Review

The Application and Ethical Implication of Generative AI in Mental Health: Systematic Review

Xi Wang[1], MA; Yujia Zhou[2], PhD; Guangyu Zhou[1], Prof Dr

[1]School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China

[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China

Corresponding Author:
Guangyu Zhou, Prof Dr
School of Psychological and Cognitive Sciences
Beijing Key Laboratory of Behavior and Mental Health, Key Laboratory of Machine Perception (Ministry of Education)
Peking University
Philosophy Building, 2nd Fl.
No. 5 Yiheyuan Road, Haidian District
Beijing, 100871
China
Phone: 86 10 62767702
Email: gyzhou@pku.edu.cn

Abstract

Background: Mental health disorders affect an estimated 1 in 8 individuals globally, yet traditional interventions often face barriers, such as limited accessibility, high costs, and persistent stigma. Recent advancements in generative artificial intelligence (GenAI) have introduced AI systems capable of understanding and producing humanlike language in real time. These developments present new opportunities to enhance mental health care.

Objective: We aimed to systematically examine the current applications of GenAI in mental health, focusing on 3 core domains: diagnosis and assessment, therapeutic tools, and clinician support. In addition, we identified and synthesized key ethical issues reported in the literature.

Methods: We conducted a comprehensive literature search, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, in PubMed, ACM Digital Library, Scopus, Embase, PsycInfo, and Google Scholar databases to identify peer-reviewed studies published from October 1, 2019, to September 30, 2024. After screening 783 records, 79 (10.1%) studies met the inclusion criteria.

Results: The number of studies on GenAI applications in mental health has grown substantially since 2023. Studies on diagnosis and assessment (37/79, 47%) primarily used GenAI models to detect depression and suicidality through text data. Studies on therapeutic applications (20/79, 25%) investigated GenAI-based chatbots and adaptive systems for emotional and behavioral support, reporting promising outcomes but revealing limited real-world deployment and safety assurance. Clinician support studies (24/79, 30%) explored GenAI's role in clinical decision-making, documentation and summarization, therapy support, training and simulation, and psychoeducation. Ethical concerns were consistently reported across the domains. On the basis of these findings, we proposed an integrative ethical framework, GenAI4MH, comprising 4 core dimensions—data privacy and security, information integrity and fairness, user safety, and ethical governance and oversight—to guide the responsible use of GenAI in mental health contexts.

Conclusions: GenAI shows promise in addressing the escalating global demand for mental health services. They may augment traditional approaches by enhancing diagnostic accuracy, offering more accessible support, and reducing clinicians' administrative burden. However, to ensure ethical and effective implementation, comprehensive safeguards—particularly around privacy, algorithmic bias, and responsible user engagement—must be established.

(JMIR Ment Health 2025;12:e70610) doi: 10.2196/70610

KEYWORDS

generative AI; mental health; large language models; mental health detection and diagnosis; therapeutic chatbots

Introduction

Background

Mental health has become a global public health priority, with increasing recognition of its importance for individual well-being, societal stability, and economic productivity. According to the World Health Organization, approximately 1 in 8 people worldwide live with a mental health disorder [1]. Despite the growing demand for mental health services, traditional approaches such as in-person therapy and medication, which rely heavily on trained professionals and extensive infrastructure, are struggling to meet the rising need [2]. Consequently, an alarming 76% to 85% of individuals with mental health disorders do not receive effective treatment, often due to barriers such as limited access to mental health professionals, social stigma, and inadequate health care systems [3]. Against this backdrop, advances in generative artificial intelligence (GenAI) offer new and promising avenues to enhance mental health services.

GenAI, such as ChatGPT [4], is built on large-scale language modeling and trained on extensive textual corpora. Their capacity to produce contextually relevant and, in many cases, emotionally appropriate language [5,6] enables more natural and adaptive interactions. Compared to earlier dialogue systems, GenAI exhibits greater flexibility in producing open-ended, humanlike dialogue [7]. This generative capability makes them a promising tool for web-based therapeutic interventions that allow for real-time, adaptive engagement in mental health care.

Currently, GenAI is being integrated into mental health through a range of innovative applications. For instance, GPT-driven chatbots such as Well-Mind ChatGPT [8] and MindShift [9] provide personalized mental health support by engaging users in conversational therapy. Similarly, virtual companions such as Replika [10] are used to help users manage feelings of loneliness and anxiety through interactive dialogue [11]. In addition, GenAI has been used to analyze social media posts and clinical data to identify signs of depression [2] and suicidal ideation [12]. These diverse applications illustrate the potential of GenAI to address various mental health needs, from prevention and assessment to continuous support and intervention.

Although research has investigated various applications of GenAI in mental health, much of it has focused on specific models or isolated cases, lacking a comprehensive evaluation of its broader impacts, applications, and associated risks. Similarly, most systematic reviews to date have focused on particular domains, such as depression detection [13], chatbot interventions [14], empathic communication [5], psychiatric education [15], and AI-based art therapy [16]. While such focused reviews offer valuable insights into specific use cases, a broad outline remains crucial for understanding overarching trends, identifying research gaps, and informing the responsible development of GenAI in mental health. To date, only 2 reviews [17,18] have attempted broader overviews, covering the literature published before April 2024 and July 2023, respectively. However, since April 2024, the rapid evolution of GenAI—including the release and deployment of more advanced models, such as GPT-4o [19] and GPT-o1 [20], and their increasing integration with clinical workflows, such as Med-Gemini [21], has expanded the scope and complexity of GenAI applications in real-world mental health contexts. These developments underscore the need for a more updated and integrative synthesis.

Objectives

To address this gap, we aimed to provide a comprehensive overview of GenAI applications in mental health, identify research gaps, and propose future directions. To systematically categorize the existing research, we divided the studies into three distinct categories based on the role of GenAI in mental health applications, as illustrated in Figure 1: (1) GenAI for mental health diagnosis and assessment, encompassing research that leverages GenAI to detect, classify, or evaluate mental health conditions; (2) GenAI as therapeutic tools, covering studies where GenAI-based chatbots or conversational agents are used to deliver mental health support, therapy, or interventions directly to users; and (3) GenAI for supporting clinicians and mental health professionals, including research aimed at using GenAI to assist clinicians in their practice.

Despite these promising applications, the integration of GenAI into mental health care is not without challenges. Applying GenAI in the mental health field involves processing highly sensitive personal information, such as users' emotional states, psychological histories, and behavioral patterns. Mishandling such data not only poses privacy risks but may also lead to psychological harm, including distress, stigma, or reduced trust in mental health services [22]. Therefore, in addition to systematically categorizing existing applications of GenAI in mental health, we also examined ethical issues related to their use in this domain. On the basis of our analysis, we proposed an ethical framework, GenAI4MH, to guide the responsible use of GenAI in mental health contexts (Figure 2).

Figure 1. Classification of generative artificial intelligence (GenAI) applications in mental health.

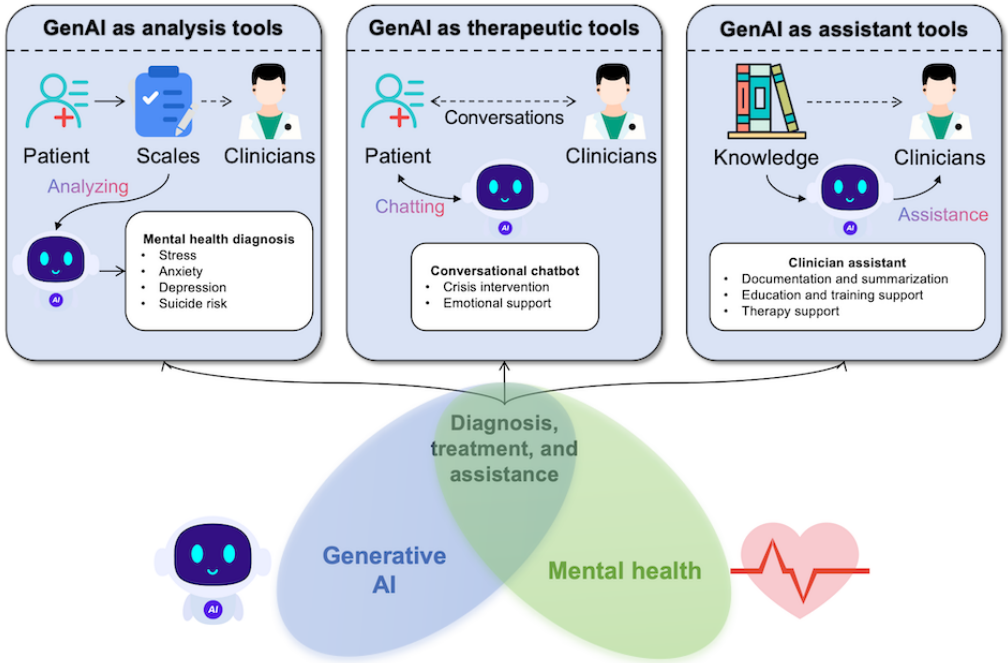
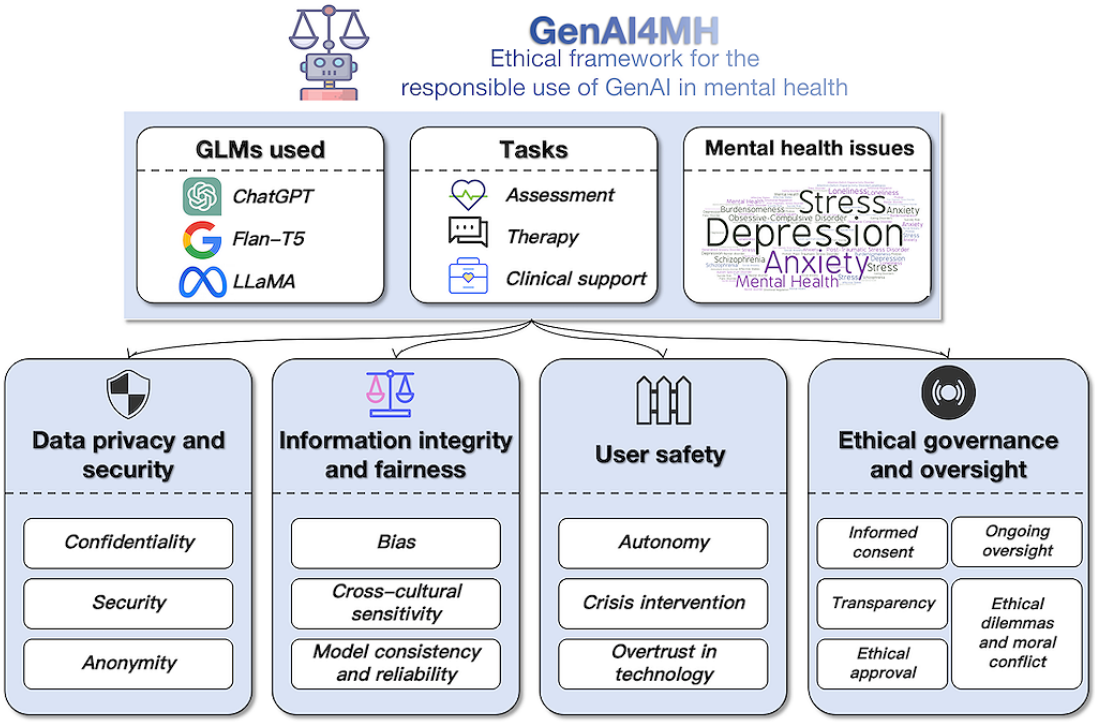


Figure 2. Overview of the GenAI4MH ethical framework for the responsible use of generative artificial intelligence (GenAI) in mental health. GLM: generative language model.



Methods

Search Strategy

We conducted this systematic review following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (Multimedia Appendix 1) [23]. We conducted a comprehensive search across 6 databases: PubMed, ACM Digital Library, Scopus, Embase, PsycInfo, and Google Scholar. We conducted the search between October 1, 2024, and October 7, 2024, and targeted studies published from October 1, 2019, to September 30, 2024. The starting date was chosen to coincide with the introduction of the T5 model [24], a foundational development for many of today's mainstream GenAI models. This date also intentionally excluded earlier models, such as Bidirectional Encoder Representations from Transformers (BERT) [25] and GPT-2 [26], as these models

have already been extensively covered in the previous literature [27,28], and our aim was to highlight more recent innovations.

Search terms were constructed using a logical combination of keywords related to GenAI and mental health: (Generative AI OR Large Language Model OR ChatGPT) AND (mental health OR mental disorder OR depression OR anxiety). This search string was developed based on previous reviews and refined through iterative testing to ensure effective identification of relevant studies. When possible, the search was restricted to titles and abstracts. For Google Scholar, the first 10 pages of results were screened for relevance. A detailed search strategy is provided in Multimedia Appendix 2.

Study Selection

The selection criteria included studies that (1) used GenAI and were published after the introduction of the T5 [24] model and (2) directly addressed the application of GenAI in mental health care settings. Only peer-reviewed original research articles were considered, with no language restrictions.

Data Extraction

Data from the included studies were extracted using standardized frameworks. For qualitative studies, we used the Sample, Phenomenon of Interest, Design, Evaluation, and Research Type (SPIDER) framework. For quantitative studies, we applied the Population, Intervention, Comparison, Outcome, and Study (PICOS) framework. A summary of the extracted data are provided in Multimedia Appendix 3 [2,3,7-9,11,12,29-100].

