior, and broader psychosocial functioning—and what remains unknown?

By addressing these questions, this paper aims to provide clinicians, researchers, and policymakers with an evidence-based foundation for evaluating the integration of generative AI into mental health care practice.

To keep this breadth tractable and transparent, the review is explicitly bounded. Studies applying, evaluating, or analyzing the impact of generative LLMs—and transformer-based models used in generative or interactive clinical roles—in mental health and contexts across clinical, research, direct-to-consumer, and online settings are in scope (Table 1). Nongenerative or purely discriminative machine learning systems and application domains outside mental health are out of scope; studies using LLMs solely as research or measurement instruments (rather than as clinical or user-facing components) were excluded from the quantitative outcome synthesis but retained for qualitative mapping where relevant to the review questions. The wide thematic breadth reflects the evidence-mapping purpose of a scoping review; where the evidence base is thin or preliminary, we mark it as such rather than narrowing the map to only mature findings.

Table 1. Eligibility criteria structured by the Population–Concept–Context (PCC) framework.

Dimension	Included	Excluded
Population	Mental health contexts: patients, clients, and service users, at-risk and general populations, and mental health professionals; or clinical and psychological data used for mental health tasks (eg, clinical notes, therapy transcripts, and social media text).	Studies unrelated to mental health.
Concept	Application, evaluation, or impact of generative LLMs[a] (eg, OpenAI's GPT-3/4, Meta's LLaMA, Anthropic's Claude, and Google's Gemini) and transformer-based models used in generative or interactive clinical roles (screening, conversational or therapeutic support, structured psychotherapy delivery, assessment, clinician training, and governance).	General AI or traditional or discriminative machine learning without a generative LLM architecture (eg, random forests, support vector machines, and standalone sentiment classifiers) unless integrated into a generative or interactive system. LLMs used solely as research instruments were excluded from the quantitative outcome synthesis only and retained for qualitative mapping.
Context	Any clinical, subclinical, or digital mental health setting: clinical care, research evaluation, direct-to-consumer or self-help, and online or social platforms; global; no restriction on acuity.	Application domains outside mental health.
Types of sources	Peer-reviewed journal articles, conference proceedings, and preprints (eg, arXiv and medRxiv) published in English, including primary empirical studies, evidence syntheses (reviews, bibliometric analyses, and review protocols), and substantive theoretical or framework papers.	Non-English publications; purely speculative opinion or editorial pieces without empirical or substantive theoretical grounding.

[a]LLM: large language model.

Methods

Protocol and Registration

This review is reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Checklist 1) [13] and draws on established scoping review methodology [14-16]. The review questions and core eligibility concept were specified in advance of screening; stage-specific screening criteria were operationalized as described below. Consistent with the purpose of a scoping review—to map the breadth, characteristics, applications, risks, and research gaps of an emerging evidence base rather than to estimate a pooled intervention effect or determine comparative effectiveness—formal study-level critical appraisal (risk of bias) was not undertaken; this is optional under PRISMA-ScR and Joanna Briggs Institute scoping review methodology (see Evidence Characterization section below). No prospectively registered or publicly posted protocol was available for this review, and the review was not registered; the absence of a prospectively registered protocol is acknowledged as a limitation. Stage-specific operationalization of the eligibility criteria is described below (see Study Selection section).

Methodological Note

To manage the rapidly expanding volume of AI literature, the review used a semiautomated computational pipeline: API-based bulk record export at identification, LLM-assisted title and abstract screening, and stream-specific downstream processing and data extraction as described below. Record identification itself relied on conventional database queries rather than AI-based discovery (see Search Strategy section). RB managed the study selection workflow and the final included corpus; the quantitative outcome statistics carried into the narrative synthesis were verified against source PDFs; and all final interpretive and analytical decisions were made by the human research team. This staged use of automation is consistent with current cross-organizational guidance on the responsible use of AI in evidence synthesis [17] and with prior AI-assisted reviews in this field [18].

Search Strategy

Records were identified by executing explicit keyword and Boolean queries—using free-text terms for LLMs (eg, large language model, ChatGPT, generative AI, and GPT) combined with mental health and psychology terms—against 5 established databases with a cutoff date of February 20, 2026. No lower publication date limit was specified in any query string (Multimedia Appendix 1); because the API-accessed streams were capped and ordered by recency or relevance, their effective coverage was concentrated on the most recent literature. To be explicit about the role of automation: record identification used only these conventional database queries; AI and LLM tools were not used to discover, recommend, or semantically retrieve records. AI assistance was confined to the later screening, eligibility, and data extraction stages (see Study Selection and Data Extraction section).

Two retrieval channels were used. First, 4 databases spanning the biomedical, computer science, and preprint literature—PubMed, arXiv, Semantic Scholar, and OpenAlex—were queried through their public APIs; here, "programmatic" or "API-based" denotes only the access and export method (bulk download of records matching a predefined query), not AI-based or semantic search. To keep the yield tractable in a high-volume field, API retrieval was capped and ordered by source-specific recency or relevance ranking (PubMed: 100 by date; arXiv: 50 by submission date; Semantic Scholar: up to 200 by relevance; OpenAlex: 50 by publication year). Because the API-based streams were capped, these source-specific sort and ranking functions influenced which records were returned within each capped set. Second, PsycINFO was searched manually through our institutional (University of Oklahoma) ProQuest (Clarivate) library interface using a targeted title and abstract Boolean string, and the complete set of 419 records was exported to a CSV file for screening. The database set was chosen

for complementary coverage of a field at the intersection of mental health and computer science: much of the LLM literature appears first in computer science and preprint venues (arXiv, Semantic Scholar, and OpenAlex), while PubMed and PsycINFO provide biomedical and psychological coverage, with PsycINFO providing dedicated psychology coverage. The exact query strings, field sets, caps, and sort orders for every source are reported verbatim in Multimedia Appendix 1. Raw retrieval comprised 384 records from the 4 API-accessed databases and 419 from PsycINFO, for 803 records before deduplication.

Inclusion and Exclusion Criteria

Studies were included if published in English and focused on the application, evaluation, or impact of LLMs in mental health contexts; peer-reviewed journal articles, conference proceedings, and preprints (eg, arXiv and medRxiv) were eligible, including primary empirical studies, evidence syntheses (reviews, bibliometric analyses, and review protocols), and substantive theoretical or framework papers. Studies were excluded if they focused on traditional machine learning models (eg, random forests and support vector machines) without generative LLM architectures, were non-English, purely speculative without empirical grounding, or unrelated to mental health. "LLMs" encompassed both generative models (OpenAI's GPT-3 and GPT-4, Meta's LLaMA, Anthropic's Claude, and Google's Gemini) and transformer-based models used in generative or interactive clinical contexts; purely discriminative classification-only applications were excluded unless integrated into a generative or interactive system. Studies using LLMs solely as research instruments rather than as clinical intervention components were excluded from outcome synthesis but retained for qualitative synthesis where relevant. Following scoping review reporting conventions, these criteria are additionally summarized using the Population–Concept–Context (PCC) framework in Table 1 [13,16].

Study Selection

Records were deduplicated by normalized title (case-folded, punctuation and whitespace removed). Within the API-accessed export (n=384 raw), 2 duplicate records were removed by normalized title matching (exact and fuzzy, similarity ≥0.90), yielding 382 records. PsycINFO records (n=419 raw) whose normalized title matched an API-accessed title were then removed as cross-stream duplicates (9 records), and any remaining within-PsycINFO title duplicates were removed (none were present), yielding 410 records—for 792 unique records entering title and abstract screening. Because cross-stream matching used exact normalized titles, a small number of near-duplicate records across streams may not have been detected. All 792 records then underwent LLM-assisted application of stage-specific operationalizations of the eligibility criteria using Google Gemini models (principally Gemini 2.5 Flash; records passing an initial PsycINFO relevance screen underwent an additional empirical-relevance refinement, using Gemini 2.5 Pro, before full-text retrieval—so the 2 streams did not have identical screening mechanics). This stage advanced 317 records to the full-text phase and excluded 475. Of the 317 advanced, 277 full texts were obtained (40 could not be retrieved or stably mapped to a PDF). Retrieved reports were assessed against the eligibility criteria; 36 were excluded at full text, yielding 241 studies for qualitative synthesis and data extraction. Counts refer to reports (sources of evidence): 2 pairs of included reports are closely related work from the same groups (a preprint and its later journal version, and 2 conference papers) and are listed separately in Multimedia Appendix 2.

Reviewer Roles and Validation

Original screening was conducted with the LLM-assisted pipeline under RB's management; it was not performed independently and in duplicate by human reviewers, and we state this plainly rather than imply otherwise. To characterize the reliability of the AI-assisted decisions, we subsequently conducted an additional, post hoc blinded human validation (a reliability exercise that did not alter the original included corpus). For title and abstract screening, a balanced random sample of 50 records (25 model-Include, 25 model-Exclude, split evenly across the 2 streams) was rescreened blinded to the model's decision and applying the review eligibility criteria: 4 coauthors (SC, SZ, JC, and DL) rescreened the sample under a paired design (each record rated by 2 reviewers, all 6 reviewer pairs represented), and RB independently rescreened all 50 records. Agreement among the independent human reviewers was 68% (Fleiss κ=0.35, 95% bootstrap CI 0.08-0.60), and agreement between RB and the model was 66% (Cohen κ=0.32, 95% bootstrap CI 0.06-0.56); where the two coauthor reviewers reached consensus, the model matched that consensus in 76% (26/34) of records. At the full-text stage, 2 authors (RB and SC) independently reassessed a balanced sample of 10 reports (5 previously included and 5 previously excluded), blinded to prior status and to each other; agreement was 80% (Cohen κ=0.62, 95% bootstrap CI 0.14-1.00; wide given n=10 and not indicative of precise agreement), and the disagreements between the reviewers were resolved by discussion within the validation exercise. Agreement was thus modest at title and abstract screening—where eligibility judgments were more sensitive to borderline scope distinctions (eg, whether mental health was the central focus, or whether an LLM was itself under evaluation vs used only as an incidental tool)—and higher at full-text reassessment, in a range broadly consistent with prior evaluations of GPT-class models against human screeners [17,19].

Data Extraction, Thematic Synthesis, and Evidence Characterization

For the programmatic (API-accessed) stream, a standardized LLM-assisted extraction schema (Google Gemini with a structured TypedDict output schema enforced through the model's structured-output mode; response_schema parameter) was used to populate study metadata, clinical themes, supporting verbatim source text, and quantitative outcome fields from available full texts; the schema required verbatim source sentences for key fields to reduce fabrication. PsycINFO records were characterized using the exported

and tracked ProQuest metadata and available study materials rather than through the same automated full-text extraction pipeline. The quantitative outcome statistics carried into the narrative synthesis were verified against the source PDFs (see Quantitative Synthesis section); extraction otherwise supported qualitative mapping and characterization. Data charting was performed once by this pipeline under the first author's management rather than independently in duplicate. Thematic synthesis followed an inductive approach adapted from Thomas and Harden (2008) [20], proceeding through free line-by-line coding, organization into descriptive clusters, and development into 5 analytically derived overarching themes.

Consistent with scoping review methodology, formal study-level risk-of-bias appraisal was not undertaken; the objective was to map the breadth, characteristics, applications, risks, and gaps of a heterogeneous evidence base rather than to estimate a pooled effect or determine comparative effectiveness. Instead, for transparency, each included source was characterized by a publication venue evidence category (journal article, conference proceedings, or preprint; Multimedia Appendix 2). Publication venue and peer review status are treated strictly as descriptive characteristics of the evidence base—a flag for additional scrutiny during synthesis—and are not equated with methodological quality or risk of bias. Interpretation was accordingly design-sensitive: primary clinical claims were anchored on peer-reviewed, controlled evidence, while findings from preprints, uncontrolled or noncomparative designs, or small samples were qualified with explicit linguistic markers (eg, preliminary evidence and available as a preprint) and treated as trend-identifying rather than confirmatory. For the quantitative narrative, inclusion required that a study report a tangible clinical or psychological outcome (eg, symptom change, perceived empathy, task performance, and validated scale score) with interpretable directional data from a between-group or pre-post design; studies were then read through 3 interpretive lenses—study design, psychological construct measured, and AI-authorship disclosure conditions—that function as analytic aids rather than formal statistical moderators.

Ethical Considerations

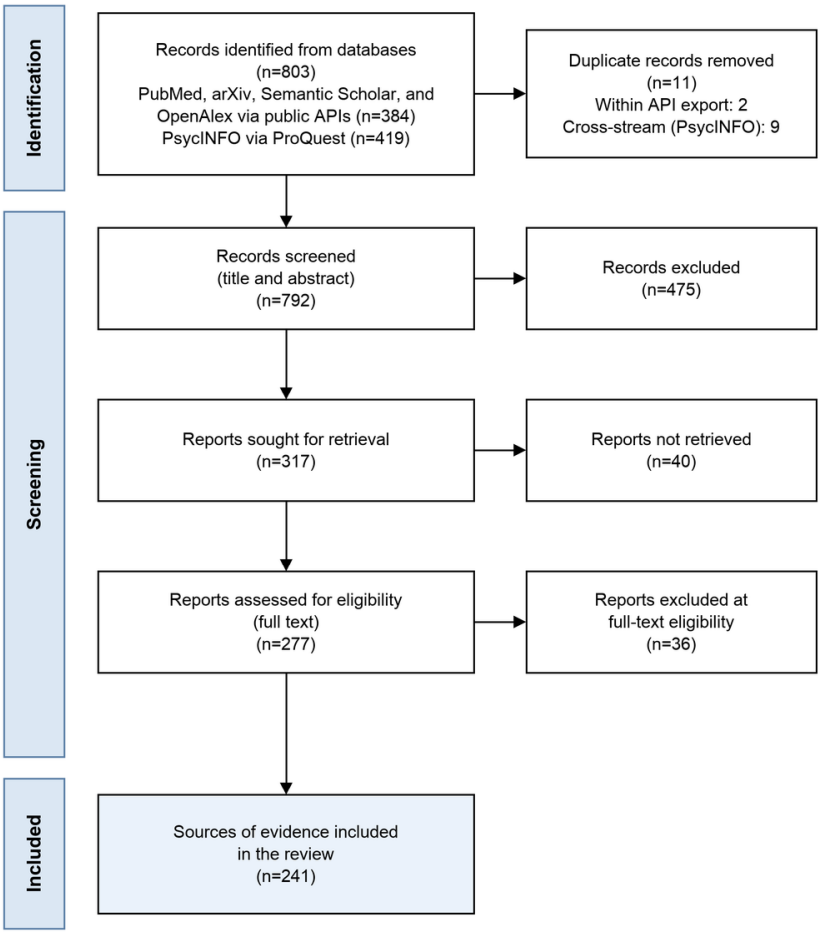
This study involved secondary analysis of published literature and did not involve human participants; accordingly, institutional ethics board approval was not required.

