npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01611-4

A scoping review of large language models for generative tasks in mental health care

Check for updates

Yining Hua[1,2], Hongbin Na[3], Zehan Li[4], Fenglin Liu[5], Xiao Fang[6], David Clifton[5,7] & John Torous[2,8] ✉

Large language models (LLMs) show promise in mental health care for handling human-like conversations, but their effectiveness remains uncertain. This scoping review synthesizes existing research on LLM applications in mental health care, reviews model performance and clinical effectiveness, identifies gaps in current evaluation methods following a structured evaluation framework, and provides recommendations for future development. A systematic search identified 726 unique articles, of which 16 met the inclusion criteria. These studies, encompassing applications such as clinical assistance, counseling, therapy, and emotional support, show initial promises. However, the evaluation methods were often non-standardized, with most studies relying on ad-hoc scales that limit comparability and robustness. A reliance on prompt-tuning proprietary models, such as OpenAI's GPT series, also raises concerns about transparency and reproducibility. As current evidence does not fully support their use as standalone interventions, more rigorous development and evaluation guidelines are needed for safe, effective clinical integration.

Mental health issues have been a concern of global health ever since they recognized the profound impact on individuals and societies, and the urgency has only grown in recent years. Nearly 1% of all global deaths annually are now due to suicide, with approximately 800,000 people dying by suicide each year[1]. In the United States alone, the annual public mental health expenditure exceeded $16.1 billion, including a $2.21 billion budget for the National Institute of Mental Health (NIMH) and $13.9 billion on mental healthcare[2]. Still, even in the United States, the psychiatry workforce is projected to face a pressing shortage through 2024, with a potential shortfall of 14,280 to 31,091 psychiatrists[3,4]. And in low-and-middle income countries, the situation is even worse, with up to 85% of people there still receiving no treatment for their mental health[5].

In response to the growing mental health crisis and the projected shortage of mental health professionals, artificial intelligence (AI)-driven mental health applications like chatbots are emerging as vital tools to bridge the treatment gap. These technologies offer scalable, accessible, and cost-effective support, particularly in areas where traditional mental health services, including psychiatric care, are insufficient or unavailable. As of 2023, the global market for mental health apps has grown rapidly, with over 10,000 apps collectively serving millions of users[6]. AI-driven platforms are increasingly incorporating psychiatric assessments, medication management reminders, and monitoring tools that assist in the management of conditions such as depression, anxiety, and bipolar disorder. Studies suggest these tools can help reduce symptoms and improve patient outcomes, making them a promising avenue for addressing mental health challenges, especially in regions with limited access to psychiatric professionals, and they are increasingly being integrated into broader mental health care strategies to help meet the growing demand[7,8].

The introduction of large language models (LLMs) like OpenAI's ChatGPT[9], Google's Bard[10], and Anthropic's Claude[11] marks a transformative advancement in AI-driven mental health care, offering capabilities far beyond those of earlier AI tools. Unlike previous models, which were limited to scripted interactions and specific tasks, LLMs can engage in dynamic, context-aware conversations that feel more natural and personalized via generating human-like conversations. This allows them to provide tailored emotional support, detect subtle cues indicating changes in mental health, and adjust their guidance to meet individual user needs in generative tasks. Increasingly, research is exploring anthropomorphic features such as empathy, politeness, and other human-like traits in these models to enhance their effectiveness in delivering more realistic and supportive mental health care[12].

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [2]Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, USA. [3]Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, NSW, Australia. [4]McWilliams School of Biomedical Informatics, UTHealth Houston, Houston, TX, USA. [5]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK. [6]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Oxford-Suzhou Centre for Advanced Research, Suzhou, China. [8] Department of Psychiatry, Harvard Medical School, Boston, MA, USA. ✉e-mail: jtorous@bidmc.harvard.edu

Despite the promising potential, these tools are still in the early stages of development and evaluation. Users often do not understand the models they are interacting with, including the limitations and biases inherent in the AI's design. Unfortunately, there is currently no standardized framework for evaluating the effectiveness and safety of these models in mental health applications. Many studies, including those focused on evaluating LLMs, often develop their own metrics and methods, leading to inconsistent and sometimes unreliable results. The lack of standardized evaluation hinders the comparison of models or assess their true impact on mental health outcomes. Concerns about data privacy, the potential for misuse, and the ethical implications of relying on AI for sensitive mental health care decisions further underscore the need for rigorous oversight. Considering these promises and challenges, a scoping review of the current applications of LLMs in mental health care is essential from the perspective of psychiatrists and clinical informaticians. Our review aims to synthesize existing research with a focus on clinical relevance, identify gaps in understanding from a mental health practice standpoint, and provide clear guidelines for future development and evaluation of these technologies in real-world settings.

Background

Subfields of Mental health care and the potential of generative AI

The potential of generative AI in mental health care is broad given the many different treatment approaches employed today for care delivery. These approaches generally fall into three main categories: psychotherapy, psychiatry, and general mental health support. Psychotherapy is one of the most common forms of mental health care. However, access to psychotherapy is often limited by factors like a shortage of therapists, long wait times, and high costs. Generative AI could help address these issues by offering on-demand support, providing education about mental health, and guiding people through therapeutic exercises when they can't see a therapist in person. Psychiatry focuses on the medical side of mental health care, including diagnosing, treating, and preventing mental disorders. But like psychotherapy, psychiatry also faces challenges, particularly a shortage of psychiatrists. Generative AI could support psychiatrists by helping monitor patients' symptoms, reminding them to take their medication, and providing initial assessments, which could reduce the strain on the healthcare system and improve patient outcomes. General mental health support includes a wide range of services designed to promote mental well-being and prevent mental health problems. This might include community programs, self-help resources, peer support networks, and public health initiatives. These services are important for early intervention, managing stress, and preventing more serious mental health issues from developing. However, many people don't take advantage of these resources, often because of stigma, lack of awareness, or insufficient availability. Generative AI could help make these resources more accessible by providing anonymous, personalized support through chatbots and apps that offer mental health education, coping strategies, and encouragement to seek help in a way that feels safe and non-judgmental.

Large language models (LLMs)

Although LLMs gained widespread attention with the release of OpenAI's ChatGPT-4, the concept has existed for some time, though there is no single unified definition. In the natural language processing (NLP) community, LLMs are generally understood as large generative AI models capable of producing text by predicting the next word or phrase based on vast amounts of training data. NLP has evolved drastically over time, with early models being task-specific and limited in their ability to understand context and nuance. The introduction of advanced deep learning frameworks marked a major improvement, as these models are designed to better capture contextual language meaning. However, they still struggled with generating coherent, contextually appropriate text over longer conversations, which is crucial for mental health applications. LLMs have advanced this further by leveraging large datasets and transformer architectures to predict and generate highly coherent and context-aware text. This enables them to mimic human conversation, making them valuable for creating therapeutic content, offering psychoeducation, and simulating therapy sessions—important tools for expanding access to mental health care. For clinicians, LLMs offer promising tools to support mental health services by providing personalized, scalable interactions. However, it's important to recognize that most current LLMs are general models and do not perform as well as specialized pre-trained models for domain-specific tasks such as prediction and classification. For example, Bidirectional Encoder Representations from Transformers (BERT) models, which model word segments (tokens) using both the segments before and after them, are more accurate and efficient for these purposes. As a result, pretraining and fine-tuning becomes a crucial step as it provides the model with contextual knowledge and linguistic patterns specific to the mental health applications. This finetuning and pretraining process can incorporate emotional cues and expert-written examples to enhance the model's interpretability and responsiveness to improve the performance of LLMs in specific generative tasks.