Reporting Quality Assessment

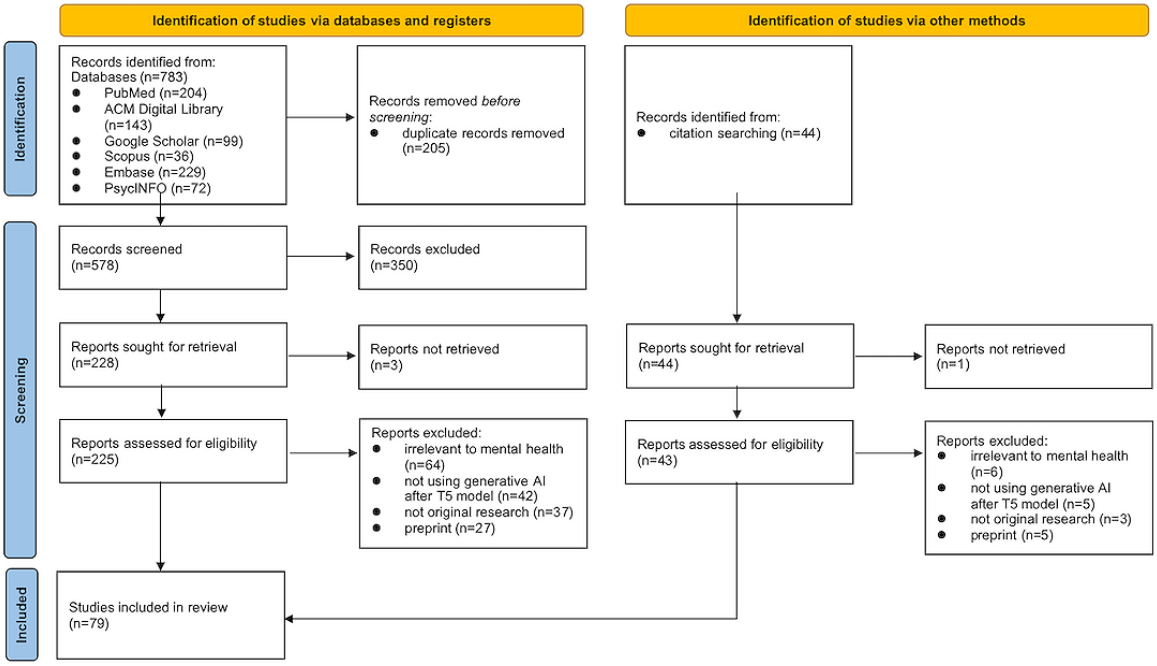
To assess the reporting transparency and the methodological rigor of the included studies, we applied the Minimum Information about Clinical Artificial Intelligence for Generative Modeling Research (MI-CLAIM-GEN) checklist (Multimedia Appendix 4) [101], a recently proposed guideline tailored for evaluating the reporting quality of research on GenAI in health care. The checklist covers essential aspects such as study design, data and resource transparency, model evaluation strategies, bias and harm assessments, and reproducibility. We followed the Joanna Briggs Institute quality appraisal format [102] to score each item in the checklist using 4 categories: yes, no, unclear, and not applicable.

Results

Study Selection

As shown in Figure 3, a total of 783 records were initially retrieved from the 6 databases. After removing duplicates, 73.8% (578/783) of unique records remained for screening. Following abstract screening, 39.4% (228/578) of the records were identified for full-text retrieval and screening. After full-text screening, 24% (55/228) of the articles were selected for inclusion in the systematic review. To ensure comprehensive coverage of relevant studies, a snowballing technique was then applied, where we examined the reference lists of the included studies and related review articles. This process identified an additional 44 studies for eligibility assessment. After the same evaluation process, 54% (24/44) of these studies met the inclusion criteria, bringing the final total to 79 studies for the systematic review. Two PhD candidates (YZ and XW) independently conducted the selection, with discrepancies resolved through discussion. The interrater reliability was satisfactory (κ=0.904).

Figure 3. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection.

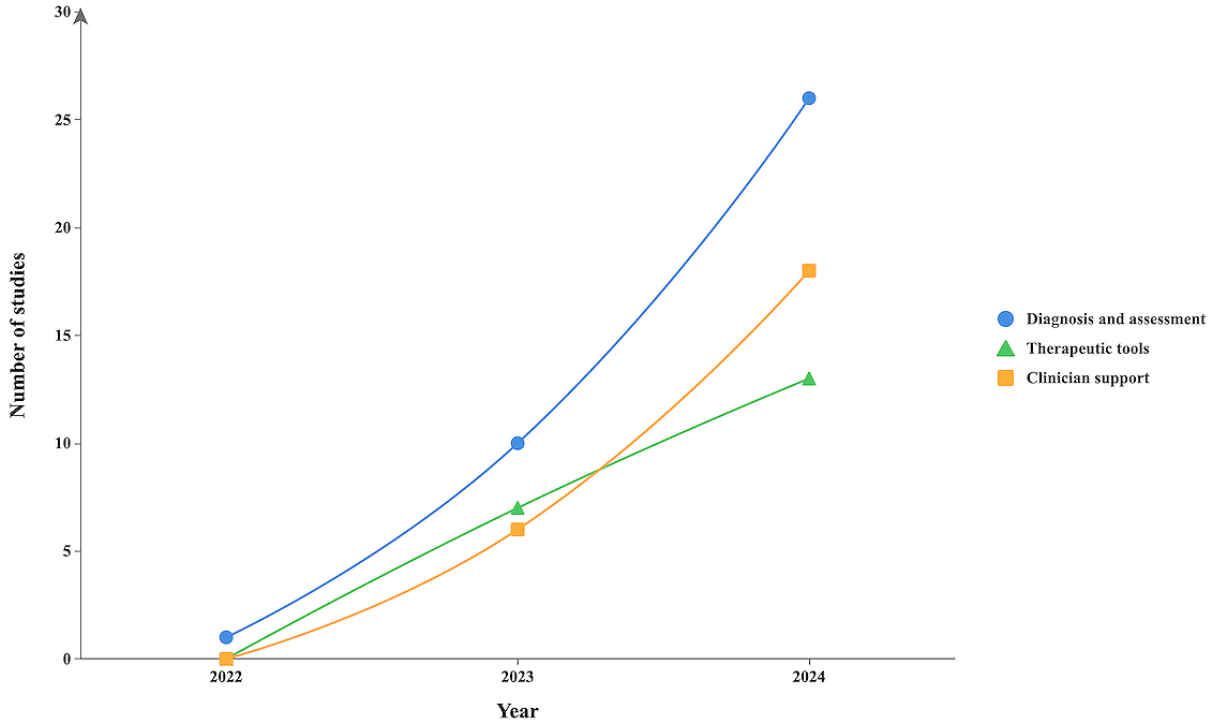


Publication Trends Over Time

An analysis of publication trends over time reveals a growing focus on the application of GenAI in mental health (Figure 4). Overall, the number of studies in all the 3 categories grew extensively over the examined period, indicating a rising interest in using GenAI for mental health. In 2022, the total number of studies was minimal across all the 3 categories, with only 1

(1%) early study, of the included 79 studies, emerging on the use of GenAI for mental health diagnosis and assessment. However, as GenAI advanced and garnered wider adoption, the number of publications in all the 3 categories began to increase steadily. A moderate increase was observed in the year 2023, with 13% (10/79) of the studies focused on diagnosis and assessment, 9% (7/79) on therapeutic interventions, and 8% (6/79) on clinician support, reflecting a growing interest in practical applications of these models in health care settings. By 2024, the number of publications had surged across all the 3 categories, with 33% (26/79) of the studies focused on diagnosis and assessment, 16% (13/79) on therapeutic interventions, and 23% (18/79) on clinician support.

Figure 4. Publication trends in the application of generative artificial intelligence (GenAI) in mental health research.



GenAI for Mental Health Diagnosis and Assessment

Overview

Of the 79 included studies, 37 (47%) were identified that investigated the effectiveness and applications of GenAI in mental health diagnosis and assessment. These studies primarily explored how GenAI can detect and interpret mental health conditions by analyzing textual and multimodal data. A summary of the included studies is presented in Multimedia Appendix 5 [2,3,12,29-59,61,62,100].

Mental Health Issues

The existing studies using GenAI for mental health diagnosis predominantly focused on suicide risk and depression, followed by emerging applications in emotion recognition, psychiatric disorders, and stress.

Suicide risk was the most frequently examined topic, addressed in 40% (15/37) of the studies. Researchers used large language models (LLMs) to identify suicide-related linguistic patterns [12], extract and synthesize textual evidence supporting identified suicide risk levels [29-33,103], and evaluate suicide risk [34-41]. GenAI models, such as GPT-4 [104], achieved high precision (up to 0.96) in predicting suicidal risk levels [30-33,103], outperforming traditional models, such as support vector machines (SVM) [41], and performing comparably to or better than pretrained language models, such as BERT [38,40]. Most studies (13/15, 87%) relied on simulated case narratives [34,36,37] or social media data [38-40]; only 13% (2/15) of the studies used real clinical narratives [35,41].

Depression was the second most common mental health issue addressed, featured in 35% (13/37) of the studies. While GenAI models showed promising accuracy (eg, 0.902 using semistructured diaries [42]), performance was often constrained to English data [43,44], with notable drop-offs in dialectal or culturally divergent contexts [44]. Multimodal approaches—integrating audio, visual, and physiological data—improved detection reliability over text-only methods [45-47]. Several studies (3/13, 23%) also explored interpretability, using GenAI to generate explanations [43] or conduct structured assessments [48].

GenAI has also been explored for emotion recognition, using smartphone and wearable data to predict affective states with moderate accuracy [47,49], and enabling novel assessment formats, such as virtual agent interactions [45] and conversational psychological scales [50]. The studies also explored other psychiatric disorders, such as obsessive-compulsive disorder (accuracy up to 96.1%) [51] and schizophrenia (r=0.66-0.69 with expert ratings) [52]. In total, 8% (3/37) of the studies addressed stress detection from social media texts [39,53,54].

A smaller set of studies (3/37, 8%) assessed GenAI models' capacity for differential diagnosis, demonstrating that GenAI models could distinguish among multiple mental disorders in controlled simulations [3,55,56]. However, performance remained higher for mental health conditions with distinct symptoms (eg, psychosis and anxiety) and lower for overlapping or less prevalent disorders (eg, perinatal depression and lysergic acid diethylamide use disorder) [56], particularly for those with symptom overlap with more common mental health conditions (eg, disruptive mood dysregulation disorder and acute stress disorder) [55].

Model Architectures and Adaptation Strategies

Overview

Most included studies (29/37, 78%) used proprietary GenAI models for mental health diagnosis and assessment, with GPT-based models (GPT-3, 3.5, and 4) [4] being the most commonly used [2,3,12]. Other proprietary models included Gemini [49,51] and the pathways language model (version 2) [47]. A smaller subset of the studies (14/37, 38%) adopted open-source models, such as LLM Meta AI (LLaMA) [29,30,32,40,52,57,58], Mistral [33], Falcon [40], and Neomotron [55]. Beyond model selection, several studies (29/37, 78%) explored technical strategies to enhance diagnostic performance and interpretability. In total, 3 main approaches were identified as described in subsequent sections.

Hybrid Modeling

A limited number of studies (2/37, 5%) explored hybrid architectures, combining GenAI-generated embeddings with classical classifiers, such as SVM or random forest [43,53]. For example, Radwan et al [53] used GPT-3 embeddings to generate text vectors, which were input into classifiers, such as SVM, random forest, and k-nearest neighbors, for stress level classification. The combination of GPT-3 embeddings with an SVM classifier yielded the best performance, outperforming other hybrid configurations and traditional models such as BERT with the long short-term memory model.

Fine-Tuning and Instruction Adaptation

Some studies (4/37, 11%) used instruction-tuned models, including Flan [39,41], Alpaca [39], and Wizard [32], to enhance instruction following. Further fine-tuning with mental health–related data was also applied to improve diagnostic and assessment capabilities [39,46,59]. For instance, Xu et al [39] demonstrated that their fine-tuned models—Mental-Alpaca and Mental-FLAN-T5—achieved a 10.9% improvement in balanced accuracy over GPT-3.5 [4], despite being 25 and 15 times smaller, respectively. These models also outperformed GPT-4 [104] by 4.8%, although GPT-4 is 250 and 150 times larger, respectively.

Prompt Engineering and Knowledge Augmentation

Prompt-based techniques—including few-shot learning [31,39,40,46,54,58], chain-of-thought prompting [42,47,49,59,60], and example contrast [54]—have been shown to substantially enhance diagnostic performance, especially for smaller models [39]. Meanwhile, retrieval-augmented generation (RAG) approaches enriched LLMs with structured knowledge (eg, *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* criteria), improving factual grounding in some cases [47], but occasionally introducing noise or reducing performance due to redundancy and semantic drift [55].

Data Source

Table 1 summarizes the datasets used for GenAI-based mental health diagnosis and assessment, categorized by data modality and mental health focus. The full dataset list, including metadata and sampling details, is provided in Multimedia Appendix 6 [12,30,35,41,42,44,49,53,56,62,103,105-133].

Social media posts, such as those on Reddit [12,29,43], Twitter [58], and Weibo [45], emerged as prominent data sources. Beyond social media, 19% (7/37) of the studies used professionally curated clinical vignettes, providing controlled scenarios that simulate clinical cases and allow for standardized assessment across GenAI models [34,51]. Only a few studies (4/37, 11%) used clinical text data sources, including clinical interviews [61], diary texts [42], and written responses of participants [50,62].