Results

Study Selection and Characteristics

Exports contained 803 records before deduplication (2 within-export and 9 cross-export duplicates removed); 792 unique records were screened, 317 advanced to full-text pursuit, 277 full texts were obtained and assessed, and 241 sources met final inclusion criteria (Figure 1). All 241 included studies are listed in Multimedia Appendix 2; not every included study is discussed individually in the narrative, and the main reference list also contains methodological and contextual sources. Of the 241 included sources, 176 (73%) were journal articles, 29 (12%) conference proceedings, and 35 (14.5%) preprints; the venue of one source could not be identified (Multimedia Appendix 2). The preprint share reflects the rapid AI publication cycle and warrants cautious interpretation of non–peer-reviewed findings. Across the 241 included papers, 77 flagged quantitative human participant outcomes; a subset with interpretable clinical endpoints, between-group or pre-post contrasts, and sufficient descriptive reporting was carried into the quantitative synthesis below.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram of source selection. Duplicates were removed by normalized-title matching, within the API-accessed export (2 records) and across streams (9 PsycINFO records matching an API-accessed title). Counts: 803 identified → 792 screened → 317 sought at full text → 277 assessed → 241 included.



Quantitative Synthesis: Clinical and Psychological Outcomes

Design Strength and Causal Inference

Across quantitatively reporting studies, effect magnitudes systematically covary with design rigor: studies with less stringent causal controls report larger apparent improvements, whereas designs using concurrent active comparators more often report modest or null between-group differences. This pattern contextualizes individual findings rather than diminishing them. Sharma et al (2024) [21] provide real-world field study evidence at a scale of 43,347 participants who initially consented on a mental health website: only 15,531 (35.83%) completed the outcome survey, and the pre-post emotion-intensity analysis was limited to the 1,922 participants who provided both ratings—just 4.4% of the original enrolled sample—among whom 1,300 (67.64%) reported a reduction in emotion intensity [21]. Because this figure derives from roughly 1 in 23 enrollees, it substantially overestimates population-level benefit. Hatch et al (2025) [22] reported ChatGPT passages rated higher on common therapy factor dimensions than human-written passages (Cohen*d*=1.63) in a randomized blind comparison of written text [22]—a finding reflecting evaluator judgments of static prose quality rather than clinical outcomes in dynamic client care. When considered alongside active-comparator trials, interpretation becomes more nuanced: Vowels et al (2025) [23] found no significant group-by-time advantage on the World Health Organization-Five Well-Being Index for a GPT-4o chatbot versus an active writing control (n=258) [23], while Mo et al (2025) [24] reported a significant between-group advantage for a GPT-4 problem-solving chatbot over an active control arm in 100 help seekers [24]. These contrasting randomized controlled trial (RCT) findings underscore that effects are context- and design-dependent and that analogue ratings or completer-only analyses should not be read as controlled evidence of clinical benefit.

Transparency, Disclosure, and Perceived Empathy

When studies experimentally varied AI-authorship disclosure, a consistent and policy-relevant signal emerged: perceived empathy declined once AI authorship was made salient. Liu et al (2026) [25] randomized 638 patients in a radiology clinic subtrial to radiology health care providers alone or supported by a fine-tuned LLM chatbot; patients in the AI condition reported significantly lower empathy—a finding requiring cautious interpretation given unblinded self-report in an open-label comparison, meaning replication in disclosure-controlled designs is needed. Shen et al (2024) [26], in crowd-sourced comparative studies (total n=985), found that participants reported significantly lower state empathy toward

AI-written than human-written stories across most conditions. Together, these disclosure-manipulation designs create an important unresolved governance question: modern ethical frameworks require transparency, yet transparency appears to reduce perceived warmth in some contexts—a tension that remains under-modeled in the current literature. Factorial trials crossing disclosure with content quality in ecologically richer mental health tasks are needed before this signal is treated as robust.

Task Support, Well-Being, and Symptom-Level Outcomes

The clearest causal signal in the RCT literature comes from Mo et al (2025) [24], whose GPT-4 problem-solving chatbot demonstrated a significant between-group advantage on problem recognition and resolution—replication across populations and timeframes would further strengthen confidence in generalizability. Held et al (2025) [27] reported modest pre-post 4-item Patient Health Questionnaire (PHQ-4) reductions in a single-arm GPT-4 cognitive behavioral therapy (CBT) feasibility study (n=61) without a control arm; a within-subject comparison of ChatGPT-4o and Llama-3–based problem-solving therapy chatbots for family caregivers similarly reported gains in empathy and therapeutic alliance (n=28) [5]; Izumi et al (2024; available as a preprint) [28], in a crowd-sourced evaluation (n=105) of generated CBT responses, found that GPT-4–generated responses were rated as producing significantly greater mood change than those of a baseline model—a rating of generated text rather than a clinical outcome. Process and outcome signals from CBT and internet-delivered CBT further specify the picture: Furukawa et al [29] integrated a T5 model into a smartphone CBT platform (n=1626; 4369 thought-feeling records), achieving 73.5% accuracy in predicting self-reported feelings from automatic thoughts (F_1-score≈0.74), with users engaging in correctly predicted cognitive restructuring pairs experiencing significantly greater negative-feeling reduction. Garcia et al [30] analyzed internet-delivered CBT language from 44 adolescents, reporting pre-post Beck Anxiety Inventory improvements (25.4-16.3; Cohen d=0.71) and Beck Depression Inventory-II (BDI-II) improvements (26.5-15.68; Cohen d=1.10), with language-derived measures correlating with rating-scale change at modest-to-moderate levels (valence-derived BDI change: r=0.44, 95% CI 0.17-0.66). Across this literature, a consistent pattern holds: large uncontrolled deployments report strong statistics among completers, and rigorous controlled trials more often report modest effects—each design type contributing unique evidentiary value.

Assessment, Measurement, and Prediction

A substantial portion of quantitative LLM research in mental health concerns measurement validity and predictive modeling rather than symptom change in trials. Shin et al [31] validated a fine-tuned GPT-3.5 on 428 diary entries, achieving 90.2% accuracy and 95.5% specificity for depression risk against 9-item Patient Health Questionnaire (PHQ-9) and Beck Scale for Suicide Ideation thresholds (GPT-4 showed very high recall at 0.972 but lower overall accuracy at 0.743)—illustrating screening feasibility while underscoring the need for external validation, careful base-rate consideration, and clinically chosen decision thresholds. Substance use disorder severity has been decoded from clinical notes using an LLM [32]. Related rule-based approaches combining classifications with GPT-2–generated CBT-aligned responses have also supported early detection and personalized mental health intervention [33]. McCoy et al [34] deployed GPT-4 to extract research domain criteria–based dimensional psychopathology from clinical notes of 3059 youth in psychiatric emergency care, with logistic regression models demonstrating that LLM-derived scores significantly predicted hospital admission likelihood. Assessment extends to higher-acuity questions: Rubin et al [35] found that structured LLM-driven follow-up questioning increased the proportion of adults (n=110) judged to meet the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) posttraumatic stress disorder (PTSD) criterion A from 65% to 86%, while LLM-clinician agreement remained limited (GPT-4 Cohen $\varkappa$=0.27 for first descriptions; range 0-0.44)—supporting LLMs as elicitation aids rather than autonomous adjudicators. In a related trauma application, Giuliani et al [36] used transformer models to predict PTSD symptom severity from children's and adolescents' trauma narratives. Pugh et al [37] reported moderate correlations between model-generated and expert ratings of thought-disorder dimensions in schizophrenia transcripts (supervised baseline r=0.60; GPT-4 mean r≈0.49), with cross-run consistency critically dependent on model and temperature (Llama 3: r=0.94-1.00 at temperature=0; GPT-3.5: r=0.70 at temperature=0.5). Hur et al [38] demonstrated that ChatGPT-derived sentiment scores from brief narratives predicted 3-week PHQ-9 change comparably to human raters; Dougherty et al [39] showed LLM-derived sentiment features from psilocybin therapy transcripts predicted Montgomery-Åsberg Depression Rating Scale responder status with 85% to 88% accuracy (area under the curve 85%-88%). Predictive and classification modeling extends to online and social media text: Zhang et al [40] applied ChatGPT-3.5 to 580 online help-seeking narratives, achieving 91.2% accuracy in identifying violence and controlling behaviors against human expert coding; van Rijswijk et al [41] showed that LLM-scored self-determination theory deficits from 2214 Reddit life stories statistically mediated the link between neurodivergence and negative well-being; Oluwalade et al [42] used a fine-tuned transformer model to characterize sentiments and emotions in social media narratives of pediatric patients with multiple long-term conditions; and Perogamvros et al [43] demonstrated that as LLM-identified cathartic dream frequency increased during therapy, clinical depression scores significantly decreased.

Psychometric Scale Development and Automated Scoring

LLMs are increasingly deployed within the psychometric lifecycle. Erdemir and Atik [44] used GPT-4 to develop and validate the 40-item AI-Digital Life Balance Scale across 3 samples (n=1184), confirming a 6-factor model explaining 60.83% of variance (Cronbach α=0.68-0.87). ChatGPT-4's

capacity to estimate personality traits has also been evaluated [45]. Perrin et al [46] demonstrated a hybrid GPT-4 and community-stakeholder methodology to draft the Internalized Ableism Inventory, validated on 409 adults with a stable 8-subscale structure (Cronbach α=0.98). Schroeders and Achaa-Amankwaa [47] constructed a 110-item vocabulary assessment validated via 3-parameter item response theory on 1052 adults, with linguistic features explaining 61.2% of variance in empirical item difficulty. For human-AI interaction, Yang and Oshio [48] validated attachment anxiety and attachment avoidance as primary dimensions; Yu et al [49] developed the Problematic ChatGPT Use Scale on 1040 adults, linking high scores to usage time and depression; and Maral et al [50] validated a Turkish adaptation (n=864), revealing psychological distress and self-control as significant mediators of well-being. Krumm et al [51] found a refined 8-item AI-generated situational judgment test achieved acceptable fit comparable to a human-authored parallel scale. For automated scoring, Wang et al [52] introduced ScaleLLM, an unsupervised architecture improving health scale interpretation accuracy relative to conventional methods; Sokol and Goodman [53] extracted future self-continuity from 164 transcripts using Claude 3.5 Sonnet (convergent validity r=0.57; intermodel reliability r=0.84-0.91, predicting perceived suicide likelihood beyond the traditional scale); Lyu et al [54] automated scoring of the Ambiguous Intentions Hostility Questionnaire, accurately replicating human expert scores and known traumatic brain injury–control group differences; Li and Zhao [55] found LLMs evaluating body image from 194 student texts achieved r=0.722 on perception and reduced extreme-score error by 48.4% versus human raters; and Strojny et al [56] demonstrated HerBERT-based gaming disorder prediction from open-ended responses of 417 participants (r=0.476).

Computational Benchmarking and Diagnostic Evaluation

Benchmarking establishes that state-of-the-art LLMs carry extensive clinical knowledge: Hanss et al [57] documented a performance leap across generations on 150 psychiatry questions (GPT-3.5: 87/150, 58.0%; GPT-4: 126/150, 84.0%; GPT-4o: 131/150, 87.3%; P<.001), with response consistency strongly predicting factual accuracy (r=0.682); Neubauer et al [58] found Gemini 2.0 Flash achieving 76.3% accuracy on child and adolescent psychiatry items with statistically significant domain variation; and Siddiqui et al [59] found ChatGPT-4o largely met high-quality documentation thresholds on diverse medical and mental health queries. Depression assessment from interview transcripts achieved state-of-the-art 8-item Patient Health Questionnaire (PHQ-8) root mean square error of 4.67 (mean absolute error [MAE] 3.80) via prefix tuning in 189 subjects [60]; GPT-4 achieved 95.3% sentiment classification accuracy on 6169 Cantonese web-counseling messages [61]; and a fine-tuned GPT-4.1 classified microcounseling skills with an accuracy of 0.78 and Cohen ϰ value of 0.69 [62]. Social-emotional intelligence benchmarks are striking: ChatGPT-4 matched human performance on the Reading the Mind in the Eyes Test (scores 26-27; significantly better than random; P<.001) [63]; in a direct comparison with 180 psychology students, ChatGPT-4 social intelligence scores exceeded 100% of the human cohort [64], and multilabel depression-emotion classification from Reddit posts achieved macroaveraged F_1-score of 0.76 with transformer architectures [65]. ChatGPT has also been explored for psychodynamic case formulation, generating outputs that expert psychiatrists rated as clinically appropriate when prompts incorporated specific theoretical concepts [66].

Critical limitations temper these results. ChatGPT analytic thinking scores showed only small meta-analytic correlations with validated linguistic inquiry and word count metrics (0.058<r<0.304) and committed mathematical errors in 22% (3561/16192) of cases [67]; a supervised Bidirectional Encoder Representations from Transformers (BERT) model outperformed GPT-4o on a sentence-level misunderstanding task (Matthews correlation coefficient=0.69 vs 0.47) [68]; and LLMs systematically underperformed traditional regression models in well-being prediction for 64,000 individuals across 64 nations, with the largest errors for underrepresented regions [69]. In high-acuity contexts, an open-source LLM achieved 74.5% diagnostic accuracy on English psychiatric notes but only 59% on Korean originals, with a substantially higher hallucination rate (30.2% vs 13.4%) [70]. Models evaluating suicidal ideation scenarios performed at levels comparable to master's-level counselors but demonstrated a statistically significant upward bias in evaluating clinician responses [71]. Encouragingly, a simulation of 449 conversations found no statistically significant bias in an LLM agent's emotional tone or engagement across patient age, sex, race, or income [72], and constraining an LLM to generate exclusively from validated clinical guidelines yielded high reliability and evidence traceability [73]—illustrating that architectural safeguards can meaningfully mitigate performance gaps.

Experimental and RCT-Based Evidence

Experimental vignette studies reveal generation-related performance differentials with direct clinical implications. Levkovich and Elyoseph [74] found ChatGPT-4 suicide risk evaluations closely mirrored those of 379 mental health professionals (average Z=0.01), while GPT-3.5 systematically underestimated risk (Z=–0.83); a follow-up study showed GPT-4 integrating culturally specific epidemiological data, assigning higher risk scores to South Korean versus Greek vignettes [75]; suicidal ideation detection has also been addressed through LLM-based data augmentation techniques [76]. A related case-vignette study similarly compared AI, clinician, and public perspectives on schizophrenia recovery prognosis [77]. Perlis et al [78] found a guideline-augmented GPT-4 identified the optimal bipolar depression treatment in 50.8% of 1000 vignettes—superior to both unaugmented LLM (23.4%) and community clinicians (23.1%; McNemar P<.001). Campellone et al [79] found a generative AI arm achieved 98% empathetic detection accuracy (vs 69% for a rules-based system) with 100% technical guardrail success in a 160-adult exploratory RCT. Wang et al [80] showed in a multicenter double-blind RCT (n=388) that an AI-assisted chatbot reduced health care worker follow-up time

from 7.85 to 2.34 minutes without compromising patient satisfaction, while a narrative AI intervention significantly lowered Generalized Anxiety Disorder-7 scores (7.06 vs 9.95). Terblanche et al [81] found a CBT-based chatbot outperformed solution-focused and Goal, Reality, Options, and Will–model alternatives on goal attainment and working alliance in 592 participants, challenging psychotherapy's principle of model equivalence. Wang and Farb [82] found mindfulness- and value-focused chatbots improved well-being in postsecondary students primarily through decentering. A 41-counselor RCT showed ChatGPT practice sessions significantly increased counseling self-efficacy and reduced state anxiety [83]. A 543-participant RCT of a GPT-4o goal-setting chatbot found significant improvements in goal specificity and implementation intentions, though long-term maintenance was not assessed [84].