Results

Mental disorders, conditions, and subconstructs

Mental disorders referenced in the included studies vary widely in terms of definitions, measurement instruments, and the use of standards. While some studies focus on clinically confirmed diagnoses, relying on established criteria like those found in the DSM-5[13], others take a less structured approach. In such cases, mental health constructs are often defined arbitrarily using user-expressed keywords or affects rather than expert knowledge or validated measures. This is especially common in studies conducted outside the medical or clinical domain, where mental health constructs may be interpreted more loosely or tailored to the context of the AI models. Such inconsistencies in the use and understanding of validated measures highlight a potential gap when applying AI models to various targeted mental health constructs—including affect, symptoms, diagnosis, and treatment—reflecting a broader issue in this interdisciplinary field. Therefore, we categorized the targeted mental health disorders, conditions, and subconstructs into two groups: 1) those measured or defined with validated approaches, relying on standard diagnostic criteria and validated clinical knowledge; and 2) those assessed with non-validated measures, lacking a clear definition, standard, or validated method for assessment or diagnosis.

As shown in Table 1, eight studies out of the sixteen reviewed included validated measures for mental health constructs[14–21], while nine relied only on ad-hoc (less well established) approaches[16,17,21–27], and three studies included constructs with a mix of both types of measurements[16,17,21]. Across both groups, depression[14,16–19,21,24–26] was the most frequently studied mental health construct. The Patient Health Questionnaire-9 (PHQ-9)[16,19] and the Center for Epidemiologic Studies Depression Scale for Children (CES-DC)[16] were adopted as inclusion criteria and outcome measures[16,19], while another study used PHQ-9 as an exclusion criterion[21]. Other clinically valid constructs include anxiety[14,16,18], positive and negative affects (PANAS)[14], Attention-Deficit/Hyperactivity Disorder (ADHD)[15,20], bipolar disorder[19], loneliness[17], and stress[16].

One study evaluated GPT's performance on 100 clinical case vignettes of different disorders, comparing GPT against psychiatrists across different clinical constructs, covering a wide range of disorders[28]. However, not all studies referenced clinical mental health constructs provided specific criteria. For example, one study diagnosed study subjects with clinical interviews "using screening instruments over different disorders"[15]. Other studies also incorporated expert judgments from mental health providers without mentioning the specific process or referring to well-established criteria used in the study[22,23,28]. Depression[17,24–26] and suicidality[16,17,22,23,27] have also been frequently studied with less well-established and customized constructs. For instance, one study associated the construct of depression with self-identified feelings of depression[17] or simply with the word "sad"[23]. Another study filtered social media posts related to suicidal ideation and self-harm using regular expressions (e.g., "*.(commit suicide).*", "*.(cut).*")[29]. More specific subconstructs of mental health care include psychological challenges due to social emotions[28,29], cognitive distortion and negative

Table 1 | Mental disorders, conditions, and subconstructs in generative applications of LLMs for mental health care

Group	Condition/Construct	Criteria/Content	References
With validated measures	Affects	The Positive Affect and Negative Affect Scale (PANAS)[46]	14
	Attention-deficit/hyperactivity disorder (ADHD)	Clinical interview; Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)[30]	15,20
	Anxiety	Generalized Anxiety Disorder 7 (GAD-7)[47]	7,26
	Bipolar depression	Expert clinician validated vignettes	19
	Major depressive disorder (MDD)	Patient Health Questionnaire-9 (PHQ-9)[48], Center for Epidemiologic Studies Depression Scale for Children (CES-DC)[47]	14,16,18,19,21
	Life satisfaction	The Satisfaction with Life Scale (SWLS)[49]	14
	Loneliness	Interpersonal Support Evaluation List (ISEL)[50], the De Jong Gierveld Loneliness Scale[51]	17
	Stress	Coping Strategies Scale[52]	16
	Psychological well-being	The Scales of Psychological Well-being (SPWB)[53], the Subjective Vitality Scale (SVS)[54]	14
With non-validated measures	Abuse	Users expressed keywords	17
	Thinking trap/Cognitive distortion	Self-identified exaggerated thoughts, thinking in extremes, jumping to conclusions based on one experience	22,23
	Depression	Self-identified feelings of depression; "I am sad and have a history of depression. How can I be happier?"	17,24–26
	Negative thoughts	"What emotion does this thought make you feel? And how strong 1–10"	22,23
	Social emotions (personality, mood, and attitudes)	Neutral, happy, sad, relaxed, and angry	21,26,27
	Suicidality or self-harm	Keywords defined by regular expressions. E.g., ".*(commit suicide).*", ".*(cut).*"; "feeling suicidal", "want to die", and "harm myself "; Custom open-ended question	16,17,22,23,27

Table 2 | Overview of input/output modalities, models, and target users in generative applications of LLMs in mental health care

Application category	Input modality	Model	Output modality	Embodiment	Open source	Language	Target user	References
Clinical Assistant	Written	ChatGPT[a]	Written	No	No	English	Healthcare Providers	14
	Written	PanGu	Written	No	No	Chinese	Healthcare Providers	26
	Written	GPT4-Turbo	Written	No	No	English	Healthcare Providers	19
Counseling	Written	GPT-4	Written	No	No	English	General Public	24
	Spoken	GPT-3	Spoken,Visual	Yes	Yes	Spanish	General Public	21
	Written	GPT-4	Written	No	No	English	General Public	34
Therapy	Written, Spoken, Visual	Customized GPTs	Written, Spoken	Yes	No	English/ Spanish	Patients	35
	Spoken	GPT-4	Spoken, Visual	Yes	No	English	Patients	18
	Written, Spoken, Visual	GPT4-Turbo Claude-3	Written, Spoken, Visual	Yes	No	Multilingual[b]	Patients	20
	Written	GPT-4	Written	No	No	Korean	Patients	16
	Written	GPT-3.5-Turbo	Written	No	No	English	General Public	32
	Written, Spoken, Visual	Not specified	Written, Spoken, Visual	Yes	No	English/ Japanese	General Public	17
Positive Psychology Intervention	Written	GPT-3.5-Turbo	Written	No	No	Chinese	Patients	14
	Written	GPT-3	Written	No	Yes	English	General Public	22
	Written	GPT-3/T5/ DialoGPT	Written	No	Yes	English	Patients	23
Education	Written	GPT-3	Written	No	No	Spanish	General Public	15

[a] Version not specified.