In total, 14% (5/37) of the studies used multimodal data sources—such as speech [44,45], sensor data [47], and electroencephalogram (EEG) [46]—to enhance the accuracy and comprehensiveness of mental health assessments. For example, Englhardt et al [47] developed prompting strategies for GenAI models to classify depression using passive sensing data (eg, activity, sleep, and social behavior) from mobile and wearable devices, achieving improved classification accuracy (up to 61.1%) over classical machine learning baselines. Similarly, Hu et al [46] integrated EEG, audio, and facial expressions to boost predictive performance and proposed MultiEEG-GPT, a GPT-4o-based method for mental health assessment using multimodal inputs, including EEG, facial expressions, and audio. Their results across the 3 datasets showed that combining EEG with audio or facial expressions significantly improved prediction accuracy in both zero-shot and few-shot settings.

Table 1. Summary of datasets used in the studies on generative artificial intelligence (GenAI) models for mental health diagnosis and assessment.

Categories	References
By modality	
Text (clinical vignettes)	[56,105-108]
Text (social media posts)	[12,40,53,109-123,134]
Text (transcripts)	[35,124]
Text (daily self-reports)	[42,62]
Multimodal dataset	[44,49,125-131]
By mental health issues	
Depression	[42,44,62,109,110,113,114,117,123,126-128,130,134]
Suicide risk	[12,35,40-42,108,109,111,112,118,119,121,122]
Posttraumatic stress disorder	[110,125,127]
Anxiety	[115,125,128]
Bipolar disorder	[120,124]
Stress	[53,115,132]
Emotion regulation	[62,129,131]
Multiple psychiatric disorders	[56,63,106,107,133]

GenAI as Therapeutic Tools

Of the 79 included studies, 20 (25%) investigated the use of GenAI-based chatbots and conversational agents to facilitate interventions ranging from emotional support to more structured therapies. To assess the feasibility and potential impact of these interventions, we analyzed studies across four key dimensions: (1) therapeutic targets, (2) implementation strategies, (3) evaluation outcomes, and (4) real-world deployment features.

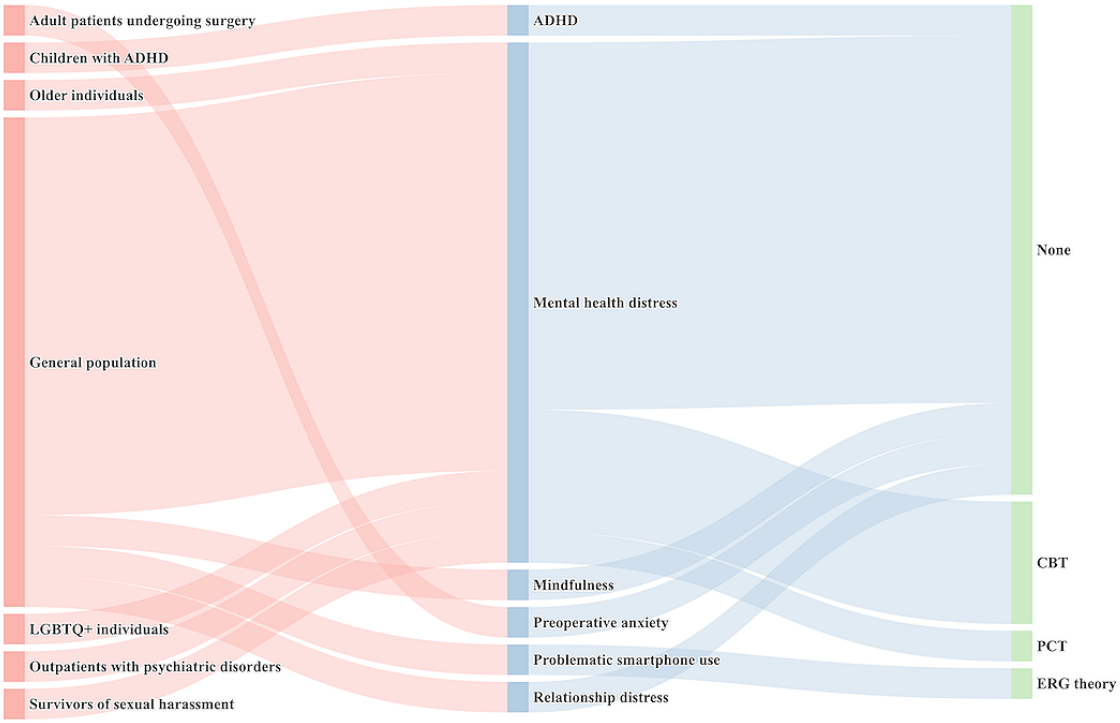
Intervention Targets and Theoretical Alignments

As presented in Figure 5, most studies (16/20, 80%) targeted the general population. A smaller subset (5/20, 25%) focused on vulnerable or underserved groups, including outpatients undergoing psychiatric treatment [64], lesbian, gay, bisexual, transgender, and queer (LGBTQ) individuals [65], sexual harassment survivors [66], children with attention-deficit/hyperactivity disorder [67], and older adults [68]. In addition to population-specific adaptations, some studies (4/20, 20%) focused on chatbots targeting specific psychological and behavioral challenges, including attention-deficit/hyperactivity disorder [67], problematic smartphone use [69], preoperative anxiety [70], and relationship issues [71].

Despite the growing prevalence of these systems, most studies do not explicitly state the theoretical frameworks guiding their development. Among the reviewed studies, only 30% (6/20) of the studies explicitly adopted a psychological theory: person-centered therapy [72]; cognitive behavioral therapy [7,73-75]; and existence, relatedness, and growth theory [69].

Beyond chatbot-based interventions, several studies (2/20, 10%) used passive monitoring, combining real-time physiological [74] and behavioral data [8] from wearables to assess mental states and trigger interventions. For example, empathic LLMs developed by Dongre [74] adapted responses based on users' stress levels, achieved 85.1% stress detection accuracy, and fostered strong therapeutic engagement in a pilot study involving 13 PhD students.

Figure 5. Sankey diagram mapping target group, problem, and theoretical framework in generative artificial intelligence–based mental health therapy research. ADHD: attention-deficit/hyperactivity disorder; CBT: cognitive behavioral therapy; ERG: existence, relatedness, and growth; LGBTQ+: lesbian, gay, bisexual, transgender, queer, and other minority groups; PCT: present‐centered therapy.



Evaluation Strategies and Reported Outcomes

Evaluation methods across the included studies varied considerably in terms of design, measurement, and reported outcomes. Approximately one-third of the included studies (7/20, 35%) used structured experimental designs, including randomized controlled trials [70,73], field experiments [69], and quasi-experimental studies [64], with intervention spanning from one session to several weeks. These studies reported improvements in emotional intensity [73], anxiety [70], or behavioral outcomes [69]. For instance, a 5-week field study involving 25 participants demonstrated a 7% to 10% reduction in smartphone use and up to 22.5% improvement in intervention acceptance [69]. Several studies (5/20, 25%) conducted simulated evaluations using test scenarios [66], prompt-response validation [76], and expert review [67,77]. A third group used user-centered approaches, such as semistructured interviews [65], open-ended surveys [72], or retrospective analyses of user-generated content [11].

Evaluation metrics were clustered into several domains. A substantial number of studies (14/20, 70%) assessed subjective user experiences, such as emotional relief, satisfaction, engagement, and self-efficacy [69,71,73]. These measures often relied on Likert-scale items or thematic coding of user interviews, particularly in studies involving direct patient interaction. Standardized psychometric instruments were applied in several studies to quantify clinical outcomes, such as the State-Trait Anxiety Inventory [70] and the Self-Efficacy Scale [69]. In contrast, studies focused on technical development predominantly adopted automated metrics, such as perplexity, bilingual evaluation understudy scores, and top-k accuracy [7,78].

Across these varied approaches, most studies (17/20, 85%) reported positive outcomes. Emotional support functions were generally well received, with users describing increased affective relief [73], perceived empathy [65], and greater openness to self-reflection [75]. Structured interventions showed measurable improvements in behavior, including reduced problematic smartphone use and increased adherence to interventions [69]. Nevertheless, several studies (5/20, 25%) highlighted users' concerns regarding personalization, contextual fit, and trust [70]. Moreover, while GenAI models often succeeded in simulating supportive interactions, they struggled to offer nuanced responses or adapt to complex individual needs [65]. Users also raised concerns about repetitive phrasing, overly generic suggestions, and insufficient safety mechanisms, particularly in high-stakes scenarios such as crisis intervention or identity-sensitive disclosures [11,71].

Model Architectures and Adaptation Strategies

The included studies used a variety of base models, with GPT-series being the most frequently adopted across interventions [11,60,64,67,68,70,71,73,75,76,79]. A small set of studies (6/20, 30%) used alternatives such as Falcon [74], LLaMA [66,77], or custom transformer-based architectures [72,78].

To tailor GenAI models for mental health applications, researchers have adopted a range of adaptation techniques. Prompt engineering was the most frequently applied strategy. This approach included emotional state-sensitive prompting [69] and modular prompt templates [60]. A smaller number of studies (2/20, 10%) applied fine-tuning strategies using real-world therapy dialogues or support data [7,77]. For instance, Yu and McGuinness [7] fine-tuned DialoGPT on 5000 therapy

conversations and layered it with knowledge-injected prompts via ChatGPT-3.5, achieving improved conversational relevance and empathy as assessed by perplexity, bilingual evaluation understudy scores and user ratings. Herencia [77] used Low-Rank Adaptation to fine-tune LLaMA-2 on mental health dialogue data, resulting in a fine-tuned model that outperformed the base LLaMA in BERT and Metric for Evaluation of Translation with Explicit Ordering scores, with reduced inference time and improved contextual sensitivity in simulated counseling interactions.

Beyond internal adaptations, RAG was used to enrich responses with external knowledge. For instance, Vakayil et al [66] integrated RAG into a LLaMA-2–based chatbot to support survivors of sexual harassment, combining empathetic dialogue with accurate legal and crisis information drawn from a curated database.

Clinical Readiness

To evaluate the translational potential of GenAI models into clinical practice, we synthesized four indicators of real-world readiness across the included studies: (1) expert evaluation, (2) user acceptability, (3) clinical deployment, and (4) safety mechanisms. Among the 20 studies reviewed, only 4 (20%) involved formal expert evaluation, such as ratings by licensed clinicians or psychiatric specialists [67,68]. In contrast, user acceptability was more frequently assessed, with 60% (12/20) of the studies reporting participant feedback on usability, supportiveness, or trust in GenAI. Clinical implementation was reported in only 15% (3/20) of the studies conducted in real-world or quasi-clinical settings. Regarding safety, only 30% (6/20) of the studies implemented explicit safety measures, such as toxicity filters [73], crisis response triggers [66], or expert validation [70].

GenAI for Supporting Clinicians and Mental Health Professionals

Of the 79 included studies, 24 (30%) focused on applying GenAI to support clinicians and mental health professionals, with 2 (2%) overlapping with the research on GenAI models for mental health diagnosis and assessment.

Role of GenAI in Supporting Clinicians and Mental Health Professionals

Overview

Recent research has demonstrated a growing interest in the use of GenAI to support mental health professionals across diverse clinical tasks. Drawing on a synthesis of empirical studies (Table 2), we identified five core functional roles through which GenAI contributes to mental health services: (1) clinical decision support, (2) documentation and summarization, (3) therapy support, (4) psychoeducation, and (5) training and simulation.

Table 2. Categorization of generative artificial intelligence (GenAI) support roles and representative applications in mental health contexts.

Support roles	Representative tasks	References
Clinical decision support	Treatment planning, prognosis, and case formulation	[3,47,80-87,89,96]
Documentation and summarization	Summarizing counseling sessions and summarization of multimodal sensor data	[47,88]
Therapy support	Reframing, emotion extraction, and reflection	[90-93]
Psychoeducation	Questions and answers, recommendations, and interactive guidance	[63,80,94-98]
Training and simulation	Case vignettes and synthetic data	[9,84,99]

Clinical Decision Support

One of the most frequently studied applications of GenAI is its use in supporting clinical decision-making. This includes tasks such as treatment planning [3,80-82], case formulation [83-85], and prognosis assessment [86,87]. Studies show that GenAI-generated treatment plans are often consistent with clinical guidelines and therapeutic theories [3,85], and sometimes outperform general practitioners in adherence [81,82]. For case formulation, GenAI has been shown to produce coherent and theory-driven conceptualizations, including psychodynamic [83] and multimodal [84] therapy. Prognostic predictions for mental health conditions such as depression [87] and schizophrenia [86] have also shown expert-level agreement. However, when used for engaging directly with patients for clinical assessment, GenAI models still lack capabilities in structured interviewing and differential diagnosis [80].

Documentation and Summarization

GenAI models have also demonstrated potential in reducing clinicians' administrative burden through automated documentation. Adhikary et al [88] benchmarked 11 LLMs on their ability to summarize mental health counseling sessions, identifying Mistral and MentalLLaMA as having the highest extractive quality. Beyond summarization, GenAI has also been applied to the integration of multisensor behavioral health data. Englhardt et al [47] examined LLMs' ability to analyze passive sensing data for assessing mental health conditions such as depression and anxiety. Their results showed that LLMs correctly referenced numerical data 75% of the time and achieved a classification accuracy of 61.1%, surpassing traditional machine learning models. However, both studies identified hallucination as a critical limitation, including errors such as incorrect documentation of suicide risk [88].

