Review

Generative Large Language Models in Mental Health Care Settings: Systematic Review and Meta-Analysis

Janni Leung, PhD; Benjamin Johnson, BPsych (Hons); Kelsey McRae, BPsych (Hons); Stephanie Fong, BSc (Hons); Paula Cardona Gonzalez, BPsych (Hons); Caitlin McClure-Thomas, BPsych (Hons); Tianze Sun, PhD; Naomi Kern, BPsych (Hons); Yuen Ming Wong, BPsych (Hons); Richard Liu, MSW; Gary Chung Kai Chan, PhD

National Centre for Youth Substance Use Research, The University of Queensland, Brisbane, Queensland, Australia

Corresponding Author:

Benjamin Johnson, BPsych (Hons)
National Centre for Youth Substance Use Research
The University of Queensland
31 Upland Road
Brisbane, Queensland 4067
Australia
Phone: 61 429894154
Email: ben.johnson@uq.net.au

Abstract

Background: General-purpose large language models (LLMs) are increasingly being tested in mental health care, where language is central to assessment, diagnosis, risk evaluation, therapeutic interaction, monitoring, and patient education. However, their clinical usefulness, safety, and readiness for implementation remain uncertain. Existing reviews have largely been descriptive or scoping in nature, and broad health care reviews have not examined in detail the distinctive risks and applications of LLMs in mental health care.

Objective: We aim to systematically review empirical evidence on the use of general-purpose LLMs in mental health care; characterize the clinical tasks, study designs, models, outcomes, and methodological quality of the evidence; and synthesize findings across clinically meaningful task domains, including quantitative synthesis where sufficiently comparable studies were available.

Methods: We followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and searched PubMed, Embase, ACM Digital Library, IEEE Xplore, and Google Scholar from November 2022 to March 2026. Eligible studies reported quantitative data evaluating general-purpose LLMs for direct mental health care tasks. Methodological quality was assessed using the Mixed Methods Appraisal Tool and certainty of evidence using GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) domains. Findings were narratively synthesized, and random-effects meta-analyses were conducted on studies that tested LLMs on screening and diagnosis by ChatGPT-4 (OpenAI), ChatGPT-3.5 (OpenAI), and GPT-3 (OpenAI) models for mental health outcomes that reported on specificity and sensitivity. Hartung-Knapp adjustments were applied, and prediction intervals (PIs) estimated.

Results: We included 66 studies, comprising 37 vignette or simulation studies, 22 retrospective studies, and 7 prospective studies. Applications included screening and diagnosis (n=29), clinical decision support (n=14), treatment support (n=10), documentation and monitoring (n=6), patient education (n=4), and patient engagement (n=3). In screening and diagnosis, meta-analysis of 8 studies found that GPT-4 had the higher pooled sensitivity (0.83, 95% CI 0.38-0.97; 95% PI 0.02-1.00) than GPT-3.5 (0.70, 95% CI 0.13-0.97; 95% PI 0.00-1.00) and GPT-3 (0.61, 95% CI 0.33-0.82; 95% PI 0.10-0.96). However, GPT-4 specificity was lower at 0.77 (95% CI 0.52-0.91; 95% PI 0.10-0.99). Narrative synthesis suggested that LLMs performed most consistently in structured and linguistically explicit tasks. Certainty of evidence was generally low across domains, although documentation and monitoring reached moderate certainty. Major limitations included indirectness from vignette-based designs, uncertain representativeness, inconsistent outcome reporting, and sparse prospective real-world evaluation.

Conclusions: General-purpose LLMs show promise for selected mental health care applications. However, current evidence remains too heterogeneous, indirect, and uncertain to support routine unsupervised use, particularly for diagnosis, risk assessment, crisis response, or therapeutic interaction. Broad accessibility should not be mistaken for clinical readiness. Future

studies should move beyond simulations and retrospective evaluations toward prospective, real-world research assessing safety, reliability, equity, acceptability, clinical outcomes, and implementation in diverse mental health care settings.

JMIR AI 2026;5:e87730; doi: 10.2196/87730

Keywords: large language models; generative AI; mental health; primary health care; screening; clinical decision support; psychotherapy; digital health; eHealth; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Background

Mental health conditions are a major contributor to global disease burden [1]; yet, access to timely and appropriate care remains limited by persistent shortages in the mental health workforce [2-4]. This is particularly concerning in low-resource settings, where demand for mental health support often substantially exceeds available service capacity. As health systems seek scalable ways to expand access, recent advances in generative large language models (LLMs; eg, ChatGPT [5,6]) have attracted growing interest as tools that may assist selected aspects of mental health care delivery.

Mental health care is a particularly important setting for the evaluation of generative LLMs. For mental health care, spoken and written language underpin diagnostic assessment, risk evaluation, case formulation, therapeutic alliance, and monitoring of treatment response [7]. Several key symptoms of mental disorders are themselves expressed through language, including disorganized speech in psychosis and increased talkativeness in mania [8]. As a result, mental health care is especially exposed to both the opportunities and risks of language-based AI systems. This creates clear potential for LLMs to support mental health care, but also important risks. Errors in interpretation, hallucinated content, limited contextual understanding, and unsafe conversational responses may directly affect clinical judgment, patient safety, and therapeutic trust [8,9].

Research on LLMs in mental health has expanded rapidly since the public release of ChatGPT in late 2022. Previous reviews show that these systems have already been explored across a broad range of applications, including early detection and screening from text, conversational agents, psychoeducation, clinician-facing decision support, therapy-related tasks, and documentation-related uses [9-13]. However, these reviews also indicate that the field remains at an early stage and that much of the existing evidence is not yet well suited to informing clinical implementation. Many studies have relied on prompt-based experiments, vignettes, simulated conversations, cross-sectional comparisons, or evaluations of chatbot responses, with fewer studies using real patient care data or prospective designs embedded in clinical settings [9,14,15]. More broadly, reviews of LLMs in health care have similarly found that evaluations have focused heavily on question answering and accuracy, with limited use of real patient care data, substantial heterogeneity in tasks and outcomes, and inconsistent evaluation methods [5,6].

The rapid evolution of LLM technology further strengthens the rationale for reassessing the evidence base. Earlier reviews identified several limitations in the use of LLMs for mental health care, including inconsistent performance across extended interactions due to limited memory, limited clinical relevance of single-turn question-answer evaluations, and broader concerns regarding transparency and explainability [9,14,16]. Since then, newer generations of LLMs have introduced much larger memory capacity [17], stronger multiturn interaction capacity, and reasoning-oriented variants that can produce more explicit stepwise responses (eg, OpenAI's o1 [18]). As such, some limitations identified in earlier studies may now be less pronounced, but the updated model may pose new risks. Consequently, conclusions based largely on older generations of models may not fully capture the capabilities and risks of newer LLM systems.

Study Aims

This systematic review aimed to synthesize empirical evidence on the use of general-purpose LLMs in mental health care. Specifically, we aimed to identify the mental health care tasks for which these models have been evaluated; characterize the study designs, data sources, models, prompting approaches, comparators, mental health conditions, and outcome measures used; synthesize findings across clinically meaningful task domains; assess methodological quality and certainty of evidence; and consider implications for clinical practice and future research. Where groups of studies were sufficiently comparable in task, model, outcome domain, and reported performance metrics, we conducted quantitative synthesis. We focused on general-purpose LLMs because they are widely accessible and therefore among the models most likely to be used ad hoc by clinicians, patients, and health systems, making their evaluation especially relevant to implementation, safety, governance, and equity.

This review adds to existing broad reviews of LLMs and generative AI in health care by focusing specifically on direct mental health care applications, where language is both the medium of care and a central source of clinical information [5,19]. It also extends previous mental health-specific reviews, which have generally been descriptive or scoping in nature, by organizing the evidence into clinically meaningful task domains, distinguishing vignette-based, retrospective, and prospective evaluations, assessing certainty of evidence, and conducting quantitative synthesis only where sufficiently comparable evidence was available [9]. This approach clarifies not only where general-purpose LLMs appear promising, but also where the evidence remains too heterogeneous, indirect, or uncertain to support routine clinical implementation.

Methods

Search Strategy

This systematic review and meta-analysis was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (see Checklist 1 for PRISMA checklist; see Checklist 2 for PRISMA-S [Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension] checklist) [20]. The review protocol was registered on PROSPERO (International Prospective Register of Systematic Reviews) [CRD420251056593], which was a broader review on AI applications in any primary health care setting. This current paper deviates from the broader PROSPERO in that we focus on clinical applications in mental health care settings. This narrower focus was adopted for the current review because of the rapid expansion of original studies in mental health and the need for a dedicated synthesis of this emerging evidence base.

We searched PubMed, Embase, ACM Digital Library, IEEE Xplore, and Google Scholar for eligible studies. We also manually searched the reference lists of relevant systematic reviews and included papers to identify additional records. The search combined terms related to LLMs (eg, "large language model," "ChatGPT," "GPT," "Claude," and "Gemini") with terms related to mental health (eg, "mental health," "mental disorder," "depression," "anxiety," "suicide," "psychiatry," and "psychotherapy"). Full search strings are in Section S2 (Multimedia Appendix 1). For Google Scholar, results were screened in order of relevance until no further potentially eligible studies were identified.

The search was first conducted on January 11, 2025, and updated on June 26, 2025, and again on March 16, 2026. No language restrictions were applied at the search stage or in exclusion criteria, but all included studies were published in English.

Selection Criteria

The selection criteria were developed based on population or setting, intervention or exposure, outcomes, and study types (Section S1, Multimedia Appendix 1). The population and setting criteria included studies conducted in primary mental health care settings or scenarios (eg, general practice, family medicine, community health centers, and integrated care) for use by individual or group users, such as patients, primary care providers, and mental health practitioners. We excluded studies on populations or settings not relevant to primary mental health care, for example, individual self-help contexts outside clinical care, social media datasets not linked to a defined clinical or decision-making context, health promotion, academic, psychiatry student examinations, and mental health research applications [21-23].