[b] Specific languages not specified.

thoughts[19,20], and abuse[19]. These studies used less well-established and more arbitrary standards for definitions and assessment. (Tables 2, 3).

Cognitive Behavior Therapy (CBT)[30] is the most referenced treatment method for anxiety, cognitive distortion, depression, and loneliness[15–18,22]. It is an evidence-based, well-established psychological treatment. Elements and techniques from CBT[30], such as cognitive restructuring[22,23] and mindfulness[18], have been incorporated into LLM models to provide digital self-guided interventions. Other evidence-based treatment approaches include occupational therapy[20], which is used to support children with ADHD, and peer support[27], where the chat agent simulates individuals with similar experiences to provide empathetic emotional support.

Table 3 | Summary of unified evaluation constructs

Step	Higher-order construct	Lower-order construct	Definition	Examples	Article references
1	Safety, Privacy, and Fairness	Safety	Prevent worse outcomes for the patient, provider, or health system from occurring as a result of the use of an ML algorithm.	Outcome proxy appropriateness, Data provenance, Harm control, Reducing automation bias, Critical help, Ethics, etc.	20,34
	Safety, Privacy, and Fairness	Privacy	Protect privacy according to standards like HIPAA and GDPR, ensuring user autonomy and dignity.	Data exchange, Data collection and storage, Data usage, Privacy Policy, Data protection, etc.	35
	Safety, Privacy, and Fairness	Fairness and bias management	Ensure the chatbot operate with minimized and acknowledged biases to ensure fair outcomes.	Systemic Bias, Computational and Statistical Bias, Human-cognitive biases, Population bias, etc.	20
2	Trustworthiness and Usefulness	Beneficence	Ensure the chatbot positively impacts its intended outcomes, emphasizing measurable benefits over potential risks	Health Outcomes, Clinical Evidence, User Behaviors, Intervention, Healthcare System, etc.	14–16,18,20–22,35
	Trustworthiness and Usefulness	Generalizability	Apply learned patterns to new, unseen data.	Contextual Adaptability, Novel Data Performance, etc.	20,34
	Trustworthiness and Usefulness	Reliability	Ensure that the chatbot consistently performs as intended under various conditions and maintains dependable operation over time.	Failure Prevention, Robustness, Workflow Integration, Reproducibility, Monitoring, Up-to-dateness, etc.	19,48
	Trustworthiness and Usefulness	Validity	Ensure the chatbot performs as expected in real-world conditions	Data Relevance and Credibility, Language Understanding, Information Retrieval Accuracy, Outcome Accuracy, Task Completion, etc.	20,21,26,34
3	Design and Operational Effectiveness	Accessibility	Ensure those involved in the chatbot's lifecycle uphold standards of auditability and harm minimization.	Versatile access, User literacy required, User experience, User Interface Design, Simplicity/Ease of Use, etc.	15,16,18,20,21,26,28,32,35
	Design and Operational Effectiveness	Personalized Engagement	Tailor responses based on patient data and preferences.	Personalization, Anthropomorphism/relationship, User Adherence, Feedback Incorporation, Progress awareness, etc.	18,20,23,31–35
	Design and Operational Effectiveness	Cost-Effectiveness	Assess whether the chatbot delivers beneficial outcomes at a reasonable cost, providing a better or more economical solution compared to existing methods.	Comparative Effectiveness, Economical Viability, Environmental Viability, Task Efficiency, Workflow Considerations, etc.	20,26,34

Table 3 summarizes the mapped primary and second-level constructs across the reviewed studies. We have also included examples of sub-constructs for each mapped second-level construct for the readers to understand the mapped constructs. Further details of evaluation subjects, evaluation methods, sample sizes, scale names, original constructs, mapped second-level constructs, and levels associated with each article can be found in Supplementary Table 3. Practical evaluation questions related to each construct can be found in the original article.

Constructs have been mapped to the second level to avoid excessive scarcity and granularity.

Applications and model information

Existing generative applications of LLMs in mental health care can be categorized into six main types based on model functionalities: Clinical Assistant, Counselling[17,29], Therapy[17,23], Emotional Support[16,17,31,32] Positive Psychology Intervention[14,22,23], and Education[15,33]. Among them, the Clinical Assistant application includes attempts to develop and evaluate LLMs for supporting mental health professionals by generating management strategies and diagnoses for psychiatric conditions. In the Counselling category, LLMs are used to interact with participants, such as engaging Spanish teenagers in discussions about mental health disorders[15] and providing relationship advice in single-session interventions[34]. Emotional Support applications have focused on offering empathetic responses and support in various contexts, such as mitigating loneliness and suicide risk among students[17]. In the Therapy category, LLMs are integrated into treatments for conditions like ADHD, enhancing care through simulated therapy scenarios[35] and immersive therapy experiences using virtual reality(VR)[18]. Positive Psychology Interventions involve using LLMs to personalize recommendations and facilitate cognitive restructuring, thereby reducing negative thoughts and emotional intensity[14,22]. Finally, in Education, LLMs have been employed to train medical students in communication skills, providing a realistic and positive simulated patient experience[33], as well as promoting awareness of mental health among young people[15]. Most of these studies only support text-based input/output modalities[14,15,19,22–24,26,27,31,32,34]. A subset of systems[17,18,20,35] supports multimodal input/output, incorporating speech, images, or video for a richer user experience. Some applications incorporate physical embodiment through VR[17,18] or robotics[20,35]. These applications are seen across various target user groups, including healthcare providers[19,26], patients[14,16,18,20,22–24,31,32,34], and the general public[15,32,33].

OpenAI's GPT series models are the most studied, see in 14 studies[14,18–20,22–24,28,31,32,34,35], with 11 using the latest advanced models like GPT-3.5, ChatGPT, GPT-4, and customized GPTs, while four studies used the earlier GPT-3 model. Other LLMs used[23,26] include Huawei's PanGu[26], T5[20], and DialoGPT[36] are open-source. Some studies did not specify the platforms they employed, while many studies used digital platforms such as websites and mobile phones. Some studies developed agents with physical embodiments[22], and some others[21,35] used Raspberry PI, a type of single-board computer (Supplementary Table 2). Among those that used OpenAI's models, three were based on OpenAI's web interface[24,28,34], one did not directly state their platform but appeared to use the API based on the structure of their methods[19], and only eight (57.1%) explicitly referenced API use or temperature parameters[14,18,20,28,32]. Language support by these models varied, covering more than English, with three applications supported by multiple languages[17,20,35], and 14 applications supporting a single language—seven in English[18,19,22–24,32,34], three in Chinese[14,26,29], two in Korean[16,31], and two in Spanish[15,33].