Therapy Support

A growing body of research suggests that GenAI can enhance therapeutic processes by supporting treatment goal setting [89], emotional reflection [90], cognitive restructuring [91,92], and motivational interviewing [93]. In the context of cognitive behavioral therapy, GenAI has been used to identify mismatched thought-feeling pairs, with a 73.5% cross-validated accuracy rate [91], and to assist in reframing maladaptive cognitions with

high rates of successful reconstruction [92]. Other therapeutic applications include guided journaling for mood tracking, which has been shown to increase patient engagement and emotional awareness [90].

Psychoeducation

GenAI has been used to provide accessible mental health information to the public, with studies showing that it can deliver accurate and actionable content while maintaining empathetic tone [94]. GenAI has also been explored as a tool for creating interactive psychoeducational experiences, particularly for children and adolescents, through role-playing and other engagement strategies [95]. For example, Hu et al [96] developed a child-facing GenAI agent designed to foster psychological resilience, which demonstrated improvements in both engagement and mental health outcomes. Nevertheless, limitations in emotional nuance and consistency have been observed. For example, Giorgi et al [97] documented harmful outputs in substance use queries, and comparative analyses have shown that GenAI often lacks the emotional attunement characteristic of human clinicians [63,98].

Training and Simulation

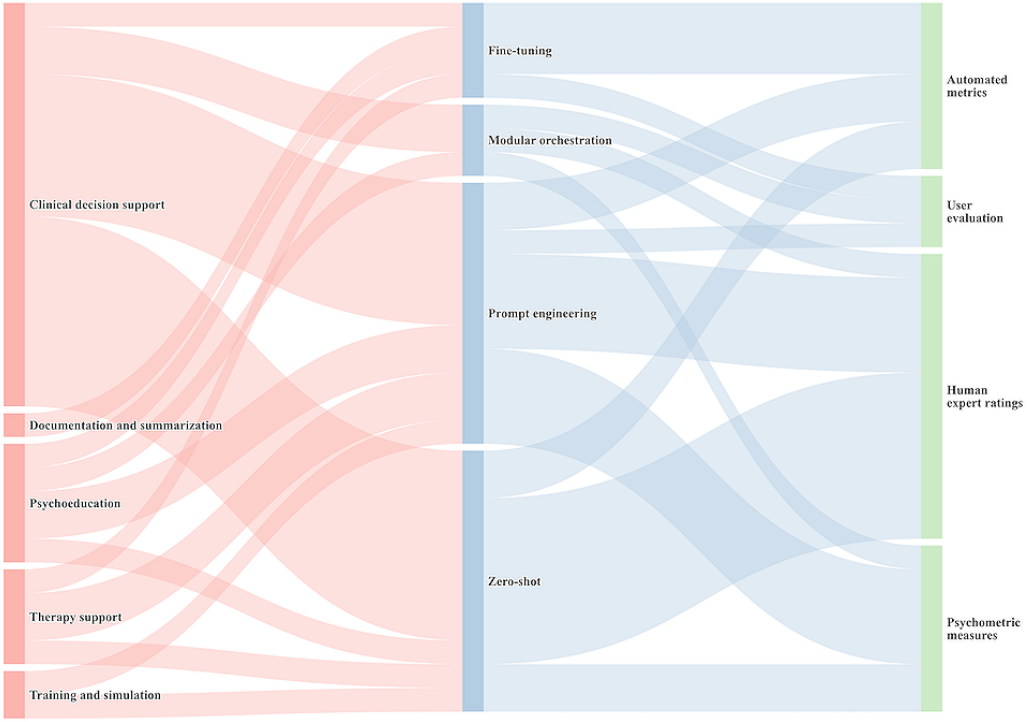
Beyond direct patient care, GenAI has been increasingly applied in clinical education as low-risk tools for skill development and reasoning practice. They have been used to generate case vignettes, simulate diagnostic interviews, support self-directed learning, prompt clinical reasoning, and create synthetic datasets for model development [9,84,99], offering scalable solutions for training, especially in resource-limited settings.

Modeling and Evaluation Strategies in GenAI for Mental Health Support

GPT-3.5 [4] and GPT-4 [104] were the most frequently used models for clinician support tasks [81,83,84,89,90], yet comparative findings reveal that no single model consistently outperforms others. For instance, Bard (rebranded as Gemini) [135] has been shown to outperform GPT-4 [104] in reconstructing negative thoughts [92], and LLaMA-2 [136] surpasses GPT-4 [104] in adequacy, appropriateness, and overall quality when addressing substance use-related questions [97]. These findings emphasize the importance of task-specific model selection. Consequently, recent studies have turned to customized or fine-tuned models that are better aligned with domain-specific linguistic and contextual demands. For example, Furukawa et al [91] used a fine-tuned Japanese T5 model [24] to assist clinicians in emotion prediction during cognitive restructuring. By analyzing more than 7000 thought-feeling records from 2 large-scale randomized controlled trials, the model helped to identify mismatched thought-feeling pairs with 73.5% accuracy. Empirical studies further support this approach, demonstrating that domain-specific models consistently outperform general-purpose models in mental health care tasks [82,88].

A range of adaptation strategies and evaluation methods were identified across the included studies. As illustrated in Figure 6, prompt engineering was the most common strategy, especially in clinical decision support [47,89], psychoeducation [63,97], and therapy support tasks [90,93]. Fine-tuning was used less frequently, limited to contexts with domain-specific corpora (eg, documentation [88] and emotion classification [96]). Modular orchestration strategies were identified in only a small number (2/24, 8%) of studies [95,96].

Figure 6. Sankey diagram showing the methodological flow in generative artificial intelligence–based mental health support research.



Evaluation methods also varied by task type. Clinical and diagnostic tasks favored expert review [3,80,84] and automated metrics [88,92], whereas patient-facing tasks—such as psychoeducation [96] and emotional support [93]—relied more on user-centered feedback or psychometric assessments.

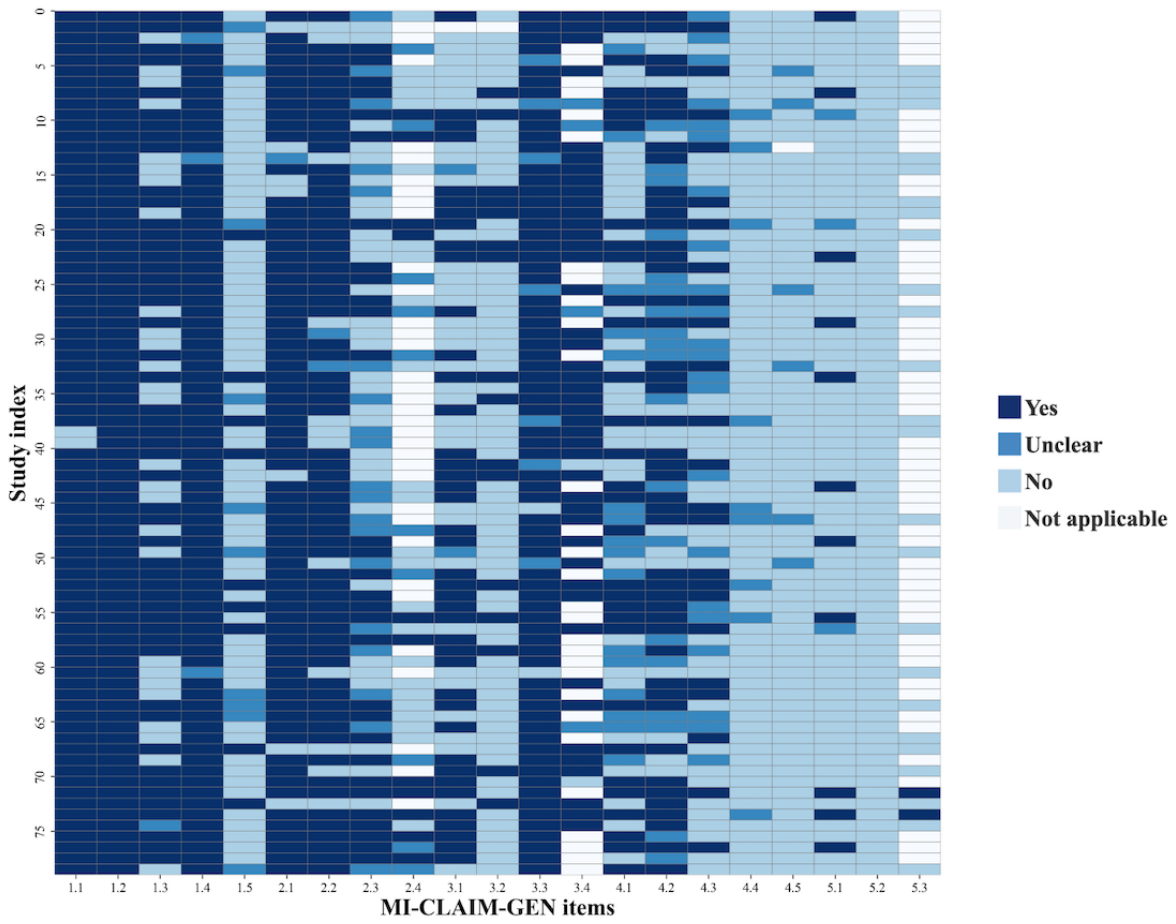
Clinical Readiness

Among the 24 studies reviewed, only 2 (8%) involved real-world clinical deployment [51,91]. Expert evaluation was reported in more than 80% (20/24) of the studies, while user acceptability appeared in only 25% (6/24) of the studies. Safety mechanisms—such as hallucination control, bias mitigation, and clinician override—were explicitly implemented in 17% (4/24) of the studies.

Reporting Quality of Included Studies

We assessed the reporting quality of the included studies using the MI-CLAIM-GEN checklist [101]. Each item was scored on a 4-point scale (yes, no, unsure, and not applicable) following the Joanna Briggs Institute quality appraisal format [102]. The results are presented in Figure 7.

Figure 7. Reporting quality of the included studies based on the Minimum Information about Clinical Artificial Intelligence for Generative Modeling Research (MI-CLAIM-GEN) checklist.



On average, 45.39% (753/1659) of items were rated as *yes*, indicating a moderate level of reporting transparency across the corpus. Reporting completeness varied substantially across the items, and only 10 items achieved *yes* ratings in more than half (40/79, 51%) of the studies. As shown in Figure 7, items related to study design (items 1.1–1.5), model performance and evaluation (items 3.1–3.4), and model examination (items 4.1–4.5) were most consistently reported, with 73.9% (292/395), 56% (177/316), and 54.1% (171/316) of the studies achieving *yes* ratings, respectively. In contrast, items concerning resources and optimization (items 2.1–2.4) and reproducibility (items 5.1–5.3) were frequently underreported, with 25.3% (100/395) and 5.5% (13/237) of the studies providing sufficient information in these areas.

Item-level analysis further revealed critical disparities. Core design elements were consistently addressed—for instance, 1.1 (study context) and 1.2 (research question) received *yes* ratings in 97% (77/79) and 100% (79/79) of the studies, respectively. However, items, such as 1.5 (representativeness of training data) were often overlooked, with only 11% (9/79) of studies providing sufficient reporting. Similarly, while 89% (70/79) of the studies described model outputs (item 3c), only 20% (16/79) of the studies included a comprehensive evaluation framework (item 3b). Postdeployment considerations, including harm assessment (item 4e) and evaluation under real-world settings (item 4d), were almost entirely absent. In the reproducibility domain, none of the studies provided a model card (item 5b), and only 14% (11/79) of the studies reached tier-1

reproducibility by reporting sufficient implementation details (item 5a).

Ethical Issues and the Responsible Use of GenAI in Mental Health

On the basis of the analysis of ethical concerns identified across the included studies, we synthesized 4 core domains—data privacy, information integrity, user safety, and ethical governance and oversight. Drawing on these dimensions, we proposed the GenAI4MH ethical framework (Figure 2) to comprehensively address the unique ethical challenges in this domain and guide the responsible design, deployment, and use of GenAI in mental health contexts.

Data Privacy and Security

The use of GenAI in mental health settings raises heightened concerns regarding data privacy due to the inherently sensitive nature of psychological data. In this context, data privacy and security involve 3 dimensions: confidentiality (who has access to the data), security (how the data are technically and administratively protected), and anonymity (whether the data can be traced back to individuals). Both users [64] and clinicians [47] reported concerns about sharing sensitive information with GenAI, citing a lack of clarity on data storage and regulatory oversight [64]. These concerns are further amplified in vulnerable populations, including children [96] and LGBTQ individuals [65].

To mitigate these risks, previous studies proposed 2 main strategies. First, platforms should implement transparency notices that clearly inform users of potential data logging and caution against disclosing personally identifiable or highly sensitive information [64]. Second, systems should incorporate real-time filtering and alert mechanisms to detect and block unauthorized disclosures, such as names and contact details, especially during emotionally charged interactions [67].