Cross-Sectional and Observational Evidence

Cross-sectional surveys reveal rapid, often clinically meaningful adoption alongside psychosocial risks. Rousmaniere et al [85] found 48.7% of US adults with ongoing mental health conditions actively used LLMs for psychological support (targeting anxiety: 73.3%; depression: 59.7%), with 63.4% reporting explicit mental health improvement and 37.8% finding AI more beneficial than human therapy. Li et al [86] confirmed that performance expectancy, price value, and social norms predicted both initial adoption and continued use across 393 symptomatic adults, while effort expectancy and AI hesitancy served as deterrents. Aral et al [87] documented significant disciplinary divides across 166 child and adolescent mental health professionals: psychiatrists scored higher than psychologists on viewing LLMs as patient-facing tools (Cohen d=0.68) and for crisis prevention (Cohen d=0.69), whereas psychologists rated bias concerns more favorably (Cohen d=1.56). Kelly et al [88] demonstrated statistically significant reluctance to disclose personal mental health information to ChatGPT compared to general health inquiries, presenting a critical bottleneck for training data quality. Adverse behavioral correlates are consistently documented: in a survey of 2709 students, compulsive ChatGPT use positively correlated with loneliness, social avoidance, and psychological distress, with distress mediating reduced life satisfaction and academic performance [89]. Gu and Zhang [90] found AI self-efficacy and perceived media empathy contributed to AI addiction and cyberchondria, and Salah et al [91] demonstrated that trust, general perceptions, and stereotype perceptions of ChatGPT directly impacted user self-esteem, with job anxiety moderating the well-being effect.

Thematic Synthesis

Overview

The following synthesis integrates perspectives from computational engineering, informatics, and mental health to situate technical findings within behavioral and public health contexts.

Theme 1: Clinical Applications—An Evolving Paradigm

Overview

The integration of AI into mental health reflects a phased progression—from passive symptom detection to interactive emotional support to structured evidence-based psychotherapy delivery—with early frameworks emphasizing embedded ethical oversight from the outset [1,2,4,92].

Symptom Identification and Risk Detection

Early applications leveraged transformer architectures—particularly BERT variants—for passive text analysis of psychiatric symptoms in digital footprints and clinical transcripts. Natural language processing (NLP) models mapped linguistic markers to psychiatric taxonomies (eg, tracking 4-fold pandemic-related stress increases on Twitter [93] and identifying linguistic dimensions of disconnection and burdensomeness in suicidality forums [94]). Benchmarking under controlled laboratory conditions yields high reported accuracy (85%-95% F_1-score across multiple studies, largely reported in conference proceedings or lower-visibility venues [95-100]); these figures reflect controlled evaluation and should not be assumed to represent real-world clinical performance, where benchmark-deployment gaps are well documented. Fine-tuned models achieved 91.53% average precision in emotion recognition from psychotherapy transcripts [101], and a LLaMA 3.1-70B framework achieved state-of-the-art depression assessment at MAE 2.91 against clinical benchmarks [102]. However, passive detection alone is insufficient for clinical intervention, motivating the shift toward interactive systems.

Conversational Support and Psychoeducation

The interactive era—ushered in by the release of ChatGPT—raised critical questions about therapeutic alliance, trust, and perceived empathy, including its use as a direct psychoeducational tool [103]. Empirical evaluations yielded surprising results: advanced LLMs achieved near-human or superhuman-perceived empathy on the Therapist Empathy Scale in preliminary evaluations (one available as a preprint) [104,105], and a preregistered study found ChatGPT rated higher on common therapy-factor dimensions than human responses (Cohen d=1.63) [22]—reflecting evaluator judgments of written text quality, not dynamic clinical outcomes. Users frequently value chatbots as nonjudgmental 24/7 spaces that reduce stigma-related access barriers [106,107], and a large-scale deployment study found higher-empathy AI reframes preferred 55.7% more than lower-empathy alternatives [108]. Qualitative analyses, however, consistently reveal a "lack of emotional depth," "reassurance loops" that fail to challenge maladaptive behaviors, and dismissive responses [109-113]. Emerging platforms such as MindBridge report promising early engagement figures (up to 80% of users reporting reduced anxiety and depression symptoms after 15- to 20-minute sessions) [114], and similar lightweight implementations are also emerging (eg, BeYouBot) [115,116], though these rely on uncontrolled self-report

and require replication. Hallucinations and inadequate crisis escalation remain critical dangers: LLMs frequently fail to respond appropriately to suicidality or psychosis, with human therapists responding correctly 93% of the time versus LLMs exceeding 20% inappropriate response rates [117,118]. Generative AI applied in co-design workshops with adolescents in low-resource settings has shown feasibility for developing digital mental health content [119], and its use as a standardized patient in clinical training simulations has been shown to reduce student negative affect and increase perceived professional competence [120].

Structured Psychotherapy Delivery

Translating LLMs from generic emotional support to structured interventions (CBT, problem-solving therapy, and motivational interviewing) typically requires supervised fine-tuning (SFT), retrieval-augmented generation (RAG), and chain-of-thought (CoT) prompting to maintain protocol fidelity [121-123]. Domain-specific models (CBT-LLM, CBT-augmented Counseling Chat Corpus–trained agents, DiaCBT) outperform baselines on specific CBT skills such as guided discovery and cognitive restructuring [124-126]. Process-based therapy has been proposed as a complementary conceptual framework for structuring such interventions [127]. RCT evidence supports measurable benefits in structured task-focused contexts [24,84] alongside null findings on broad well-being indices [23]. LLMs perform well at reciting CBT knowledge but lag behind expert clinicians on complex real-world scenarios requiring deep patient-specific cognitive analysis (level III deliberate practice) [128,129], and while AI can train therapists by acting as accurate simulated patients [130], experts caution that LLMs currently lack capacity for the transformative unconscious work required in sophisticated psychotherapy [1,4,131]. Related work has extended early maladaptive schema prediction for mental health forum questions [132] and scoped AI-delivered motivational interviewing [133].

Specialized and Multimodal Adaptations

As clinical utility expands, specialized adaptations address niche populations and modalities including postpartum mental health care support [134], vocational guidance for people with schizophrenia spectrum disorder [135], sexual health information delivery [136], companionship-oriented use addressing loneliness and social anxiety [137], and AI-generated bedside art interventions [138,139]. For neurodivergent users (particularly attention-deficit/hyperactivity disorder), LLMs provide valued metacognitive scaffolding and task-breakdown support but carry documented risks of cognition outsourcing and dependency [110,140]. Cultural sensitivity remains critically underdeveloped, with severe underrepresentation of non-Western contexts in LLM-based CBT research [141]; targeted cultural prompting and fine-tuning for specific languages consistently improve motivational and triage relevance [142,143]. Multimodal integrations combining text with physiological signals, speech emotion recognition, and music therapy demonstrate initial feasibility but unproven clinical efficacy [144,145]. In complex medical environments such as psycho-oncology and radiology, fine-tuned LLMs reduce provider burnout and improve dialogue quality [25], and broader narrative reviews document AI's emerging role in breast cancer psychological support [146], though they still struggle with grief-laden interactions requiring deep emotional regulation. Ultimately, the clinical evolution of LLMs—from novelty toward structured adjunct—remains tethered to human oversight, rigorous governance, and continuous cultural calibration [11,147-152].

Theme 2: Technical Frameworks—Fine-Tuning, RAG, and Prompt Engineering

General-Purpose Model Performance

Initial evaluations of zero- and few-shot generalist capabilities found baseline empathy and clinical reasoning to be surprisingly robust [153]. Comprehensive evaluation of 127 open-source models found that newer architectures and larger parameter scales consistently yielded superior simulated clinical performance, while mental health–specific fine-tuning conferred no consistent statistical benefit and occasionally degraded performance (available as a preprint) [154]; lightweight generalists (Alibaba's Qwen2.5-3B and Microsoft's Phi-3-Mini) outperformed domain-specific systems (MentalHealthBot-7B) in perceived empathy and contextual understanding in a preprint evaluation [155]. A recurring user-level pattern is that less heavily tuned, out-of-the-box models are perceived as more conversationally natural than rigid specialist systems—despite generalists' vulnerability to hallucination and generic textbook-style responses under complex clinical pressure.

Clinical SFT

When generalists showed insufficient reliability for structured protocols, SFT was applied to open-source foundation models (LLaMA, Mistral AI's Mistral, and Qwen) using synthetic therapy transcripts. Mental-Alpaca and Mental-FLAN-T5, despite being 15 to 25 times smaller, outperformed GPT-3.5 best-prompt designs and edged GPT-4 by 4.8% on balanced accuracy for mental health prediction [156]. Subsequent domain-specific models—MeChat (PlayMe Studio; fine-tuned on 55,165 multiturn dialogues) [157] and DiaCBT (Qwen-based, improving session length and CBT skill mastery) [126]—illustrate how SFT encodes clinical structure at the cost of "overmedicalization," strategy bias, and reduced conversational flexibility. Complementary approaches include foundational GPT-2 CBT proof-of-concept work [158], BERT and RoBERTa joint pipelines for emotion detection and CBT-aligned response generation [159], multiclass mental health condition classification systems [160,161], motivational interviewing [MI]–aligned fine-tuned architectures [162,163], specialized tokenization reducing generation costs by 60% [164], bespoke open-domain transformer designs for multiturn clinical dialogue [165], parameter-efficient fine-tuning methods (low-rank adaptation and quantized low-rank adaptation) achieving high BERTScore metrics [166,167], emotion-focused therapy multiagent CoT frameworks demonstrating superior empathetic depth and somatic awareness [168], and emotion-aware therapeutic music generation [169]. The AFlow framework introduced

fine-grained affective supervision directly into the training loop, anchoring token generation to a continuously updated affective state variable to improve multiturn emotional support quality [170].

Retrieval-Augmented Clinical Grounding and Prompt Engineering

RAG constrains LLM outputs to verified external knowledge, functioning as a de facto safety layer: Patil et al [171] demonstrated that augmenting LLaMA3-8B with an Elasticsearch (Elastic) vector database improved response relevance (Hit@5: ~78.9% vs ~50.4% with baseline lexical search). The MIND-SAFE framework anchors outputs to CBT and MI manuals while cross-referencing user state data [172]; a comparative RAG-based CBT system has also reported response-quality improvements of 15.4% over baseline [173]. Constraining an LLM to generate exclusively from validated national clinical guidelines yielded high reliability and evidence traceability [73], and a human-guided empathetic agent combining reinforcement learning with RAG has been evaluated technically, without human participants [174,175]. Prompt engineering complements RAG through CoT approaches—forcing the model to map clinical reasoning before generating user-facing text—improving analytical and empathic response quality [176-178]. Adaptive prompting systems dynamically adjusting therapeutic persona to user real-time state outperformed basic prompts in structured comparisons [179], "Semantic anchoring" prevents LLMs from drifting into stereotypic tangents during personality-matching tasks [180], and problem-solving therapy–specific prompting strategies illuminate user-experience trade-offs in GPT-delivered problem-solving [5,181]. Multiple-choice prompting and yes-or-no prompting structures extract early maladaptive schemas with higher fidelity [182], semantic network analysis ensures AI framing of "anxiety" aligns with human cognitive structures [183], and LLM-assisted Motivational Interviewing frameworks successfully manipulate psychological safety to increase user self-efficacy [184]. AI-authored cognitive bias modification scenarios produced ratings in the intended emotional direction, though not statistically equivalent to human-authored versions [185]. Commercial LLMs possess distinct measurable "personalities" that influence therapeutic style (eg, Claude 3 Opus as introverted, intuitive, thinking, and judging; Gemini Advanced as introverted, intuitive, feeling, and judging) [186]. LLM Therapy Governance Model (LTGM) architectures add a "Prompt Coach" and "Disclosure Monitor" layer to intercept and refine prompts before LLM processing [187], and cultural and spiritual nuance can be injected via CoT prompting without expensive retraining [188,189].

Theme 3: Evaluation and Metrics

Limits of Traditional NLP Metrics and Automated Evaluation

Traditional NLP metrics (bilingual evaluation understudy, recall-oriented understudy for gisting evaluation [ROUGE], and BERTScore) do not measure empathy, safety, or clinical appropriateness; a system may score highly on ROUGE while producing contextually unsafe responses [190-194]. Scoping reviews found the vast majority of studies relied on non-standardized ad hoc scales, with only a fraction evaluating foundational constructs like safety, privacy, and fairness [1,9]. Depression detection deep learning approaches achieved high statistical performance (r=0.741) that does not translate directly to clinical safety [195]. Adjacent applied evaluations extend to psychometric tool development for anomalous perceptual experiences [196], AI-assisted cross-cultural scale translation [197], LLM-assisted thematic analysis of caregiver online narratives [198], and qualitative study of LLM-based conversational support for body image concerns [199]. "LLM-as-a-Judge" frameworks offer scalability—the MindEval benchmark, using Claude-Sonnet as judge and Claude-Haiku as simulated patient, found, in a preprint evaluation, that all frontier systems scored below 4/6 on average, with performance deteriorating for severe symptom profiles or conversations exceeding 20 turns [200]—but introduce systematic evaluator bias: LLMs exhibit distinct value profiles (prioritizing universalism and self-direction; de-emphasizing security and tradition) that predict their clinical judgments rather than reflecting objective clinical standards [201].

Human-Centered Assessment

Blind vignette experiments found no statistically significant differences between AI-generated and human-written mental health vignettes on any perceived quality dimension—including realism, clarity, emotional impact, or diagnostic likelihood [202]. However, analysis of live multiturn interactions reveals important limitations: chatbots overcompensate with substantially more words (mean 1,414.9 [SD 878.1] vs human mean 392.5 [SD 163.3]) and rely heavily on reassurance and psychoeducation rather than the elaboration-evoking strategies and strategic silence characteristic of human therapists [203]; ChatGPT generates higher interaction involvement and politeness scores than human providers but lacks the reciprocal adaptability that human response quality shows in response to deeper user self-disclosure [204]. While 54% (13/24) of participants found ChatGPT nonjudgmental and empathic, diagnostic accuracy and deep emotional engagement remained limited [193], and cognitive network analyses comparing students with LLM-based digital twins suggest that LLMs lack the context-sensitive, experience-based semantic networks of human respondents (available as a preprint) [205]. Anthropomorphization fosters overreliance and erodes critical judgment [10]; commercially available companion AIs frequently fail to recognize user crises and generate negative consumer reactions [206]; and

co-design of evaluation metrics with service users—operationalized in the Adult Session Wants And Needs Outcome Measure scale (scale-level content validity index=0.91) [147]—and ChatGPT assistance for clinician assessment of medical and psychiatric questions [207] highlight the need to center patient-defined outcomes. User trust, though initially high, does not translate into behavioral compliance; studies consistently document high initial engagement followed by significant adherence decline attributed to insufficient emotional depth and the absence of relational accountability central to therapeutic change [208-210].

Theme 4: Challenges and Limitations

Clinical Safety and Crisis Response

LLMs operate on statistical pattern recognition rather than clinically grounded reasoning, generating fluent text that may mask critical failures in high-stakes situations [211,212]. The most immediate danger is hallucination: GPT-4o fabricated citations in nearly 20% of psychiatric literature synthesis outputs (28%-29% for specialized topics such as binge eating disorder and body dysmorphic disorder) [213]. An LLM tasked with generating a mental health nursing care plan for self-harm produced output containing significant clinical errors and ethical issues, underscoring the potential for iatrogenic harm without expert oversight [214]. In simulated adolescent crisis vignettes, therapy and companion bots endorsed harmful suggestions in 32% of opportunities [215]; human therapists responded appropriately 93% of the time versus LLMs exceeding 20% inappropriate responses [117]. Contemporary ChatGPT variants were less effective than older fine-tuned models for dialogue safety detection in mental health support contexts [118]. The "black box" nature of deep learning models hinders clinical trust; explainable AI frameworks applying Shapley additive explanations and local interpretable model-agnostic explanations to interpret transformer-based detection address but do not resolve this barrier [216]. A key structural concern is that LLM-generated empathy, while behaviorally convincing, may lack sufficient semantic grounding to represent the urgency of suicidal or psychotic crises, leading to missed critical cues and clinically unsafe responses.