For intervention or exposure, we included general-purpose LLMs, referring to large-scale, pretrained models trained broadly and designed for multipurpose use, without domain-specific training or restriction to clinical mental health applications, without fine-tuning, and without specific use of retrieval augmentation tools (including but not limited to the ChatGPT series, Claude [Anthropic PBC] series, Llama [Meta] series, PaLM/Gemini [Google LLC] series, and Mistral series). We were open to including papers that applied the general-purpose models to any mental health care setting. We excluded studies focusing only on image generation models, domain-specific or fine-tuned models, or small specialized systems developed for specific projects. Chatbot-based studies were eligible if the chatbot functioned primarily as an interface to a general-purpose LLM. Studies were excluded if they evaluated fine-tuned, custom-trained, or otherwise specialized chatbot systems whose performance was not representative of a general-purpose LLM.

We included any outcomes examined, which, based on previous reviews, included accuracy or performance metrics, time efficiency, objectively measured user satisfaction, impact on clinical decision-making, or patient mental health outcomes. Studies without outcomes quantified were excluded.

Study types included were empirical human studies with quantitative data, such as experimental, observational, implementation, evaluation, or mixed-methods studies with measurable outcomes. Opinion pieces, commentaries, reviews, theoretical frameworks, and purely qualitative studies without quantified outcomes were excluded. Preprints, conference papers, and abstracts were included if they met all the criteria, but those with no quantitative data were excluded.

We included studies published from November 2022 onward, corresponding to the public release and widespread accessibility of ChatGPT, the first widely accessible and publicly available general-purpose LLM, in any language and publication type (including peer-reviewed papers, preprints, and conference papers).

Study Selection

All records identified through database and supplementary searches were imported into Covidence for deduplication and screening. Title and abstract screening and full-text screening were conducted independently by two reviewers. Interreviewer agreement across screening decisions was 70.5% (10,957/15,542). Disagreements were resolved through discussion, and a third reviewer was consulted when consensus could not be reached. Screening decisions were guided by the predefined eligibility criteria, and common reasons for disagreement were discussed within the review team to support consistent interpretation of the criteria.

To minimize double counting, we examined studies with overlapping authors, datasets, or study designs and descriptions for possible duplication. Where multiple reports appeared to use the same or partially overlapping samples, input data, or outcomes, duplicate reports were excluded, and the most relevant or complete report was retained. Studies based on the same datasets were retained if different LLMs were tested or if there were different mental health outcomes.

Data Extraction

A standardized data extraction form was developed and piloted on 3 included studies before formal extraction commenced. Data were extracted by one reviewer and independently checked by a second reviewer. Any discrepancies were resolved through discussion, with consultation from a third reviewer when needed.

The extracted information included study characteristics (eg, country or region, study design, clinical setting or scenario, data source or population, and target mental health condition), intervention characteristics (eg, model name, model type, prompting approach, and primary mental health care task), and evaluation characteristics (eg, comparator or reference standard, outcome measures, and key findings).

Data Analysis

Methodological quality and risk of bias were appraised using the Mixed Methods Appraisal Tool (MMAT), which was selected because it provides design-specific criteria applicable to different study designs [24]. MMAT ratings were completed by 1 reviewer and independently checked by a second reviewer. Any disagreements were resolved through discussion until consensus was reached.

A narrative synthesis was used to summarize findings across the full set of included studies because substantial heterogeneity was anticipated in study design, clinical context, model type, prompting approach, and outcome measurement. For synthesis, studies were grouped according to the primary mental health care task evaluated based on studies identified, which were screening and diagnosis, clinical decision support, treatment support, documentation and monitoring, patient education, and patient engagement. These groupings were developed post hoc based on the characteristics of the included studies and were used to support structured comparison across task domains. Similarly, we also grouped the studies based on what designs the included papers had used (eg, prospective, retrospective, or vignette-based design). Prospective studies involved real-time data collection or live interaction with LLMs, including both real-world implementation and controlled research interview settings. Retrospective studies analyzed or tested LLMs on preexisting datasets without live deployment. Vignette-based studies evaluated LLMs using constructed scenarios or standardized prompts rather than real-world data.

Whether to conduct quantitative synthesis and meta-analyses was based on criteria used to determine if comparable estimates were available for pooling. This reflected both clinical and statistical considerations. Studies were grouped according to (1) mental health care task type; (2) LLM model tested; (3) mental disorder, symptom domain, or clinical outcome examined, guided where applicable by *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) or *ICD* (*International Classification of Diseases*) classifications; and (4) the quantitative metrics reported. Meta-analysis was conducted only when at least four estimates were available within the same task group [25].

Based on these criteria, meta-analysis was restricted to studies in the screening and diagnosis task group that reported sensitivity and specificity, or provided sufficient information to derive these estimates. As too few studies were available to conduct disorder-specific meta-analyses, we grouped studies within a broader internalizing distress and suicide-risk outcome family. This grouping included depressive and anxiety disorders, posttraumatic stress disorder (PTSD)–related outcomes, and suicidality-related outcomes. These outcomes were not treated as equivalent diagnoses; rather, they were grouped because they represent clinically overlapping mental health presentations commonly assessed in screening and risk-detection contexts.

As the outcomes to be pooled were sensitivity and specificity estimates, we conducted random-effects bivariate diagnostic test accuracy meta-analyses using the *MIDAS* package in Stata/SE 18 (StataCorp LLC). A bivariate approach was chosen because sensitivity and specificity are correlated and may vary jointly across studies, so modeling both outcomes together in the same model is appropriate. Separate meta-analyses were conducted by LLM models with at least four estimates for pooling, which included GPT-4, GPT-3.5, and GPT-3. Other models did not have sufficiently comparable data for pooling. It should be noted that in this review, we used the term "GPT-version" instead of referring to the generative pretrained transformer (GPT) family model generically as ChatGPT because we considered ChatGPT to be a web interface that allows connection to different versions of the model. We use the term ChatGPT without the version number when the original study authors did not report the version of GPT used. We did not calculate an overall pooled estimate across all models because combining different LLMs into a single summary estimate was not considered meaningful. We planned to conduct sensitivity analyses by excluding outliers, but there were no outliers identified; therefore, the sensitivity analyses were not conducted. From these models, we estimated pooled sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, and diagnostic odds ratio with 95% CIs. Between-study heterogeneity was assessed using Cochran Q and I^2 statistics, and model adequacy was examined using diagnostic plots [26]. We estimated prediction intervals (PIs) and CIs, and Hartung-Knapp adjustments were applied [27,28].

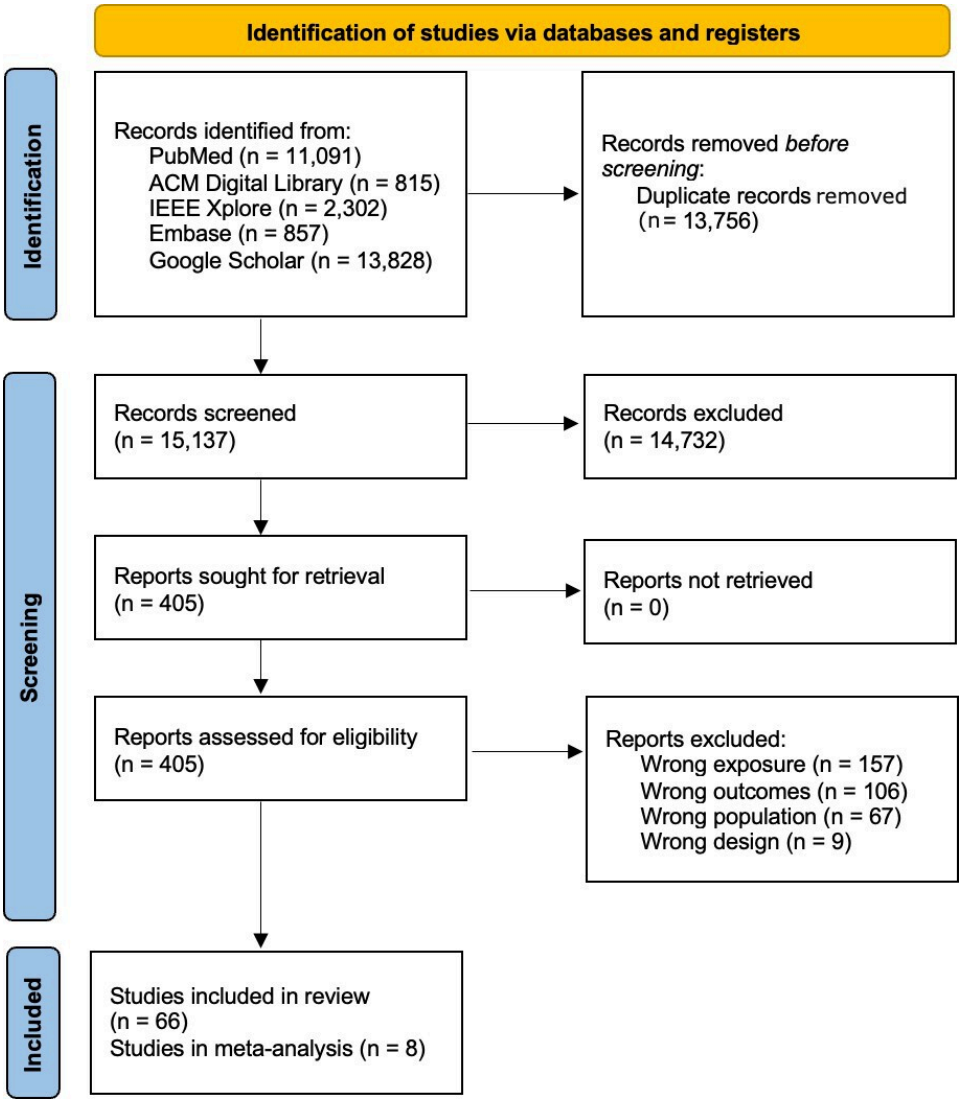
In addition, we used evidence from the MMAT, narrative review, and meta-analyses to determine the current level of evidence for each of the mental health care tasks that LLMs had been tested on, based on the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) domains [29].

Results

Study Selection

We identified a total of 28,893 records through database and register searching (PubMed n=11,091; ACM Digital Library n=815; IEEE Xplore n=2302; Embase n=857; Google Scholar n=13,828). After removing 12,777 duplicate records and 979 records for other reasons, 15,137 records were screened, of which 14,732 were excluded at title and abstract screening. We then screened 405 full-texts for eligibility. Finally, a total of 66 studies met the inclusion criteria and were included in the systematic review, with 8 studies included in the meta-analysis (Figure 1).