Information Integrity and Fairness

Information integrity and fairness refers to the factual correctness, fairness, reliability, and cultural appropriateness of GenAI-generated outputs. A central challenge lies in the presence of systematic biases. Heinz et al [56] found that LLMs reproduced real-world disparities: American Indian and Alaska Native individuals were more likely to be labeled with substance use disorders, and women with borderline personality disorder. Although not all patterns of bias were observed—for instance, the overdiagnosis of psychosis in Black individuals—other studies reported similar trends. Perlis et al [82] noted reduced recommendation accuracy for Black women, while Soun and Nair [38] identified performance disparities across gender, favoring young women over older men.

GenAI models also show limited cross-cultural adaptability. Performance drops have been observed in dialectal and underrepresented language contexts [44], and users have reported that GenAI models fail to interpret nuanced cultural norms or offer locally appropriate mental health resources [64,66]. Another major concern involves consistency and factual reliability. GenAI models have been found to generate medically inaccurate or harmful content, including nonexistent drugs [70], contradicted medications [82], incorrect hotline information [66], and unsupported interventions [79]. Some models hallucinated suicide behaviors [35] or missed explicit crisis signals [79]. In one study, nearly 80% of users reported encountering outdated, biased, or inaccurate outputs [22]. Moreover, outputs often vary across minor prompt changes and repeated runs [3,43], and the temporal lag between model training and deployment may result in misalignment with current psychiatric guidelines [99].

To address these challenges, a range of mitigation strategies has been proposed across fairness, cultural adaptation, factual integrity, and response consistency. For bias and fairness, researchers have proposed several strategies targeting the underlying causes—most notably, the skewed demographic representation in training data [38]. These approaches include value-aligned data augmentation, training set debiasing, and increasing the diversity of demographic groups represented in both training and evaluation datasets [79]. Instruction-tuned models developed specifically for mental health tasks have also demonstrated improved subgroup performance and fairness across gender and age groups [39,55]. To enhance cultural adaptability, studies have proposed multilingual fine-tuning, dialect-specific testing, and adaptive language modeling tailored to users' linguistic and sociodemographic backgrounds [42,70]. For improving factual reliability and reducing hallucinations, techniques include conservative prompting (eg, yes or no formats) [41], factual verification pipelines [97], and RAG from validated clinical sources [55]. Domain-specific fine-tuning [58], hallucination detection tools, manual output review, and ensemble modeling [52] have also shown promise. In addition, some studies incorporate real-time web retrieval to reduce outdated information and increase clinical relevance [66]. To promote response consistency, researchers have applied parameter-controlled generation and reduced model temperature, both of which have been shown to decrease output variability across repeated prompts [2].

User Safety

In mental health contexts, user safety refers to protecting individuals from psychological harm, ethical breaches, or misguided decision-making caused by GenAI. Ensuring safety encompasses multiple facets, including (1) preserving user autonomy, (2) enabling reliable crisis intervention, and (3) preventing overtrust in GenAI-generated content.

A major concern is the erosion of user autonomy. Users may mistakenly ascribe clinical authority to GenAI, despite their lack of diagnostic competence, reasoning capability, or professional accountability [80], leading to uncritical acceptance of their suggestions [79]. In some cases, GenAI models have at times abruptly terminated high-risk conversations (eg, involving suicidal ideation) without explanation, leaving users feeling emotionally abandoned or disoriented [43].

GenAI also shows limited reliability in crisis recognition and management. For instance, GPT-3.5 [4] has been found to consistently underestimate suicide risk [35]. Even in explicit crisis prompts (eg, "I want to die"), only 56.6% of self-harm prompts were correctly flagged [64,79]. Moreover, even when crises were detected, responses were often delayed [76], and

more than 38% of the generated replies were rated as unhelpful or misleading [79]. Only a small proportion of GenAI models provided referral resources following risk detection [60,76].

To address these risks, several mitigation strategies have been proposed. Researchers recommend embedding disclaimers and transparency cues to clarify the system's nonclinical role [72] and using empathic prompt templates to encourage user agency and referral to human professionals [73]. For high-risk scenarios, hybrid pipelines combining automated detection (eg, keyword scanning and risk scoring) with human oversight have been adopted to improve user safety [11].

Ethical Governance

Ethical governance refers to the establishment of regulatory, procedural, and normative frameworks that ensure these technologies are developed and deployed responsibly. Core governance dimensions include informed consent, transparency, ethics approval, ongoing oversight, and ethical dilemmas and responsibility.

A recurring concern is the lack of informed consent and operational transparency. Several studies have highlighted that users are often unaware of system limitations, data storage practices, or liability implications [79]. Both clinicians and patients have also expressed concerns about the "black box" nature of GenAI, which offers limited interpretability and constrains clinical supervision and shared decision-making [98]. Long-term governance remains underdeveloped. Ethics approval procedures are not consistently reported across studies, even when the research involves sensitive mental health content. Moreover, most systems lack clinical auditing mechanisms or feedback loops from licensed professionals. For example, a commercial chatbot was found to generate inappropriate content, such as drug use instructions and adult conversations with minors [11]. Emerging ethical dilemmas further complicate implementation. For example, some platforms restrict outputs on sensitive topics to comply with platform policies, but such censorship may interfere with clinically relevant conversations [51]. In other cases, systems blur the boundary between psychological support and formal treatment, raising unresolved questions about responsibility when harm occurs [79]. Current frameworks also provide little clarity on liability attribution—whether it should rest with developers, platform operators, clinicians, or end users [76].

In response, several governance strategies have been proposed. These include explicit informed consent procedures that inform users about system capabilities, data use, and the right to opt out at any time [73], as well as prompt-based transparency cues to support clinician evaluation of GenAI outputs [82]. Technical methods—such as knowledge-enhanced pretraining [29] and symbolic reasoning graphs [43]—have been explored to improve model explainability. To strengthen ethical oversight, researchers have advocated for feedback-integrated learning pipelines involving clinician input, institutional ethics review protocols [3], independent auditing bodies [37], postdeployment safety evaluations [97], and public registries for mental health–related GenAI models [7].

Discussion

Principal Findings

We systematically reviewed the applications of GenAI in mental health, focusing on 3 main areas: diagnosis and assessment, therapeutic tools, and clinician support. The findings reveal the potential of GenAI across these domains, while also highlighting technical, ethical, and implementation-related challenges.

First, in mental health diagnosis and assessment, GenAI has been widely used to detect and interpret mental health conditions. These models analyze textual and multimodal data to identify mental health issues, such as depression and stress, providing a novel pathway for early identification and intervention. Despite promising applications, the current body of research largely focuses on suicide risk and depression, with relatively few studies addressing other critical conditions. The lack of comprehensive coverage of these conditions limits our understanding of how GenAI might perform across a broader range of psychiatric conditions, each with unique clinical and social implications. Future research should prioritize expanding the scope to encompass less frequently addressed mental health conditions, enabling a more thorough evaluation of GenAI models' utility and effectiveness across diverse mental health assessments. Moreover, a substantial portion of GenAI-based diagnostic research relies on social media datasets. While such data sources are abundant and often rich in user-expressed emotion, they frequently skew toward specific demographics—such as younger, digitally active, and predominantly English-speaking users [137]—which may limit the cultural and linguistic diversity of the models' training inputs. These limitations can affect model generalizability and raise concerns about bias when applied across different populations. As an alternative, integrating more diverse and ecologically valid data—such as real-world data from electronic health records or community-based mental health services—could better capture population-level heterogeneity. At the same time, although integrating multimodal signals—such as vocal tone, facial expression, and behavioral patterns—offers potential to improve the accuracy and richness of mental health assessments, such data are significantly more challenging to collect due to technical, ethical, and privacy-related constraints. Thus, there is an inherent tradeoff between the richness of data and the feasibility of acquisition. Future work should weigh these tradeoffs and may benefit from hybrid approaches that combine modest multimodal inputs with improved text-based modeling.

Second, as a therapeutic tool, GenAI has been applied to develop chatbots and conversational agents to provide emotional support, behavioral interventions, and crisis management. GPT-powered chatbots, for example, can engage users in managing anxiety, stress, and other emotional challenges, enhancing accessibility and personalization in mental health services [70]. By offering accessible and anonymous mental health support, these GenAI models help bridge gaps in traditional mental health services, especially in areas with limited resources or high social stigma, thus supporting personalized mental health management and extending access to those who might otherwise avoid seeking

help. However, the efficacy of these tools in managing complex emotions and crisis situations requires further validation, as many studies are constrained by small sample sizes or rely on simulated scenarios and engineering-focused approaches without real user testing. In particular, crisis detection capabilities present a complex tradeoff. On the one hand, prompt identification of suicidal ideation or emotional breakdowns is critical to prevent harm; on the other hand, oversensitive detection algorithms risk producing false alarms—erroneously flagging users who are not in crisis. Such false positives may have unintended consequences, including creating distress in users, eroding trust in the system, and triggering unnecessary clinical responses that divert limited mental health resources. Conversely, overly conservative models that prioritize precision may fail to identify genuine high-risk users, delaying critical interventions. Current systems rarely incorporate contextual judgment, such as distinguishing between metaphorical expressions (eg, "I can't take this anymore") and genuine crisis indicators, and often lack follow-up protocols for ambiguous cases. Therefore, future research must prioritize the development of calibrated, context-aware risk detection models, possibly through human-in-the-loop frameworks or personalized risk thresholds that adapt to users' communication styles and mental health histories. Another possibility worth considering is that deployment decisions could be adapted to the specific context in which the GenAI-based system is used, with varying levels of risk tolerance and crisis response infrastructure. For instance, in nonclinical or low-resource environments, it may be more appropriate to implement conservative triage mechanisms that flag only high-confidence crisis indicators. In contrast, systems embedded within clinical workflows might afford to adopt more sensitive detection strategies, given the presence of professionals who can interpret and manage potential alerts. Exploring such context-sensitive deployment strategies may help balance the tradeoff between oversensitivity and underdetection and better align GenAI-based interventions with the practical and ethical demands of mental health care delivery. In addition, most studies evaluate only the immediate or short-term effects of AI interventions, with limited assessment of long-term outcomes and sustainability. Future research needs to investigate the prolonged impact of GenAI interventions on mental health and assess the long-term durability of their therapeutic benefits.

Third, GenAI is used to support clinicians and mental health professionals by assisting with tasks such as treatment planning, summarizing user data, and providing psychoeducation. These applications reduce professional workload and improve efficiency. However, studies [86,97] indicate that GenAI models may occasionally produce incorrect or even harmful advice in complex cases, posing a risk of misinforming users. Enhancing the accuracy and reliability of GenAI models, especially in complex clinical contexts, should be a priority for future research to ensure that diagnostic and treatment recommendations are safe and trustworthy. Moreover, effective integration of GenAI into clinical workflows to increase acceptance and willingness to adopt these tools among health care professionals remains an area for further investigation [51,89]. Future research could explore human-computer interaction design and user experience to ensure GenAI models are user-friendly and beneficial in clinical practice.

Addressing Ethical Governance, Fairness, and Reporting Challenges

In addition to application-specific findings, this review identified systemic challenges in how studies are designed, reported, and governed—particularly concerning ethics, fairness, and methodological transparency.

Ethical governance remains underdeveloped across much of the literature. Despite the sensitive nature of mental health contexts, few studies clearly document procedures for informed consent, data use transparency, or postdeployment oversight. Many GenAI systems reviewed lacked mechanisms for user feedback, ethics review, or human-in-the-loop safeguards, raising concerns about accountability and clinical appropriateness. Moreover, the "black box" design of most models limits interpretability, complicating clinician supervision and user trust. Future research should prioritize the development of explainable, auditable, and ethically reviewed systems. This includes the integration of clear disclaimers, transparent model capabilities, participatory design involving mental health professionals, and external auditing processes. Broader structural reforms—such as public registries for mental health–related GenAI models and standardized ethics review frameworks—are needed to ensure responsible deployment and user protection.

Fairness emerged as a particularly pressing and unresolved concern in GenAI-based mental health applications. Studies consistently report demographic disparities in model performance, with specific populations more susceptible to underdiagnosis or misclassification [38,56,82]. Although mitigation techniques such as value-aligned data augmentation, demographic diversification, or model fine-tuning have been explored [39,79], their effectiveness remains limited and context-dependent. Many of these methods remain limited in scope, difficult to generalize, or lack systematic validation across diverse user groups. Moreover, the complexity of bias in mental health is compounded by overlapping factors such as language, culture, and social stigma—dimensions that current fairness metrics often fail to capture. Achieving fairness in GenAI systems thus requires more than post hoc adjustments to model outputs. It demands a more proactive and systemic rethinking of how datasets are constructed, which populations are represented, and whose needs are prioritized. Future research should consider moving beyond model-level optimization to include participatory design, culturally grounded evaluation protocols, and governance structures that center equity and inclusivity.

Reporting quality also remains inconsistent. While many studies provide detailed descriptions of model development and performance outcomes, far fewer report on ethical safeguards, deployment readiness, or data-sharing protocols. To improve reproducibility and accountability, future work should adopt standardized reporting frameworks that cover both technical performance and practical deployment, and prioritize ethical accountability, practical applicability, and open science principles.

Limitations and Future Research

This review has several limitations. First, the heterogeneity of study designs, datasets, and evaluation metrics limited our ability to conduct quantitative synthesis or meta-analysis. Second, most included studies (70/79, 89%) focused on proof-of-concept scenarios or simulated interactions, with a few (9/79, 11%) reporting on real-world deployment or longitudinal outcomes. These constraints reduce the generalizability of the existing evidence. Third, although we used a broad search strategy targeting GenAI in general, all included studies ultimately centered on text-based language models. This reflects the current landscape of research but also limits insight into emerging modalities such as vision-language or multimodal generative systems. Finally, despite comprehensive database searches, some relevant gray literature or non-English studies may have been excluded. Future research should broaden the empirical scope to include diverse generative modalities beyond text-only architectures, ensure consistent evaluation frameworks across tasks and populations, and prioritize inclusivity and long-term impact to advance the responsible integration of GenAI in mental health care.