Algorithmic Bias, Therapeutic Deception, and Cognitive Outsourcing

LLM outputs reflect and amplify patterns in training data. Gender stereotyping is documented: GPT-3 consistently assigned competence to male-generated text and warmth to female-generated text [217]. GPT-4 offered significantly lower empathy to Black and Asian users than White users in a preprint evaluation [218]. Llama 3.1 exhibited stigmatizing attitudes 75% of the time toward severe mental illness [117], and a significant gender effect in how ChatGPT-4 evaluates clinical vignettes has been empirically confirmed [219]. LLMs exhibit systematic "appraisal bias," consistently struggling to assess shame or fear and misinterpreting power dynamics [220]. Training datasets can embed harmful body-image stereotypes that generative models amplify [221,222], and frontier models can be prompted to malinger convincingly on validated instruments like the Minnesota Multiphasic Personality Inventory-2 while existing validity scales retain some detection capacity [223]. As LLMs become more fluent, therapeutic deception—users projecting genuine intentionality and care onto statistical systems—raises the risk of emotional dependency, overdisclosure, and retraumatization when the AI responds with a hallucinated output [12,26,215]. When an adverse outcome occurs, physicians who followed AI advice receive substantially less moral blame (diffusion of responsibility) [224], and LLMs evaluated on bipolar depression ethical dilemmas show notable gaps in nuanced clinical-ethical situations [225]. Even trained clinicians struggled to differentiate AI-generated from human therapeutic responses [218], raising unresolved legal accountability questions when AI causes harm [149,226]. An equally critical challenge is cognitive outsourcing: frictionless AI experiences may truncate the cognitive effort therapeutic growth requires, eroding metacognitive capacities—self-monitoring, reflective appraisal, and adaptive self-correction—that sustained behavioral change demands [110,140,179].

Privacy and Regulatory Compliance

Mental health interventions require divulgence of deeply sensitive information that, in LLM contexts, becomes "data exhaust" transmitted to remote corporate servers. Compliance with the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) is a primary integration bottleneck [227]; the conversational nature of chatbots encourages overdisclosure and erosion of therapeutic boundaries [211,215], and marginalized populations relying on free public AI tools face disproportionate privacy exposure due to the digital divide [12,149].

Theme 5: Governance and Future Directions

Overview

Four governance directions emerge from the thematic synthesis, and each responds to a specific class of risk evidence documented above; because this is a scoping review without a formal certainty-of-evidence assessment, we frame them as risk-derived candidate priorities rather than validated standards. First, human-in-the-loop oversight responds to evidence of inappropriate crisis responses [117,215], hallucinated or fabricated content [213], and therapeutic deception, in which users attribute genuine care to statistical systems [12,26]: it is prudent to conceptualize AI as a copilot, with diagnostic and therapeutic authority reserved for licensed professionals. Survey data indicate 72.6% of researchers report efficiency improvements using LLMs, yet 78.6% encountered inaccurate or misleading responses [7]; current models cannot reliably detect high-acuity risk in marginalized youth [2]. Second, institutional safety standards—industry-wide LTGMs spanning predeployment clinical evaluation, continuous algorithmic auditing, and disclosure protocols—respond to documented safety failures (eg, the Tessa chatbot [Cass, formerly X2AI], withdrawn after it began providing harmful body image and eating disorder–related advice [228]) and to the absence

of accountability and disclosure frameworks [149,226] and would help align deployment with bioethical principles [215]. Third, multimodal integration responds to evidence that text-only systems lack emotional depth [109-111] and remain culturally undercalibrated [141]: combining text with audio, visual, and physiological signals is a technical frontier for deeper computational empathy, with the potential to let models perceive vocal prosody, physiological arousal, and psychomotor signs inaccessible to text-only architectures [6,229]. Fourth, scalable, equitable access responds to evidence that research and model performance are concentrated in Western, well-resourced settings—with the largest well-being prediction errors for underrepresented regions [69] and severe underrepresentation of non-Western contexts in LLM-based CBT research [141]—and to evidence of unmet need [230]. This most compelling opportunity is conditional upon the above safeguards; proposed use cases include culturally adapted CBT for youth in low- and middle-income countries [2] and 24/7 localized support in regions with limited psychiatric infrastructure [6].

Thematic Patterns in the ProQuest-Indexed Subset

Records from the ProQuest PsycINFO stream were screened through the same eligibility framework (with the additional empirical-relevance refinement noted in the Methods section). Automated descriptive coding of the included PsycINFO abstracts shows 2 frequency patterns consistent with the thematic synthesis above. First, depression and general mental health are the most frequently addressed clinical targets, with comparatively fewer studies in severe or high-acuity psychopathology. Second, the ChatGPT model family (GPT-3.5, GPT-4, and GPT-4o) is the most frequently evaluated system, appearing more often than open-source or domain-specific models. These are descriptive frequency patterns in the indexed abstracts, not measures of effect or study quality.

The subset also spans a range of exploratory applications—for example, psychoeducation and communication-style evaluations, sensitive-information delivery, substance-use severity extraction from clinical text, digital companionship, and therapeutic-boundary questions—alongside methodological work addressing data scarcity. Study-level detail for every included source, including venue type, is provided in Multimedia Appendix 2. Taken together, these patterns are consistent with the broader synthesis: generative models show linguistic fluency and prediction capability in controlled, text-based studies, while their safety and efficacy in higher-risk deployment settings remain experimentally immature [215].

Discussion

Principal Findings

The primary insight from this review is a nuanced empirical picture: LLMs show substantial advances in conversational fluency while also presenting important clinical, ethical, and architectural considerations. No stable generalizable "LLM effect" emerges from the synthesized evidence; outcomes are closely tied to study design, disclosure conditions, and implementation context. Against the quantitative findings—variation by design strength, disclosure-related empathy penalties, task-specific versus broad well-being gains—the thematic synthesis clarifies behavioral and regulatory dimensions that outcome statistics alone do not capture.

RQ1: Outcomes and Efficacy Versus Traditional Interactions

When we ask whether LLM-based technologies can match or routinely outperform traditional human-to-human therapy, the converged evidence demands a cautious, highly contextualized answer [231-233]. For contextual orientation, face-to-face CBT demonstrates moderate-to-large effect sizes (d≈0.80-1.10) for depression and anxiety [234]; rigorously controlled LLM trials have not yet consistently approached these benchmarks across broad clinical populations, more often showing null results on broad well-being indices and narrower improvements on structured tasks. The thematic synthesis explains why: adapting a generalist AI into a clinically useful digital therapist requires substantial engineering (SFT, RAG, and CoT prompting), and even specialized models face unresolved trade-offs between human-like interaction and clinical safety. Comparing LLMs directly to human therapists remains premature, and the field continues to navigate these tensions rather than having resolved them.

RQ2: Relational and Process-Level Dynamics

RQ2 focuses on interaction-level mechanisms—alliance formation, perceived empathy, trust calibration, and engagement trajectories [235-242]. In the short term, the absence of judgment and continuous availability often trigger therapeutic deception: users anthropomorphize the system, projecting genuine empathy onto a statistical model, creating low-barrier initial encounters and reduced stigma [6,106]. Yet this low-friction environment may constrain long-term psychological growth. Qualitative analyses consistently report high short-term engagement followed by adherence decline due to "lack of emotional depth"; without a genuine internal emotional state, the AI cannot hold a user accountable or sit with them in shared discomfort [109-111].

We propose that the commercial design priorities favoring agreement and minimizing emotional discomfort may be misaligned with key active ingredients of psychotherapy. Therapeutic growth can involve navigating discomfort, experiencing constructive challenge, and practicing relational skills within clear therapeutic boundaries. Building on these principles, we propose the "Friction of Therapy" hypothesis: primary reliance on AI companions that prioritize agreement over constructive exploration may limit opportunities for relational learning, particularly when their use displaces human interaction. Whether users who rely primarily on a perpetually validating algorithmic companion develop reduced capacity to navigate the motivational demands and

natural frictions of real human relationships remains a pressing, largely unanswered empirical question for longitudinal evaluation [6,106,109-111].

RQ3: Integration Into the Broader Self-Help Industry

Rapid commercialization has accelerated deployment of direct-to-consumer wellness applications—including in severe mental health contexts—with limited regulatory oversight. Data privacy risk is central: mental health interventions require divulgence of deeply sensitive information that may be processed on remote servers without protections aligned to clinical standards such as HIPAA or GDPR. The commercial landscape also currently lacks accountability frameworks: if an AI system delivers harmful advice leading to patient harm (as documented with the Tessa chatbot [228]), traditional medical malpractice paradigms are ill-equipped to assign liability. Industry-wide LTGMs institutionalizing algorithmic auditing, predeployment clinical evaluation, and transparency protocols warrant development to mitigate these risks and to align commercial deployment with bioethical principles [215,243].

RQ4: Downstream Longitudinal Outcomes

Looking forward, widespread adoption of AI as a primary emotional support mechanism introduces important longitudinal risks [244]. Cognitive outsourcing—delegating emotional regulation and cognitive reframing to systems that instantly provide structured plans—may reduce personal agency and weaken ownership of the change process, undermining sustained engagement over time [106,109-111]. LLMs are not neutral systems; documented output biases—lower empathy ratings for certain demographic groups, gender-linked attribution patterns, stigmatizing framing of severe mental illness—can compound if unmitigated [2,117,228,229]. Advancing multimodal integration and ongoing technical debiasing are necessary to support more context-aware and equitable forms of computational empathy.

Limitations, Implications, and Epistemic Lag

Several limitations should be considered. First, no prospectively registered or publicly posted protocol was available for this review. Second, consistent with scoping review methodology, formal study-level risk-of-bias appraisal was not performed; we characterize the evidence base descriptively (by design, publication status, and sample size) and temper claims accordingly. Third, original title and abstract screening was not performed independently and in duplicate by 2 human reviewers; the post hoc blinded validation (see Reviewer Roles and Validation section) quantified agreement but cannot rule out missed eligible records. Fourth, API-based retrieval was capped and ordered using source-specific recency or relevance ranking, so the search was not exhaustive and may have omitted relevant studies; PsycINFO was exported as the complete 419-record result set. Fifth, the boundary on model type was not applied with complete uniformity: a small number of included sources used discriminative transformer or topic-modeling approaches (eg, [93]) or were evidence syntheses, and these were retained as contextual rather than primary evidence. Sixth, the included studies table (Multimedia Appendix 2) was generated with LLM-assisted extraction and subsequently harmonized; it is intended as a descriptive map, and one included source could be characterized only bibliographically. Finally, the quantitative narrative drew on a heterogeneous subset of studies with diverse designs and populations, limiting direct comparability, and some evidence was available only as preprints, whose inclusion reflects a field that evolves faster than traditional publication cycles, mitigated by anchoring core conclusions on peer-reviewed literature and flagging preliminary findings explicitly.

This review should also be interpreted in the context of rapid technological change. The academic literature typically lags current AI model development [7], and many models evaluated across the 241 synthesized papers are earlier versions relative to currently deployed systems. The newest model generation now incorporates test-time reasoning protocols that fundamentally alter complex prompt processing; newer entrants are not yet well represented in peer-reviewed outcome studies [1,2]. This epistemic lag is expected in a fast-moving field and does not invalidate current findings.

For clinical practice, the quantitative and qualitative evidence together support a human-in-the-loop governance approach in which clinical oversight remains central [212,245,246]. This paradigm should not be viewed as merely restrictive: clinicians may productively use LLMs as copilots—synthesizing patient histories, drafting clinical notes, supporting CBT restructuring exercises—while retaining final diagnostic and therapeutic authority. With appropriate structure, human oversight allows AI tools to augment clinical capacity while preserving core safety obligations in mental health care [6,106].

Conclusion

LLMs represent an important development in mental health, but the most actionable signal from this review emerges from the broader thematic synthesis rather than from any single outcome study. Across 241 included papers, a consistent map takes shape: the mapped literature suggests clinical promise in circumscribed settings, yet performance, safety, and acceptability are tightly conditioned by implementation context, disclosure practices, human oversight, and governance quality. The field is not facing a purely engineering problem; it confronts a sociotechnical and clinical integration challenge requiring coordinated work across computer science and AI, health informatics, mental health disciplines, human-computer interaction, implementation science, public health, ethics, law, and policy. Future progress should be

evaluated longitudinally—not only at the point of immediate symptom change—with key priorities including sustained engagement, therapeutic alliance, differential effects across demographic and cultural groups, and potential unintended consequences such as overreliance, cognitive outsourcing, and erosion of human relational support. The central goal should be careful, multidisciplinary, human-centered integration: positioning LLMs as accountable supports within care systems, not substitutes for the relational foundations of mental health care.

Acknowledgments

The authors gratefully acknowledge support from the University of Oklahoma (OU) Libraries' Subvention Funding Program and the OU Institute for Community and Society Transformation (ICAST) Faculty Fellowship Program.

Generative AI Use

Two distinct uses of generative AI are disclosed. As a research method: Google Gemini models were used within the review workflow for LLM-assisted title/abstract screening, full-text eligibility support, and structured data extraction, as detailed in the Methods. For manuscript preparation and revision: generative AI tools—including Anthropic's Claude and OpenAI's ChatGPT—were used to assist with drafting and rephrasing, structural and copy editing. The authors reviewed and edited all AI-assisted material, independently verified the scientific claims, statistics, and references, and retain full responsibility for the review methodology, interpretation, and final manuscript. No AI system is listed as an author.

Funding

No funding was received for this study.

Data Availability

This review synthesized previously published studies; no new participant-level data were generated. The search strategies, eligibility criteria, Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping reviews (PRISMA-ScR) materials, retained screening records, and available analysis code are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest. Google Gemini models were used as a methodological tool within the review workflow, as described in the Methods and the Generative AI Disclosure; this reflects a pragmatic tooling choice and does not imply endorsement of any specific LLM platform.