Task performance and clinical effectiveness

The study designs and evaluations of existing research are highly heterogeneous and often inconsistent, making it challenging to accurately assess their task performance and clinical effectiveness. Thus, we provide a high-level summary of the findings here. We offer a detailed summary of each study's task, performance/results, sample size, clinical validation method, and participant demographics can be found in Supplementary Table 3.

Several studies have explored the use of LLMs for clinical decision support in psychiatry. In one study, ChatGPT-3.5 was evaluated using 100 clinical case vignettes covering diverse psychiatric conditions[28]. The model achieved a "Grade A" rating in 61% of cases, "Grade B" in 31%, and "Grade C" in 8%, indicating different levels of diagnostic accuracy in simulated scenarios. However, this study did not involve real patients, and no clinical validation was performed. Similarly, another study assessed GPT-4's performance in clinical decision-making for bipolar depression cases. GPT-4 selected optimal treatments in 50.8% of cases, slightly outperforming community clinicians[19]. Although promising, these results are based on hypothetical cases, and the model's effectiveness in actual clinical practice remains unverified. Overall, while LLMs demonstrate potential in generating clinically relevant information, the lack of clinical validation and reliance on simulated vignettes limit the evidence for their effectiveness in real-world diagnostic support.

Several studies have investigated the application of LLMs in aspects of therapeutic interventions, particularly in cognitive restructuring and positive psychology. Liu et al.[14] conducted randomized controlled trials with 326 participants to test GPT-based chatbots delivering Positive Psychology Interventions (PPIs). The chatbot provided personalized recommendations and engaged users in multi-round dialogues with resulting improvements in mental well-being, reductions in anxiety, and increased life satisfaction metrics. This suggests that LLMs can effectively facilitate interventions aimed at enhancing psychological well-being. Another study explored the use of LLMs in self-reflective journaling among 28 psychiatric outpatients diagnosed with Major Depressive Disorder[22]. Clinicians reported that the LLM-assisted journaling system enriched patient records and provided better insights into patients' conditions. In a large-scale randomized controlled trial[34], involving over 15,000 participants, Sharma et al. evaluated an LLM's assistance in cognitive restructuring for self-guided mental health interventions. The study found that 67% of participants reported reduced emotional intensity, and 65% overcame negative thoughts after interacting with the LLM. These results indicate the potential scalability and effectiveness of LLMs in supporting cognitive-behavioral techniques.

LLMs have also been used to provide emotional support and enhance engagement, particularly among youth and marginalized populations, Mármol-Romero et al. examined a GPT-based chatbot's engagement with Spanish-speaking teenagers on mental health topics[15]. The observational study involved 102 students, and the chatbot facilitated open discussions on anxiety and depression. The engagement led to meaningful conversations with 44 participants, indicating potential for early outreach and mental health education among adolescents. Another study investigated the use of the Replika chatbot among 1006 students[17]. The study found that 3% of participants reported cessation of suicidal ideation after interacting with the chatbot, and 75% reported feeling less lonely, suggesting that LLM chatbots can provide immediate emotional support.t. However, the lack of long-term outcomes from all studies is notable.

Evaluation methods, scales, and constructs

A standardized and well-established set of constructs and scales is essential in systematically measuring mental health interventions, particularly when evaluating new technologies. Constructs refer to specific concepts or characteristics being measured, such as privacy, safety, or user experience. They provide a clear focus on what is being assessed in a study, which is crucial for ensuring that the evaluation is meaningful and relevant. Scales, in turn, offer a structured and standardized approach to quantify these constructs. This standardization is necessary for consistency across different studies, allowing researchers to compare results and draw more robust conclusions.

Given the diversity in how constructs are defined and measured across studies, it is important to use a framework that can harmonize these variations. While there are many approaches, we used a hierarchical framework[37] inspired by the American Psychiatric Association app evaluation model. A 2024 review of evaluation models[38] noted this framework "is straightforward, comprehensive, flexible, and relevant to diverse contexts" and so also provides us a promising starting point. This framework categorizes constructs into three levels: (1) Safety, Privacy, and Fairness; (2) Trustworthiness and Usefulness; and (3) Design and Operational Effectiveness. The pyramid framework ensures that each level of evaluation builds on the previous one. For example, without ensuring that an intervention is safe, it would be premature to evaluate its usability or cost-effectiveness.

Among the studies reviewed, those that involved direct participant feedback ($n = 5$)[14,17,18,22,23] generally focused on user-centric constructs. These studies typically involved larger sample sizes ranging from 28 to over 15,000 participants and assessed constructs such as accessibility, ease of use, personalized engagement, user experience, and cost-effectiveness. They provide direct insights into how user experience of LLMs is in real-world

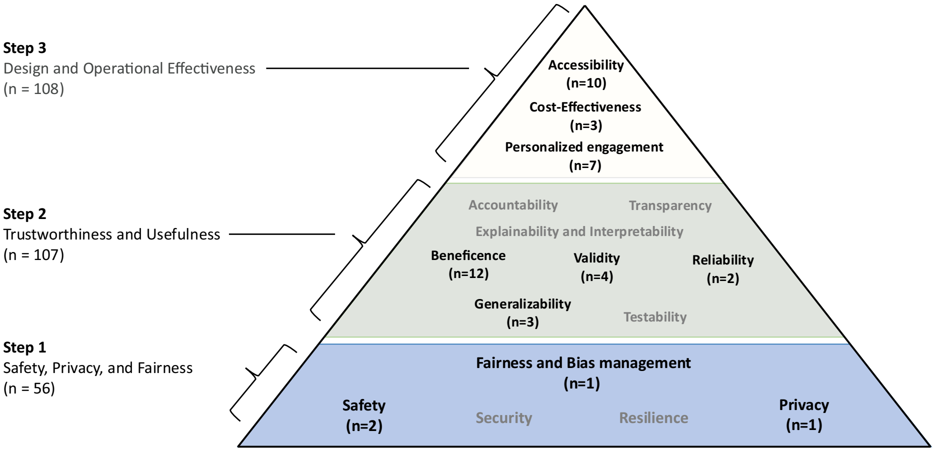

Fig. 1 | Pyramid framework of evaluation constructs in generative applications of LLMs in mental health care. Constructs in gray represent constructs with no associated articles. "*N*" represents the number of unique articles that assessed each construct. Gray text indicates constructs that were not assessed in any study. Foundational areas like "Safety, Privacy, and Fairness" are rarely evaluated, highlighting key gaps in critical aspects such as "Accountability," "Transparency," and "Security".

settings. On the other hand, studies that focused on evaluating LLM performance—typically involving expert assessments—concentrated more on foundational and core efficacy constructs. These studies often used smaller sample sizes, ranging from 12 to 100 cases, focusing on technical or functional aspects of the LLMs. Additionally, one study[23] designed and incorporated automated metrics for Rationality, Positivity, and Empathy, using NLP models to evaluate LLM outputs. These automated evaluations offer a more detailed, algorithmic perspective on the LLM's performance, complementing human judgments.

The heterogeneous use of scales remains a problem in the mental health field. We observe that 12 studies developed their own scales[15,18–21,23,26,28,32,34,35] or adapted existing ones for their evaluations. Most of the studies using validated scales were those directly measuring patient outcomes, such as anxiety, where the General Anxiety Disorder-7 (GAD-7) was employed[14,32]. However, many articles that created their own scales without clear rationale, and often lacked references to support their methods, raising challenges with the validity and reliability of their methods.

Figure 1 presents a pyramid shaped schematic of the current status of evaluated constructs in the generative applications of LLMs for mental health care, based on the health AI-chatbot evaluation framework[37]. The figure includes the number of articles counted for each level 2 construct, with gray texts indicating constructs never evaluated by existing research. The foundational levels are less frequently assessed: only three studies evaluated the fundamental construct "Safety, Privacy, and Fairness"; Thirteen studies assessed the second-level construct "Trustworthiness and Usefulness"; and another 11 articles evaluated the third-level construct "Design and Operational Effectiveness." Although "Trustworthiness and Usefulness" is the most evaluated category, more than half of its subconstructs remain unassessed. Across the framework, constructs such as "Accountability," "Transparency," "Explainability and Interpretability," "Testability," "Security," and "Resilience" have never been evaluated.

Discussion

Our review suggests that there is great enthusiasm for LLM-based mental health interventions and that many teams are creating interesting and unique applications. We found these chatbots already developed to serve as clinical assistants, counselors, emotional support vehicles, and positive psychology interventions. However, despite the enthusiasm for applying LLMs in mental health care, the current evidence regarding their task performance and clinical effectiveness is limited and varies across studies. Many studies lack rigorous clinical validation, standardized outcome measures, and adequate sample sizes, which hampers the ability to draw definitive conclusions. Furthermore, the inconsistent use and understanding of well established measurement methods across studies complicate the evaluation of these interventions. We observed that mental health constructs were often referenced without accompanying well established instruments and measurements, and in some cases, researchers tailored the definition or assessment to fit their specific AI models, leading to challenges in consistent categorization. This inconsistency underscores a broader issue within the interdisciplinary field of AI and mental health—the variation in how constructs like affect, mood, diagnosis, and treatment are applied complicates efforts to maintain clear distinctions between mental health constructs with and without validated measurements.

The evaluation of LLM-based mental health interventions is hindered by the lack of unified guidelines for scale development and reporting. While this is appropriate for feasibility testing, it belies the ability to understand the actual clinical potential of these new chatbots. With the majority of studies using non-well-established ad-hoc scales without addressing their validity and reliability, there is an opportunity for the next wave of research to better support the credibility and the need for guidelines to standardize reporting and scales used in this field. While effective evaluation is still nascent, results, as shown in the table, highlight that the current focus ignores foundational privacy and safety concerns. LLM-based mental health chatbots are multifaceted with privacy, technical, engagement, legal, and clinical considerations. Our team recently introduced a simplified framework to unify these many evaluations, suggesting that safety and privacy should be the foundation of any evaluation[37]. This is not to minimize the value of evaluation of design and effectiveness (level 3) and usefulness and trustworthiness (level 2), but rather that such should not be at the expense of priority over safety, privacy, and fairness (level 1). Without these level 1 considerations, LLM-based mental health interventions may be impressive but unfit for healthcare or clinical use.

Our results also show that the focus of current LLMs today is directed more at patients and less at clinicians. This approach is logical as direct to consumer/patient approaches often avoid complex healthcare regulations and clinical workflow barriers. However, this approach also risks fragmenting the potential of LLM-based mental health interventions to influence care as there is strong evidence that clinician engagement is required for more sustained and impactful patient use with any digital technology[10]. There is strong data that clinicians are interested in using LLMs in care, but first require and are asking for more training and support on how to use these in care[39].

The LLMs reviewed in this paper target a wide variety of disorders. Over half of the studies reviewed included clinically valid disorders, with other studies targeting general mental health constructs. However, we found that many studies did not offer sufficient details on the target population, and the difference between mental health risk factors versus mental health conditions was poorly delineated. We acknowledge that psychiatric nosology is challenging, as highlighted in recent literature[40], but this challenge highlights how the evaluation of AI systems in mental health may quickly reach an impasse. For example, constructs like depression were often mentioned in a broad and non-specific manner, without reference to diagnostic criteria or standardized and well-established metrics such as the PHQ-9 or GAD-7. This was particularly pronounced in studies conducted by researchers outside the medical or clinical domains. Such inconsistent use of constructs and measurement methods complicates efforts to maintain a clear distinction between mental health constructs with and without validated measures, calling attention to a broader issue within the interdisciplinary field of AI and mental health. For example, one study specified a population of children and adolescents, ages between 12 and 18 years old[15], but overall, most studies lacked detailed demographic information. Given that only one study emphasized data security, with conversations proceeding through a HIPAA-compliant environment[18], the lack of more clinical use cases is perhaps appropriate. Another issue is the dependence on proprietary models, such as OpenAI's GPT-3.5 and GPT-4, in many mental health applications. This reliance raises concerns about transparency and customization, as the use of closed-source models limits external validation of reliability and safety, crucial in mental health research. To improve measurement specificity for specific populations or disorders, model pretraining and fine-tuning are key aspects to be considered[41]. More models and studies should include domain and audience-specific models pretrained on clinical data with more rigorous applications of standardized diagnostic tools. Promoting the use of open-source models and improving transparency can enhance the scientific and ethical standards of these applications.

To advance the scalability and scientific rigor of LLM-based mental health interventions, the research community must also adopt more controlled methodologies. Some studies, particularly those utilizing ChatGPT, rely on the website interface for research purposes. While this approach is convenient, it should be discouraged by rigorous scientific investigations. Research should be conducted using the API, where hyperparameters such as the "temperature" can be controlled, ensuring replicability of the results. The website interface should primarily be used for testing third-level constructs such as Design and Operational Effectiveness and potentially assessing the safety and transparency of the user-facing system. However, researchers must also address factors like backend model updates and stochastic elements in the sampling process to ensure consistent reproducibility and reliability.

Finally, the global applicability of LLM-based mental health tools warrants careful consideration. Public health, especially mental health care, is a global issue, and it's crucial to develop and deploy mental health chatbots in countries and regions where resources are limited and where stigma may be higher. These areas often do not primarily speak English. It's encouraging that 10 out of the 17 studies (58.8%) support non-English languages, either in a single other language or as multilingual chatbots, which is a positive step toward language equity and global health. But this also raises an issue, beyond the scope of this paper, of whether these chatbots offer the same level of correctness, consistency, and verifiability as English-trained chatbots, given that research suggests this is often not the case[42].

Future directions for LLMs in mental health care should prioritize expanding their applications beyond narrow prediction tasks, especially given that only 17 studies over the past five years have explored generative tasks prospectively involving human participants for evaluation. Human-centered studies provide critical insights into how LLMs interact with individuals, particularly in sensitive contexts like mental health care, where nuances in communication and emotional understanding are vital. Addressing current limitations such as small sample sizes and lack of diverse participant demographics, future research should employ larger, more representative samples to enhance the generalizability of findings. To improve the rigor and credibility of LLM-based mental health interventions, studies should prioritize the development of standardized evaluation guidelines. These guidelines should include the creation of validated and reliable scales that can be universally applied across studies, ensuring consistent and accurate assessments of clinical potential. By standardizing evaluation metrics, researchers can overcome the variability that currently impedes comparability and synthesis of results across different studies. To enhance transparency and overcome the limitations of proprietary models, researchers should move away from using web interfaces like ChatGPT for rigorous scientific studies, as these platforms lack the necessary controls for reproducibility. Instead, APIs and locally deployable models that allow for control over hyperparameters should be used to ensure the replicability of the results. This approach will mitigate concerns about reproducibility and allow for more precise manipulation of model parameters, leading to more reliable outcomes. Finally, studies focused on critical constructs such as beneficence, validity, and reproducibility should adopt rigorous evaluation methods and well-established scales, moving beyond metrics like recall and F1 scores, to establish a more comprehensive understanding of model accuracy and clinical relevance. Incorporating ethical considerations and addressing privacy and safety concerns in study designs will also enhance the trustworthiness of LLM applications in mental health care. Equally important is the advancement of novel methodologies and rigorous standards to ensure fairness. A recent study has demonstrated strategies to mitigate biases and promote equity in LLM applications, including assessing demographic disparities in empathy, the implementing demographic-aware prompting, and evaluating subgroup performance in mental health contexts. Future studies should explore new fairness metrics tailored specifically to mental health contexts, such as cultural adaptability or intersectional biases[43].

We would like to acknowledge the limitations of the evidence in this review, which are primarily rooted in the absence of standardized evaluation criteria across studies, resulting in challenges for comparison and synthesis of findings. Many studies depend on non-well-established, ad-hoc scales without thorough clinical validation, which undermines the robustness and generalizability of their conclusions. Furthermore, the frequent use of proprietary LLMs, such as OpenAI's GPT series, introduces issues of transparency and reproducibility, as closed-source settings hinder independent verification and limit replicability. The review processes used also have limitations, as inconsistent reporting practices lead to gaps in essential metrics, demographic detail, and evaluation frameworks, all of which are critical for cross-study analysis. Collectively, these factors highlight an urgent need for a unified, rigorous framework to assess and validate LLM applications in mental health systematically. Addressing these gaps through standardization will be essential for improving the reliability of findings and ensuring that LLMs contribute meaningfully and safely to mental health care.

Methods

We adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines[44] to ensure a transparent and reproducible search process (Fig. 2). Our search included four databases: APA PsycNet, Scopus, PubMed, and Web of Science. To ensure comprehensiveness, we employed a combination of generative AI keywords and LLM keywords, and used the shortest matching string to capture all lexical variations. Our search query was as follows, with different variations used across database platforms (detailed in Supplementary Table 1):

> ("generative artificial intelligence" OR "large language models" OR "generative model" OR "chatbot") AND ("mental" OR "psychiatr" OR "psycho" OR "emotional support")

We conducted the search in the title or abstract of articles, covering the period from January 1, 2020, to July 19, 2024, without language restrictions. The search results included 259 articles from PubMed, 444 articles from Scopus, 1 article from APA PsycNet (PsychInfo and PsycArticles), and 500 articles from Web of Science. The initial search yielded 1,204 articles, with 14 additional articles identified from sources such as Google Scholar, the ACM Digital Library, and reverse referencing. After removing 492 duplicates, we were left with a total of 726 unique articles.

We applied the following inclusion criteria to select studies for our review: first, the study must involve using an LLM to generate responses (generative task); second, the study must focus specifically on mental health care, distinguishing it from studies in related fields like psycholinguistics; third, the study must have human validations rather than relying purely on automated evaluation. An LLM is defined as "transformer-based models

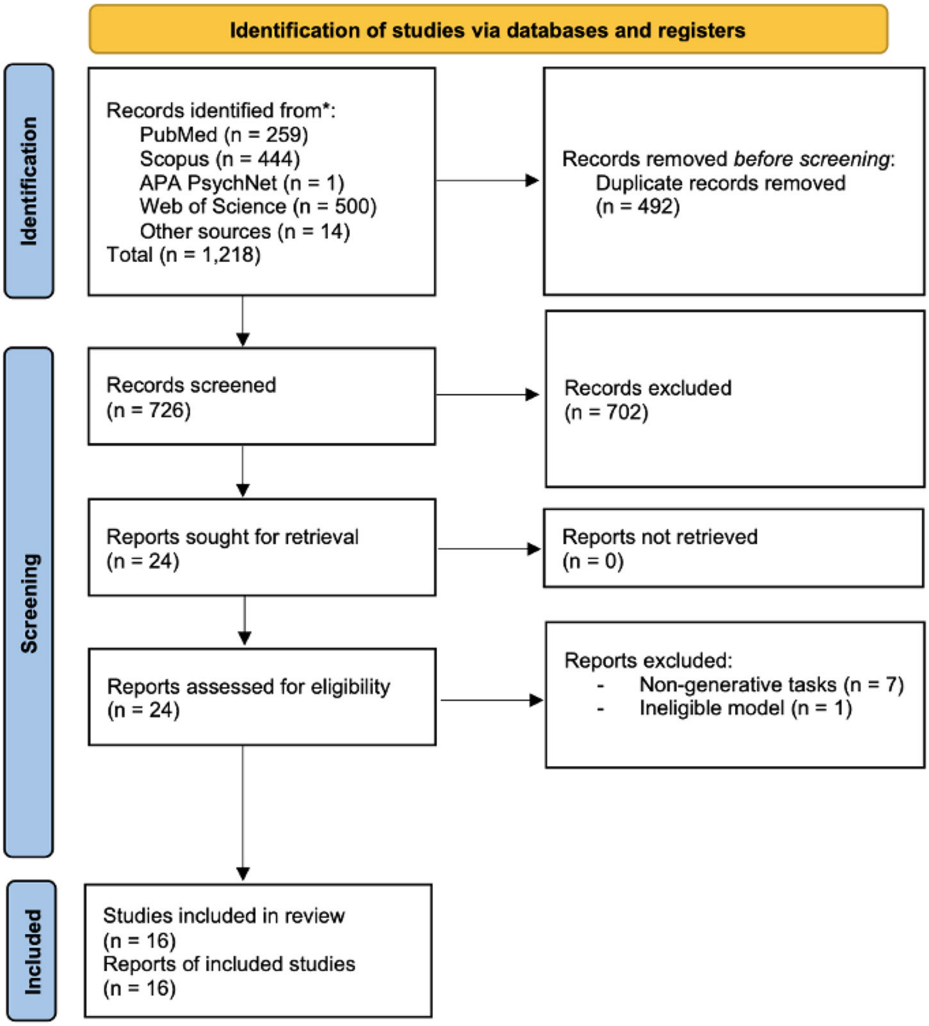
Identification of studies via databases and registers
Identification
Screening
Included
Records identified from*:
PubMed (n = 259)
Scopus (n = 444)
APA PsychNet (n = 1)
Web of Science (n = 500)
Other sources (n = 14)
Total (n = 1,218)
Records removed *before screening*:
Duplicate records removed
(n = 492)
Records screened
(n = 726)
Records excluded
(n = 702)
Reports sought for retrieval
(n = 24)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 24)
Reports excluded:
- Non-generative tasks (n = 7)
- Ineligible model (n = 1)
Studies included in review
(n = 16)
Reports of included studies
(n = 16)

Fig. 2 | The PRISMA figure of the search and screening process. The PRISMA diagram shows the systematic process of study selection. Of 1204 articles initially identified across databases, 726 unique records remained after duplicates were removed. Further screening yielded 16 articles meeting the inclusion criteria.

with more than ten billion parameters, which are trained on massive text data and excel at a variety of complex generation tasks." in this study, following a highly cited review from the NLP community[45]. We excluded reviews, meta-analyses, and clinical trials from our selection. Then, we further removed seven studies not meet our inclusion criteria upon full-text review. The result analysis review includes 16 articles, with 15 full-text-length papers and one brief communication paper.

Data extraction was conducted by one or two authors for each section, with a second author independently reviewing for accuracy. For mental health conditions, data were extracted to categorize disorders, symptoms, care settings, interventions, assessments, and diagnostic sources, with a distinction made between clinically validated disorders and general mental health constructs. For applications and model details, we extracted data on input/output modalities, model types, embodiment, open-source availability, and target user populations. Regarding tasks and clinical effectiveness, we collected data on the primary tasks involving LLMs, sample sizes, demographic characteristics, and methods of clinical validation. Evaluation methods were categorized, with constructs mapped to a hierarchical evaluation framework, producing a harmonized pyramid to systematically assess LLMs across various levels of evidence. Further details on the screening process, data extraction, and synthesis are provided in Supplementary Note 1.

Data availability

All data associated with this study has been made available in appendices.

Code availability

Not applicable.

Received: 20 August 2024; Accepted: 2 April 2025;
Published online: 30 April 2025

References

1. World Health Organization. One in 100 deaths is by suicide - WHO guidance to help the world reach the target of reducing suicide rate by 1/3 by 2030. https://www.who.int/news/item/17-06-2021-one-in-100-deaths-is-by-suicide (2021).
2. National Institutes of Health. FY 2023 Budget - Congressional Justification - National Institute of Mental Health (NIMH). https://www.nimh.nih.gov/about/budget/fy-2023-budget-congressional-justification (2023).
3. Satiani, A., Niedermier, J., Satiani, B. & Svendsen, D. P. Projected Workforce of Psychiatrists in the United States: A Population Analysis. *Psychiatr. Serv.* **69**, 710–713 (2018).
4. Mongelli, F., Georgakopoulos, P. & Pato, M. T. Challenges and Opportunities to Meet the Mental Health Needs of Underserved and Disenfranchised Populations in the United States. *FOC* **18**, 16–24 (2020).
5. World Health Organization. WHO Special Initiative for Mental Health. https://www.who.int/initiatives/who-special-initiative-for-mental-health.

6. Goodings, L., Ellis, D. & Tucker, I. *Understanding Mental Health Apps: An Applied Psychosocial Perspective*. (Springer Nature, 2024).
7. Torous, J. et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* **20**, 318–335 (2021).
8. Zhang, M. et al. The Adoption of AI in Mental Health Care–Perspectives From Mental Health Professionals: Qualitative Descriptive Study. *JMIR Form. Res.* **7**, e47847 (2023).
9. OpenAI et al. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
10. Google. An important next step on our AI journey. https://blog.google/technology/ai/bard-google-ai-search-updates/ (2023).
11. Anthropic. Introducing Claude. https://www.anthropic.com/news/introducing-claude.
12. Hua, Y. et al. Large Language Models in Mental Health Care: a Scoping Review. Preprint at https://doi.org/10.48550/arXiv.2401.02984 (2024).
13. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5^{TM}, 5th Ed*. xliv, 947 (American Psychiatric Publishing, Inc., Arlington, VA, US, 2013). https://doi.org/10.1176/appi.books.9780890425596.
14. Liu, I. et al. Investigating the Key Success Factors of Chatbot-Based Positive Psychology Intervention with Retrieval- and Generative Pre-Trained Transformer (GPT)-Based Chatbots. *Int. J. Human–Comput. Interact.* (2024).
15. Mármol-Romero, A. M., García-Vega, M., García-Cumbreras, M. Á. & Montejo-Ráez, A. An Empathic GPT-Based Chatbot to Talk About Mental Disorders With Spanish Teenagers. *Int. J. Human–Comput. Interact.* 1–17. https://doi.org/10.1080/10447318.2024.2344355.
16. Kim, T. et al. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. in *Proceedings of the CHI Conference on Human Factors in Computing Systems* 1–20 (Association for Computing Machinery, New York, NY, USA, 2024). https://doi.org/10.1145/3613904.3642937.
17. Maples, B., Cerit, M., Vishwanath, A. & Pea, R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Ment. Health Res* **3**, 1–6 (2024).
18. Spiegel, B. M. R. et al. Feasibility of combining spatial computing and AI for mental health support in anxiety and depression. *npj Digit. Med.* **7**, 1–5 (2024).
19. Perlis, R. H., Goldberg, J. F., Ostacher, M. J. & Schneck, C. D. Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacol* **49**, 1412–1416 (2024).
20. Berrezueta-Guzman, S., Kandil, M., Martín-Ruiz, M.-L., de la Cruz, I. P. & Krusche, S. Exploring the Efficacy of Robotic Assistants with ChatGPT and Claude in Enhancing ADHD Therapy: Innovating Treatment Paradigms. in *2024 International Conference on Intelligent Environments (IE)* 25–32 https://doi.org/10.1109/IE61493.2024.10599903 (2024).
21. Llanes-Jurado, J., Gómez-Zaragozá, L., Minissi, M. E., Alcañiz, M. & Marín-Morales, J. Developing conversational Virtual Humans for social emotion elicitation based on large language models. *Expert Syst. Appl*. **246**, (2024).
22. Sharma, A., Rushton, K., Lin, I. W., 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 CHI Conference on Human Factors in Computing Systems* 1–29 (Association for Computing Machinery, New York, NY, USA, 2024). https://doi.org/10.1145/3613904.3642761.
23. Sharma, A. et al. Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds. Rogers, A., Boyd-Graber, J. & Okazaki, N.) 9977–10000 (Association for Computational Linguistics, Toronto, Canada, 2023). https://doi.org/10.18653/v1/2023.acl-long.555.
24. Grabb, D. The impact of prompt engineering in large language model performance: a psychiatric example. *J. Med. Artif. Intell.* **6**, (2023).
25. Liu, Y., Ding, X., Peng, S. & Zhang, C. Leveraging ChatGPT to optimize depression intervention through explainable deep learning. *Front. Psychiatry* **15**, (2024).
26. Lai, T. et al. Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs). *BioMedInformatics* **4**, 8–33 (2024).
27. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
28. Franco D'Souza, R., Amanullah, S., Mathew, M. & Surapaneni, K. M. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. *Asian J. Psychiatr.* **89**, 103770 (2023).
29. 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* 429–435 (Association for Computing Machinery, New York, NY, USA, 2023). https://doi.org/10.1145/3594806.3596580.
30. Beck, J. S. *Cognitive Behavior Therapy: Basics and beyond, 2nd Ed*. xix, 391 (Guilford Press, New York, NY, US, 2011).
31. Kim, M., Hwang, K., Oh, H., Kim, H. & Kim, M. A. Can a Chatbot be Useful in Childhood Cancer Survivorship? Development of a Chatbot for Survivors of Childhood Cancer. in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* 4018–4022 (Association for Computing Machinery, New York, NY, USA, 2023). https://doi.org/10.1145/3583780.3615234.
32. Qian, Y., Zhang, W. & Liu, T. Harnessing the Power of Large Language Models for Empathetic Response Generation: Empirical Investigations and Improvements. in *Findings of the Association for Computational Linguistics: EMNLP 2023* (eds. Bouamor, H., Pino, J. & Bali, K.) 6516–6528 (Association for Computational Linguistics, Singapore, 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.433.
33. Holderried, F. et al. A Generative Pretrained Transformer (GPT)-Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study. *JMIR Med Educ.* **10**, e53961 (2024).
34. Vowels, L. M., Francois-Walcott, R. R. R. & Darwiche, J. AI in relationship counselling: Evaluating ChatGPT's therapeutic capabilities in providing relationship advice. *Computers Hum. Behav.: Artif. Hum.* **2**, 100078 (2024).
35. Berrezueta-Guzman, S., Kandil, M., Martín-Ruiz, M.-L., Pau de la Cruz, I. & Krusche, S. Future of ADHD Care: Evaluating the Efficacy of ChatGPT in Therapy Enhancement. *Healthcare* **12**, 683 (2024).
36. Zhang, Y. et al. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 270–278 (Association for Computational Linguistics, Online, 2020). https://doi.org/10.18653/v1/2020.acl-demos.30.
37. Hua, Y. et al. Standardizing and Scaffolding Healthcare AI-Chatbot Evaluation. 2024.07.21.24310774 Preprint at https://doi.org/10.1101/2024.07.21.24310774 (2024).
38. Cozad, M. J. et al. Mobile Health Apps for Patient-Centered Care: Review of United States Rheumatoid Arthritis Apps for Engagement and Activation. *JMIR mHealth uHealth* **10**, e39881 (2022).
39. Mirzaei, T., Amini, L. & Esmaeilzadeh, P. Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications. *BMC Med Inf. Decis. Mak.* **24**, 250 (2024).
40. Leucht, S., van Os, J., Jäger, M. & Davis, J. M. Prioritization of Psychopathological Symptoms and Clinical Characterization in Psychiatric Diagnoses: A Narrative Review. *JAMA Psychiatry* https://doi.org/10.1001/jamapsychiatry.2024.2652 (2024).

41. Ji, S., Zhang, T., Yang, K., Ananiadou, S. & Cambria, E. Rethinking Large Language Models in Mental Health Applications. Preprint at http://arxiv.org/abs/2311.11267 (2023)
42. Jin, Y. et al. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. in *Proceedings of the ACM on Web Conference 2024* 2627–2638 (ACM, Singapore Singapore, 2024). https://doi.org/10.1145/3589334.3645643.
43. Gabriel, S., Puri, I., Xu, X., Malgaroli, M. & Ghassemi, M. Can AI Relate: Testing Large Language Model Response for Mental Health Support. in *Findings of the Association for Computational Linguistics: EMNLP 2024* (eds. Al-Onaizan, Y., Bansal, M. & Chen, Y.-N.) 2206–2221 (Association for Computational Linguistics, Miami, Florida, USA, 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.120.
44. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
45. Zhao, W. X. et al. A Survey of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2303.18223 (2023).
46. Watson, D., Clark, L. A. & Tellegen, A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Personal. Soc. Psychol.* **54**, 1063–1070 (1988).
47. William Li, H. C., Chung, O. K. J. & Ho, K. Y. Center for Epidemiologic Studies Depression Scale for Children: psychometric testing of the Chinese version. *J. Adv. Nurs.* **66**, 2582–2591 (2010).
48. Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern Med* **16**, 606–613 (2001).
49. Diener, E. Satisfaction With Life Scale. https://doi.org/10.1037/t01069-000 (2011).
50. Merz, E. L. et al. Validation of interpersonal support evaluation list-12 (ISEL-12) scores among English- and Spanish-speaking Hispanics/Latinos from the HCHS/SOL Sociocultural Ancillary Study. *Psychol. Assess.* **26**, 384–394 (2014).
51. De Jong Gierveld, J. & Van Tilburg, T. The De Jong Gierveld short scales for emotional and social loneliness: tested on data from 7 countries in the UN generations and gender surveys. *Eur. J. Ageing* **7**, 121–130 (2010).
52. Monticone, M. et al. The 27-item coping strategies questionnaire-revised: confirmatory factor analysis, reliability and validity in Italian-speaking subjects with chronic pain. *Pain. Res Manag* **19**, 153–158 (2014).
53. Akin, A. The Scales of Psychological Well-Being: A Study of Validity and Reliability. *Educ. Sci.: Theory Pract.* **8**, 741–750 (2008).
54. Ryan, R. M. & Frederick, C. On energy, personality, and health: Subjective vitality as a dynamic reflection of well-being. *J. Personal.* **65**, 529–565 (1997).

Acknowledgements

This study received no funding.

Author contributions

Study design: Y.H., J.T.; Screening, data extraction, data analysis: Y.H., H.N., Z.L., F.L.; Validation: Y.H. and J.T.; Manuscript drafting, feedback, revision: Y.H., H.N., Z.L., F.L., X.F., D.C., J.T. Supervision: J.T.

Competing interests

J.T. has research support from Otsuka and is an adviser to Precision Mental Wellness. He is also an assistant editor for NPJ Digital Medicine. All other authors have no conflict of interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01611-4.

Correspondence and requests for materials should be addressed to John Torous.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.