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



Study Characteristics

The characteristics of the 66 included studies are presented in Multimedia Appendix 2. Across the included studies, 48.5% (32/66) of studies were conducted in North America, 15.2% (10/66) of studies in Europe, 19.7% (13/66) of studies in East Asia, 18.2% (12/66) of studies in the Middle East, 4.5% (3/66) of studies in South Asia, and 3% (2/66) of studies in Oceania; some studies spanned multiple regions.

By study design, 56.1% (37/66) of studies were vignette or simulation studies, 33.3% (22/66) of studies were retrospective studies, and 10.6% (7/66) of studies were prospective studies. By application type, 43.9% (29/66) of studies focused on screening and diagnosis, 21.2% (14/66) of studies on clinical decision support, 15.2% (10/66) of studies on treatment support, 9.1% (6/66) of studies on documentation and monitoring, 6.1% (4/66) of studies on patient education, and 4.5% (3/66) of studies on patient engagement.

Target conditions were most commonly general or nonspecific mental health presentations, examined in 31.8% (21/66) of studies, followed by depression in 28.8% (19/66) of studies and suicidality or suicide risk in 16.7% (11/66) of studies. Anxiety-related conditions and schizophrenia or psychosis were each examined in 7.6% (5/66) of studies, bipolar or other mood disorders in 6.1% (4/66) of studies, and PTSD in 4.5% (3/66) of studies. Autism spectrum disorder, obsessive-compulsive disorder (OCD), and psychiatric emergencies or acute psychiatric crises were each

examined in 3% (2/66) of studies. Attention-deficit/hyperactivity disorder and substance use or addiction psychiatry were each examined in 1.5% (1/66) of studies. As some studies addressed more than one target condition, these categories were not mutually exclusive.

Quality Assessment

Based on the MMAT appraisal, methodological quality was generally acceptable, although study quality varied across designs (Section S6 in Multimedia Appendix 1). Among the quantitative nonrandomized studies, 60.4% (29/48) met at least 4 of the 5 MMAT criteria. The most common limitation was sample representativeness: this criterion was rated as unclear in 31.3% (15/48) of studies and not met in 12.5% (6/48) of studies, suggesting potential limitations in generalizability. Control for confounding was reported in 31.3% (15/48) of studies. Among the quantitative descriptive studies, 75% (6/8) of studies met at least 4 of the 5 MMAT criteria, with the main concerns relating to sample representativeness and nonresponse bias. Among the mixed methods studies, 90% (9/10) met at least 4 of the 5 MMAT criteria. Where mixed methods criteria were not fully met, this was usually due to limited justification for the mixed methods design, insufficient integration of qualitative and quantitative findings, or inadequate discussion of divergences between components. Overall, the evidence base was methodologically heterogeneous, and the principal concern was uncertainty regarding representativeness rather than pervasive major flaws in study conduct. The full assessment can be seen in Section S5 (Multimedia Appendix 1). In the following section, studies are summarized by application types.

Screening and Diagnosis

Screening and diagnosis referred to tasks in which LLMs were used to classify or identify a mental health condition or estimate symptom risk or severity. We identified 29 studies on screening and diagnosis, including 14 retrospective, 13 vignette, and 2 prospective studies.

The 29 studies used heterogeneous methods to evaluate LLM performance in mental health screening and diagnostic classification, including retrospective analyses of patient narrative [30], clinical interview transcripts [31-36], diary entries [37], hospital clinical notes and discharge summaries [38-40], sentence completion narratives [41], inpatient psychiatric admission narratives [42], and structured questionnaire responses [43]. In addition, prospective studies used LLM-derived sentiment ratings from brief written responses [44] and depression screening through conversations [45]. The remaining studies were vignette-based or other simulated evaluations, including *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) or *DSM-5-TR* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition, Text Revision]) psychiatric diagnostic vignettes [46-48], suicidality and suicide-risk scenarios [48-53], childhood anxiety vignettes [54], and OCD diagnostic vignettes [55,56]. Reference standards and comparators were similarly diverse, including validated symptom scales and thresholds [30-33,35-37,41,43], psychiatrist adjudication or clinical diagnosis [37,45], human annotation with physician validation [38], benchmark *DSM*-based vignette diagnoses [46-48], and mental health professional comparators [51,53,54,57,58].

Across the vignette-based screening and diagnosis studies, LLM performance was mixed. Stronger results were reported in more narrowly defined vignette tasks, such as childhood anxiety disorders, where LLMs often matched or exceeded human comparators or achieved high diagnostic accuracy [54,55,57]. Performance was also strong in OCD studies, where LLMs achieved high diagnostic accuracy and outperformed human comparators [55,56]. In broader psychiatric vignette studies, GPT-3.5 produced broadly acceptable responses across 100 psychiatric case vignettes [59], while newer frontier models showed moderate-to-strong diagnostic performance, with better reasoning associated with greater diagnostic accuracy [60]. Similarly, in a broader psychiatric vignette set, performance was high for depression, social phobia, and PTSD, but weaker for schizophrenia [58].

Structured approaches also improved performance in some cases, such as self-verification prompting, which increased positive predictive value in *DSM-5-TR* case diagnosis in a study [48], but reduced sensitivity [61]. However, important limitations were mentioned in some studies, including weaker performance for disorders in which the diagnosis depends on the symptoms as well as the timing and course of the symptom presentations, such as peripartum depression and acute stress disorder [47]. In addition, limitations mentioned included variability across diagnostic categories and demographic bias [46], delayed escalation and limited crisis referral in simulated suicidality scenarios [49], and systematic deviation from professional norms in suicide-risk judgments, with risks over- or underestimated depending on model type [50,51]. Other vignette-based studies further highlighted variability in referral decisions, intervention stability, and benchmark performance across broader psychiatric or suicide-related tasks [54,62].

Across the retrospective studies, findings were similarly mixed. LLMs performed well for autism spectrum disorder, depression, and PTSD estimation from interviews and inpatient psychiatric diagnosis from admission narratives, although conventional models still outperformed LLMs in some settings [32,34,42,63]. Some studies also reported promising performance for symptom scoring from interview transcripts and suicide-risk estimation from discharge summaries, although discrimination for suicide-risk prediction remained only modest [39,40]. A retrospective study also showed strong performance for binary condition identification from hospital clinical notes, as illustrated by high sensitivity and specificity for depression extraction in 1 small emergency health record study [38]. However, performance was weaker in other applications, including childbirth-related PTSD screening, where sensitivity was low despite high specificity, and adolescent suicidal-ideation estimation, where traditional machine-learning models slightly outperformed LLM-based approaches [30,43]. Other retrospective studies also showed only moderate performance for depression and suicide-risk prediction from sentence-completion narratives

and for extraction of mental health causes, although prompting refinements improved results in some settings [33,41,64].

In the 2 prospective studies, a GPT-based fast screen showed high agreement with clinical diagnosis and outperformed the validated Patient Health Questionnaire scale for screening for depression, but in a very small sample [45]. In a separate prospective study, LLM-derived sentiment ratings from brief written responses were associated with both current and future depression severity and significantly predicted short-term worsening, suggesting potential utility for language-based mood monitoring rather than direct diagnosis [44].

For consideration in the meta-analysis, of the 29 studies on screening and diagnosis summarized above, 8 provided sufficiently comparable data on depression and anxiety disorders, PTSD-related outcomes, and suicidality-related outcomes included [30,31,33,37,38,43,45,46]. Reference standards varied across studies and included validated symptom scales [30,31,33,43], symptom scale thresholds with psychiatrist adjudication [37], psychiatrist clinical diagnosis [45], manual human annotation with physician validation [38], and *DSM-5*-based vignette diagnoses [46].

Visual inspection of the meta-analyses' diagnostic outputs did not suggest concerns about the model fit, and funnel plot asymmetry tests were not statistically significant (GPT-4 P=.79, GPT-3.5 P=.31, GPT-3 P=.42), noting the small number of studies (Table 1; Sections S2-S4, Multimedia Appendix 1). The bivariate meta-analysis found that GPT-4 had the highest pooled sensitivity (0.83, 95% CI 0.38-0.97; 95% PI 0.02-1.00), but lower specificity (0.77, 95% CI 0.52-0.91; 95% PI 0.10-0.99) than GPT-3.5 and GPT-3, but with large CIs and PIs. GPT-3.5 showed a pooled sensitivity of 0.70 (95% CI 0.13-0.97; 95% PI 0.00-1.00) and specificity of 0.96 (95% CI 0.95-0.96; 95% PI 0.95-0.96), while GPT-3 showed a pooled sensitivity of 0.61 (95% CI 0.33-0.82; 95% PI 0.10-0.96) and specificity of 0.94 (95% CI 0.88-0.96; 95% PI 0.87-0.96). GPT-3.5 had the highest positive likelihood ratio (15.90, 95% CI 9.30-27.30), whereas GPT-4 had the lowest negative likelihood ratio (0.23, 95% CI 0.11-0.47). Heterogeneity was large. Overall, the meta-analyses showed that GPT-4 had the highest pooled sensitivity, but poor specificity with wide adjusted CIs, and wide PIs indicating substantial between-study variability in expected performance across settings.

Table 1. Bivariate meta-analysis of diagnostic test accuracy of LLMs[a] on internalizing and distress-related mental conditions. There were not enough like-for-like studies for meta-analyses for other LLMs, other areas of mental health care applications, and other mental health outcomes. Forest plots and diagnostic plots are available in Sections S2-S4 in Multimedia Appendix 1.