Multimedia Appendix 1

Search strings and retrieval parameters.
[DOCX File (Microsoft Word File), 13 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Included sources table.
[DOCX File (Microsoft Word File), 46 KB-Multimedia Appendix 2]

Checklist 1

PRISMA-ScR checklist.
[DOCX File (Microsoft Word File), 14 KB-Checklist 1]

References

1. 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]
2. Lawrence HR, Schneider RA, Rubin SB, Matarić MJ, McDuff DJ, Jones Bell M. The opportunities and risks of large language models in mental health. JMIR Ment Health. Jul 29, 2024;11:e59479. [doi: 10.2196/59479] [Medline: 39105570]
3. OmkarS SKK, K A, Raghuwanshi K, Sahu M, Gupta P. AI-powered drug discovery and mental health therapy: uniting technologies for holistic healthcare. IJFMR. 2024;6(5). [doi: 10.36948/ijfmr.2024.v06i05.28667]
4. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. Oct 18, 2024;11:e57400. [doi: 10.2196/57400] [Medline: 39423368]
5. Wang L, Carrington D, Filienko D, et al. Large language model-powered conversational agent delivering problem-solving therapy (PST) for family caregivers: enhancing empathy and therapeutic alliance using in-context learning. AMIA Annu Symp Proc. 2024;2024:1315-1324. [Medline: 41726496]
6. Pandey HM. Harnessing large language models for mental health: opportunities, challenges, and ethical considerations. arXiv. Preprint posted online on Dec 2024. [doi: 10.48550/arXiv.2501.10370]

7. Linardon J, Messer M, Anderson C, et al. Role of large language models in mental health research: an international survey of researchers' practices and perspectives. BMJ Ment Health. Jun 12, 2025;28(1):e301787. [doi: 10.1136/bmjment-2025-301787] [Medline: 40514050]
8. Hua Y, Liu F, Yang K, et al. Large language models in mental health care: a scoping review. Curr Treat Options Psych. ;12(1). [doi: 10.1007/s40501-025-00363-y]
9. 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]
10. Bucher A, Egger S, Vashkite I, Wu W, Schwabe G. "It's not only attention we need": systematic review of large language models in mental health care. JMIR Ment Health. Nov 4, 2025;12:e78410. [doi: 10.2196/78410] [Medline: 41186978]
11. Balan R, Gumpel TP. ChatGPT clinical use in mental health care: scoping review of empirical evidence. JMIR Ment Health. Dec 24, 2025;12:e81204. [doi: 10.2196/81204] [Medline: 41442647]
12. De Choudhury M, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health. ArXiv. Preprint posted online on Nov 7, 2023. [doi: 10.31234/osf.io/y8ax9]
13. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 2, 2018;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
14. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol. Feb 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
15. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci. Sep 20, 2010;5:69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
16. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. Oct 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
17. Flemyng E, Noel-Storr A, Macura B, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence 2025. Environ Evid. Oct 31, 2025;14(1):20. [doi: 10.1186/s13750-025-00374-5] [Medline: 41174758]
18. Scherbakov DA, Hubig NC, Lenert LA, Alekseyenko AV, Obeid JS. Natural language processing and social determinants of health in mental health research: AI-assisted scoping review. JMIR Ment Health. Jan 16, 2025;12:e67192. [doi: 10.2196/67192] [Medline: 39819656]
19. Landschaft A, Antweiler D, Mackay S, et al. Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. Int J Med Inform. Sep 2024;189:105531. [doi: 10.1016/j.ijmedinf.2024.105531] [Medline: 38943806]
20. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol. Jul 10, 2008;8(1):1-10. [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
21. Sharma A, Rushton K, Lin IW, Nguyen T, Althoff T. Facilitating self-guided mental health interventions through human-language model interaction: a case study of cognitive restructuring. Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24); Honolulu, HI. May 11, 2024.[doi: 10.1145/3613904.3642761]
22. Hatch SG, Goodman ZT, Vowels L, et al. When ELIZA meets therapists: a Turing test for the heart and mind. PLOS Ment Health. 2025;2(2):e0000145. [doi: 10.1371/journal.pmen.0000145] [Medline: 41661855]
23. Vowels LM, Vowels MJ, Sweeney SK, Hatch SG, Darwiche J. The efficacy, feasibility, and technical outcomes of a GPT-4o-based chatbot Amanda for relationship support: a randomized controlled trial. PLOS Ment Health. 2025;2(9):e0000411. [doi: 10.1371/journal.pmen.0000411] [Medline: 41662052]
24. Mo L, Li H, Zhang Y, et al. Self-help psychological intervention for young individuals during the post-COVID-19 era: development of a PST chatbot using GPT-4. Front Digit Health. 2025;7:1627268. [doi: 10.3389/fdgth.2025.1627268] [Medline: 41064794]
25. 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. 2026;24(1):314. [doi: 10.1186/s12967-026-07738-6]
26. Shen J, DiPaola D, Ali S, Sap M, Park HW, Breazeal C. Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: comparative study. JMIR Ment Health. Sep 25, 2024;11:e62679. [doi: 10.2196/62679] [Medline: 39321450]
27. Held P, Pridgen SA, Szoke DR, Chen Y, Akhtar Z, Amin D. AI-facilitated cognitive reappraisal via Socrates 2.0: mixed methods feasibility study. JMIR Ment Health. Dec 5, 2025;12:e80461. [doi: 10.2196/80461] [Medline: 41348953]

28. Izumi K, Tanaka H, Shidara K, et al. Response generation for cognitive behavioral therapy with large language models: comparative study with socratic questioning. arXiv. Preprint posted online on Jan 29, 2024. [doi: 10.48550/arXiv.2401.15966]
29. Furukawa TA, Iwata S, Horikoshi M, et al. Harnessing AI to optimize thought records and facilitate cognitive restructuring in smartphone CBT: an exploratory study. Cogn Ther Res. Dec 2023;47(6):887-893. [doi: 10.1007/s10608-023-10411-7]
30. Garcia D, Granjard A, Vanhée L, et al. AI-driven analyzes of open-ended responses to assess outcomes of internet-based cognitive behavioral therapy (ICBT) in adolescents with anxiety and depression comorbidity. J Affect Disord. Jul 15, 2025;381:659-668. [doi: 10.1016/j.jad.2025.04.003] [Medline: 40187428]
31. Shin D, Kim H, Lee S, Cho Y, Jung W. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. J Med Internet Res. Sep 18, 2024;26:e54617. [doi: 10.2196/54617] [Medline: 39292502]
32. Mahbub M, Dams GM, Srinivasan S, et al. Decoding substance use disorder severity from clinical notes using a large language model. Npj Ment Health Res. Feb 7, 2025;4(1):5. [doi: 10.1038/s44184-024-00114-6] [Medline: 39915681]
33. A NK, S M, A R, V J. Early detection and personalized intervention in mental health. Presented at: 2025 International Conference on Computing and Communication Technologies (ICCCT); Apr 16-17, 2025; Chennai, India. [doi: 10.1109/ICCCT63501.2025.11019669]
34. McCoy TH, Perlis RH. Dimensional measures of psychopathology in children and adolescents using large language models. Biol Psychiatry. Dec 15, 2024;96(12):940-947. [doi: 10.1016/j.biopsych.2024.05.008] [Medline: 38866172]
35. Rubin M, Watters K, Snyder K, Dicker E, Cordova M. Can large language models enhance assessment of Criterion A for PTSD from self-report? Psychol Trauma. Dec 1, 2025. [doi: 10.1037/tra0002087] [Medline: 41325158]
36. Giuliani A, Sharp T, Chideya Y, Meiser-Stedman R, Tomlinson M, Halligan SL. Can machine learning predict PTSD symptoms from trauma narratives of children and adolescents? Eur J Psychotraumatol. Dec 2025;16(1):2589709. [doi: 10.1080/20008066.2025.2589709] [Medline: 41328798]
37. Pugh SL, Chandler C, Cohen AS, Diaz-Asper C, Elvevåg B, Foltz PW. Assessing dimensions of thought disorder with large language models: the tradeoff of accuracy and consistency. Psychiatry Res. Nov 2024;341:116119. [doi: 10.1016/j.psychres.2024.116119] [Medline: 39226873]
38. Hur JK, Heffner J, Feng GW, Joormann J, Rutledge RB. Language sentiment predicts changes in depressive symptoms. Proc Natl Acad Sci U S A. Sep 24, 2024;121(39):e2321321121. [doi: 10.1073/pnas.2321321121] [Medline: 39284070]
39. Dougherty RF, Clarke P, Atli M, et al. Psilocybin therapy for treatment resistant depression: prediction of clinical outcome by natural language processing. Psychopharmacology (Berl). Jul 2025;242(7):1553-1561. [doi: 10.1007/s00213-023-06432-5] [Medline: 37606733]
40. Zhang Y, Fang J, Luo X, et al. Exploring the efficacy of ChatGPT in understanding and identifying intimate partner violence. Fam Relat. Jul 2025;74(3):1233-1249. [doi: 10.1111/fare.13176]
41. van Rijswijk J, Curșeu PL. Neurodivergence and well-being: the fulfilment of fundamental psychological needs, work-related stress and life satisfaction. Br J Psychol. Aug 2026;117(3):1187-1207. [doi: 10.1111/bjop.70057] [Medline: 41578606]
42. Oluwalade TI, Ahmadi H, Huo L, Sharpe R, Zhou SM. Comorbidities and emotions - unpacking the sentiments of pediatric patients with multiple long-term conditions through social media feedback: a large language model-driven study. J Affect Disord. Nov 1, 2025;388:119752. [doi: 10.1016/j.jad.2025.119752] [Medline: 40550274]
43. Perogamvros L, Rochas V, Beau JB, Sterpenich V, Bayer L. The cathartic dream: using a large language model to study a new type of functional dream in healthy and clinical populations. J Sleep Res. Dec 2025;34(6):e70001. [doi: 10.1111/jsr.70001] [Medline: 39924340]
44. Erdemir N, Atik S. Validity and reliability analysis of the artificial intelligence-digital life balance scale. Psychiatr Q. Mar 2026;97(1):1-21. [doi: 10.1007/s11126-025-10167-1] [Medline: 40464881]
45. Piastra M, Catellani P. On the emergent capabilities of ChatGPT 4 to estimate personality traits. Front Artif Intell. 2025;8:1484260. [doi: 10.3389/frai.2025.1484260] [Medline: 40017486]
46. Perrin PB, Christ BR, Vargas TA, et al. The Internalized Ableism Inventory: scale development using a hybrid artificial intelligence and community-based participatory research design. Rehabil Psychol. Aug 2026;71(3):370-382. [doi: 10.1037/rep0000639] [Medline: 40991817]
47. Schroeders U, Achaa-Amankwaa P. Developing NOVA – a Next-Generation Open Vocabulary Assessment. Eur J Psychol Assess. 2026. [doi: 10.1027/1015-5759/a000937]
48. Yang F, Oshio A. Using attachment theory to conceptualize and measure the experiences in human-AI relationships. Curr Psychol. Jun 2025;44(11):10658-10669. [doi: 10.1007/s12144-025-07917-6]

49. Yu SC, Chen HR, Yang YW. Development and validation the problematic ChatGPT use scale: a preliminary report. Curr Psychol. Aug 2024;43(31):26080-26092. [doi: 10.1007/s12144-024-06259-z]
50. Maral S, Naycı N, Bilmez H, Erdemir Eİ, Satici SA. Problematic ChatGPT use scale: AI-human collaboration or unraveling the dark side of ChatGPT. Int J Ment Health Addiction. Jun 2026;24(3):2369-2395. [doi: 10.1007/s11469-025-01509-y]
51. Krumm S, Thiel AM, Reznik N, Freudenstein JP, Schäpers P, Mussel P. Creating a psychological test in a few seconds: can ChatGPT develop a psychometrically sound situational judgment test. Eur J Psychol Assess. Dec 9, 2024. [doi: 10.1027/1015-5759/a000878]
52. Wang X, Zhou Y, Zhou G. Enhancing health assessments with large language models: a methodological approach. Appl Psychol Health Well Being. Feb 2025;17(1):e12602. [doi: 10.1111/aphw.12602] [Medline: 39392122]
53. Sokol Y, Goodman M. From Likert scales to large language models: validating a computational approach to psychological assessment of future self-continuity. J Pers Assess. 2026;108(3):437-447. [doi: 10.1080/00223891.2025.2576664] [Medline: 41166750]
54. Lyu Y, Combs D, Neumann D, Leong YC. Automated scoring of the Ambiguous Intentions Hostility Questionnaire with fine-tuned large language models. Psychol Assess. Sep 2026;38(9):635-649. [doi: 10.1037/pas0001466] [Medline: 41843453]
55. Li F, Zhao N. The assessment of body image based on large language model. Psych J. Oct 2025;14(5):669-684. [doi: 10.1002/pchj.70048] [Medline: 40884170]
56. Strojny P, Kapela K, Lipp N, Sikström S. Use of 4 open-ended text responses to help identify people at risk of gaming disorder: preregistered development and usability study using natural language processing. JMIR Serious Games. Dec 31, 2024;12:e56663. [doi: 10.2196/56663] [Medline: 39739378]
57. Hanss K, Sarma KV, Glowinski AL, et al. Assessing the accuracy and reliability of large language models in psychiatry using standardized multiple-choice questions: cross-sectional study. J Med Internet Res. May 20, 2025;27:e69910. [doi: 10.2196/69910] [Medline: 40392576]
58. Neubauer JC, Kaiser A, Lettermann L, Volkert T, Häge A. Performance of large language models ChatGPT and Gemini in child and adolescent psychiatry knowledge assessment. PLoS One. 2025;20(9):e0332917. [doi: 10.1371/journal.pone.0332917] [Medline: 40971849]
59. Siddiqui ZH, Azeez MA, Sohail SS, Ahmed J, Madsen DØ. A preliminary exploration of ChatGPT's potential in medical reasoning and patient care. Crit Public Health. Dec 31, 2025;35(1). [doi: 10.1080/09581596.2025.2484344]
60. Lau C, Zhu X, Chan WY. Automatic depression severity assessment with deep learning using parameter-efficient tuning. Front Psychiatry. 2023;14:1160291. [doi: 10.3389/fpsyt.2023.1160291] [Medline: 37398577]
61. Fu Z, Hsu YC, Chan CS, Lau CM, Liu J, Yip PSF. Efficacy of ChatGPT in Cantonese sentiment analysis: comparative study. J Med Internet Res. Jan 30, 2024;26:e51069. [doi: 10.2196/51069] [Medline: 38289662]
62. Hammerfald K, Schmidt F, Vlassov V, Haaland Jahren H, Solbakken OA. Leveraging large language models to identify microcounseling skills in psychotherapy transcripts. Psychother Res. Jul 2026;36(6):1058-1076. [doi: 10.1080/10503307.2025.2539405] [Medline: 40817802]
63. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. JMIR Ment Health. Feb 6, 2024;11:e54369. [doi: 10.2196/54369] [Medline: 38319707]
64. Sufyan NS, Fadhel FH, Alkhathami SS, Mukhadi JYA. Artificial intelligence and social intelligence: preliminary comparison study between AI models and psychologists. Front Psychol. 2024;15:1353022. [doi: 10.3389/fpsyg.2024.1353022] [Medline: 38379623]
65. Rahman ABS, Ta HT, Najjar L, Azadmanesh A, Gönul AS. DepressionEmo: a novel dataset for multilabel classification of depression emotions. J Affect Disord. Dec 1, 2024;366:445-458. [doi: 10.1016/j.jad.2024.08.013] [Medline: 39214375]
66. Hwang G, Lee DY, Seol S, et al. Assessing the potential of ChatGPT for psychodynamic formulations in psychiatry: an exploratory study. Psychiatry Res. Jan 2024;331:115655. [doi: 10.1016/j.psychres.2023.115655] [Medline: 38056130]
67. Markowitz DM. Can generative AI infer thinking style from language? Evaluating the utility of AI as a psychological text analysis tool. Behav Res Methods. Apr 2024;56(4):3548-3559. [doi: 10.3758/s13428-024-02344-0] [Medline: 38277084]
68. Goddard A, Gillespie A. The repeated adjustment of measurement protocols method for developing high-validity text classifiers. Psychol Methods. Oct 6, 2025. [doi: 10.1037/met0000787] [Medline: 41051829]
69. Pataranutaporn P, Powdthavee N, Archiwaranguprok C, Maes P. Simulating human well-being with large language models: systematic validation and misestimation across 64,000 individuals from 64 countries. Proc Natl Acad Sci U S A. Dec 2, 2025;122(48):e2519394122. [doi: 10.1073/pnas.2519394122] [Medline: 41296717]