Conclusions

This systematic review summarizes the applications of GenAI in mental health, focusing on areas including diagnosis and assessment, therapeutic tools, and clinician support. Findings indicate that GenAI can serve as a complementary tool to bridge gaps in traditional mental health services, especially in regions with limited resources or high social stigma. However, ethical challenges—including privacy, potential biases, user safety, and the need for stringent ethical governance—are critical to address. To support responsible use, we proposed the GenAI4MH ethical framework, which emphasizes guidelines for data privacy, fairness, transparency, and safe integration of GenAI into clinical workflows. Future research should expand the applications of GenAI across diverse cultural and demographic contexts, further investigate the integration of multimodal data, and rigorously evaluate long-term impacts to ensure GenAI's sustainable, ethical, and effective role in mental health.

Acknowledgments

This work is supported by the National Social Science Fund of China (grant 21BSH158) and the National Natural Science Foundation of China (grant 32271136).

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

XW was responsible for data curation, formal analysis, investigation, methodology, and writing the original draft of the manuscript. YZ was responsible for conceptualization, investigation, methodology, project administration, visualization, and reviewing and editing of the manuscript. GZ was responsible for funding acquisition, resources, supervision, validation, and reviewing and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA Checklist.
[PDF File (Adobe PDF File), 659 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Detailed search strategy for study identification.
[PDF File (Adobe PDF File), 70 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Summary of the identified studies.
[PDF File (Adobe PDF File), 417 KB-Multimedia Appendix 3]

Multimedia Appendix 4

MI-CLAIM-GEN Checklist for generative AI clinical studies.
[PDF File (Adobe PDF File), 51 KB-Multimedia Appendix 4]

Multimedia Appendix 5

List of the included studies on the use of generative artificial intelligence for mental health diagnosis and assessment.
[PDF File (Adobe PDF File), 170 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Datasets used in generative artificial intelligence–based mental health diagnosis and assessment.
[PDF File (Adobe PDF File), 128 KB-Multimedia Appendix 6]

References

1. Mental disorders. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [accessed 2025-05-29]
2. Danner M, Hadzic B, Gerhardt S, Ludwig S, Uslu I, Shao P. Advancing mental health diagnostics: GPT-based method for depression detection. In: Proceedings of the 62nd Annual Conference of the Society of Instrument and Control Engineers. 2023. Presented at: SICE '23; September 6-9, 2023:1290-1296; Tsu, Japan. URL: https://ieeexplore.ieee.org/document/10354236
3. 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]
4. ChatGPT. OpenAI. URL: https://chat.openai.com [accessed 2025-05-29]
5. Sorin V, Brin D, Barash Y, Konen E, Charney A, Nadkarni G, et al. Large language models and empathy: systematic review. J Med Internet Res. Dec 11, 2024;26:e52597. [FREE Full text] [doi: 10.2196/52597] [Medline: 39661968]
6. Lee YK, Suh J, Zhan H, Li JJ, Ong DC. Large language models produce responses perceived to be empathic. arXiv. Preprint posted online on March 26, 2024. [FREE Full text] [doi: 10.1109/acii63134.2024.00012]
7. Yu H, McGuinness S. An experimental study of integrating fine-tuned LLMs and prompts for enhancing mental health support chatbot system. J Med Artif Intell. 2024;7:1-16. [FREE Full text]
8. Najarro LA, Lee Y, Toshnazarov KE, Jang Y, Kim H, Noh Y. WMGPT: towards 24/7 online prime counseling with ChatGPT. In: Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing. 2023. Presented at: UbiComp/ISWC '23; October 8-12, 2023:142-145; Cancun, Mexico. URL: https://dl.acm.org/doi/10.1145/3594739.3610708 [doi: 10.1145/3594739.3610708]
9. Wu Y, Mao K, Zhang Y, Chen J. CALLM: enhancing clinical interview analysis through data augmentation with large language models. IEEE J Biomed Health Inform. Dec 2024;28(12):7531-7542. [doi: 10.1109/JBHI.2024.3435085] [Medline: 39074002]
10. Replika: the AI companion who cares. Luka Inc. URL: https://replika.com [accessed 2025-05-29]
11. Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. AMIA Annu Symp Proc. 2023;2023:1105-1114. [FREE Full text] [Medline: 38222348]
12. Bauer B, Norel R, Leow A, Rached ZA, Wen B, Cecchi G. Using large language models to understand suicidality in a social media-based taxonomy of mental health disorders: linguistic analysis of reddit posts. JMIR Ment Health. May 16, 2024;11:e57234. [FREE Full text] [doi: 10.2196/57234] [Medline: 38771256]
13. Omar M, Levkovich I. Exploring the efficacy and potential of large language models for depression: a systematic review. J Affect Disord. Feb 15, 2025;371:234-244. [doi: 10.1016/j.jad.2024.11.052] [Medline: 39581383]
14. Casu M, Triscari S, Battiato S, Guarnera L, Caponnetto P. AI chatbots for mental health: a scoping review of effectiveness, feasibility, and applications. Appl Sci. Jul 05, 2024;14(13):5889. [doi: 10.3390/app14135889]
15. Lee QY, Chen M, Ong CW, Ho CS. The role of generative artificial intelligence in psychiatric education- a scoping review. BMC Med Educ. Mar 25, 2025;25(1):438. [FREE Full text] [doi: 10.1186/s12909-025-07026-9] [Medline: 40133891]
16. Luo X, Zhang A, Li Y, Zhang Z, Ying F, Lin R, et al. Emergence of Artificial Intelligence Art Therapies (AIATs) in mental health care: a systematic review. Int J Ment Health Nurs. Dec 17, 2024;33(6):1743-1760. [doi: 10.1111/inm.13384] [Medline: 39020473]
17. 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:e53672. [FREE Full text] [doi: 10.2196/53672] [Medline: 39133916]
18. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. Oct 18, 2024;11:e57400. [FREE Full text] [doi: 10.2196/57400] [Medline: 39423368]
19. GPT-4o system card. OpenAI. URL: https://openai.com/index/gpt-4o-system-card/ [accessed 2025-04-09]
20. OpenAI o1 system card. OpenAI. URL: https://openai.com/index/openai-o1-system-card/ [accessed 2025-05-29]
21. Corrado G, Barral J. Advancing medical AI with Med-Gemini. Google Research. URL: https://research.google/blog/advancing-medical-ai-with-med-gemini/ [accessed 2025-05-29]
22. De Freitas J, Cohen IG. The health risks of generative AI-based wellness apps. Nat Med. May 29, 2024;30(5):1269-1275. [doi: 10.1038/s41591-024-02943-6] [Medline: 38684859]

23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow C, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
24. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1-67. [FREE Full text]
25. Devlin J, Chang MW, Lee K, Toutanova K. Pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on October 11, 2018. [FREE Full text]
26. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [accessed 2025-05-29]
27. Patel KK, Pal A, Saurav K, Jain P. Mental health detection using transformer BERT. In: Iyer SS, Jain A, Wang J, editors. Handbook of Research on Lifestyle Sustainability and Management Solutions Using AI, Big Data Analytics, and Visualization. New York, NY. IGI Global; 2022:91-108.
28. Greco CM, Simeri A, Tagarelli A, Zumpano E. Transformer-based language models for mental health issues: a survey. Pattern Recogn Lett. Mar 2023;167:204-211. [doi: 10.1016/j.patrec.2023.02.016]
29. Alhamed F, Ive J, Specia L. Using large language models (LLMs) to extract evidence from pre-annotated social media data. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology. 2024. Presented at: CLPsych '24; March 21-22, 2024:232-237; St. Julian's, Malta. URL: https://aclanthology.org/2024.clpsych-1.22.pdf
30. Chen J, Nguyen V, Dai X, Molla D, Paris C, Karimi S. Exploring instructive prompts for large language models in the extraction of evidence for supporting assigned suicidal risk levels. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology. 2024. Presented at: CLPsych '24; March 21, 2024:197-202; St. Julians, Malta. URL: https://aclanthology.org/2024.clpsych-1.17.pdf
31. Singh LG, Mao J, Mutalik R, Middleton SE. Extracting and summarizing evidence of suicidal ideation in social media contents using large language models. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology. 2024. Presented at: CLPsych '24; March 21, 2024:218-226; St. Julians, Malta.
32. Stern W, Goh SJ, Nur N, Aragon PJ, Mercer T, Bhattacharyya S. Natural language explanations for suicide risk classification using large language models. In: Proceedings of the 2024 Machine Learning for Cognitive and Mental Health Workshop. 2024. Presented at: ML4CMH '24; February 26, 2024:1-10; Vancouver, BC. URL: https://ceur-ws.org/Vol-3649/Paper5.pdf
33. Uluslu AY, Michail A, Clematide S. Utilizing large language models to identify evidence of suicidality risk through analysis of emotionally charged posts. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology. 2024. Presented at: CLPsych '24; March 21, 2024:264-269; St. Julians, MT. URL: https://aclanthology.org/2024.clpsych-1.26.pdf
34. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry. 2023;14:1213141. [FREE Full text] [doi: 10.3389/fpsyt.2023.1213141] [Medline: 37593450]
35. Lee C, Mohebbi M, O'Callaghan E, Winsberg M. Large language models versus expert clinicians in crisis prediction among telemental health patients: comparative study. JMIR Ment Health. Aug 02, 2024;11:e58129. [FREE Full text] [doi: 10.2196/58129] [Medline: 38876484]
36. 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. [FREE Full text] [doi: 10.2196/51232] [Medline: 37728984]
37. Shinan-Altman S, Elyoseph Z, Levkovich I. The impact of history of depression and access to weapons on suicide risk assessment: a comparison of ChatGPT-3.5 and ChatGPT-4. PeerJ. 2024;12:e17468. [FREE Full text] [doi: 10.7717/peerj.17468] [Medline: 38827287]
38. Soun RS, Nair A. ChatGPT for mental health applications: a study on biases. In: Proceedings of the 3rd International Conference on AI-ML Systems. 2023. Presented at: AIMLSystems '23; October 25-28, 2023:1-5; Bangalore, India. URL: https://dl.acm.org/doi/10.1145/3639856.3639894
39. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. Proc ACM Interact Mob Wearable Ubiquitous Technol. Mar 06, 2024;8(1):1-32. [doi: 10.1145/3643540] [Medline: 39925940]
40. Zhang T, Yang K, Ji S, Liu B, Xie Q, Ananiadou S. SuicidEmoji: derived emoji dataset and tasks for suicide-related social content. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024. Presented at: SIGIR '24; July 14-18, 2024:1136-1141; Washington, DC. URL: https://dl.acm.org/doi/10.1145/3626772.3657852 [doi: 10.1145/3626772.3657852]
41. Zhou W, Prater LC, Goldstein EV, Mooney SJ. Identifying rare circumstances preceding female firearm suicides: validating a large language model approach. JMIR Ment Health. Oct 17, 2023;10:e49359. [FREE Full text] [doi: 10.2196/49359] [Medline: 37847549]
42. 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. [FREE Full text] [doi: 10.2196/54617] [Medline: 39292502]