	GPT-4 (k=4)	GPT-3.5 (k=4)	GPT-3 (k=5)
Pooled sensitivity, pooled estimate (95% CI)	0.83 (0.38-0.97)	0.70 (0.13-0.97)	0.61 (0.33-0.82)
Prediction intervals	0.02-1.00	0.00-1.00	0.10-0.96
Q for sensitivity	35.12[b]	41.44[b]	12.25[c]
I-sq for sensitivity, pooled estimate (95% CI)	91.46 (84.75-98.16)	92.76 (87.34-98.19)	67.36 (36.28-98.43)
Specificity, pooled estimate (95% CI)	0.77 (0.52-0.91)	0.96 (0.95-0.96)	0.94 (0.88-0.96)
Prediction intervals	0.10-0.99	0.95-0.96	0.87-0.96
Q for specificity	58.10[b]	16.50[b]	6.23[d]
I-sq for specificity, pooled estimate (95% CI)	94.84 (91.33-98.35)	81.82 (64.51-99.13)	35.80 (0.00-98.86)
Positive likelihood ratio, pooled estimate (95% CI)	3.60 (2.50-5.30)	15.90 (9.30-27.30)	10.50 (6.20-17.90)
Negative likelihood ratio, pooled estimate (95% CI)	0.23 (0.11-0.47)	0.31 (0.11-0.91)	0.42 (0.25-0.70)
Asymmetry test (P value)	.79	.31	.42

[a]LLM: large language model.
[b]P<.01.
[c]P=.02.
[d]P=.18.

Clinical Decision Support

Clinical decision support studies included those that tested LLMs in supporting mental health-related clinical judgments, such as prognosis, triage, medication support, or evaluation of therapeutic responses. We identified 14 studies, including 13 vignette-based studies and 1 prospective simulated training study.

The clinical decision support studies covered prognosis, treatment selection, medication support, suicide-response scoring, emergency triage, psychotherapy case-response tasks, counseling benchmarks, and cognitive behavioral therapy (CBT) knowledge and therapist-response tasks. In studies examining schizophrenia and depression vignettes, models generally recommended treatment appropriately, but prognostic judgments varied across models, with GPT-3.5 tending to be more pessimistic than newer models and human comparators [54,65].

Studies of clinical decision support showed promise but also important safety limitations. GPT-4 often selected appropriate antidepressant treatments, but also included contraindicated or poor options in many evaluations [66]. Guideline augmentation improved bipolar treatment selection substantially, although some inappropriate treatments remained [67]. Retrieval-augmented psychiatric medication support also performed well, especially with GPT-4o-based systems [68], and *ICD-11* (*International Classification of Diseases, 11th Revision*) criteria could be

translated into highly accurate executable logic after expert correction [55].

Performance was more mixed on complex response-generation tasks. Models showed bias when scoring suicide-intervention responses [69], no model achieved consistently acceptable performance on psychotherapy case-response tasks [70], and LLMs underperformed human reference responses on CBT therapist-response generation despite strong knowledge performance [62]. In psychiatric emergency triage, GPT-4 models showed substantial agreement with clinicians, but some false positives suggested over-triage [71]. Overall, these findings suggest that LLMs have considerable potential for structured mental health clinical decision support, but still require careful oversight for prognosis, safety-sensitive decisions, prescribing, and therapeutically nuanced tasks.

Treatment Support

Treatment support referred to tasks in which LLMs were used to deliver, simulate, or evaluate therapeutic or supportive interactions, such as psychotherapy-style conversations, CBT-based responses, journaling support, or feedback on helping skills, rather than to screen for a condition or make a formal clinical decision. We identified 10 studies on treatment support, including 5 vignette-based studies, 1 retrospective study, and 4 prospective studies.

The treatment support studies examined a range of applications, including simulated CBT sessions [72], anxiety support through repeated chatbot conversations [73], AI-generated feedback for suicide-prevention role-play training [74], CBT-style cognitive distortion and reframing tasks [75], AI-supported journaling for mental well-being [76], depression treatment recommendations from vignette scenarios [53], psychotherapy knowledge and behavioral activation tasks [58,77], responses to common psychological questions [78], and retrospective CBT-style dialogue generation from psychotherapy transcripts [79]. Reference standards and comparators were similarly varied, including human CBT therapists [72,75], primary care physicians [53], psychotherapists in training [58,77], a licensed clinical psychologist [78], trained human psychology raters [74], human psychotherapy dialogue datasets [79], and user-reported evaluations without a formal comparator [73,76].

Across the 5 vignette-based treatment support studies, findings were generally positive but mixed [50,72,77,78]. LLMs performed moderately to strongly on several structured treatment-support tasks, including CBT-style exercises [75], behavioral activation and psychotherapy case-scenario responses [77], depression treatment recommendation [53], and psychological support question answering [78]. However, important limitations were also identified. Human CBT therapists outperformed GPT-3.5 across multiple CBT competence domains in simulated therapy sessions [72]. In depression vignettes, GPT-3.5 and GPT-4 more often recommended psychotherapy, whereas physicians more often recommended pharmacotherapy [53]. Performance also varied across models, with GPT-4 outperforming GPT-3.5 on common psychological prompts in 1 study [78].

In the 1 retrospective study, LLMs showed feasible CBT-style dialogue generation from psychotherapy transcripts, with stronger performance in multiturn than single-turn settings and further gains when a CBT knowledge base was added [79]. Responses were generally more positive than those of human therapists, suggesting a possible overly positive bias despite reasonable dialogue quality [79].

Across the 4 prospective studies, users and evaluators generally reported favorable experiences with LLM-supported interventions, including anxiety support [73], AI journaling [76], suicide-prevention role-play feedback [74], and psychotherapy training support [58]. Among adults with anxiety disorders, most participants reported that GPT-3.5 understood their anxiety accurately, and ratings of helpfulness, empathy, trustworthiness, and effectiveness were generally favorable, although privacy, ethics, and lack of human connection remained common concerns [73]. AI-supported journaling also received high ratings across counseling skill, behavior, and learning domains [76]. In suicide-prevention role-play training, AI-generated feedback correlated strongly with human ratings, although it tended to score performances more favorably than human raters [74]. In psychotherapy training tasks, LLMs performed similarly to or better than psychotherapists in training on several knowledge and response-quality indicators [53]. Overall, these findings suggest that LLMs may be useful as supportive or training-adjunct tools, but their outputs still require cautious interpretation in sensitive therapeutic contexts [53,73,74].

Documentation and Monitoring

Documentation and monitoring referred to tasks in which LLMs were used to summarize sessions, generate discharge summaries, standardize notes, extract structured information from free-text records, score written documentation, or monitor symptom change from routine clinical text. We identified 6 studies on documentation and monitoring, including 5 retrospective studies and 1 vignette-based study.

These 6 studies used heterogeneous methods to evaluate LLM performance in documentation and monitoring tasks, including structured summarization of counseling-session transcripts [80], classification of mental health concepts from emergency electronic health record (EHR) text [81], scoring of suicide safety-plan documentation [82], proofreading and structured information extraction from addiction psychiatry notes [83], symptom monitoring from psychiatric EHR language during clozapine treatment [84], and generation of psychiatric discharge summaries from structured case information [85]. Reference standards and comparators were similarly diverse, including human-annotated reference summaries [80], consensus coding by expert clinicians [81], trained clinical coders using a structured scoring algorithm [82], human-annotated gold-standard proofread notes and extraction labels with comparison against non-LLM tools [83], human-rated mental health symptom scores and conventional natural language processing features [84], and human-written discharge summaries by physicians and psychotherapists [85].

Across the 5 retrospective studies, findings were overall positive, with variation by task. LLMs performed well on structured summarization, concept classification, documentation scoring, information extraction, and symptom monitoring from routine clinical language [80-84]. In counseling-session summarization, hallucinations were uncommon across outputs [80]. In emergency EHRs, where clinicians have to file cases under mental health or physical health categories, performance was strongest for the binary mental-vs-physical distinction, with κ=0.77, precision 0.93, recall 0.93, and F_1-score 0.93, but dropped for finer-grained categorization across specific mental and physical health classes [81]. In suicide safety-plan scoring, best F_1-scores ranged from 0.77 to 0.92 across sections [82]. In addiction psychiatry notes, larger LLMs outperformed simpler non-LLM tools for proofreading, and GPT-4o performed strongly on structured information extraction, with a mean F_1-score of 0.97 for substance-class detection, an exact match of 0.96 for time since last use, and mean accuracy of 0.97 for adequacy of time-since-last-use information [83]. In monitoring applications, LLM-derived psychiatric symptom severity scores captured symptom improvement during clozapine treatment and correlated positively with human-rated measures for key psychotic and behavioral features [84]. Overall, performance appeared strongest for more structured extraction and classification tasks, whereas results were mixed for more specific categorization and documentation tasks, suggesting that LLMs performed more consistently in tasks with low complexity.

In the 1 vignette-based study, GPT-4 generated psychiatric discharge summaries from structured case information that were of similar quality to human-written summaries, but did so substantially faster [85]. These findings suggest that LLMs may be used for documentation support to save time, although the evaluation was based on only 2 cases.

Patient Education

Patient education included tasks in which LLMs were used to answer patient or caregiver questions, provide psychoeducational information, or respond to common mental health information needs and questions. We identified 4 studies on patient education, including 3 vignette-based studies and 1 retrospective study.

These studies examined responses to autism-related consultation questions in a real-world online care setting [86], real-world mental health questions from an online counseling forum evaluated in a simulated format [87], psychosis-related psychoeducation questions [88], and common antidepressant-related questions paired with short patient scenarios [89]. Reference standards included physician responses [86], licensed mental health professional ratings together with comparison against online human therapist responses and LLM-as-judge evaluations [87], psychiatrist and psychologist ratings [88], and blinded psychiatrist ratings comparing LLM and psychiatrist responses [89].

Across the 3 vignette-based studies, GPT-4 provided highly accurate, clear, and clinically useful psychoeducation for psychosis-related questions, although inclusivity was the weakest-rated domain and readability was only moderate [88]. In responses to general mental health questions, a study showed that performance varied across models, with LLaMA-3.3 showing the strongest overall performance, while GPT-4 had lower overall quality and more refusal-style safety disclaimers, and Gemini showed lower empathy despite similar factual consistency [87]. Human therapist responses scored lower overall than LLM responses in that study, although 7%-14% of LLM outputs still contained unauthorized medical advice [87]. For antidepressant-related questions, GPT-4o performed similarly to psychiatrists in accuracy, was more concise, but produced fewer clear responses, with no significant difference in readability [89].

In the retrospective study, physicians were preferred overall to GPT-4 for autism-related consultation questions, with physicians scoring higher on relevance and usefulness, whereas ChatGPT scored highest on empathy and slightly exceeded physicians on correctness [86]. Overall, these findings on patient education and question answering suggest that LLMs can provide useful patient education responses in mental health settings, but performance remains dependent on the model used and the evaluation domain, with ongoing concerns around clarity, safety, and response quality relative to clinicians [86-89].

Patient Engagement

Patient engagement referred to tasks in which LLMs were used as interactive, user-facing systems to engage people in mental health conversations or crisis-oriented exchanges, rather than primarily to provide formal diagnosis or structured treatment. Among the 3 patient engagement studies, all 3 were conducted in simulated or vignette-based settings [90-92], and none used retrospective or prospective real-world clinical data.

LLMs were generally perceived as capable of providing supportive or contextually appropriate responses, particularly in lower-risk or general mental health scenarios, but important limitations were identified in more complex or crisis-sensitive situations [69,90,92]. In a psychiatric crisis role-play study, participants rated GPT-3.5 positively overall for helpfulness, pleasantness, and appropriateness, although ratings were lower in psychosis scenarios than in depression or adjustment disorder scenarios [90]. In a study of suicide-related queries, LLMs including Claude 3.5, GPT-4o, and Gemini 1.5 appropriately avoided directly answering very-high-risk questions, but showed limited ability to consistently distinguish between intermediate levels of suicide risk [69]. In a comparison with licensed therapists, LLMs demonstrated some therapeutic elements such as reassurance and psychoeducation, but were judged to be more generic, overly directive, and potentially unsafe in crises [92].

Certainty of Evidence

Overall, taking into account the quality appraisal, meta-analysis findings, and narrative synthesis, the certainty of evidence was generally low across mental health care tasks (Table 2). The use of LLMs for documentation and monitoring of mental health care had the highest level of evidence,

which was moderate. The certainty of evidence was downgraded for other mental health tasks, where evidence was sparse, mixed, or based on simulated designs. A substantial proportion of existing evidence was based on vignette-based studies, limiting the directness and clinical applicability of the evidence. Many other studies were retrospective, which was moderate, but a higher level of evidence would be achieved by having prospective or trial-based studies conducted. Across domains, there were important methodological limitations, including unclear confounding assessment, uncertain sample representativeness, inconsistent outcome measures, unclear publication bias, and imprecision in reported estimates. Heterogeneity was also substantial in the pooled screening and diagnosis studies, with wide CIs and PIs for some outcomes, indicating considerable uncertainty and variation across settings.

Table 2. Certainty of evidence based on GRADE[a] assessments.

Mental health care task	Studies conducted	Summary	Certainty of the evidence
Screening and diagnosis	29 studies; 13 vignette, 14 retrospective, 2 prospective	There was a relatively larger body of evidence available, but findings were mixed across mental health outcomes and task types. Most studies relied on vignette-based or retrospective datasets, with relatively few prospective evaluations. Performance was often promising for some structured classification tasks, but results varied by condition, prompting strategy, and model, and several studies identified important safety or bias-related concerns.	Low
Clinical decision support	14 studies: 13 vignette, 0 retrospective, 1 prospective	Nearly all evidence came from simulated or vignette-based tasks, limiting direct clinical applicability. Findings varied by task and model, and several studies identified clinically important weaknesses, including biased prognostic judgments, over- or undertriage, and inclusion of contraindicated or poor treatment suggestions.	Low
Treatment support	10 studies: 5 vignette, 1 retrospective, 4 prospective	Evidence suggested potential usefulness of LLMs[b] in supportive or adjunctive treatment roles, but findings varied across contexts, outcomes, and prompting strategies. Prospective studies reported positive experiences with LLMs, although concerns remained regarding privacy, human nuance, and safety in sensitive settings.	Low
Documentation and monitoring	6 studies: 1 vignette, 5 retrospective, 0 prospective	Findings were more consistent in showing positive results, especially for structured extraction, classification, summarization, and documentation tasks. However, all evidence came from retrospective analyses or simulated cases, with no prospective implementation studies.	Moderate
Patient education	4 studies: 3 vignette, 1 retrospective, 0 prospective	Evidence was limited to a small number of mainly simulated studies. Findings generally suggested that LLMs could produce accurate and well-rated psychoeducational or question-answering responses, although safety, clarity, readability, and appropriateness concerns remained. There was a lack of direct real-world evaluation.	Low
Patient engagement	3 studies: 3 vignette, 0 retrospective, 0 prospective	Evidence was sparse and based entirely on simulated interactive scenarios. Findings suggested that LLMs could be engaging and acceptable in some contexts, but concerns remained regarding suitability in complex or crisis-related situations, and the small evidence base limits confidence.	Low

[a]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.
[b]LLM: large language model.

Discussion

Principal Findings

This systematic review provides a clinically focused synthesis of empirical evidence on general-purpose LLMs in mental health care. In contrast to broad reviews of LLMs across medicine, this review focuses on a setting in which language is central to clinical assessment, risk evaluation, therapeutic interaction, monitoring, and patient engagement [5,19]. We found that general-purpose LLMs have been evaluated across a wide range of mental health care tasks, including screening and diagnosis, clinical decision support, treatment support, documentation and monitoring, patient education, and patient engagement. The main contribution is a structured assessment of what has been tested, how it has been evaluated, where evidence is beginning to accumulate, and where the evidence remains too heterogeneous or indirect to support clinical implementation. Quantitative synthesis was feasible only for a subset of screening and diagnostic studies with comparable diagnostic test accuracy metrics; all other task domains required narrative synthesis because of substantial variation in tasks, models, data sources, comparators, outcome measures, and reporting. Overall, the findings suggest that general-purpose LLMs show promise for selected mental health care applications, but the evidence remains early, uneven, and insufficient to support routine unsupervised use.

A consistent pattern across the included studies was that general-purpose LLMs performed better in tasks that were relatively structured, constrained, and linguistically explicit

[44,80,85,93]. This was most apparent in documentation and monitoring, where models often performed well in summarization, information extraction, concept classification, and other forms of structured text processing [83]. Similar strengths were also evident in some screening, diagnostic, and decision-support tasks when the input format and expected outputs were clear. Many of these tasks, such as classification, extraction, summarization, and question answering, are well-established areas of language model research in which LLMs have shown strong performance across a wide range of domains outside of mental health care [94,95]. The relative strength observed in these more bounded mental health care applications is therefore broadly consistent with the known capabilities of contemporary LLMs as general-purpose language-processing systems [96].

Performance was less reliable in tasks that required nuanced clinical judgment, sustained interpersonal responsiveness, or safe management of ambiguity and risk. In screening and diagnostic applications, results varied across conditions, models, and input formats, and even when overall classification performance appeared encouraging, with better detection rates in more recent LLMs, sensitivity was often less reassuring than specificity, raising concern about missed cases [30,32]. Further, the large PIs indicated that performance cannot be predicted across settings. In treatment support, patient education, patient engagement, and some clinical decision-support applications, LLMs often produced plausible, coherent, and well-structured responses [63,73,76], but this did not always translate into clinically appropriate or trustworthy performance. Reported limitations included generic or overly positive therapeutic responses, inconsistent crisis escalation, weaker handling of temporally nuanced diagnoses, variable prognostic judgments, and concerns about privacy, personalization, and cultural sensitivity [97,98]. Similar concerns have been raised in relation to LLM-supported substance use disorder interventions, where potential benefits such as low-threshold support need to be weighed against risks relating to stigma, demographic bias, hallucinated or unsafe advice, privacy, governance, and the need for robust human oversight [99]. Particularly in patient engagement and crisis-related settings, the evidence remained limited and indirect, being based entirely on simulated scenarios [90-92]. These findings suggest that generating plausible mental health language is not equivalent to demonstrating dependable clinical judgment in more complex or high-stakes settings.

Although this review focused on general-purpose LLMs rather than specialized or fine-tuned mental health models, the broader literature suggests a similar overall pattern. Customized ChatGPT variants and fine-tuned LLMs have shown promise in simulated counseling interactions and in structured tasks such as summarizing counseling notes, discharge summaries, and medication logs [44,73,80,85]. However, these models also continued to raise concerns regarding privacy, ethics, intervention depth, and missed emotional nuance, limiting their reliability as standalone tools [44,80]. Taken together, these findings suggest that both general-purpose and more specialized LLMs share similar limitations in mental health care settings [100].

Our findings are broadly aligned with earlier reviews showing that LLM research in mental health has grown rapidly but has been dominated by early-stage and indirect evaluations [10,12,15]. Prior mental health reviews found that much of the literature consisted of prompt experiments, vignette studies, or evaluations of chatbot responses, with very few prospective studies involving participants [10,12,15]. Broader health care reviews reached similar conclusions, reporting that only a small proportion of studies used real patient care data and that most evaluations relied on examination questions, vignettes, or expert-generated prompts [101]. Further, ethical concerns raised included fairness, bias, being too positive, transparency, and privacy [100]. Compared with these earlier syntheses, the present review suggests an important, although still incomplete, progress in the evidence base. Retrospective clinical data and some prospective real-world evaluations are now beginning to appear in mental health applications of general-purpose LLMs [32,34,42,63]. This review also extends prior work by quantitatively synthesizing a subset of screening and diagnostic studies, whereas many earlier reviews were primarily descriptive, narrative, or scoping in nature [9,13]. Our focus on general-purpose LLMs is also important because these are the most widely accessible models, and therefore the ones most accessible in practice-like settings or used ad hoc by clinicians and patients [102]. This brings practical value to the field by clarifying not only where these models appear most promising, but also where the evidence remains too uncertain, heterogeneous, unpredictable, or indirect to support routine use.

Limitations

This review should be interpreted in light of several limitations in the available evidence base and synthesis. First, much of the literature was still based on vignette, simulated, or retrospective studies rather than prospective evaluations embedded in routine care, as mentioned above [54,63]. Although these designs are useful for early-stage assessment, they do not fully capture the complexity, uncertainty, and longitudinal nature of real mental health care encounters, in which presentations are more heterogeneous, and interactions unfold over time [10,12,15]. As a result, ecological validity remains limited, and the safety, clinician trust, and real-world uptake of these systems remain insufficiently tested.

Second, the evidence base was highly heterogeneous, with substantial variation in clinical tasks applied, target mental health conditions, LLM model versions, prompting strategies, comparators, data sources, and outcome measures. This limited direct comparison across studies, constrained generalizability, and reduced the extent to which findings could be synthesized quantitatively. These issues also contributed to reduced confidence in the certainty of evidence. There were still only a small number of studies that were based on human data. Future studies will need better causal design, such as well-designed cohort studies

with comprehensive confounding adjustment, to estimate the effect of LLMs on various clinical tasks [103,104].

Third, the meta-analysis has important limitations. Quantitative synthesis was possible only after identifying groups of studies that were sufficiently comparable in task, model, outcome domain, and reported performance metric. In practice, this meant that the meta-analysis was restricted to a small number of screening and diagnostic classification studies reporting sensitivity and specificity. The included studies were not sufficiently numerous to support robust subgroup analyses by individual mental disorder, data source, study design, prompting approach, or reference standard. As a result, the pooled estimates should not be interpreted as disorder-specific diagnostic accuracy estimates. The inclusion of related but distinct outcomes, such as depression, anxiety, PTSD-related outcomes, and suicidality-related outcomes, was a pragmatic decision made because disorder-specific pooling was not feasible. Although these outcomes are clinically related and commonly evaluated in screening and risk-detection contexts, they are not interchangeable diagnoses. The wide CIs, wide PIs, and substantial heterogeneity indicate that model performance is likely to vary considerably across settings. The meta-analysis should therefore be viewed as an early and cautious synthesis of the most comparable available evidence, not as definitive evidence of clinical effectiveness or implementation readiness, and also quantifies the lack of data and uncertainty in the evidence.

Fourth, important reporting gaps limited interpretation. Prompting was often incompletely described, even though prompt design can substantially influence model outputs [105]. Without transparent reporting of prompts and prompt refinement, it is difficult to determine whether observed performance differences reflected the model itself, the way it was instructed, or both. Some studies also did not adequately describe recruitment strategies or sampling frames, making it difficult to assess whether included participants or datasets were representative of the populations of interest. In addition, individuals with more complex presentations may have been underrepresented, which may have inflated estimates of model performance [106].

Fifth, most studies were conducted in English, even though LLM performance may differ in non-English settings [107]. This is particularly important because many non-English-speaking settings, including low- and middle-income countries, face greater shortages in mental health care resources and may have the most to gain from scalable digital tools [108]. The limited linguistic and geographic diversity of the evidence therefore constrains the extent to which these findings can be assumed to apply across settings.

Finally, the evidence base is evolving rapidly. Only a limited number of studies evaluated the most recent model variants, including newer GPT-4 and GPT-4o [51,69,71]. Accordingly, the conclusions of this review would need updating as further research becomes available, and AI-assisted living systematic review approaches may help keep pace with the rapid expansion of this literature [109].

Conclusions

This review provides a clinically focused synthesis of general-purpose LLMs in mental health care. The evidence suggests that these models are most promising for structured, language-based tasks such as documentation, summarization, information extraction, and monitoring, where the input and expected output are relatively constrained. In contrast, evidence remains much less convincing for high-stakes tasks that require nuanced clinical judgment, including diagnosis, risk assessment, prognosis, crisis response, and therapeutic interaction.

The central implication is that broad accessibility should not be mistaken for clinical readiness. General-purpose LLMs may become useful adjunctive tools in mental health care, particularly where they support clinicians rather than replace them. However, current evidence remains too heterogeneous, indirect, and uncertain to justify routine unsupervised use. Future research needs to move beyond vignette-based and retrospective evaluations toward prospective, real-world studies that assess safety, reliability, equity, acceptability, clinical outcomes, and implementation in diverse care settings.

Acknowledgments

Generative AI was used solely for basic spelling and language checking. ChatGPT was used for limited editorial and formatting assistance during manuscript preparation. Its use was restricted to table-formatting support, including concatenation of cells, and language suggestions to improve flow and help identify possible drafting errors. The authors remain fully responsible for this paper, wrote the substantive content, critically reviewed and verified all AI-generated suggestions, and made all final decisions regarding content, interpretation, wording, citations, and revisions. All authors reviewed and edited the manuscript and take full responsibility for its content.

Funding

This work was supported by Winter Research Programs, which provided funding for student research assistance. JL, BJ, and GCKC were supported by the Australian National Health and Medical Research Council. SF, KM, PCG, YMW, and RL were supported by the University of Queensland Research and Training Program scholarship, which provided funding for student research assistance. The funders had no role in study design, data collection, analysis, interpretation, writing, or publication decisions.

Data Availability

All data associated with this study are present in this paper or the supplementary information. The corresponding author can be contacted for any additional data.

Authors' Contributions

JL and GCKC conceptualized this study. JL, SF, and KM conducted the literature search. All authors contributed to title and abstract screening. JL, BJ, SF, KM, CM-T, and NK screened the full-text papers. BJ, JL, SF, KM, and PCG extracted the data. BJ, RL, YMW, SF, KM, PCG, CM-T, and NK performed the quality assessment. JL and BJ conducted the formal analysis and visualization. JL, BJ, GCKC, SF, KM, and PCG prepared the original draft. All authors contributed to subsequent drafts, review and editing, and interpretation of the findings. JL, GCKC, BJ, and TS provided supervision. All authors approved the final version and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Review methods, meta-analysis plots, and methodological quality assessment.
[DOCX File (Microsoft Word File), 2353 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Summary of studies on the applications of large language models in mental health care (N=66).
[DOCX File (Microsoft Word File), 62 KB-Multimedia Appendix 2]

Checklist 1

PRISMA Checklist.
[DOCX File (Microsoft Word File), 22 KB-Checklist 1]

Checklist 2

PRISMA-S Checklist.
[DOCX File (Microsoft Word File), 18 KB-Checklist 2]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry. Feb 2022;9(2):137-150. [doi: 10.1016/S2215-0366(21)00395-3]
2. Kakuma R, Minas H, van Ginneken N, et al. Human resources for mental health care: current situation and strategies for action. Lancet. Nov 5, 2011;378(9803):1654-1663. [doi: 10.1016/S0140-6736(11)61093-3] [Medline: 22008420]
3. Rameez S, Nasir A. Barriers to mental health treatment in primary care practice in low- and middle-income countries in a post-COVID era: a systematic review. J Family Med Prim Care. Aug 2023;12(8):1485-1504. [doi: 10.4103/jfmpc.jfmpc_391_22] [Medline: 37767443]
4. Hoffmann JA, Attridge MM, Carroll MS, Simon NJE, Beck AF, Alpern ER. Association of youth suicides and county-level mental health professional shortage areas in the US. JAMA Pediatr. Jan 1, 2023;177(1):71-80. [doi: 10.1001/jamapediatrics.2022.4419] [Medline: 36409484]
5. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
6. Wang L, Wan Z, Ni C, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res. 2024;26:e22769. [doi: 10.2196/22769]
7. First MB, Tasman A. Clinical Guide to the Diagnosis and Treatment of Mental Disorders. John Wiley & Sons; 2010. ISBN: 978-0-470-74520-5
8. Hoffman RE, Stopek S, Andreasen NC. A comparative study of manic vs schizophrenic speech disorganization. Arch Gen Psychiatry. Sep 1986;43(9):831-838. [doi: 10.1001/archpsyc.1986.01800090017003] [Medline: 3753163]
9. Hua Y, Na H, Li Z, et al. A scoping review of large language models for generative tasks in mental health care. npj Digital Med. Apr 30, 2025;8(1):230. [doi: 10.1038/s41746-025-01611-4] [Medline: 40307331]
10. 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(1):e70014. [doi: 10.2196/70014] [Medline: 40373033]
11. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. 2024;11(1):e57400. [doi: 10.2196/57400]

12. Wang X, Zhou Y, Zhou G. The application and ethical implication of generative AI in mental health: systematic review. JMIR Ment Health. 2025;12:e70610. [doi: 10.2196/70610]
13. Xian X, Chang A, Xiang YT, Liu MT. Debate and dilemmas regarding generative AI in mental health care: scoping review. Interact J Med Res. Aug 12, 2024;13(1):e53672. [doi: 10.2196/53672] [Medline: 39133916]
14. Kolding S, Lundin RM, Hansen L, Østergaard SD. Use of generative artificial intelligence (AI) in psychiatry and mental health care: a systematic review. Acta Neuropsychiatr. 2025;37:e37. [doi: 10.1017/neu.2024.50]
15. Jin Y, Liu J, Li P, et al. The applications of large language models in mental health: scoping review. J Med Internet Res. May 5, 2025;27(1):e69284. [doi: 10.2196/69284] [Medline: 40324177]
16. Chung NC, Dyer G, Brocki L. Challenges of large language models for mental health counseling. arXiv. Preprint posted online on Nov 23, 2023. [doi: 10.48550/arXiv.2311.13857]
17. Zhang D, Li W, Song K, et al. Memory in large language models: mechanisms, evaluation and evolution. arXiv. Preprint posted online on Sep 23, 2025. [doi: 10.48550/arXiv.2509.18868]
18. OpenAI. OpenAI o1 system card. arXiv. Preprint posted online on Apr 30, 2026. [doi: 10.48550/arXiv.2412.16720]
19. Chen SF, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine. Nat Med. Mar 2026;32(3):1152-1159. [doi: 10.1038/s41591-026-04229-5] [Medline: 41776077]
20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
21. Chan GCK, He E, Leung J, Verspoor K. A comprehensive systematic review dataset is a rich resource for training and evaluation of AI systems for title and abstract screening. Res Synth Methods. Mar 2025;16(2):308-322. [doi: 10.1017/rsm.2025.1] [Medline: 41626972]
22. Leung J, Sun T, Stjepanovic D, et al. Generative artificial intelligence with youth codesign to create vaping awareness advertisements. JAMA Netw Open. Jul 1, 2025;8(7):e2514040. [doi: 10.1001/jamanetworkopen.2025.14040] [Medline: 40742592]
23. Erinoso O. Generative artificial intelligence for tobacco health promotion. JAMA Netw Open. Jul 1, 2025;8(7):e2514047. [doi: 10.1001/jamanetworkopen.2025.14047] [Medline: 40742599]
24. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI. 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
25. Diaz M. Performance measures of the bivariate random effects model for meta-analyses of diagnostic accuracy. Comput Stat Data Anal. Mar 2015;83:82-90. [doi: 10.1016/j.csda.2014.09.021]
26. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. Jul 22, 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
27. IntHout J, Ioannidis JPA, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. J Clin Epidemiol. Aug 2015;68(8):860-869. [doi: 10.1016/j.jclinepi.2015.03.017] [Medline: 25959635]
28. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. Res Synth Methods. Mar 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
29. Schünemann HJ, Higgins JPT, Vist GE, et al. Completing 'summary of findings' tables and grading the certainty of the evidence. In: Cochrane Handbook for Systematic Reviews of Interventions. Cochrane; 2019:375-402. [doi: 10.1002/9781119536604.ch14]
30. Bartal A, Jagodnik KM, Chan SJ, Dekel S. AI and narrative embeddings detect PTSD following childbirth via birth stories. Sci Rep. Apr 11, 2024;14(1):8336. [doi: 10.1038/s41598-024-54242-2] [Medline: 38605073]
31. Danner M, Hadzic B, Gerhardt S, et al. Advancing mental health diagnostics: GPT-based method for depression detection. Presented at: 2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE); Sep 6-9, 2023; Tsu, Japan. [doi: 10.23919/SICE59929.2023.10354236]
32. Hu C, Li W, Ruan M, et al. Exploiting ChatGPT for diagnosing autism-associated language disorders and identifying distinct features. Research Square. Preprint posted online on May 20, 2024. [doi: 10.21203/rs.3.rs-4359726/v1] [Medline: 38826194]
33. Lorenzoni G, Velmovitsky PE, Alencar P, Cowan D. GPT-4 on clinic depression assessment: an LLM-based pilot study. Presented at: 2024 IEEE International Conference on Big Data (BigData); Dec 15-18, 2024:5043-5049; Washington, DC. [doi: 10.1109/BigData62323.2024.10825184]
34. Kaliosis P, Ganesan AV, Kjell ONE, et al. A systematic evaluation of large language models for PTSD severity estimation: the role of contextual knowledge and modeling strategies. Research Square. Preprint posted online on Dec 24, 2025. [doi: 10.21203/rs.3.rs-8376581/v1]
35. Teferra BG, Perivolaris A, Hsiang WN, et al. Leveraging large language models for automated depression screening. PLOS Digital Health. Jul 2025;4(7):e0000943. [doi: 10.1371/journal.pdig.0000943] [Medline: 40720397]

36. Lee JJ, Han J, Woo CW. Interpretable depression assessment using a large language model. PLOS Digital Health. Feb 2026;5(2):e0001205. [doi: 10.1371/journal.pdig.0001205] [Medline: 41662257]
37. 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]
38. Bhagat N, Mackey O, Wilcox A. Large language models for efficient medical information extraction. AMIA Jt Summits Transl Sci Proc. 2024;2024:509-514. [Medline: 38827084]
39. McCoy TH, Perlis RH. Applying large language models to stratify suicide risk using narrative clinical notes. J Mood Anxiety Disord. Jun 2025;10:100109. [doi: 10.1016/j.xjmad.2025.100109] [Medline: 40657592]
40. McCoy TH, Perlis RH. Reasoning language models for more transparent prediction of suicide risk. BMJ Ment Health. May 11, 2025;28(1):e301654. [doi: 10.1136/bmjment-2025-301654] [Medline: 40350181]
41. Lho SK, Park SC, Lee H, et al. Large language models and text embeddings for detecting depression and suicide in patient narratives. JAMA Netw Open. May 1, 2025;8(5):e2511922. [doi: 10.1001/jamanetworkopen.2025.11922] [Medline: 40408109]
42. Sun M, Yu J, Long Z, et al. Large language models for psychiatric diagnosis based on multicenter real-world clinical records: comparative study. JMIR Med Inf. Jan 13, 2026;14:e77699. [doi: 10.2196/77699] [Medline: 41408781]
43. Marengo D, Longobardi C. Detecting suicidal ideation in adolescence using self-reported emotional and behavioral patterns: comparing machine learning and large language model predictions. Assessment. Dec 31, 2025;0:10731911251406405. [doi: 10.1177/10731911251406405] [Medline: 41472619]
44. 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]
45. Jin Z, Hu J, Bi D, Zhao K, Yu H. Evaluating the efficacy of AI-based interactive assessments using large language models for depression screening: development and usability study. JMIR Form Res. Jan 13, 2026;10:e78401. [doi: 10.2196/78401] [Medline: 41529832]
46. Heinz MV, Bhattacharya S, Trudeau B, et al. Testing domain knowledge and risk of bias of a large-scale general artificial intelligence model in mental health. Digital Health. 2023;9:20552076231170499. [doi: 10.1177/20552076231170499] [Medline: 37101589]
47. Gargari OK, Fatehi F, Mohammadi I, Firouzabadi SR, Shafiee A, Habibi G. Diagnostic accuracy of large language models in psychiatry. Asian J Psychiatr. Oct 2024;100:104168. [doi: 10.1016/j.ajp.2024.104168] [Medline: 39111087]
48. Sarma KV, Hanss KE, Halls AJM, Becker DF, Glowinski AL, Krystal A. Simulated reasoning and self-verification in generalist large language models for psychiatric diagnostic performance: cross-sectional study. medRxiv. Preprint posted online on Sep 9, 2025. [doi: 10.1101/2025.09.05.25335196] [Medline: 40963745]
49. Heston TF. Safety of large language models in addressing depression. Cureus. 2023;15(12):e50729. [doi: 10.7759/cureus.50729]
50. 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]
51. Lauderdale SA, Schmitt R, Wuckovich B, Dalal N, Desai H, Tomlinson S. Effectiveness of generative AI-large language models' recognition of veteran suicide risk: a comparison with human mental health providers using a risk stratification model. Front Psychiatry. 2025;16:1544951. [doi: 10.3389/fpsyt.2025.1544951] [Medline: 40248601]
52. Zeng Q, Li X, Wang S, Liu K. Adversarial evaluation algorithm for detecting extreme behaviors of LLMs in psychological counseling scenarios. Presented at: 2025 2nd International Conference on Algorithms, Software Engineering and Network Security (ASENS); Mar 21-23, 2025:412-415; Guangzhou, China. [doi: 10.1109/ASENS64990.2025.11011189]
53. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health. Sep 2023;11(4):e002391. [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
54. Levkovich I, Rabin E, Brann M, Elyoseph Z. Large language models outperform general practitioners in identifying complex cases of childhood anxiety. Digital Health. 2024;10:20552076241294182. [doi: 10.1177/20552076241294182] [Medline: 39687523]
55. Kim J, Leonte KG, Chen ML, et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. npj Digital Med. Jul 19, 2024;7(1):193. [doi: 10.1038/s41746-024-01181-x] [Medline: 39030292]
56. Levkovich I. Harnessing large language models for identification and treatment of obsessive-compulsive disorder. Comput Hum Behav: Artif Hum. Dec 2025;6:100212. [doi: 10.1016/j.chbah.2025.100212]

57. 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 Invest Health Psychol Educ. Jan 18, 2025;15(1):9. [doi: 10.3390/ejihpe15010009] [Medline: 39852192]
58. Levkovich I, Haber Y, Levi-Belz Y, Elyoseph Z. A step toward the future? Evaluating GenAI QPR simulation training for mental health gatekeepers. Front Med (Lausanne). 2025;12:1599900. [doi: 10.3389/fmed.2025.1599900] [Medline: 40568211]
59. D'Souza RF, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr. Nov 2023;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
60. Jin KW, Rostam-Abadi Y, Chaudhary P, et al. Evaluating diagnostic accuracy and clinical reasoning of multiple large language models in psychiatry. medRxiv. Preprint posted online on Feb 11, 2026. [doi: 10.64898/2026.02.03.26345402] [Medline: 41728283]
61. Sarma KV, Hanss KE, Halls AJM, et al. Integrating expert knowledge into large language models improves performance for psychiatric reasoning and diagnosis. Psychiatry Res. Jan 2026;355:116844. [doi: 10.1016/j.psychres.2025.116844] [Medline: 41270691]
62. Zhang M, Yang X, Zhang X, et al. CBT-bench: evaluating large language models on assisting cognitive behavior therapy. Presented at: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics; Apr 29 to May 4, 2025; Albuquerque, New Mexico. [doi: 10.18653/v1/2025.naacl-long.196]
63. Aleem M, Zahoor I, Naseem M. Towards culturally adaptive large language models in mental health: using ChatGPT as a case study. Presented at: CSCW Companion '24: Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing; Nov 9-13, 2024:240-247; San Jose Costa Rica. [doi: 10.1145/3678884.3681858]
64. Liu C, Wang WYC, Khan G. Understanding medical information and emotional support needs in mental health questions with large language models. Ind Manage Data Syst. Jun 22, 2026;126(7):2205-2230. [doi: 10.1108/IMDS-05-2025-0609]
65. 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. 2024;11:e53043-e53043. [doi: 10.2196/53043]
66. Perlis RH. Research letter: application of GPT-4 to select next-step antidepressant treatment in major depression. medRxiv. Preprint posted online on Apr 18, 2023. [doi: 10.1101/2023.04.14.23288595] [Medline: 37131648]
67. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology. Aug 2024;49(9):1412-1416. [doi: 10.1038/s41386-024-01841-2] [Medline: 38480911]
68. Silva R, Gomes L. An adaptive language model-based intelligent medication assistant for the decision support of antidepressant prescriptions. Comput Biol Med. May 2025;190:110065. [doi: 10.1016/j.compbiomed.2025.110065] [Medline: 40147190]
69. McBain RK, Cantor JH, Zhang LA, et al. Evaluation of alignment between large language models and expert clinicians in suicide risk assessment. Psychiatr Serv. Nov 1, 2025;76(11):944-950. [doi: 10.1176/appi.ps.20250086] [Medline: 41174947]
70. Tan KS, Cervin M, Leman P, Nielsen K, Kumar PV, Medvedev O. AI meets psychology: an exploratory study of large language models' competence in psychotherapy contexts. J Psychol AI. Dec 31, 2025;1(1):2545258. [doi: 10.1080/29974100.2025.2545258]
71. Thotapalli S, Yilanli M, McKay I, et al. Potential of ChatGPT in youth mental health emergency triage: comparative analysis with clinicians. PCN Rep. Sep 2025;4(3):e70159. [doi: 10.1002/pcn5.70159] [Medline: 40673126]
72. Acevedo S, Aneja E, Opler DJ, Valera P, Jarmon E. Evaluating the efficacy of ChatGPT-3.5 versus human-delivered text-based cognitive-behavioral therapy: a comparative pilot study. Am J Psychother. Mar 1, 2026;79(1):4-11. [doi: 10.1176/appi.psychotherapy.20240070] [Medline: 40820507]
73. Alanzi TM, Alharthi A, Alrumman S, et al. ChatGPT as a psychotherapist for anxiety disorders: an empirical study with anxiety patients. Nutr Health. Sep 2025;31(3):1111-1123. [doi: 10.1177/02601060241281906] [Medline: 39370914]
74. Haber Y, Levi-Belz Y, Elbak YS, Elyoseph Z, Levkovich I. Validating GenAI feedback in suicide prevention training: a mixed-methods study of QPR skill assessment. Front Med (Lausanne). 2025;12:1709743. [doi: 10.3389/fmed.2025.1709743] [Medline: 41625774]
75. Hodson N, Williamson S. Can large language models replace therapists? Evaluating performance at simple cognitive behavioral therapy tasks. JMIR AI. Jul 30, 2024;3:e52500. [doi: 10.2196/52500] [Medline: 39078696]
76. Jain A, Sandhu R, Singh G, Rakhra M. The role of AI counselling in journaling for mental health improvement. Presented at: 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT); Aug 29-31, 2024:1-6; Greater Noida, India. [doi: 10.1109/ICEECT61758.2024.10739128]

77. Napiwotzki I, Laue J, Caldarone F, et al. Comparing human and AI therapists in behavioral activation for depression: cross-sectional questionnaire study. JMIR Form Res. Dec 4, 2025;9:e78138. [doi: 10.2196/78138] [Medline: 41343763]
78. Moell B. Comparing the efficacy of GPT-4 and Chat-GPT in mental health care: a blind assessment of large language models for psychological support (preprint). JMIR Ment Health. Preprint posted online on Mar 20, 2023. [doi: 10.2196/preprints.47439] [Medline: 37114093]
79. Shen H, Li Z, Yang M, et al. Are large language models possible to conduct cognitive behavioral therapy? Presented at: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2024:3695-3700; Lisbon, Portugal. [doi: 10.1109/BIBM62325.2024.10821773]
80. 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]
81. Cardamone NC, Olfson M, Schmutte T, et al. Classifying unstructured text in electronic health records for mental health prediction models: large language model evaluation study. JMIR Med Inf. Jan 21, 2025;13:e65454. [doi: 10.2196/65454] [Medline: 39864953]
82. Donnelly HK, Brown GK, Green KL, et al. Exploring the potential of large language models for automated safety plan scoring in outpatient mental health settings (preprint). JMIR Ment Health. Preprint posted online on Sep 3, 2025. [doi: 10.2196/preprints.79010] [Medline: 40196270]
83. Gireesh H, Shukla L, Shivaprakash P, Mukherjee A, Chand P, Murthy P. Language models for standardising clinical notes and information extraction in addiction psychiatry-an empirical study. Drug Alcohol Rev. Jan 2026;45(1):e70059. [doi: 10.1111/dar.70059] [Medline: 41158037]
84. Matsumura M, Nishida K, Toyoda K, et al. Quantifying improvement of psychotic symptoms in clozapine-treated schizophrenia: clinical note analysis with large language models. Sci Rep. Feb 13, 2026;16(1):8835. [doi: 10.1038/s41598-026-39676-0] [Medline: 41680410]
85. Janota B, Janota K. Application of artificial intelligence (AI) in the creation of discharge summaries in psychiatric clinics. Int J Psychiatry Med. May 2025;60(3):330-337. [doi: 10.1177/00912174241284730] [Medline: 39285727]
86. He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician versus large language model chatbot responses to web-based questions from autistic patients in Chinese: cross-sectional comparative analysis. J Med Internet Res. Apr 30, 2024;26:e54706. [doi: 10.2196/54706] [Medline: 38687566]
87. Li Y, Wang G, Huang Z, et al. Development and preliminary evaluation of a virtual standardized patient system for psychiatric interview training. BMC Psychiatry. 2026;26(1):264. [doi: 10.1186/s12888-026-07896-3]
88. Yilanli M, McKay I, Jackson DI, Sezgin E. Large language models for individualized psychoeducational tools for psychosis: a cross-sectional study. medRxiv. Preprint posted online on Jul 29, 2024. [doi: 10.1101/2024.07.26.24311075]
89. Senturk E, Koparal B. ChatGPT-4o vs psychiatrists in responding to common antidepressant concerns. Am J Health Promot. Jan 2026;40(1):10-17. [doi: 10.1177/08901171251348208] [Medline: 40440446]
90. Barabas L, Novotny M, Jung D, Müller T, Mertse NN. Exploring the potential of ChatGPT as a digital advisor in acute psychiatric crises: a feasibility study. Nervenarzt. May 2026;97(3):265-271. [doi: 10.1007/s00115-025-01837-3] [Medline: 40478290]
91. 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]
92. Scholich T, Barr M, Stirman SW, 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]
93. Bannett Y, Gunturkun F, Pillai M, et al. Leveraging a large language model to assess quality-of-care: monitoring ADHD medication side effects. medRxiv. Preprint posted online on Apr 24, 2024. [doi: 10.1101/2024.04.23.24306225] [Medline: 38712037]
94. Rao SJ, Isath A, Krishnan P, et al. ChatGPT: a conceptual review of applications and utility in the field of medicine. J Med Syst. Jun 5, 2024;48(1):59. [doi: 10.1007/s10916-024-02075-x] [Medline: 38836893]
95. Giacobbe DR, Marelli C, Guastavino S, et al. Artificial intelligence and prescription of antibiotic therapy: present and future. Expert Rev Anti-Infect Ther. Oct 2024;22(10):819-833. [doi: 10.1080/14787210.2024.2386669] [Medline: 39155449]
96. Ferrag MA, Tihanyi N, Debbah M. From LLM reasoning to autonomous AI agents: a comprehensive review. IEEE Access. 2025;14:84237-84285. [doi: 10.1109/ACCESS.2026.3698694] [Medline: 40292759]
97. Dergaa I, Fekih-Romdhane F, Hallit S, et al. ChatGPT is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry. 2023;14:1277756. [doi: 10.3389/fpsyt.2023.1277756] [Medline: 38239905]

98. 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 2024;12(Suppl 1):e002583. [doi: 10.1136/fmch-2023-002583]
99. de Vries M, Schaub MP. Opportunities and risks of large language models in digital interventions for substance use disorders. Curr Opin Psychiatry. Jul 1, 2026;39(4):308-313. [doi: 10.1097/YCO.0000000000001088] [Medline: 42083979]
100. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). npj Digital Med. Jul 8, 2024;7(1):183. [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
101. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed. Mar 2024;245:108013. [doi: 10.1016/j.cmpb.2024.108013] [Medline: 38262126]
102. Liang W, Zhang Y, Codreanu M, Wang J, Cao H, Zou J. The widespread adoption of large language model-assisted writing across society. Patterns (N Y). Dec 12, 2025;6(12):101366. [doi: 10.1016/j.patter.2025.101366] [Medline: 41472824]
103. Chan GCK, Lim C, Sun T, et al. Causal inference with observational data in addiction research. Addiction. Oct 2022;117(10):2736-2744. [doi: 10.1111/add.15972]
104. Chan GCK, Sun T, Stjepanović D, et al. Designing observational studies for credible causal inference in addiction research-directed acyclic graphs, modified disjunctive cause criterion and target trial emulation. Addiction. Jun 2024;119(6):1125-1134. [doi: 10.1111/add.16442] [Medline: 38343103]
105. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. npj Digital Med. Feb 20, 2024;7(1):41. [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
106. Wright E, Pagliaro C, Page IS, Diminic S. A review of excluded groups and non-response in population-based mental health surveys from high-income countries. Soc Psychiatry Psychiatr Epidemiol. Sep 2023;58(9):1265-1292. [doi: 10.1007/s00127-023-02488-y] [Medline: 37212903]
107. Jin Y, Chandra M, Verma G, Hu Y, De Choudhury M, Kumar S. Better to ask in English: cross-lingual evaluation of large language models for healthcare queries. Presented at: WWW '24: Proceedings of the ACM Web Conference 2024; May 13-17, 2024:2627-2638; Singapore, Singapore. [doi: 10.1145/3589334.3645643]
108. Nadkarni A, Hanlon C, Patel V. Mental health care models in low-and middle-income countries, in Tasman's Psychiatry. In: Tasman A, RibaMB, AlarcónRD, Alfonso CA, editors. Tasman's Psychiatry. Springer; 2023:1-47. [doi: 10.1007/978-3-030-42825-9_156-1]
109. Reynolds SA, Christie AP, Dicks LV, et al. Will AI speed up literature reviews or derail them entirely? Nature. Jul 10, 2025;643(8071):329-331. [doi: 10.1038/d41586-025-02069-w]

Abbreviations

CBT: cognitive behavioral therapy
***DSM*:** *Diagnostic and Statistical Manual of Mental Disorders*
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
***DSM-5-TR*:** *Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition, Text Revision)*
EHR: electronic health record
GPT: generative pretrained transformer
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
***ICD*:** *International Classification of Diseases*
***ICD-11*:** *International Classification of Diseases, 11th Revision*
LLM: large language model
MMAT: Mixed Methods Appraisal Tool
OCD: obsessive-compulsive disorder
PI: prediction interval
PRISMA : Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROSPERO : International Prospective Register of Systematic Reviews
PTSD : posttraumatic stress disorder

Edited by Ivan Steenstra; peer-reviewed by Liying Wang, Markus W Haun; submitted 13.Nov.2025; final revised version received 23.Jun.2026; accepted 24.Jun.2026; published 31.Aug.2026

Please cite as:
Leung J, Johnson B, McRae K, Fong S, Gonzalez PC, McClure-Thomas C, Sun T, Kern N, Wong YM, Liu R, Chan GCK
Generative Large Language Models in Mental Health Care Settings: Systematic Review and Meta-Analysis
JMIR AI 2026;5:e87730

URL: https://ai.jmir.org/2026/1/e87730
doi: 10.2196/87730