70. Kim MG, Hwang G, Chang J, Chang S, Roh HW, Park RW. Performance of open-source large language models in psychiatry: usability study through comparative analysis of non-English records and English translations. J Med Internet Res. Aug 18, 2025;27:e69857. [doi: 10.2196/69857] [Medline: 40825309]
71. McBain RK, Cantor JH, Zhang LA, et al. Competency of large language models in evaluating appropriate responses to suicidal ideation: comparative study. J Med Internet Res. Mar 5, 2025;27:e67891. [doi: 10.2196/67891] [Medline: 40053817]
72. Yeo YH, Peng Y, Mehra M, et al. Evaluating for evidence of sociodemographic bias in conversational AI for mental health support. Cyberpsychol Behav Soc Netw. Jan 2025;28(1):44-51. [doi: 10.1089/cyber.2024.0199] [Medline: 39446671]
73. Kharitonova K, Pérez-Fernández D, Gutiérrez-Hernando J, Gutiérrez-Fandiño A, Callejas Z, Griol D. Incorporating evidence into mental health Q&A: a novel method to use generative language models for validated clinical content extraction. Behav Inf Technol. Jun 15, 2025;44(10):2333-2350. [doi: 10.1080/0144929X.2024.2321959]
74. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health. Sep 20, 2023;10:e51232. [doi: 10.2196/51232] [Medline: 37728984]
75. Levkovich I, Shinan-Altman S, Elyoseph Z. Can large language models be sensitive to culture suicide risk assessment? J Cult Cogn Sci. Dec 2024;8(3):275-287. [doi: 10.1007/s41809-024-00151-9]
76. Zhang Y, Huang F, Mo L, Liu X, Zhu T. Suicidal ideation data augmentation and recognition technology based on large language models. Acta Psychologica Sinica. 2025;57(6):987. [doi: 10.3724/SP.J.1041.2025.0987]
77. Elyoseph Z, Levkovich I. Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: case vignette study. JMIR Ment Health. Mar 18, 2024;11:e53043. [doi: 10.2196/53043] [Medline: 38533615]
78. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacol. Aug 2024;49(9):1412-1416. [doi: 10.1038/s41386-024-01841-2]
79. 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]
80. 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]
81. Terblanche N, Rutschmann R, Reitz J. Comparing GROW, solution-focused and CBT coaching chatbots: a self-determination and transformative learning perspective. Coach Int J Theory Res Pract. Jul 3, 2026;19(2):250-271. [doi: 10.1080/17521882.2026.2625707]
82. Wang Y, Farb NAS. An exploratory study of chatbot-based mindfulness for mental health support. Mindfulness (N Y). Nov 2025;16(11):3143-3162. [doi: 10.1007/s12671-025-02685-7]
83. Jeong D, Lee H, Yang H, Park JC, Lee SM. AI integration in counseling training: aiding counselors-in-training in self-efficacy enhancement and anxiety reduction. J Couns Dev. Jul 2025;103(3):346-360. [doi: 10.1002/jcad.12560]
84. Schimpf M, Maier S, Wyrowski A, Christoforakos L, Feuerriegel S, Bohné T. Supporting effective goal setting with LLM-based chatbots. In: Oliver N, Shamma DA, Candello H, et al, editors. CHI '26: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. 2026. URL: https://dl.acm.org/doi/proceedings/10.1145/3772318 [doi: 10.1145/3772318.3791064]
85. Rousmaniere T, Zhang Y, Li X, Shah S. Large language models as mental health resources: patterns of use in the United States. Pract Innov (Wash D C). ;11(2):139-155. [doi: 10.1037/pri0000292]
86. Li L, Peng W, Rheu MMJ. Factors predicting intentions of adoption and continued use of artificial intelligence chatbots for mental health: examining the role of UTAUT model, stigma, privacy concerns, and artificial intelligence hesitancy. Telemed J E Health. Mar 2024;30(3):722-730. [doi: 10.1089/tmj.2023.0313] [Medline: 37756224]
87. Aral A, Gerdan G, Usta MB, Aral AE. From promise to practice: insights into ChatGPT-4o use in child and adolescent mental health from professionals. Front Psychiatry. 2025;16:1668814. [doi: 10.3389/fpsyt.2025.1668814] [Medline: 41080775]
88. Kelly S, White KM, Kaye SA, Oviedo-Trespalacios O. Tapping into key drivers: self-disclosure in sensitive health conversations with ChatGPT. Int J Hum-Comput Interact. Dec 17, 2025;41(24):15668-15678. [doi: 10.1080/10447318.2025.2499656]
89. Duong CD, Vu TN, Ngo TVN, Do ND, Tran NM. Reduced student life satisfaction and academic performance: unraveling the dark side of ChatGPT in the higher education context. Int J Human Comput Inter. Apr 18, 2025;41(8):4948-4963. [doi: 10.1080/10447318.2024.2356361]
90. Gu C, Zhang J. How does "Dr. AI" trigger cyberchondria? An empirical study based on the CMIS framework using a hybrid SEM–ANN approach. Curr Psychol. Jan 2026;45(1). [doi: 10.1007/s12144-025-08899-1]

91. Salah M, Alhalbusi H, Ismail MM, Abdelfattah F. Chatting with ChatGPT: decoding the mind of Chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. Curr Psychol. Mar 2024;43(9):7843-7858. [doi: 10.1007/s12144-023-04989-0]
92. Lotfy AY, Elaziz MA, Dahou A, Mahmoud N. Generative ai for psychology and mental health: review study. Presented at: 2024 International Conference on Smart-Digital-Green Technologies and Artificial Intelligence Sciences (CSDGAIS); Dec 9-11, 2024; Suez, Egypt. [doi: 10.1109/CSDGAIS64098.2024.11064846]
93. Baird A, Xia Y, Cheng Y. Consumer perceptions of telehealth for mental health or substance abuse: a Twitter-based topic modeling analysis. JAMIA Open. Jul 2022;5(2):ooac028. [doi: 10.1093/jamiaopen/ooac028] [Medline: 35495736]
94. Bauer B, Norel R, Leow A, Rached ZA, Wen B, Cecchi G. Using large language models to understand suicidality in a social media-based taxonomy of mental health disorders: linguistic analysis of Reddit posts. JMIR Ment Health. May 16, 2024;11:e57234. [doi: 10.2196/57234] [Medline: 38771256]
95. Wicaksono ADA. Classifcation of social media posts X for mental health symptoms identification using NLP techniques and transformers model. J Comput Sci. Jan 1, 2026;22(1):244-259. [doi: 10.3844/jcssp.2026.244.259]
96. Hussain D, Fauz F, Abbas M. An AI-powered monitoring system for employee mental health and wellbeing in the workplace. J Workplace Behav Health. 2025;6(1):41-55. [doi: 10.70580/jwb.06.01.0264]
97. Yadav N. AI for mental health: the use of chatbots and NLP to support therapy and early detection. Int Sci J Eng Manag. 2025;04(5):1-7. [doi: 10.55041/ISJEM03402]
98. Bookanakere P, S S, Saniya S, R A, S SK. Textual therapies: harnessing language models for mental health insights. Presented at: 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET); Sep 27-28, 2024; Indore, India. [doi: 10.1109/ACROSET62108.2024.10743711]
99. Kiruthika R, Ajay R, Balakrishnan P, Giridharan S. AI chatbot for mental health. ShodhKosh J Vis Per Arts. 2024;5(5). [doi: 10.29121/shodhkosh.v5.i5.2024.2853]
100. Chauhan VP, Goel V. Serenify: an AI-powered virtual therapist for accessible mental health support. nt J Res Appl Sci Eng Technol. 2025;13(5):2412-2418. [doi: 10.22214/ijraset.2025.70773]
101. Rasool A, Aslam S, Hussain N, Imtiaz S, Riaz W. nBERT: harnessing NLP for emotion recognition in psychotherapy to transform mental health care. Information. 2025;16(4):301. [doi: 10.3390/info16040301]
102. Lee JJ, Han J, Woo CW. Interpretable depression assessment using a large language model. PLOS Digit Health. Feb 2026;5(2):e0001205. [doi: 10.1371/journal.pdig.0001205] [Medline: 41662257]
103. Maurya RK, Montesinos S, Bogomaz M, DeDiego AC. Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. Couns and Psychother Res. Mar 2025;25(1). [doi: 10.1002/capr.12759]
104. Karki A, Kamble C, Chavan R, Chapke N. Mental health meets machine learning: the rise of chatbots and LLMs in therapy. IJRTI. 2025;10(5). [doi: 10.56975/ijrti.v10i5.203281]
105. Kampman OP, Phang YS, Han S, et al. A multi-agent dual dialogue system to support mental health care providers. arXiv. Preprint posted online on Nov 28, 2024. [doi: 10.48550/arXiv.2411.18429]
106. Jung K, Lee G, Huang Y, Chen Y. "I've talked to ChatGPT about my issues last night.": examining mental health conversations with large language models through Reddit analysis. Presented at: 28th ACM SIGCHI Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '25); Oct 18-22, 2025:1-25; Bergen, Norway. [doi: 10.1145/3757537]
107. Yu HQ, McGuinness S. An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. J Med Artif Intell. 2024;7:16-16. [doi: 10.21037/jmai-23-136]
108. Sharma A, Rushton K, Lin I, et al. Cognitive reframing of negative thoughts through human-language model interaction. Presented at: 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023); Jul 9-14, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.acl-long.555]
109. Sun F, Guo X. Exploring the use of ChatGPT-4o in cognitive behavioural therapy for university students: enhancing mental health with AI-powered voice interaction. ASBR. 2025;16(3):99-107. [doi: 10.54254/2753-7102/2025.23159]
110. Aghakhani E, Rezapour R. Like a therapist, but not: Reddit narratives of AI in mental health contexts. Presented at: The 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026); Jul 2-7, 2026; San Diego, CA. [doi: 10.18653/v1/2026.findings-acl.569]
111. Zhang Y, Li X, Zhu J, Sheng Z, Rousmaniere T. "What happens, what helps, what hurts": a qualitative analysis of user experiences with large language models for mental health support. Pract Innov (Wash D C). 2026. [doi: 10.1037/pri0000319]
112. Kalam KT, Rahman JM, Islam MR, Dewan SMR. ChatGPT and mental health: friends or foes? Health Sci Rep. Feb 2024;7(2):e1912. [doi: 10.1002/hsr2.1912] [Medline: 38361805]
113. Chiang S, Laban G, Gunes H. Do we talk to robots like therapists, and do they respond accordingly? language alignment in AI emotional support. arXiv. Preprint posted online on Jun 19, 2025. [doi: 10.48550/arXiv.2506.16473]
114. Gowroju S, Choudhary S, Nakirekanti M, Kalidas MP, Swapna A, Ramya G. MindBridge: AI-enhanced virtual mental health platform for emotional analysis and levaraging. In: 2025 4th OPJU International Technology Conference

(OTCON) on Smart Computing for Innovation and Advancement in Industry 50. IEEE; 2025. [doi: 10.1109/OTCON65728.2025.11070611]
115. Sahana MB, Anil Kumar KN. BeYouBot: a caring AI chatbot for your mental well-being. Int J Res Eng Sci Manage. 2025. [doi: 10.55041/IJSREM51793]
116. Anusha S, Hiranmayee N, Kamala PS, et al. AI for mental health support. IJISRT. 1047-1053. [doi: 10.38124/ijisrt/25apr831]
117. Moore J, Grabb D, Agnew W, et al. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. Presented at: 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25); Jun 23-26, 2025; Athens, Greece. [doi: 10.1145/3715275.3732039]
118. Qui H, Zhao T, Li A, Zhang S, He H, Lan Z. A benchmark for understanding dialogue safety in mental health support. In: Liu F, Duan N, Xu Q, Hong Y, editors. Natural Language Processing and Chinese Computing (NLPCC 2023). [doi: 10.1007/978-3-031-44696-2_1]
119. Dallison S, Moffett B, Makhubela P, et al. Using generative AI to co-design digital mental health interventions with adolescents in rural South Africa: qualitative thematic analysis of participatory workshops. J Med Internet Res. Dec 5, 2025;27:e73535. [doi: 10.2196/73535] [Medline: 41348954]
120. Sanz A, Tapia JL, García-Carpintero E, Rocabado JF, Pedrajas LM. ChatGPT simulated patient: use in clinical training in psychology. Psicothema. Jun 2025;37(3):23-32. [doi: 10.70478/psicothema.2025.37.21] [Medline: 40549860]
121. Yahia A, Bakry M, Radwan N, et al. MindHaven: retrieval-augmented large language models for AI-powered mental health support. Presented at: 16th Student Research Conference on Applied Computing (SRC 2025); Sep 24-25, 2025; Dubai, United Arab Emirates. [doi: 10.1109/SRC65875.2025.11263620]
122. Jaiswal S, Lee J, Berria J, et al. Building personality-adaptive conversational AI for mental health therapy. Presented at: 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB 2024); Nov 22-25, 2024; Shenzhen, China. [doi: 10.1145/3698587.3701489]
123. Abeesh SV, Srinitya G. AI-powered enhanced medical QA chatbot system with sentiment analysis mood tracking and personalized mental health recommendations. In: 2025 5th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS). IEEE; 2025. [doi: 10.1109/ICUIS67429.2025.11380270]
124. Na H. CBT-LLM: a Chinese large language model for cognitive behavioral therapy-based mental health question answering. Presented at: Joint International Conference on Computational Linguistics, Language Resources and Evaluation; May 22-24, 2024; Turin, Italy. [doi: 10.63317/3bgf5o6ze4k6]
125. Lee S, Kim S, Kim M, et al. Cactus: towards psychological counseling conversations using cognitive behavioral theory. arXiv. Preprint posted online on 2024. [doi: 10.48550/arXiv.2407.03103]
126. Zhou Y, Zhou N, Chen Q, Zhou J, Zhou A, He L. DiaCBT: a long-periodic dialogue corpus guided by cognitive conceptualization diagram for CBT-based psychological counseling. arXiv. Preprint posted online on 2025. [doi: 10.48550/arXiv.2509.02999]
127. Ciarrochi J, Hernández C, Hill D, et al. Process-based therapy: a common ground for understanding and utilizing therapeutic practices. J Psychother Integr. ;34(3):265-290. [doi: 10.1037/int0000348]
128. Zhang M, Yang X, Zhang X. CBT-bench: evaluating large language models on assisting cognitive behavior therapy. arXiv. Preprint posted online on Jan 2025. [doi: 10.48550/arXiv.2410.13218]
129. Moon P, Bhattacharyya P. We care: multimodal depression detection and knowledge infused mental health therapeutic response generation. arXiv. Preprint posted online on 2024. [doi: 10.48550/arXiv.2406.10561]
130. Wang R, Milani S, Chiu JC, et al. Patient-ψ: using large language models to simulate patients for training mental health professionals. arXiv. Preprint posted online on 2024.
131. Iftikhar Z. Re-imagining mental health access: the role of human, AI and design. Presented at: CSCW Companion '24: Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing; San Jose Costa Rica. Nov 11, 2024.
132. Gollapalli SD, Ang BH, Ng M. Counseling responses for mental health forum questions with early maladaptive schema prediction. Front Artif Intell Appl. 2024;392:2556-2563. [doi: 10.3233/FAIA240785]
133. Karve Z, Calpey J, Machado C, Knecht M, Mejia MC. New doc on the block: scoping review of AI systems delivering motivational interviewing for health behavior change. J Med Internet Res. Sep 16, 2025;27:e78417. [doi: 10.2196/78417] [Medline: 40957014]
134. Alam S, Sohail SS. Integrating ChatGPT: enhancing postpartum mental healthcare with artificial intelligence (AI) support. Digit HEALTH. 2024;10:20552076241295565. [doi: 10.1177/20552076241295565] [Medline: 39655059]
135. Lange M, Koutsouleris N, Carter B, Twumasi R. Generating job recommendations for people with schizophrenia spectrum disorder using Gemini 2.0 Flash and Claude Sonnet 4: an exploratory analysis. Healthc Technol Lett. 2026;13(1):e70058. [doi: 10.1049/htl2.70058] [Medline: 41695595]
136. Mondal H, Mishra S, Jangra J, Agrawal AK, Nath S, Sarkhel S. Effectiveness of artificial intelligence chatbots in providing sexual health information: an exploratory study. J Sex Med. Feb 6, 2026;23(3):qdag028. [doi: 10.1093/jsxmed/qdag028] [Medline: 41701078]