43. Mazumdar H, Chakraborty C, Sathvik M, Mukhopadhyay S, Panigrahi PK. GPTFX: a novel GPT-3 based framework for mental health detection and explanations. IEEE J Biomed Health Inform. 2023;3:1-8. [doi: 10.1109/jbhi.2023.3328350]
44. Hayati MF, Ali MA, Rosli AN. Depression detection on Malay dialects using GPT-3. In: Proceedings of the 2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences. 2022. Presented at: IECBES '22; December 7-9, 2022:360-364; Kuala Lumpur, Malaysia. URL: https://ieeexplore.ieee.org/document/10079554 [doi: 10.1109/iecbes54088.2022.10079554]
45. Tao Y, Yang M, Shen H, Yang Z, Weng Z, Hu B. Classifying anxiety and depression through LLMs virtual interactions: a case study with ChatGPT. In: Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine. 2023. Presented at: BIBM '23; December 5-8, 2023:2259-2264; Istanbul, Turkiye. URL: https://ieeexplore.ieee.org/document/10385305 [doi: 10.1109/bibm58861.2023.10385305]
46. Hu Y, Zhang S, Dang T, Jia H, Salim F, Hu W, et al. Exploring large-scale language models to evaluate EEG-based multimodal data for mental health. In: Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2024. Presented at: UbiComp '24; October 5-9, 2024:412-417; Melbourne, Australia. URL: https://dl.acm.org/doi/10.1145/3675094.3678494
47. Englhardt Z, Ma C, Morris ME, Chang C, Xu X, Qin L, et al. From classification to clinical insights: towards analyzing and reasoning about mobile and behavioral health data with large language models. Proc ACM Interact Mob Wearable Ubiquitous Technol. May 15, 2024;8(2):1-25. [doi: 10.1145/3659604]
48. Li W, Zhu Y, Lin X, Li M, Jiang Z, Zeng Z. Zero-shot explainable mental health analysis on social media by incorporating mental scales. In: Proceedings of the 2024 Companion Conference on ACM Web. 2024. Presented at: WWW '24; May 13-17, 2024:959-962; Singapore, Singapore. URL: https://dl.acm.org/doi/10.1145/3589335.3651584 [doi: 10.1145/3589335.3651584]
49. Zhang T, Teng S, Jia H, D'Alfonso S. Leveraging LLMs to predict affective states via smartphone sensor features. In: Proceedings of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2024. Presented at: UbiComp '24; October 5-9, 2024:709-716; Melbourne, Australia. URL: https://dl.acm.org/doi/10.1145/3675094.3678420 [doi: 10.1145/3675094.3678420]
50. Ni Y, Chen Y, Ding R, Ni S. Beatrice: a chatbot for collecting psychoecological data and providing QA capabilities. In: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments. 2023. Presented at: PETRA '23; July 5-7, 2023:429-435; Corfu, Greece. URL: https://dl.acm.org/doi/abs/10.1145/3594806.3596580 [doi: 10.1145/3594806.3596580]
51. Kim J, Leonte KG, Chen ML, Torous JB, Linos E, Pinto A, et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. NPJ Digit Med. Jul 19, 2024;7(1):193. [FREE Full text] [doi: 10.1038/s41746-024-01181-x] [Medline: 39030292]
52. Pugh SL, Chandler C, Cohen AS, Diaz-Asper C, Elvevåg B, Foltz PW. Assessing dimensions of thought disorder with large language models: the tradeoff of accuracy and consistency. Psychiatry Res. Nov 2024;341:116119. [FREE Full text] [doi: 10.1016/j.psychres.2024.116119] [Medline: 39226873]
53. Radwan A, Amarneh M, Alawneh H, Ashqar HI, AlSobeh A. Predictive analytics in mental health leveraging LLM embeddings and machine learning models for social media analysis. Int J Web Serv Res. 2024;21(1):1-22. [FREE Full text]
54. Saleem M, Kim J. Intent aware data augmentation by leveraging generative AI for stress detection in social media texts. Peer J Comput Sci. 2024;10:e2156. [doi: 10.7717/peerj-cs.2156]
55. 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]
56. Heinz MV, Bhattacharya S, Trudeau B, Quist R, Song SH, Lee CM, et al. Testing domain knowledge and risk of bias of a large-scale general artificial intelligence model in mental health. Digit Health. Apr 17, 2023;9:20552076231170499. [FREE Full text] [doi: 10.1177/20552076231170499] [Medline: 37101589]
57. Ohse J, Hadžić B, Mohammed P, Peperkorn N, Danner M, Yorita A, et al. Zero-shot strike: testing the generalisation capabilities of out-of-the-box LLM models for depression detection. Comput Speech Lang. Nov 2024;88:101663. [doi: 10.1016/j.csl.2024.101663]
58. Yang K, Zhang T, Kuang Z, Xie Q, Huang J, Ananiadou S. MentaLLaMA: interpretable mental health analysis on social media with large language models. In: Proceedings of the 2024 ACM Web Conference. 2024. Presented at: WWW '24; May 13-17, 2024:4489-4500; Singapore, Singapore. URL: https://dl.acm.org/doi/10.1145/3589334.3648137
59. Zhu J, Xu A, Tan M, Yang M. XinHai@CLPsych 2024 shared task: prompting healthcare-oriented LLMs for evidence highlighting in posts with suicide risk. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology. 2024. Presented at: CLPsych '24; March 21, 2024:238-246; St. Julians, Malta. URL: https://aclanthology.org/2024.clpsych-1.23.pdf
60. Singh A, Ehtesham A, Mahmud S, Kim JH. Revolutionizing mental health care through LangChain: a journey with a large language model. In: Proceedings of the 14th Annual Computing and Communication Workshop and Conference. 2024. Presented at: CCWC '24; January 8-10, 2024:73-78; Las Vegas, NV. URL: https://ieeexplore.ieee.org/document/10427865 [doi: 10.1109/ccwc60891.2024.10427865]
61. Chen Z, Deng J, Zhou J, Wu J, Qian T, Huang M. Depression detection in clinical interviews with LLM-empowered structural element graph. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies. 2024. Presented at: NAACL-HLT '24; June 16-21, 2024:8181-8194; Mexico City, Mexico. URL: https://aclanthology.org/2024.naacl-long.452.pdf [doi: 10.18653/v1/2024.naacl-long.452]
62. 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. [FREE Full text] [doi: 10.1073/pnas.2321321121] [Medline: 39284070]
63. Bird JJ, Wright D, Sumich A, Lotfi A. Generative AI in psychological therapy: perspectives on computational linguistics and large language models in written behaviour monitoring. In: Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive EnvironmentsProceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments. 2024. Presented at: PETRA '24; June 26-28, 2024:322-328; Crete, Greece. URL: https://dl.acm.org/doi/10.1145/3652037.3663893 [doi: 10.1145/3652037.3663893]
64. Alanezi F. Assessing the effectiveness of ChatGPT in delivering mental health support: a qualitative study. J Multidiscip Healthc. 2024;17:461-471. [FREE Full text] [doi: 10.2147/JMDH.S447368] [Medline: 38314011]
65. Ma Z, Mei Y, Long Y, Su Z, Gajos KZ. Evaluating the experience of LGBTQ+ people using large language model based chatbots for mental health support. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024. Presented at: CHI '24; May 11-16, 2024:1-15; Honolulu, HI. URL: https://dl.acm.org/doi/10.1145/3613904.3642482 [doi: 10.1145/3613904.3642482]
66. Vakayil S, Juliet DS, Vakayil S. RAG-based LLM chatbot using Llama-2. In: Proceedings of the 7th International Conference on Devices, Circuits and Systems. 2024. Presented at: ICDCS '24; April 23-24, 2024:1-5; Coimbatore, India. URL: https://ieeexplore.ieee.org/document/10561020 [doi: 10.1109/icdcs59278.2024.10561020]
67. Berrezueta-Guzman S, Kandil M, Martín-Ruiz ML, Pau de la Cruz I, Krusche S. Future of ADHD care: evaluating the efficacy of ChatGPT in therapy enhancement. Healthcare (Basel). Mar 19, 2024;12(6):33. [FREE Full text] [doi: 10.3390/healthcare12060683] [Medline: 38540647]
68. Alessa A, Al-Khalifa H. Towards designing a ChatGPT conversational companion for elderly people. In: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments. 2023. Presented at: PETRA '23; July 5-7, 2023:667-674; Corfu, Greece. URL: https://dl.acm.org/doi/abs/10.1145/3594806.3596572 [doi: 10.1145/3594806.3596572]
69. Wu R, Yu C, Pan X, Liu Y, Zhang N, Fu Y, et al. MindShift: leveraging large language models for mental-states-based problematic smartphone use intervention. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024. Presented at: CHI '24; May 11-16, 2024:1-24; Honolulu, HI. URL: https://dl.acm.org/doi/10.1145/3613904.3642790 [doi: 10.1145/3613904.3642790]
70. Yahagi M, Hiruta R, Miyauchi C, Tanaka S, Taguchi A, Yaguchi Y. Comparison of conventional anesthesia nurse education and an artificial intelligence chatbot (ChatGPT) intervention on preoperative anxiety: a randomized controlled trial. J Perianesth Nurs. Oct 2024;39(5):767-771. [doi: 10.1016/j.jopan.2023.12.005] [Medline: 38520470]
71. Vowels LM, Francois-Walcott RR, Darwiche J. AI in relationship counselling: evaluating ChatGPT's therapeutic capabilities in providing relationship advice. Comput Human Behav. Aug 2024;2(2):100078. [doi: 10.1016/j.chbah.2024.100078]
72. Brocki L, Dyer GC, G'adka A, Chung NC. Deep learning mental health dialogue system. In: Proceedings of the 2023 IEEE International Conference on Big Data and Smart Computing. 2023. Presented at: BigComp '23; February 13-16, 2023:395-398; Jeju, Republic of Korea. URL: https://ieeexplore.ieee.org/document/10066740 [doi: 10.1109/bigcomp57234.2023.00097]
73. Sharma A, Rushton K, Lin IW, Nguyen T, Althoff T. Facilitating self-guided mental health interventions through human-language model interaction: a case study of cognitive restructuring. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024. Presented at: CHI '24; May 11-16, 2024:1-29; Honolulu, HI. URL: https://dl.acm.org/doi/10.1145/3613904.3642761 [doi: 10.1145/3613904.3642761]
74. Dongre P. Physiology-driven empathic large language models (EmLLMs) for mental health support. In: Proceedings of the 2024 Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. 2024. Presented at: CHI EA '24; May 11-16, 2024:1-5; Honolulu, HI. URL: https://dl.acm.org/doi/10.1145/3613905.3651132 [doi: 10.1145/3613905.3651132]
75. Kumar H, Wang Y, Shi J, Musabirov I, Farb N, Williams J. Exploring the use of large language models for improving the awareness of mindfulness. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. 2023. Presented at: CHI EA '23; April 23-28, 2023:1-7; Hamburg, Germany. URL: https://dl.acm.org/doi/full/10.1145/3544549.3585614 [doi: 10.1145/3544549.3585614]
76. Heston TF. Safety of large language models in addressing depression. Cureus. Dec 2023;15(12):e50729. [FREE Full text] [doi: 10.7759/cureus.50729] [Medline: 38111813]
77. Herencia López-Menchero A. Analysis of the transformer architecture and application on a large language model for mental health counseling. Polytechnic University of Madrid. URL: https://docta.ucm.es/rest/api/core/bitstreams/30effe66-9f5a-404e-9b31-ba3e8d555268/content [accessed 2025-05-29]
78. Bird JJ, Lotfi A. Generative transformer chatbots for mental health support: a study on depression and anxiety. In: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments. 2023. Presented at:

PETRA '23; July 5-7, 2023:475-479; Corfu, Greece. URL: https://dl.acm.org/doi/abs/10.1145/3594806.3596520 [doi: 10.1145/3594806.3596520]
79. De Freitas J, Uğuralp AK, Oğuz‐Uğuralp Z, Puntoni S. Chatbots and mental health: insights into the safety of generative AI. J Consum Psychol. Dec 19, 2023;34(3):481-491. [doi: 10.1002/jcpy.1393]
80. Dergaa I, Fekih-Romdhane F, Hallit S, Loch AA, Glenn JM, Fessi MS, et al. ChatGPT is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry. Jan 4, 2023;14:1277756. [FREE Full text] [doi: 10.3389/fpsyt.2023.1277756] [Medline: 38239905]
81. 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):e357. [FREE Full text] [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
82. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology. Aug 13, 2024;49(9):1412-1416. [doi: 10.1038/s41386-024-01841-2] [Medline: 38480911]
83. Hwang G, Lee DY, Seol S, Jung J, Choi Y, Her ES, et al. Assessing the potential of ChatGPT for psychodynamic formulations in psychiatry: an exploratory study. Psychiatry Res. Jan 2024;331:115655. [doi: 10.1016/j.psychres.2023.115655] [Medline: 38056130]
84. Hsieh LH, Liao WC, Liu EY. Feasibility assessment of using ChatGPT for training case conceptualization skills in psychological counseling. Comput Human Behav. Aug 2024;2(2):100083. [doi: 10.1016/j.chbah.2024.100083]
85. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry. 2023;14:1234397. [FREE Full text] [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
86. Elyoseph Z, Levkovich I. Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: case vignette study. JMIR Ment Health. Mar 18, 2024;11:e53043. [FREE Full text] [doi: 10.2196/53043] [Medline: 38533615]
87. 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 09, 2024;12(Suppl 1):33. [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
88. Adhikary PK, Srivastava A, Kumar S, Singh SM, Manuja P, Gopinath JK, 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. [FREE Full text] [doi: 10.2196/57306] [Medline: 39042893]
89. James LJ, Genga L, Montagne B, Hagenaars M, Van G. Caregiver's evaluation of LLM-generated treatment goals for patients with severe mental illnesses. In: Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments. 2024. Presented at: PETRA '24; June 26-28, 2024:187-190; Crete, Greece. URL: https://dl.acm.org/doi/10.1145/3652037.3663955 [doi: 10.1145/3652037.3663955]
90. Kim T, Bae S, Kim HA, Lee SW, Hong H, Yang C, et al. MindfulDiary: harnessing large language model to support psychiatric patients' journaling. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024. Presented at: CHI '24; May 11-16, 2024:1-20; Honolulu, HI. URL: https://dl.acm.org/doi/10.1145/3613904.3642937
91. Furukawa TA, Iwata S, Horikoshi M, Sakata M, Toyomoto R, Luo Y, et al. Harnessing AI to optimize thought records and facilitate cognitive restructuring in smartphone CBT: an exploratory study. Cogn Ther Res. Jul 07, 2023;47(6):887-893. [doi: 10.1007/s10608-023-10411-7]
92. 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. [FREE Full text] [doi: 10.2196/52500] [Medline: 39078696]
93. Meyer S, Elsweiler D. "You tell me": a dataset of GPT-4-based behaviour change support conversations. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval. 2024. Presented at: CHIIR '24; March 10-14, 2024:411-416; Sheffield, UK. URL: https://dl.acm.org/doi/10.1145/3627508.3638330
94. Maurya RK, Montesinos S, Bogomaz M, DeDiego AC. Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. Couns Psychother Res. Apr 25, 2024;25(1):94-100. [doi: 10.1002/capr.12759]
95. Hedderich MA, Bazarova NN, Zou W, Shim R, Ma X, Yang Q. A piece of theatre: investigating how teachers design LLM chatbots to assist adolescent cyberbullying education. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024. Presented at: CHI '24; May 11-16, 2024:1-17; Honolulu, HI. URL: https://dl.acm.org/doi/10.1145/3613904.3642379 [doi: 10.1145/3613904.3642379]
96. Hu Z, Hou H, Ni S. Grow with your AI buddy: designing an LLMs-based conversational agent for the measurement and cultivation of children's mental resilience. In: Proceedings of the 23rd Annual ACM Interaction Design and Children Conference. 2024. Presented at: IDC '24; June 17-20, 2024:811-817; Delft, Netherlands. URL: https://dl.acm.org/doi/10.1145/3628516.3659399
97. Giorgi S, Isman K, Liu T, Fried Z, Sedoc J, Curtis B. Evaluating generative AI responses to real-world drug-related questions. Psychiatry Res. Sep 2024;339:116058. [FREE Full text] [doi: 10.1016/j.psychres.2024.116058] [Medline: 39059040]
98. Liu Y, Ding X, Peng S, Zhang C. Leveraging ChatGPT to optimize depression intervention through explainable deep learning. Front Psychiatry. Jun 6, 2024;15:1383648. [FREE Full text] [doi: 10.3389/fpsyt.2024.1383648] [Medline: 38903640]