137. Jain G, Patole E, Pareek S. Digital companionship: the interplay of conversational AI, loneliness, social anxiety, and quality of life among young adults. Acta Psychol (Amst). Apr 2026;264:106467. [doi: 10.1016/j.actpsy.2026.106467] [Medline: 41702094]
138. Smith BR, Toomer LA, Vu E, et al. A pilot study evaluating traditional and artificial intelligence (AI)-generated bedside art interventions in hospital care. J Patient Exp. 2026;13:23743735261423084. [doi: 10.1177/23743735261423084] [Medline: 41696641]
139. Han X. NeuroPal: a clinically-informed multimodal LLM assistant for mental health combining sleep chronotherapy. arXiv. Preprint posted online on May 10, 2025. [doi: 10.48550/arXiv.2505.06640]
140. Zhu Z, Yu J, Luo Y. Scaffolding metacognition with genai: exploring design opportunities to support task management for university students with ADHD. Presented at: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26); Barcelona Spain. Apr 13, 2026.
141. Igwe K, Durrhiem K. A scoping review of culturally sensitive large language models-based cognitive behavioural therapy for anxiety and depression: global lessons for African implementation. IJSS. 2025;5(1):a06. [doi: 10.38140/ijss-2025.vol5.1.06]
142. Xie SJ, Zhai S, Liang Y, et al. Cultural prompting improves the empathy and cultural responsiveness of GPT-generated therapy responses. AMIA Annu Symp Proc. 2026;2025:1384-1393. [Medline: 41726474]
143. Shi Y. Fine-tuning large language models for early mental health intervention in China: a culturally adapted triage and therapy framework. In: Advanced Intelligent Computing Technology and Applications (ICIC 2025). 2025. [doi: 10.1007/978-981-95-0033-8_9]
144. Joshi I, Jaiswal K, Singh J, Chakraborty R, Goswami RG. MindTune: AI-based mood recognition and personalised music therapy for early-stage mental health management. Presented at: 2025 International Conference on Engineering Innovations and Technologies (ICoEIT); Jun 4-5, 2025; Bhopal, India. 2025.[doi: 10.1109/ICoEIT63558.2025.11211602]
145. Huang R, Nguyen B. A personalized mobile application to generate music therapy using a large language model and storing the user's data on firebase. Presented at: 14th International Conference on Cryptography and Information Security (CRYPIS 2025); Sep 20-21, 2025. [doi: 10.5121/csit.2025.151719]
146. Penberthy JK, Penberthy DR, Bires J. A narrative review of the role of artificial intelligence in supporting the mental health of patients with breast cancer. AI Precis Oncol. Aug 2025;2(4):121-128. [doi: 10.1177/2993091X251361147]
147. Milligan G, Bernard A, Dowthwaite L, et al. Developing a single-session outcome measure using natural language processing on digital mental health transcripts. Couns Psychother Res. Sep 2024;24(3):1057-1068. [doi: 10.1002/capr.12766]
148. Levkovich I. Evaluating diagnostic accuracy and treatment efficacy in mental health: a comparative analysis of large language model tools and mental health professionals. Eur J Investig Health Psychol Educ. Jan 18, 2025;15(1):9. [doi: 10.3390/ejihpe15010009] [Medline: 39852192]
149. Elyoseph Z, Gur T, Haber Y, et al. An ethical perspective on the democratization of mental health with generative artificial intelligence. JMIR Mental Health. Preprint posted online on 2024. [doi: 10.2196/preprints.58011]
150. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health. Jan 9, 2024;12(Suppl 1):e002583. [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
151. Anser MK, Nabi AA, Ahmad I, Abro MMQ, Zaman K. Advancing mental health care: a comprehensive review of digital tools and technologies for enhancing diagnosis, treatment, and wellness. Health Care Sci. Jun 2025;4(3):163-178. [doi: 10.1002/hcs2.70018] [Medline: 40568629]
152. Adhikary PK, Srivastava A, Kumar S, et al. Exploring the efficacy of large language models in summarizing mental health counseling sessions: benchmark study. JMIR Ment Health. Jul 23, 2024;11:e57306. [doi: 10.2196/57306] [Medline: 39042893]
153. Priyadarshana Y, Senanayake A, Liang Z, Piumarta I. Prompt engineering for digital mental health: a short review. Front Digit Health. 2024;6. [doi: 10.3389/fdgth.2024.1410947]
154. Kalinich M, Luccarelli J, Maria JS, Williams G, Moss F, Torous J. Evaluating the effect of mental health fine-tuning relative to other model characteristics on LLM safety performance. medRxiv. Preprint posted online on 2026. [doi: 10.64898/2026.01.02.25343289]
155. Abdullah-Al-Kafi M, Moni R, Banshal SK. Reasoning over recall: evaluating the efficacy of generalist architectures vs. specialized fine-tunes in RAG-based mental health dialogue systems. arXiv. Preprint posted online on Jan 4, 2026. [doi: 10.48550/arXiv.2601.01341]
156. Xu X, Yao B, Dong Y, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. Proc ACM Interact Mob Wearable Ubiquitous Technol. Mar 2024;8(1):31. [doi: 10.1145/3643540] [Medline: 39925940]
157. Qiu H, He H, Zhang S, Li A, Lan Z. SMILE: single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. Presented at: 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024); Nov 12-16, 2024; Miami, FL. [doi: 10.18653/v1/2024.findings-emnlp.34]

158. Rajagopal A, Nirmala V, J A, M A. Novel AI to avert the mental health crisis in COVID-19: novel application of GPT2 in cognitive behaviour therapy. Res Sq. Preprint posted online on 2022. [doi: 10.21203/rs.3.rs-382748/v2]
159. Sambana B, Archana K, Reddy SIS, Basha SMJ, Karishma S. Data augmentation for cognitive behavioral therapy: leveraging ERNIE language models using artificial intelligence. In: 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT-2025). 2025. URL: https://arxiv.org/pdf/2506.23503 [Accessed 2026-10-30]
160. Adamu H, Azmi Murad MA, Nasharuddin NA. A hybrid stacked ensemble learning framework for multilabel text emotion detection. Sci Rep. Feb 7, 2026;16(1):7714. [doi: 10.1038/s41598-026-38172-9] [Medline: 41654610]
161. Kaur B, Chen D. Classifying and understanding mental health discourse on social media: a hybrid topic modelling approach. Presented at: 2025 International Conference on Cyber-Physical Social Intelligence (CPSI); Nov 7-10, 2025; Macau, China. [doi: 10.1109/CPSI66656.2025.11343977]
162. Galland L, Pelachaud C, Pecune F. Generating unexpected yet relevant user dialog acts. Presented at: 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2024); Sep 18-20, 2024; Kyoto, Japan. [doi: 10.18653/v1/2024.sigdial-1.17]
163. Kumar V, Ntoutsi E, Rajawat PS, Medda G, Recupero DR. Unlocking llms: addressing scarce data and bias challenges in mental health. arXiv. Preprint posted online on 2024. [doi: 10.48550/arXiv.2412.12981]
164. Liu S, Deng N, Sabour S, Jia Y, Huang M, Mihalcea R. Task-adaptive tokenization: enhancing long-form text generation efficacy in mental health and beyond. Presented at: 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023); Dec 6-10, 2023; Singapore. [doi: 10.18653/v1/2023.emnlp-main.944]
165. Dharrao D, Gite S. TherapyBot: a chatbot for mental well-being using transformers. Int J Adv Appl Sci. 2024;13:1-12. [doi: 10.11591/ijaas.v13.i1.pp1-12]
166. Yuan A, Garcia Colato E, Pescosolido B, Song H, Samtani S. Improving workplace well-being in modern organizations: a review of large language model-based mental health chatbots. ACM Trans Manage Inf Syst. Mar 31, 2025;16(1):1-26. [doi: 10.1145/3701041]
167. Xu J, Wei T, Hou B, Demiris G, Shen L. Revolutionizing dementia care: enhancing talk therapy with fine-tuned large language models using GPT self-generated data. Alzheimers Dement. Dec 2024;20(S4). [doi: 10.1002/alz.093496]
168. Du L, Li Y, Long Y, Chen S. EFT-CoT: a multi-agent chain-of-thought framework for emotion-focused therapy. arXiv. Preprint posted online on 2026. [doi: 10.48550/arXiv.2601.17842]
169. Naik PS, R RH, D. U. Cross-modal emotion-aware music generation for therapeutic healing using lora-tuned transformers. Presented at: 2025 20th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP); Nov 12-14, 2025; Phuket, Thailand. [doi: 10.1109/iSAI-NLP66160.2025.11320541]
170. Zou C, Wang N, Shen T, et al. Affective flow language model for emotional support conversation. arXiv. Preprint posted online on 2026. [doi: 10.48550/arXiv.2602.08826]
171. Patil S, Chakrabarti S, Shimpi R, Rode R, Pawar R. Mental wellness assistant using retrieval-augmented generation (RAG). Presented at: 2025 6th International Conference on Electronics and Sustainable Communication Systems (ICESC); Sep 10-12, 2025; Coimbatore, India. [doi: 10.1109/ICESC65114.2025.11212290]
172. Boit S, Patil R. A prompt engineering framework for large language model-based mental health chatbots: conceptual framework. JMIR Ment Health. Nov 7, 2025;12:e75078. [doi: 10.2196/75078] [Medline: 41202205]
173. Yahia A, Bakry M, Radwan N, et al. Empathetic AI for mental health: a comparative study of retrieval-augmented language models in CBT applications. In: 2025 Intelligent Methods, Systems, and Applications (IMSA). 2025. [doi: 10.1109/IMSA65733.2025.11166690]
174. Soman G, Judy MV, Abou AM. Human guided empathetic AI agent for mental health support leveraging reinforcement learning-enhanced retrieval-augmented generation. Cogn Syst Res. Apr 2025;90:101337. [doi: 10.1016/j.cogsys.2025.101337]
175. Stoll D, Jud A, Wehrli S, et al. Case reports unlocked: leveraging retrieval-augmented generation with large language models to advance research on psychological child maltreatment. Child Abuse Negl. Nov 2025;169(Pt 2):107653. [doi: 10.1016/j.chiabu.2025.107653] [Medline: 40884999]
176. Chen Q, Liu D. MADP: multi-agent deductive planning for enhanced cognitive-behavioral mental health question answer. arXiv. Preprint posted online on 2025. [doi: 10.48550/arXiv.2501.15826]
177. Yao X, She D, Zhang C, et al. Empathy-R1: a chain-of-empathy and reinforcement learning framework for long form mental health support. IEEE Trans Affective Comput. 1-16. [doi: 10.1109/TAFFC.2026.3732970]
178. Ferizaj D, Lalk C, Lahmann N, Strube-Lahmann S, Rubel J. Identifying Yalom's group therapeutic factors in anonymous mental health discussions on Reddit: a mixed-methods analysis using large language models, topic modeling and human supervision. Front Psychiatry. 2025;16:1503427. [doi: 10.3389/fpsyt.2025.1503427] [Medline: 40551826]
179. Lozoya D, Berazaluce A, Perches J, Lúa E, Conway M, D'Alfonso S. Generating mental health transcripts with SAPE (Spanish adaptive prompt engineering). Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 16-21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.naacl-long.285]

180. Xu S, Zhou Z, Zhao N. How topic content shapes LLM personality-tailored persuasion: semantic anchoring and topic stereotype effects. Front Psychiatry. 2026;17:1756792. [doi: 10.3389/fpsyt.2026.1756792] [Medline: 41696465]
181. Filienko D, Wang Y, Jazmi CE, et al. Toward large language models as a therapeutic tool: comparing prompting techniques to improve GPT-delivered problem-solving therapy. AMIA Annu Symp Proc. 2024;2024:417-426. [Medline: 40417590]
182. Gollapalli S, Ang B, Ng SK. Identifying early maladaptive schemas from mental health question texts. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2023; Singapore. 2023.
183. Franchino E, Gariboldi F, Grecucci A, Lattanzi G, Stella M. Complex networks map test anxiety and wellbeing levels in students and chatgpt. arXiv. Preprint posted online on 2026. [doi: 10.48550/arXiv.2602.13302]
184. He S, Chen YN. Power distance and psychological safety in LLM counseling: effects on self-efficacy with implications for mental health-relevant behavior change. Behav Sci (Basel). Feb 8, 2026;16(2):241. [doi: 10.3390/bs16020241] [Medline: 41750050]
185. Hsu CW, Cochrane M, Bambarawana S. Generative artificial intelligence in mental health: a preliminary study on automating materials development for cognitive bias modification. Int J Ment Health. 2026:1-19. [doi: 10.1080/00207411.2025.2611527]
186. Heston TF, Gillette J. Do large language models have a personality? A psychometric evaluation with implications for clinical medicine and mental health AI. medRxiv. Preprint posted online on Mar 15, 2025. [doi: 10.1101/2025.03.14.25323987]
187. Anvari SS, Wehbe RR. Therapeutic AI and the hidden risks of over-disclosure: an embedded AI-literacy framework for mental health privacy. arXiv. Preprint posted online on 2025. [doi: 10.48550/arXiv.2510.10805]
188. Kapuriya J, Singh A, Shukla J, Shah RR. Spiritual-LLM: Gita inspired mental health therapy in the era of llms. arXiv. Preprint posted online on 2025. [doi: 10.48550/arXiv.2506.19185]
189. Forane SG, Ezugwu AE, Igwe K, van den Berg K. An Ubuntu-guided large language model framework for cognitive behavioral mental health dialogue. Front Digit Health. 2026;8:1903421. [doi: 10.3389/fdgth.2026.1903421] [Medline: 42688427]
190. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Singapore. 2023.
191. Naher J. Can ChatGPT provide a better support: a comparative analysis of ChatGPT and dataset responses in mental health dialogues. Curr Psychol. Jul 2024;43(28):23837-23845. [doi: 10.1007/s12144-024-06140-z]
192. Omar M, Levkovich I. Exploring the efficacy and potential of large language models for depression: a systematic review. J Affect Disord. Feb 15, 2025;371:234-244. [doi: 10.1016/j.jad.2024.11.052] [Medline: 39581383]
193. Wang L, Bhanushali T, Huang Z, Yang J, Badami S, Hightow-Weidman L. Evaluating generative AI in mental health: systematic review of capabilities and limitations. JMIR Ment Health. May 15, 2025;12:e70014. [doi: 10.2196/70014] [Medline: 40373033]
194. Khadangi A, Marxen H, Sartipi A, Tchappi I, Fridgen G. When AI takes the couch: psychometric jailbreaks reveal internal conflict in frontier models. Preprint posted online on Jun 20, 2026. [doi: 10.48550/arXiv.2512.04124]
195. Zhang S, Zhang C, Zhang J. Text-based depression estimation using machine learning with standard labels: systematic review and meta-analysis. J Med Internet Res. Feb 11, 2026;28:e82686. [doi: 10.2196/82686] [Medline: 41671575]
196. Houran J, O'Keeffe C. Prototyping haunted people syndrome: artificial intelligence-generated narrative, structured interview, and clinical validation. Psychol Conscious Theory Res Prac. 2025. [doi: 10.1037/cns0000447]
197. Conrow AR, Swift JK, Li A, Bingham J, Claps E. Can AI provide quality translation of measures for use in cross-cultural psychotherapy? Pract Innov (Wash D C). 2025. [doi: 10.1037/pri0000309]
198. Zhao Y, Qiu X, Xiao Z, Kaufman MR. Trauma narratives and healing of offspring caregivers of breast cancer patients: a mixed qualitative methods study. BMC Psychol. Dec 20, 2025;13(1):1395. [doi: 10.1186/s40359-025-03854-y] [Medline: 41422084]
199. Zhang X, Zayed A, Rehn Hamrin J, Güneysu A, Kuoppamäki S. Exploring body image awareness with a large language model–based conversational agent: qualitative study with young adults. J Med Internet Res. ;27:e78829. [doi: 10.2196/78829]
200. Pombal J, D'Eon M, Guerreiro NM, Martins PH, Farinhas A, Rei R. MindEval: benchmarking language models on multi-turn mental health support. arXiv. Preprint posted online on 2025. [doi: 10.48550/arXiv.2511.18491]
201. Hadar-Shoval D, Asraf K, Mizrachi Y, Haber Y, Elyoseph Z. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic values. JMIR Ment Health. Apr 9, 2024;11:e55988. [doi: 10.2196/55988] [Medline: 38593424]
202. Arnold VX, Abad MEP, Young SD. Can people tell the difference between AI-generated mental health vignettes? An exploratory comparison of user evaluations. Information. 2026;17(2):188. [doi: 10.3390/info17020188]
203. Scholich T, Barr M, Wiltsey Stirman S, Raj S. A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: mixed methods study. JMIR Ment Health. May 21, 2025;12:e69709. [doi: 10.2196/69709] [Medline: 40397927]

204. Hwang LJ, McKenzie R, Feng B, Pan W. Can you read between the lines?: A comparison of human and ChatGPT as support providers. Int J Human Comput Inter. Jul 3, 2026;42(13):10344-10364. [doi: 10.1080/10447318.2025.2584724]
205. Gariboldi F, Franchino E, Haima E, Lattanzi G, Grecucci A, Stella M. Cognitive networks reconstruct mindsets about STEM subjects and educational contexts in almost 1000 high-schoolers, university students and LLM-based digital twins. arXiv. Preprint posted online on 2026. [doi: 10.48550/arXiv.2602.14749]
206. De Freitas J, Uğuralp AK, Oğuz-Uğuralp Z, Puntoni S. Chatbots and mental health: Insights into the safety of generative AI. J Consum Psychol. Jul 2024;34(3):481-491. [doi: 10.1002/jcpy.1393]
207. Luykx JJ, Gerritse F, Habets PC, Vinkers CH. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. World Psychiatry. Oct 2023;22(3):479-480. [doi: 10.1002/wps.21145] [Medline: 37713576]
208. Algumaei A, Yaacob NM, Doheir M, Al-Andoli MN, Algumaie M. Symmetric therapeutic frameworks and ethical dimensions in AI-based mental health chatbots (2020–2025): a systematic review of design patterns, cultural balance, and structural symmetry. Symmetry (Basel). 2025;17(7):1082. [doi: 10.3390/sym17071082]
209. Leightley D, Williamson C, Korbacz A, Curcin V, Marshall IJ, Greenberg N. Effectiveness of AI-based interventions in workplace mental health: a systematic review and narrative synthesis. Br Med Bull. Jan 2, 2026;157(1):ldag007. [doi: 10.1093/bmb/ldag007] [Medline: 41668332]
210. Lee S, Rheu MM, Zhuang J. The ChatGPT effect: investigating shifting discourse patterns, sentiment, and benefit-challenge framing in AI mental health support. Behav Sci (Basel). Aug 28, 2025;15(9):1172. [doi: 10.3390/bs15091172] [Medline: 41009203]
211. Horn S. Exploring ChatGPT's potential for mental health and emotional well-being support: A nano review. J Agri Environ. 2026:186-189. URL: file:///Users/rohan/Downloads/ChatGPT-7.pdf [Accessed 2026-11-04]
212. Rowshon M, Mosaddeque A, Ahmed T, Twaha U. Exploring the impact of generative AI and virtual reality on mental health: opportunities, challenges, and implications for well-being. IJMRGE. 2025;6(2):10-796. [doi: 10.54660/.IJMRGE.2022.3.1.784-796]
213. Linardon J, Jarman HK, McClure Z, Anderson C, Liu C, Messer M. Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: experimental study. JMIR Ment Health. Nov 12, 2025;12:e80371. [doi: 10.2196/80371] [Medline: 41223407]
214. Woodnutt S, Allen C, Snowden J, et al. Could artificial intelligence write mental health nursing care plans? Psychiatric Ment Health Nurs. Feb 2024;31(1):79-86. [doi: 10.1111/jpm.12965]
215. Ohu FC, Burrell DN, Jones LA. Public health risk management, policy, and ethical imperatives in the use of AI tools for mental health therapy. Health Care (Don Mills). 2025;13(21):2721. [doi: 10.3390/healthcare13212721]
216. Malhotra A, Jindal R. XAI transformer based approach for interpreting depressed and suicidal user behavior on online social networks. Cogn Syst Res. Mar 2024;84:101186. [doi: 10.1016/j.cogsys.2023.101186]
217. Lozoya DC, D'Alfonso S, Conway M. Identifying gender bias in generative models for mental health synthetic data. Presented at: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI); Jun 26-29, 2023; Houston, TX. [doi: 10.1109/ICHI57859.2023.00109]
218. Gabriel S, Puri I, Xu X, Malgaroli M, Ghassemi M. Can AI relate: testing large language model response for mental health support. arXiv. Preprint posted online on 2024. [doi: 10.48550/arXiv.2405.12021]
219. Schnepper R, Roemmel N, Schaefert R, Lambrecht-Walzinger L, Meinlschmidt G. Exploring biases of large language models in the field of mental health: comparative questionnaire study of the effect of gender and sexual orientation in anorexia nervosa and bulimia nervosa case vignettes. JMIR Ment Health. Mar 20, 2025;12:e57986. [doi: 10.2196/57986] [Medline: 40111287]
220. Tak AN, Gratch J, Scherer KR. Aware yet biased: investigating emotional reasoning and appraisal bias in large language models. IEEE Trans Affective Comput. 2025;16(4):2871-2880. [doi: 10.1109/TAFFC.2025.3581461]
221. Thibodeau DE, Gollish SM, Boyes JE, Bijvoet E, Sabiston CM. The "ideal" body according to artificial intelligence: body image implications for athletes and nonathletes. Psychol Pop Media. 2025;15(3):347-359. [doi: 10.1037/ppm0000634]
222. Linardon J. Generative AI and eating disorders: exposing stereotypes in image depictions and setting a research agenda. Int J Eat Disord. Mar 2026;59(3):450-455. [doi: 10.1111/eat.70005] [Medline: 41254955]
223. Battista F, Lanciano T, Ribatti RM, Curci A. Malingering depression: a comparative study of human and GPT-3.5 performance. Curr Psychol. Feb 2026;45(4). [doi: 10.1007/s12144-025-09009-x]
224. Zhang W, Li J, Guo F, Zhu C, Kludacz-Alessandri M, Evans R. Diffusion of responsibility: patient moral judgments of generative AI-informed clinical decisions. Soc Sci Med. Mar 2026;392:118928. [doi: 10.1016/j.socscimed.2026.118928] [Medline: 41529517]
225. D'Souza RF, Surapaneni KM, Mathew M, Amanullah S, Tandon R. Choices of artificial intelligence (AI): ChatGPT's solutions to ethical dilemmas in bipolar disorder care. Bipolar Disord. Nov 2025;27(7):511-518. [doi: 10.1111/bdi.70059] [Medline: 40891014]

226. Poudel U, Jakhar S, Mohan P, Nepal A. AI in mental health: a review of technological advancements and ethical issues in psychiatry. Issues Ment Health Nurs. Jul 2025;46(7):693-701. [doi: 10.1080/01612840.2025.2502943] [Medline: 40378075]
227. Malgir BZ, Tian YA, Milford SR, Herger E, Shaw D, Elger BS. Healthcare professionals' perspectives on conversational mental health chatbots: a systematic review protocol. JMIR Research Protocols. Preprint posted online on 2026. [doi: 10.2196/preprints.92780]
228. Craddock N, Smith HG, Linardon J, Tylka TL, Diedrichs PC. Existing and future use cases, and safety and ethical considerations for AI in body image, and eating disorder prevention. Body Image. Mar 2026;56:102042. [doi: 10.1016/j.bodyim.2026.102042] [Medline: 41655362]
229. Long Y, Ma Z, Mei Y, Su Z. AffirmativeAI: towards LGBTQ+ friendly audit frameworks for large language models. arXiv. Preprint posted online on 2024. [doi: 10.48550/arXiv.2405.04652]
230. Nagra H, Mines RA, Dana Z. Exploring the impact of digital peer support services on meeting unmet needs within an employee assistance program: retrospective cohort study. JMIR Hum Factors. Feb 25, 2025;12:e68221. [doi: 10.2196/68221] [Medline: 39998863]
231. Eryılmaz A, Başal A. Rational AIs with emotional deficits: ChatGPT vs. counselors in providing emotional reflections. Curr Psychol. Dec 2024;43(45):34962-34977. [doi: 10.1007/s12144-024-06947-w]
232. Du Q, Ren Y, Meng ZL, He H, Meng S. The efficacy of rule-based versus large language model-based chatbots in alleviating symptoms of depression and anxiety: systematic review and meta-analysis. J Med Internet Res. Dec 4, 2025;27:e78186. [doi: 10.2196/78186] [Medline: 41343858]
233. Tan JT, Gan RK, Alsua C, et al. Psychological first aid by AI: proof-of-concept and comparative performance of ChatGPT-4 and Gemini in different disaster scenarios. J Clin Psychol. Aug 2025;81(8):726-738. [doi: 10.1002/jclp.23808] [Medline: 40347026]
234. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJH. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry. Oct 2016;15(3):245-258. [doi: 10.1002/wps.20346] [Medline: 27717254]
235. Habicht J, Dina LM, McFadyen J, et al. Generative AI-enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: real-world observational study. J Med Internet Res. Mar 10, 2025;27:e60435. [doi: 10.2196/60435] [Medline: 40063074]
236. Held P, Pridgen SA, Szoke DR, Chen Y, Akhtar Z, Amin D. Clinician perceptions of Socrates 2.0: a multi-agent artificial intelligence tool to facilitate Socratic dialogue. Cogn Behav Pract. Oct 2025. [doi: 10.1016/j.cbpra.2025.10.001]
237. Tian X, Liu Y, Litt S, Shin K, Jiang T. Artificial clients, genuine interactions? Assessing ChatGPT4's process-level performance in clinical simulations using qualitative content analysis. Couns Psychother Res. Dec 2025;25(4). [doi: 10.1002/capr.70063]
238. Khozaei F, Carbon CC, Islam QU, Ravari ZK. Exploring the perceived social support (PSS) of ChatGPT among female migrants in the gulf countries: a qualitative analysis. Open Psychol J. Jun 5, 2025;18(1). [doi: 10.2174/0118743501364500250515043823]
239. Matthews EB, Lerman D, Beach N, Wiczyk D, Goldkind L. "It's like having that supervisor in the room": examining AI as a reflective partner. Psychother Res. Sep 2026;36(7):1306-1317. [doi: 10.1080/10503307.2025.2569047] [Medline: 41084310]
240. Barzkar F, Zaribaf A, Mirfazeli FS, Keshavarz-Akhlaghi AA. The machine as therapist: unpacking transference and emotional healing in AI-assisted therapy. J Contemp Psychother. Dec 2025;55(4):361-368. [doi: 10.1007/s10879-025-09677-7]
241. Chan CKY. AI as the therapist: student insights on the challenges of using generative AI for school mental health frameworks. Behav Sci (Basel). Feb 28, 2025;15(3):287. [doi: 10.3390/bs15030287] [Medline: 40150182]
242. Chan C. An exploratory study of a narrative therapy chatbot with social work students. J Technol Hum Serv. Apr 3, 2026;44(2):140-167. [doi: 10.1080/15228835.2025.2559342]
243. Zhang B, Lau LY, Chen Y. Trends in digital mental health interventions: a 20-year bibliometric analysis. Digit Health. 2025;11:20552076251403203. [doi: 10.1177/20552076251403203] [Medline: 41323092]
244. Akiba D. ChatGPT told me to say it: AI chatbots and class participation apprehension in university students. Educ Sci. 2025;15(7):897. [doi: 10.3390/educsci15070897]
245. Christiansen M, Eriksson H, Fagerström L. Generative AI chatbots as digital adjuncts for sexual health information after prostate cancer in men who have sex with men: auto-netnographic study. JMIR Cancer. Feb 9, 2026;12:e81745. [doi: 10.2196/81745] [Medline: 41662502]
246. Ma Y, Zeng Y, Liu T, Sun R, Xiao M, Wang J. Integrating large language models in mental health practice: a qualitative descriptive study based on expert interviews. Front Public Health. 2024;12. [doi: 10.3389/fpubh.2024.1475867]

Abbreviations

BDI-II: Beck Depression Inventory-II
BERT: Bidirectional Encoder Representations from Transformers
CBT: cognitive behavioral therapy
CoT: chain-of-thought
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
LTGM: large language model therapy governance model
MAE: mean absolute error
MI: model misalignment
NLP: natural language processing
PCC: Population–Concept–Context
PHQ-4: 4-item Patient Health Questionnaire
PHQ-8: 8-item Patient Health Questionnaire
PHQ-9: 9-item Patient Health Questionnaire
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping reviews
PTSD: post-traumatic stress disorder
RAG: retrieval-augmented generation
RCT: randomized controlled trial
RQ: research question
SFT: supervised fine-tuning

Edited by John Torous; peer-reviewed by Antonio Miguel-Cruz, Hemalatha Sabbineni; submitted 19.May.2026; initial comments to author 12.Aug.2026; final revised version received 23.Sep.2026; accepted 06.Oct.2026; published 27.Nov.2026

Please cite as:
Babaei R, Cheng S, Li D, Cao J, Zhao S
Large Language Models (LLMs) in Mental Health: Scoping Review of Clinical Promise, Human Factors, and Implementation Challenges
JMIR Ment Health 2026;13:e101785
URL: https://mental.jmir.org/2026/1/e101785
doi: 10.2196/101785

© Reza Babaei, Samuel Cheng, Dan Li, Jie Cao, Shangqing Zhao. Originally published in JMIR Mental Health (https://mental.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 JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.