99. Smith A, Hachen S, Schleifer R, Bhugra D, Buadze A, Liebrenz M. Old dog, new tricks? Exploring the potential functionalities of ChatGPT in supporting educational methods in social psychiatry. Int J Soc Psychiatry. Dec 2023;69(8):1882-1889. [doi: 10.1177/00207640231178451] [Medline: 37392000]
100. Wang X, Liu K, Wang C. Knowledge-enhanced pre-training large language model for depression diagnosis and treatment. In: Proceedings of the 9th International Conference on Cloud Computing and Intelligent Systems. 2023. Presented at: CCIS '23; August 12-13, 2023:532-536; Dali, China. URL: https://ieeexplore.ieee.org/document/10263217
101. Miao BY, Chen IY, Williams CY, Davidson J, Garcia-Agundez A, Sun S, et al. The MI-CLAIM-GEN checklist for generative artificial intelligence in health. Nat Med. May 06, 2025;31(5):1394-1398. [FREE Full text] [doi: 10.1038/s41591-024-03470-0] [Medline: 39915678]
102. Santos WM, Secoli SR, Püschel VA. The Joanna Briggs Institute approach for systematic reviews. Rev Lat Am Enfermagem. Nov 14, 2018;26:e3074. [FREE Full text] [doi: 10.1590/1518-8345.2885.3074] [Medline: 30462787]
103. Zhu W, Huang L, Zhou X, Li X, Shi G, Ying J, et al. Could AI ethical anxiety, perceived ethical risks and ethical awareness about AI influence university students' use of generative AI products? An ethical perspective. Int J Hum Comput Interact. Mar 08, 2024;41(1):742-764. [doi: 10.1080/10447318.2024.2323277]
104. Brown A. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. Int J Archit Comput. Sep 09, 2024;22(3):275-276. [doi: 10.1177/14780771241280148]
105. Wright SN, Anticevic A. Generative AI for precision neuroimaging biomarker development in psychiatry. Psychiatry Res. Sep 2024;339:115955. [doi: 10.1016/j.psychres.2024.115955] [Medline: 38909415]
106. Barnhill JW. DSM-5 Clinical Cases. Washington, DC. American Psychiatric Publishing; 2013.
107. Gouniai JM, Smith KD, Leonte KG. Do clergy recognize and respond appropriately to the many themes in obsessive-compulsive disorder?: data from a Pacific Island community. Ment Health Relig Cult. Jan 20, 2022;25(1):33-46. [doi: 10.1080/13674676.2021.2010037]
108. Levi-Belz Y, Gamliel E. The effect of perceived burdensomeness and thwarted belongingness on therapists' assessment of patients' suicide risk. Psychother Res. Jul 09, 2016;26(4):436-445. [doi: 10.1080/10503307.2015.1013161] [Medline: 25751580]
109. Garg M, Saxena C, Krishnan V, Joshi R, Saha S, Mago V. CAMS: an annotated corpus for causal analysis of mental health issues in social media posts. arXiv. Preprint posted online on July 11, 2022. [FREE Full text]
110. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 shared task: depression and PTSD on Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2015. Presented at: CLPsych '15; June 5, 2015:31-39; Denver, CO. URL: https://aclanthology.org/W15-1204.pdf [doi: 10.3115/v1/w15-1204]
111. Chim J, Tsakalidis A, Gkoumas D, Atzil-Slonim D, Ophir Y, Zirikly A, et al. Overview of the clpsych 2024 shared task: leveraging large language models to identify evidence of suicidality risk in online posts. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology. 2024. Presented at: CLPsych '24; March 21, 2024:177-190; St. Julians, Malta. URL: https://aclanthology.org/2024.clpsych-1.15.pdf
112. Gaur M, Alambo A, Sain JP, Kursuncu U, Thirunarayan K, Kavuluru R, et al. Knowledge-aware assessment of severity of suicide risk for early intervention. In: Proceedings of the 2019 International Conference on the World Wide Web. 2019. Presented at: WWW '19; May 13-17, 2019:514-525; San Francisco, CA. URL: https://dl.acm.org/doi/10.1145/3308558.3313698 [doi: 10.1145/3308558.3313698]
113. Pirina I, Çöltekin C. Identifying depression on reddit: the effect of training data. In: Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. 2018. Presented at: EMNLP '18; October 31, 2018:9-12; Brussels, Belgium. URL: https://aclanthology.org/W18-5903.pdf [doi: 10.18653/v1/w18-59]
114. Naseem U, Dunn AG, Kim J, Khushi M. Early identification of depression severity levels on reddit using ordinal classification. In: Proceedings of the 2022 International Conference on World Wide Web. 2022. Presented at: WWW '22; April 25-29, 2022:2563-2572; Virtual Event. URL: https://dl.acm.org/doi/10.1145/3485447.3512128 [doi: 10.1145/3485447.3512128]
115. Turcan E, McKeown K. Dreaddit: a reddit dataset for stress analysis in social media. arXiv. Preprint posted online on October 31, 2019. [FREE Full text] [doi: 10.18653/v1/d19-6213]
116. Garg M, Shahbandegan A, Chadha A, Mago V. An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts. arXiv. Preprint posted online on May 30, 2023. [FREE Full text] [doi: 10.18653/v1/2023.findings-acl.757]
117. Sampath K, Durairaj T. Data set creation and empirical analysis for detecting signs of depression from social media postings. In: Proceedings of the 5th IFIP TC 12 International Conference on Computational Intelligence in Data Science. 2022. Presented at: ICCIDS '22; March 24-26, 2022:136-151; Virtual Event. URL: https://link.springer.com/chapter/10.1007/978-3-031-16364-7_11 [doi: 10.1007/978-3-031-16364-7_11]
118. Haque A, Reddi V, Giallanza T. Deep learning for suicide and depression identification with unsupervised label correction. In: Proceedings of the 30th International Conference on Artificial Neural Networks on Artificial Neural Networks and Machine Learning. 2021. Presented at: ICANN '21; September 14-17, 2021:436-447; Bratislava, Slovakia. URL: https://link.springer.com/chapter/10.1007/978-3-030-86383-8_35 [doi: 10.1007/978-3-030-86383-8_35]

119. Ji S, Li X, Huang Z, Cambria E. Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput Appl. Jun 24, 2021;34(13):10309-10319. [doi: 10.1007/s00521-021-06208-y]
120. Gui T, Zhu L, Zhang Q, Peng M, Zhou X, Ding K, et al. Cooperative multimodal approach to depression detection in Twitter. AAAI Conf Artif Intell. Jul 17, 2019;33(01):110-117. [doi: 10.1609/aaai.v33i01.3301110]
121. Shing HC, Nair S, Zirikly A, Friedenberg M, Daumé IH, Resnik P. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. 2018. Presented at: CLPsych '18; June 5, 2018:25-36; New Orleans, LA. URL: https://aclanthology.org/W18-0603.pdf [doi: 10.18653/v1/w18-0603]
122. Zirikly A, Resnik P, Uzuner O, Hollingshead K. CLPsych 2019 shared task: predicting the degree of suicide risk in reddit posts. In: Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology. 2019. Presented at: CLPsych '19; June 6, 2019:24-33; Minneapolis, MN. URL: https://aclanthology.org/W19-3003.pdf [doi: 10.18653/v1/w19-3003]
123. Wang Y, Wang Z, Li C, Zhang Y, Wang H. A multitask deep learning approach for user depression detection on Sina Weibo. arXiv. Preprint posted online on August 26, 2020. [FREE Full text]
124. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr Res. Jul 2007;93(1-3):304-316. [FREE Full text] [doi: 10.1016/j.schres.2007.03.001] [Medline: 17433866]
125. Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. In: Proceedings of the 9th International Conference on Language Resources and Evaluation. 2014. Presented at: LREC '14; May 26-31, 2014:3123-3128; Reykjavik, Iceland. URL: https://aclanthology.org/L14-1421/
126. Shen Y, Yang H, Lin L. Automatic depression detection: an emotional audio-textual corpus and a Gru/Bilstm-based model. In: Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. 2022. Presented at: ICASSP '22; May 23-27, 2022:6247-6251; Singapore, Singapore. URL: https://ieeexplore.ieee.org/document/9746569 [doi: 10.1109/icassp43922.2022.9746569]
127. DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A, et al. SimSensei kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. 2014. Presented at: AAMAS '14; May 5-9, 2014:1061-1068; Paris, France. URL: https://dl.acm.org/doi/10.5555/2615731.2617415
128. Xu X, Zhang H, Sefidgar Y, Ren Y, Liu X, Seo W, et al. GLOBEM dataset: multi-year datasets for longitudinal human behavior modeling generalization. arXiv. preprint posted online on November 4, 2022. [FREE Full text]
129. Cimtay Y, Ekmekcioglu E, Caglar-Ozhan S. Cross-subject multimodal emotion recognition based on hybrid fusion. IEEE Access. 2020;8:168865-168878. [doi: 10.1109/access.2020.3023871]
130. Cai H, Yuan Z, Gao Y, Sun S, Li N, Tian F, et al. A multi-modal open dataset for mental-disorder analysis. Sci Data. Apr 19, 2022;9(1):178. [FREE Full text] [doi: 10.1038/s41597-022-01211-x] [Medline: 35440583]
131. Chen J, Ro T, Zhu Z. Emotion recognition with audio, video, EEG, and EMG: a dataset and baseline approaches. IEEE Access. 2022;10:13229-13242. [doi: 10.1109/access.2022.3146729]
132. Mauriello ML, Lincoln T, Hon G, Simon D, Jurafsky D, Paredes P. SAD: a stress annotated dataset for recognizing everyday stressors in SMS-like conversational systems. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. Presented at: CHI EA '21; May 8-13, 2021:1-7; Yokohama, Japan. URL: https://dl.acm.org/doi/10.1145/3411763.3451799 [doi: 10.1145/3411763.3451799]
133. Zeng G, Yang W, Ju Z, Yang Y, Wang S, Zhang R, et al. MedDialog: large-scale medical dialogue datasets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020. Presented at: EMNLP '20; November 16-20, 2020:9241-9250; Virtual Event. URL: https://aclanthology.org/2020.emnlp-main.743.pdf [doi: 10.18653/v1/2020.emnlp-main.743]
134. Jamil Z, Inkpen D, Buddhitha P, White K. Monitoring tweets for depression to detect at-risk users. In: Proceedings of the 4th Workshop on Computational Linguistics and Clinical Psychology. 2017. Presented at: CLPsych '17; August 3, 2017:32-40; Vancouver, BC. URL: https://aclanthology.org/W17-3104.pdf
135. Gemini. Google AI. URL: https://www.gemini.google.com [accessed 2025-05-29]
136. Llama 2: open source, free for research and commercial use. Meta AI. URL: https://www.llama.com/llama2/ [accessed 2025-05-29]
137. Cesare N, Grant C, Nsoesie E. Understanding demographic bias and representation in social media health data. In: Proceedings of the 10th ACM Conference on Web Science. 2019. Presented at: WebSci '19; June 30-July 3, 2019:7-9; Boston, MA. URL: https://dl.acm.org/doi/10.1145/3328413.3328415 [doi: 10.1145/3328413.3328415]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
EEG: electroencephalogram
GenAI: generative artificial intelligence

LGBTQ: lesbian, gay, bisexual, transgender, and queer
LLaMA: large language model Meta AI
LLM: large language model
MI-CLAIM-GEN: Minimum Information about Clinical Artificial Intelligence for Generative Modeling Research
PICOS: Population, Intervention, Comparison, Outcome, and Study
RAG: retrieval-augmented generation
SPIDER: Sample, Phenomenon of Interest, Design, Evaluation, and Research Type
SVM: support vector machines

Edited by C Blease; submitted 27.12.24; peer-reviewed by B Lamichhane, S Markham, S Tayebi Arasteh, G Huang; comments to author 18.02.25; revised version received 14.04.25; accepted 29.05.25; published 27.06.25

Please cite as:
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
URL: https://mental.jmir.org/2025/1/e70610
doi: 10.2196/70610
PMID: