Beyond the Buzz: A Systematic Review of Generative AI’s Capabilities in Mental Health

Liying Wang, Tanmay Bhanushali, Zhouran Huang, Jingyi Yang, Sukriti Badami, Lisa Hightow-Weidman

Submitted to: JMIR Mental Health
on: December 12, 2024

Table of Contents

Beyond the Buzz: A Systematic Review of Generative AI's Capabilities in Mental Health

Liying Wang[1, 2] PhD; Tanmay Bhanushali[3]; Zhouran Huang[4]; Jingyi Yang[5]; Sukriti Badami[3]; Lisa Hightow-Weidman[1] MPH, MD

[1]Florida State University Institute on Digital Health and Innovation Tallahassee US
[2]Florida State University Center of Population Sciences for Health Equity Tallahassee US
[3]University of Washington Seattle US
[4]Northeastern University Boston US
[5]Columbia University New York US

Corresponding Author:
Liying Wang PhD
Florida State University
Institute on Digital Health and Innovation
222 S Copeland St
Tallahassee
US

Abstract

Background: The global shortage of mental health professionals, exacerbated by increasing mental health needs post-COVID-19, has driven interest in leveraging large language models (LLMs) like ChatGPT to address these challenges through applications such as clinical note generation, personalized treatment planning, and therapeutic support.

Objective: This systematic review aims to evaluate the current capabilities of generative AI (genAI) models in the context of mental health applications.

Methods: A comprehensive search across five databases yielded 1,046 references, of which eight studies met the inclusion criteria. These criteria required original research with experimental designs (e.g., Turing tests, socio-cognitive tasks, trials, or qualitative methods), a focus on genAI models, and explicit measurement of socio-cognitive abilities (e.g., empathy, emotional awareness), mental health outcomes, and user experience (e.g., perceived trust, empathy).

Results: The studies, published between 2023 and 2024, primarily evaluated models like ChatGPT 3.5 and 4.0, Bard, and Claude in tasks such as psychoeducation, diagnosis, emotional awareness, and clinical interventions. Most studies employed zero-shot prompting and human evaluators to assess the AI responses, using standardized rating scales or qualitative analysis. However, these methods were often insufficient to fully capture the complexity of genAI capabilities. The reliance on single-shot evaluation techniques, limited comparisons, and task-based assessments isolated from a specific context may oversimplify genAI's abilities and overlook the nuances of human-AI interaction, especially in areas requiring contextual reasoning or cultural sensitivity. The findings suggest that while genAI models demonstrate strengths in psychoeducation and emotional awareness, their diagnostic accuracy, cultural competence, and ability to engage users emotionally remain limited. Users frequently reported concerns about trustworthiness, accuracy, and the lack of emotional engagement.

Conclusions: Future research could use more sophisticated evaluation methods, such as few-shot and chain-of-thought prompting to fully uncover genAI's potential. Future studies should also focus on longitudinal research, broader comparisons with human benchmarks, and exploring how AI can be better integrated into mental health care with improved socio-cognitive and ethical decision-making capabilities.

(JMIR Preprints 12/12/2024:70014)
DOI: https://doi.org/10.2196/preprints.70014

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Review

Beyond the Buzz: A Systematic Review of Generative AI’s Capabilities in Mental Health

Abstract

Background: The global shortage of mental health professionals, exacerbated by increasing mental health needs post-COVID-19, has driven interest in leveraging large language models (LLMs) like ChatGPT to address these challenges through applications such as clinical note generation, personalized treatment planning, and therapeutic support.

Objective: This systematic review aims to evaluate the current capabilities of generative AI (genAI) models in the context of mental health applications.

Methods: A comprehensive search across five databases yielded 1,046 references, of which eight studies met the inclusion criteria. These criteria required original research with experimental designs (e.g., Turing tests, socio-cognitive tasks, trials, or qualitative methods), a focus on genAI models, and explicit measurement of socio-cognitive abilities (e.g., empathy, emotional awareness), mental health outcomes, and user experience (e.g., perceived trust, empathy).

Results: The studies, published between 2023 and 2024, primarily evaluated models like ChatGPT 3.5 and 4.0, Bard, and Claude in tasks such as psychoeducation, diagnosis, emotional awareness, and clinical interventions. Most studies employed zero-shot prompting and human evaluators to assess the AI responses, using standardized rating scales or qualitative analysis. However, these methods were often insufficient to fully capture the complexity of genAI capabilities. The reliance on single-shot evaluation techniques, limited comparisons, and task-based assessments isolated from a specific context may oversimplify genAI’s abilities and overlook the nuances of human-AI interaction, especially in areas requiring contextual reasoning or cultural sensitivity. The findings suggest that while genAI models demonstrate strengths in psychoeducation and emotional awareness, their diagnostic accuracy, cultural competence, and ability to engage users emotionally remain limited. Users frequently reported concerns about trustworthiness, accuracy, and the lack of emotional engagement.

Conclusions: Future research could use more sophisticated evaluation methods, such as few-shot and chain-of-thought prompting to fully uncover genAI’s potential. Future studies should also focus on longitudinal research, broader comparisons with human benchmarks, and exploring how AI can be better integrated into mental health care with improved socio-cognitive and ethical decision-making capabilities.

Keywords: Mental Health, Diagnosis, Assessment, Psychoeducation, Intervention, Cultural Competency

Introduction

Artificial intelligence (AI) is a branch of computer science that emulates human intelligence using computational technologies to perform tasks that require an understanding of natural language, deep learning, adaptation, problem-solving, and decision-making.[1] AI applications in healthcare have surged in recent years, and the field has seen a variety of use cases, such as applying computer vision in diagnosing breast cancer from imaging and using predictive modeling to predict suicide risk. [2–4]

With the advent of large language models (LLMs) like ChatGPT by OpenAI in 2022, there has been a significant shift toward models that excel in natural language understanding and generation.[5] LLMs like ChatGPT have quickly gained widespread adoption due to their remarkable versatility. What makes LLMs particularly special is their ability to understand and generate contextually relevant responses, making them highly effective in tasks that require nuanced language comprehension and conversational depth. Therefore, these models excel at a wide range of tasks, including generating human-like text, summarizing complex information, translating languages, answering questions, and assisting with creative writing.[6,7]

The interest and exploration of generative AI applications (genAI) in mental health have been growing over the past few years. The mental health burden globally, and in the US, has increased significantly post the COVID-19 pandemic, exceeding the capacities of the slow growth of mental health professionals [8]. Health Professional Shortage Areas (HPSAs) for mental health are defined as areas with a population-to-provider ratio of at least 30,000 to 1. According to the Behavioral Health Workforce 2023, as of December 2023, 169 million people living in the United States live in a Mental Health HPSA [9]. LLMs such as ChatGPT, Claude, and Bard hold great promise in mitigating this stark situation, including using LLMs in writing clinical notes to reduce clinicians' burden, formulating differential diagnoses, personalized treatment plans, summarizing patient chart data and providing clinical insights, providing on-demand coaching and companionship, and ultimately, providing therapy.[10,11]

It may be tempting to conclude that the field of mental health is poised to benefit significantly from the numerous opportunities presented by LLMs. AI-based chatbots are readily available in app stores, often at no cost for basic features such as chat.[12] Additionally, some individuals have begun using ChatGPT as a therapeutic companion or informal therapist.[13] However, their effectiveness in mental health care and applications in enhancing clinical practice remains largely unclear.

Additionally, attention has been given to the major areas of risks in LLM application in mental health, such as instability of output, hallucination, ethical and legal risks, as well as privacy and security risks.[14] Less discussion was held around the actual capabilities of LLM in the context of mental health, analogous to the clinical skills of a human therapist. The training of a mental health professional usually starts with a foundational knowledge base of psychopathology and clinical treatment, which LLM, with its vast amount of data, presumably already possesses.[15] However, the clinical competencies go far beyond the knowledge base and include skills such as assessment, case conceptualization, diagnostic-analytic skills, intervention skills (e.g., maintaining working alliance), and cultural humility/competencies (checklist; and other references).[16] Clinical skills are essential in effectively conducting therapy and improving patient outcomes. Little is known regarding the extent to which LLM models possess the capabilities that can mimic or simulate clinical skills required for a human therapist to earn the qualifications.

To advance the field's understanding of how LLM models can contribute to mental health practice while outlining recommendations for future research. This study aims to: 1) systematically review the extent to which LLM models possess capabilities that are parallel to clinical skills of mental health professionals; 2) describe how the capabilities were evaluated; 3) identify critical gaps and provide recommendations for future research.

Methods

Study Design

This review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines.[17]

Search Strategy

We conducted a systematic review of published literature in the following databases in June 2024: PubMed, Embase, Web of Science, Engineering Village, and PsycINFO. A combination of Medical Subject Heading (MeSH) controlled vocabulary and keywords were used to conduct the search. The search terms covered the following concepts and domains: mental health, therapy/interventions, clinical skills, human-computer interaction, text-based generative AI, and study type. We did not restrict the time or language of publication. Detailed search strategy see appendix.

Eligibility Criteria

Studies included for data extraction: 1) were original research in peer-reviewed journals (not a systematic review, conference abstract, comments, or letters); 2) focused on one more generative AI model (e.g., ChatGPT, Bard, Claude); 3) evaluated one or more capabilities important for therapy or mental health-related services (e.g., psychoeducation, assessment, empathy); 4) included measures of user experience (if the study involved actual user interaction with the genAI tools).

Study Screening and Selection

The database results were imported into Rayyan software. The studies were allocated amongst four authors who were paired up and independently screened the study titles and abstracts for eligibility. Studies that met all eligibility criteria based on title and abstract review then proceeded to full-text screening. The four authors then were paired up and screened full texts to decide upon final inclusion within the review. Results at every stage of this screening were discussed as a group to ensure consensus prior to advancing through the review and final inclusion.

Data Extraction

We created a standardized data extraction form to organize relevant information from included articles, focusing on study characteristics and therapy-related constructs and domains. We extracted the following information: author, title, publication year, sample size, participant sociodemographic characteristics, clinical diagnoses made, study design and methodology, and study outcomes. Study outcomes we considered included the type of therapeutic skill or alternate outcome measured, results, user experiences, and mental health outcomes when relevant. The studies were allocated to five reviewers for data extraction and discussion.

Analysis

We summarized the characteristics of included studies using mainly descriptive statistics, given the small number of studies included. We qualitatively described the methodology and capabilities evaluated in the studies.

Results

We obtained 1,046 references after searching in five databases. After removing duplicates, 953 articles were included in the title and abstract review. After the round of full-text review (n = 46), eight articles met all inclusion criteria and were included in this review (Figure 1).

Study Characteristics

All studies (N = 8) included in this review were published within the past two years (2023-2024). Study populations, methods, and materials were heterogeneous (Appendix A.1). The countries where the studies were performed included the US (n = 2), Israel (n = 3), and the UK (n = 1). The most evaluated model was ChatGPT 3.5, [18–24] followed by ChatGPT 4.0. [22,23] Bard, [22,23] and Claude.[22]

Evaluation of GenAI capabilities

Task-based-evaluation

The majority of the studies (n = 6) directly evaluated a specific area of performance of genAI models using prompts.[18,20–24] These studies used materials to prompt AI models' responses, which were then evaluated by human evaluators with a standardized rating scale or manual for outcomes of interest. The most used prompting method was zero-shot prompting (5 out of 8), where questions, scenarios, or case vignettes were input into ChatGPT. Zero-shot prompting refers to providing the models with instructions describing the task.[25] For example, a study created prompts in the form of common questions addressed in therapy spanned across 7 categories, including depression, anxiety, general health, substance abuse, religious/spiritual/issues, lifestyle, and interpersonal (e.g., How can I tell someone something that I think might upset them?).[20] In the category of depression, questions such as "Why do I hate myself at times?" and "How do I know if I am depressed?" were input into ChatGPT. In two studies, the Levels of Emotional Awareness Scale (LEAS) was used to evaluate the model's emotional awareness. It included 20 emotionally charged scenarios, and the model was asked to describe the emotional experience of the characters in the scenarios, which was then evaluated by human evaluators.[21,24]

One study used a simplified Chain-of-Thought (CoT) prompting method and found that compared to zero-shot promoting, this method improved the performance of ChatGPT 3.5, 4, but not Bard in recognizing Alzheimer's dementia (AD) and Cognitively Normal (CN).[23] Chain-of-thought refers to the method where LLMs were provided with step-by-step reasoning examples rather than question-and-answer examples and has been shown to outperform zero-shot prompting methods in reasoning tasks.[25] One study did not use any prompting of ChatGPT.[19]

User-involved evaluation

Only two studies involved human subjects and examined the experience and perspective of genAI's performance through user experience and survey.[19,26] One study included outpatients from a hospital with anxiety, depression, or other behavioral disorders and assessed an AI model's performance by asking participants to use the bot to manage their symptoms for two weeks for at least 15 minutes a day. Of note, this study did not involve prompting to pre-set ChatGPT, and individuals accessed ChatGPT 3.5 on their own personal devices.[19] The other study surveyed seven trainees in counseling and psychotherapy to understand their perspective on genAI's application in mental health and especially its role in training.[26]

Comparison

Most studies evaluated the genAI model on its own without comparison groups.[18,20,24] Three studies involved comparisons, including human norms,[21] clinical experts,[22] public opinion, [22] and multiple genAI models.[22,23] The metrics used to evaluate the responses from ChatGPT included accuracy,[18,20,23] readability,[18] reproducibility,[18] clarity,[20] relevance,[20] empathy,[20] engagement,[20] ethical considerations,[20] contextual suitability,[21] sensitivity,[23] specificity, [23] precision,[23] number of distinct emotions identified,[24] intensity of emotions.[24] The study compared the AI model's responses with mental health professionals, treating mental health professionals as the golden criterion.[22]

Capabilities of genAI in mental health

The specific therapy skill examined in the studies includes psychoeducation,[18–20,26] assessment/prognostic assessment,[19,22] diagnosis,[23] empathy/emotional support,[19] emotional awareness,[21,24] goal setting,[19] motivation,[19] cognitive restructuring,[19] crisis intervention,[19] guided imagery,[19] journaling prompts,[19] cultural and linguistic capabilities,[19] and ethical and legal capabilities.[19,22,26]

Five studies had specific use cases when evaluating the capabilities of genAI models, including psychoeducation/counseling for Erectile Dysfunction (ED),[18] diagnosis of AD and CN, [23] prognostic assessment of schizophrenia,[22] emotional awareness of case examples with Borderline Personality Disorder and Schizoid Personality Disorder,[24] emotional support and intervention skills for anxiety, depression or other behavioral disorders.[19] The rest of the studies (n = 3) did not include a specific condition and examined more generic therapy skills such as psychoeducation [20,26] and emotional awareness.[21]

Psychoeducation

The genAI models overall performed well on tasks related to general psychoeducation with human evaluators. Psychoeducation-related responses generated by ChatGPT in the context of ED patient care were rated by two board-certified urologists as comprehensive and empathetic, with good reproducibility.[18] However, the readability of the responses was 13.8, measured by the Gunning Fog Index, indicating that the responses were suitable for audiences with at least a high school degree. ChatGPT's responses to psychoeducational questions were also found to be comprehensive, accurate, simple, clear, understandable, relevant to the prompts, and engaging, rated by two mental health professionals.[20]

The evidence supporting genAI's skill in performing psychoeducation is mixed from the end user perspective. Participants who interacted with ChatGPT generally found ChatGPT to be helpful in improving their mental health literacy (60%) and managing their mental health symptoms (80%).[19] The findings also suggested that many participants (80%) reported issues around the accuracy of information provided and reliability concerns. In contrast, counseling students reported concerns about the accuracy and trustworthiness of the quality of information generated by any generative AI tools.[26]

Diagnosis, assessment, and prognosis

The evidence for the assessment and diagnostic capabilities of LLM chatbots is mixed. The diagnostic performance of GPT-4 surpassed chance-level performance in identifying CN patients (true-positive at 56%), while Bard reached an 88.6% true-positive rate for identifying AD.[23] Both models have shortcomings as well, where ChatGPT-4 tended to avoid making a clear diagnostic decision between AD and CN, and Bard tended to misdiagnose CN as AD with high confidence. In comparing prognosis assessment of schizophrenia case vignettes, a study found that GPT-4, Bard, and Claude generated similar assessments as clinical professionals, while GPT-3.5 tended to predict more negative outcomes long term compared to the rest of the models.[22]

From the users' perspective, participants who interacted with GPT-3.5 found the chatbot to be lacking in assessment skills to understand their complaints/symptoms before making recommendations and medications. In addition, participants also noted the lack of human touch and emotion in the process of assessment, which reduced their motivation to stay engaged in the conversation.[19]

Emotional awareness and empathy

The emotional capabilities of ChatGPT 3.5 received evidence from LEAS-based testing. ChatGPT-3.5 outperformed all individuals in a population norm dataset of LEAS from the general French population (N = 750; 506 female and 244 male).[21] The responses from ChatGPT-3.5 also fit the context (as contextually appropriate) based on the ratings of two licensed psychologists, whose ratings reached high consistency and inter-rater agreement. Similarly, ChatGPT-3.5 demonstrated emotional awareness and mentalizing capabilities, reflected by its responses that differentiate BPD and SPD through different LEAS scoring, number of emotions, and emotion intensity. This indicates the model's awareness of emotional experiences underlying different psychopathologies.[24]

From the users' perspective, half of the study participants (13 out of 24) found chatGPT able to stay non-judgmental, offer empathic responses, and provide emotional support, including validating their feelings, helping them feel seen, and offering jokes to lighten their mood. These participants reported feeling happy, relaxed, peaceful, and cared for during their interactions with chatGPT.[19]

Clinical intervention skills

One study provided support for ChatGPT-3.5's capabilities in applying clinical intervention skills to help users.[19] Participants (5 out of 24) in the study reported that ChatGPt helped them set realistic goals, devise a plan, and provide motivational support to help them progress toward their goals. Participants also reported that ChatGPT helped them practice cognitive restructuring, a widely used Cognitive Behavioral Therapy technique that helps individuals identify negative thoughts and replace them with more rational thoughts to change mood. ChatGPT helped users gain a range of skills to understand and cope with their emotions by doing guided imagery with the users and providing journaling prompts. Guided imagery is a technique commonly used in mindfulness exercises and can be used for different purposes, such as relaxation and generating positive emotions.[27,28]

Cultural and linguistic capabilities

Only one study reported users' experience with ChatGPt 3.5 regarding its cultural and linguistic capabilities.[19] Participants in the study found that ChatGPT was not able to fully understand certain terms and symptoms rooted in Arabic culture. Although it can understand Arabic generally, it still lacks nuanced cultural understanding, hindering communication and connection between users and ChatGPT.

Ethical and legal considerations

The ability of AI to align its behaviors with ethical and legal standards was questioned in two studies.[19,26] Participants in both studies brought up the potential for biases that are innate in AI models, which may impact their ability to provide accurate information or assistance in decision-making for patients and trainees in counseling and psychotherapy. Although not a behavior or capability of AI per se, participants expressed concerns about data privacy and confidentiality of the interactions with ChatGPT. On a brighter note, ChatGPT3.5, 4, Bard, and Claude were able to identify the risk of being discriminated against after reading the case vignettes of persons with schizophrenia.[26]

Discussion

This is the first review that examined evidence of the capabilities of generative AI in its applications in mental health. The result shows that generative AIs are most promising in providing psychoeducation among the evaluated capabilities. The evaluation method is limited, with the majority using zero-shot prompting and human evaluators to examine LLMs' responses. There is a dearth of studies overall covering a limited range of therapeutic skills, with the majority focusing on its potential to provide psychoeducation. There is not enough evidence to derive a conclusion regarding genAI's capabilities in assessment, diagnosis, intervention skills, cultural competency, and ethical decision-making.

The psychoeducation capabilities of genAI in providing information related to mental health received the most evidence, while the rest of the skills were rarely studied. This could be related to the single approach of using zero-shot prompting that was shared among almost all studies in this review. This methodology limitation was also reflected in studies on GenAI's use cases in the medical field.[29–31] Although the zero-shot prompting method was sufficient to elicit a response from GenAI models, it is far from enough to unmask the potential of genAI. More sophisticated prompting methods, such as Few-Shot, CoT, and a combination of the two, have been shown to significantly improve model performance in multiple reasoning tasks.[32–34] Indeed, if the only tool researchers have is a hammer, we tend to see every problem as a nail. More sophisticated methods of prompting and fine-tuning are needed to unveil the capabilities of GenAI models in a broader range of use cases.

We found a scarcity of studies on the assessment and diagnostic capabilities of genAI tools. The models performed above chance level in the context of diagnosing AD and NC and exhibited distinct patterns of behaviors, with GPT4 being indecisive and Bard being overconfident. Capabilities such as assessment and diagnosis are essential to early detection and treatment planning and have serious implications in clinical practice. While medical diagnostic instruments/tools using predictive ML algorithms have long existed, research has yet to be done regarding a comparison in diagnostic and assessment performance between LLM-based and ML-based models.[35]

The capabilities were also mostly tested in isolation and without a specific context or problem at hand. Model performance may vary depending on the context. The models often perform better with higher specificity and richer information regarding the task, such as providing example problems, logical reasoning steps, or abstract solution structures.[32,34,36] Additionally, future studies may benefit from examining a combination of capabilities to increase the external validity of the results. In clinical practice, clinicians more often than not flexibly pick and choose a combination of techniques.[37] As one study pointed out, more empathy and warmth from ChatGPT during assessment may enhance their engagement.[24]

Although task-based testing may elucidate the technical capabilities of GenAI by stripping away the human factors, it is paramount to engage diverse subgroups of users both within and beyond clinicians and patients for metrics such as readability. We identified discrepancies in evaluations of GenAI's capabilities. Studies that engaged clinicians as evaluators of GenAI's performance mostly reported positive findings regarding the accuracy, clarity, and relevancy of the models' responses. However, studies that engaged users of GenAI who were mental health outpatients or trainees seem to report mixed feelings based on their prior experience or interactions with the GenAI tools during the study. This discrepancy could be a result of the methodological differences. Task-based testing tends to provide more unified responses and limit the room for errors to occur. However, user interaction and actual experience with GenAI tools bring a range of factors, such as the users' goals, expectations, demographic characteristics, and cultural and linguistic backgrounds, which may all contribute to their experience with the GenAI tools.

Current models are largely trained on data from Western, English-speaking sources, which introduces inherent biases into their responses.[38,39] Based AI systems can perpetuate inequalities while reinforcing harmful stereotypes, especially as genAI becomes more prevalent in creating content that influences public perceptions and judgments.[40] This training can limit genAI's ability to address the needs of users from diverse cultural, linguistic, and socio-economic backgrounds. Mental health conditions may manifest differently across cultures. For instance, research on culture and psychopathology has established that persons in non-Western cultural contexts, such as China, express emotional distress with somatic symptoms more frequently than those in North American cultural contexts, such as Canada, which can lead to misdiagnosis or ineffective treatments if the cultural context is not in consideration.[41] Thus, culturally sensitive approaches are often needed for effective assessment and intervention. However, genAI tools may not fully capture these nuances, leading to misinterpretation or inappropriate recommendations. To improve cultural competency, future research needs to prioritize training models on more diverse data sets and involve fine-tuning techniques that account for cultural and linguistic variations. Incorporating feedback from diverse user groups during model development can also enhance genAI's ability to provide culturally appropriate care, ultimately making it a more inclusive tool in global mental health contexts.

Using culturally appropriate interventions is a common criterion for evaluating clinicians. Mental health professionals are assessed based on their ability to recognize their cultural prejudices and understand how these biases can impact their relationships with clients from diverse cultural backgrounds.[42] GenAI models should also be evaluated to determine if they demonstrate an understanding of cultural nuances and biases in the training data. To address this issue, data from diverse cultural backgrounds is needed to create a more inclusive genAI. For genAI, this would involve evaluating the AI's ability to tailor therapeutic interventions or suggestions that are culturally sensitive. For example, Sue's model of cultural competency provides guidelines for practice in the mental health field, which includes components of cultural awareness, knowledge, and skills. It could be considered a standard for evaluating genAI's cultural competence.[42]

It is unclear the extent to which GenAI tools can perform ethical decision-making in the context of mental health. What is clear, though, is that users expressed ethical concerns that reflect a lack of trust. This may be due to the general lack of trust in technology when it comes to user data, especially in sensitive topics such as mental health. In addition to continuing to build AI infrastructure that prioritizes user privacy, more research can be done to examine specific behaviors AI tools could exhibit to address user concerns.

A limitation of this review article lies in its dependence on the small number of available studies, which primarily focus on a narrow range of capabilities, particularly psychoeducation, while other important areas, such as diagnostic skills, clinical interventions, cultural competency, and ethical decision-making, remain underexplored. Furthermore, the review highlights the methodological constraints of the studies included, many of which use simplistic evaluation techniques like zero-shot prompting that may not provide a comprehensive understanding of generative AI's potential. The lack of comparative analyses across different generative AI models and traditional human benchmarks further restricts the ability to draw conclusive insights into their clinical utility. Additionally, the review is limited by the scarcity of longitudinal studies that track the evolution of these AI models in mental health care, leaving gaps in understanding their long-term impact and effectiveness. Finally, the review may not fully account for the diverse user experiences and expectations, as most studies engaged either clinical professionals or a narrow patient population, underscoring the need for more inclusive research designs that consider different cultural, linguistic, and socio-economic contexts.

Recommendations

Based on the findings of this review, below are the recommendations we propose for future research to consider:

- Diversify Evaluation Metrics: Future researchers should evaluate additional cognitive and emotional capabilities, including empathy, emotional awareness, and treatment skills such as screening, intake, diagnosis, and treatment planning.
- Innovate Prompting Techniques: Researchers should experiment with new prompting methods, such as few-shot and COT prompting, and tailor them to specific therapeutic tasks.
- Test with Human Subject: Future studies should involve human participants or clinical samples as end-users to provide comprehensive feedback on the performance of GenAI models.
- Expand Cultural and Linguistic Research: AI models should be tested in diverse cultural contexts and languages to ensure global applicability.
- Conduct Longitudinal Studies: Long-term research, preferably in a clinical setting, is needed to assess the sustainability and effectiveness of AI interventions in mental health.
- Compare AI with Human Clinicians: Future studies should directly compare AI models with human clinicians using standardized criteria.
- Experiment with Different Models: GenAI models are rapidly evolving, and each may excel at different tasks. Future studies should test various GenAI models on the same task to identify which performs best.
- Systematic evaluation framework of genAI in mental health: As the field of mental health continues to move forward and explore the potential and use cases of AI, we are in need of a framework that guides systematic evaluation of the various AI models as well as longitudinally track the improvement of genAI in performing certain tasks and skills related to mental health.

Conclusions

This review provides a detailed description of the capabilities of genAI models in mental health applications. Although genAI showed promise in areas such as psychoeducation, there remain significant gaps in research on other therapeutic skills such as assessment, diagnosis, cultural competency, and ethical decision-making. Additionally, the methodological constraints may prevent researchers from examining the more advanced capabilities of genAI. Future research should adopt robust and diverse evaluation techniques and develop comprehensive evaluation frameworks to guide systematic evaluation of genAI’s capabilities in mental health applications. Fully unleashing AI’s potential in addressing mental health challenges would require concerted efforts to address the limitations and gaps identified in this review.

Acknowledgements

We would like to extend our sincere gratitude to the colleagues who provided invaluable feedback and guidance throughout this study.

Conflicts of Interest

The authors declare no conflict of interest.

Abbreviations

LLMs: Large Language Models
genAI: Generative AI
AI: Artificial Intelligence
US: United States
HPSAs: Health Professional Shortage Areas
AD: Alzheimer's Dementia
CN: Cognitively Normal
CoT: Chain-of-Thought
ED: Erectile Dysfunction
BPD: Borderline Personality Disorder
SPD: Schizoid Personality Disorder
LEAS: Levels of Emotional Awareness Scale
ML: Machine Learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
MeSH: Medical Subject Heading

Multimedia Appendix 1

N/A

References

1. What Is Artificial Intelligence (AI)? Google Cloud. Accessed September 9, 2024. https://cloud.google.com/learn/what-is-artificial-intelligence
2. Malerbi FK, Nakayama LF, Gayle DR, et al. Digital Education for the Deployment of Artificial Intelligence in Health Care. *Journal of medical Internet research*. 2023;25. doi:10.2196/43333
3. Zahoor S, Lali IU, Khan MA, Javed K, Mehmood W. Breast cancer detection and classification using traditional Computer vision techniques: A comprehensive review. *Curr Med Imaging Rev*. 2020;16(10):1187-1200. doi:10.2174/1573405616666200406110547
4. Ji S, Pan S, Li X, Cambria E, Long G, Huang Z. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Trans Comput Soc Syst*. 2021;8(1):214-226. doi:10.1109/tcss.2020.3021467
5. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the

new AI-driven infodemic threat in public health. *Front Public Health*. 2023;11:1166120. doi:10.3389/fpubh.2023.1166120

6. Semrl N, Feigl S, Taumberger N, et al. AI language models in human reproduction research: exploring ChatGPT's potential to assist academic writing. *Hum Reprod*. 2023;38(12):2281-2288. doi:10.1093/humrep/dead207
7. Lee TK. Artificial intelligence and posthumanist translation: ChatGPT versus the translator. *Appl Linguist Rev*. Published online August 1, 2023. doi:10.1515/applirev-2023-0122
8. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *Lancet*. 2021;398(10312):1700-1712. doi:10.1016/S0140-6736(21)02143-7
9. National Center for Health Workforce Analysis. *Behavioral Health Workforce 2023 Brief*. https://bhw.hrsa.gov/sites/default/files/bureau-health-workforce/Behavioral-Health-Workforce-Brief-2023.pdf
10. Singh OP. Artificial intelligence in the era of ChatGPT - Opportunities and challenges in mental health care. *Indian J Psychiatry*. 2023;65(3):297-298. doi:10.4103/indianjpsychiatry.indianjpsychiatry_112_23
11. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *Npj Ment Health Res*. 2024;3(1):12. doi:10.1038/s44184-024-00056-z
12. Khawaja Z, Bélisle-Pipon JC. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Front Digit Health*. 2023;5:1278186. doi:10.3389/fdgth.2023.1278186
13. Maples B, Cerit M, Vishwanath A, Pea R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *Npj Ment Health Res*. 2024;3(1):4. doi:10.1038/s44184-023-00047-6
14. King DR, Nanda G, Stoddard J, et al. An introduction to generative artificial intelligence in mental health care: Considerations and guidance. *Curr Psychiatry Rep*. 2023;25(12):839-846. doi:10.1007/s11920-023-01477-x
15. Obradovich N, Khalsa SS, Khan WU, et al. Opportunities and risks of large language models in psychiatry. *NPP-Digit Psychiatry Neurosci*. 2024;2(1):1-8. doi:10.1038/s44277-024-00010-z
16. Falender C, Shafranske EP. Clinical Supervision. Published online 2015. doi:10.4135/9781473934870
17. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med*. 2009;151(4):264-269, W64. doi:10.7326/0003-4819-151-4-200908180-00135
18. Razdan S, Siegal AR, Brewer Y, Sljivich M, Valenzuela RJ. Assessing ChatGPT's ability to answer questions pertaining to erectile dysfunction: can our patients trust it? *Int J Impot Res*. Published online November 20, 2023:1-7. doi:10.1038/s41443-023-00797-z
19. Alanezi F. Assessing the effectiveness of ChatGPT in delivering mental health support: A qualitative study. *J Multidiscip Healthc*. 2024;17:461-471. doi:10.2147/JMDH.S447368
20. Maurya RK, Montesinos S, Bogomaz M, DeDiego AC. Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. *Couns Psychother Res*. Published online April 25, 2024. doi:10.1002/capr.12759
21. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. *Front Psychol*. 2023;14:1199058. doi:10.3389/fpsyg.2023.1199058
22. Elyoseph Z, Levkovich I. Comparing the Perspectives of Generative AI, Mental Health Experts, and the General Public on Schizophrenia Recovery: Case Vignette Study. *JMIR Ment Health*. 2024;11:e53043-e53043. doi:10.2196/53043
23. B T B, Chen JM. Performance assessment of ChatGPT versus Bard in detecting Alzheimer's dementia. *Diagnostics (Basel)*. 2024;14(8). doi:10.3390/diagnostics14080817
24. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities:

personalization for personality structures. *Front Psychiatry*. 2023;14:1234397. doi:10.3389/fpsyt.2023.1234397

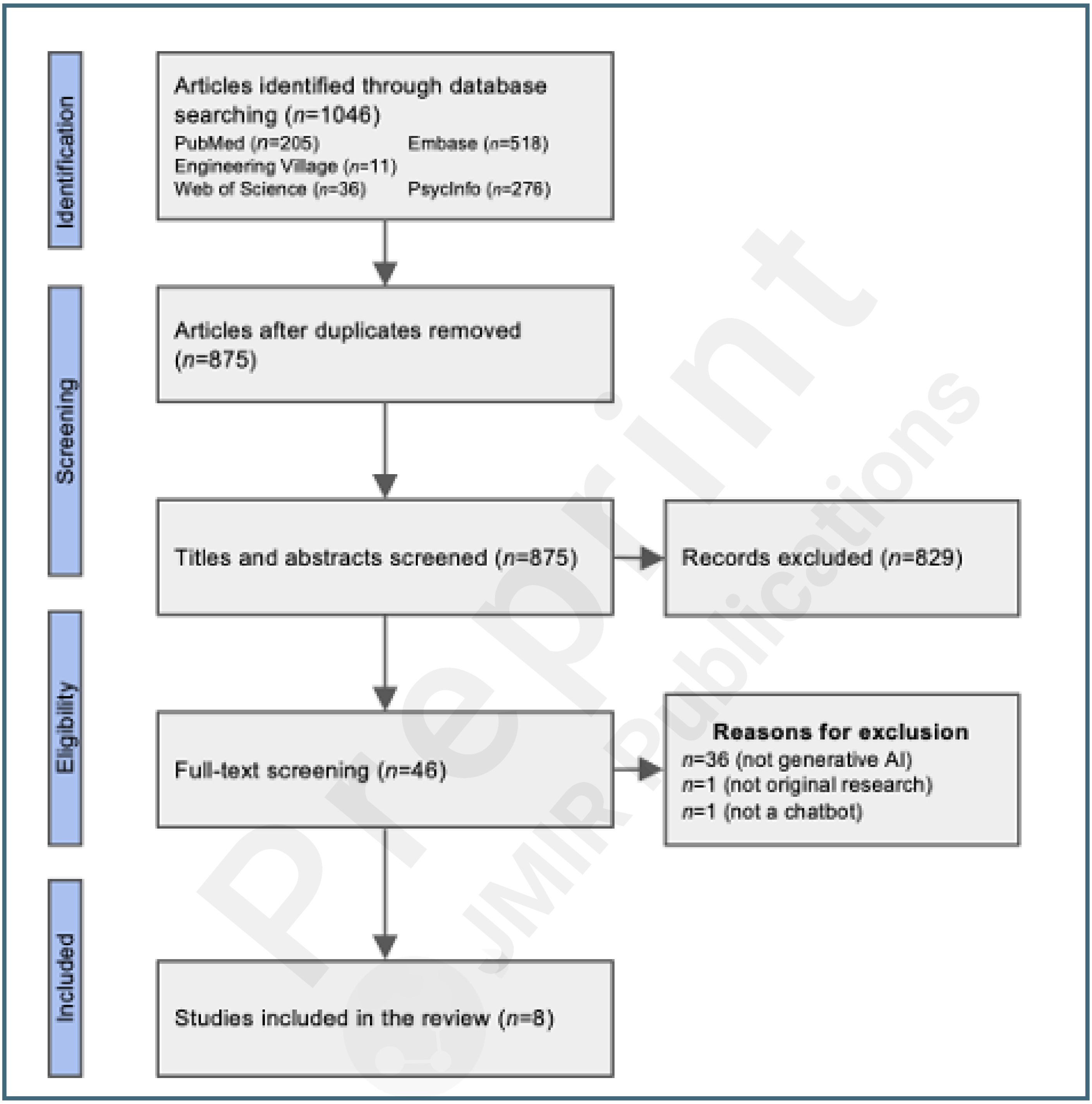
25. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. *arXiv [csCL]*. Published online May 24, 2022. Accessed September 3, 2024. http://arxiv.org/abs/2205.11916
26. Gore S, Dove E. Ethical considerations in the use of artificial intelligence in counselling and psychotherapy training: A student stakeholder perspective—A pilot study. *Couns Psychother Res*. Published online May 15, 2024. doi:10.1002/capr.12770
27. Utay JADM. Guided Imagery as an Effective Therapeutic Technique: A Brief Review of its History and Efficacy Research. 2006;33(1):40-43. Accessed September 10, 2024. https://www.proquest.com/openview/26bdce6b33c205ccdccdc5969b38bfbc/1?cbl=48173&pq-origsite=gscholar
28. *Acute Effects of Stress-Reduction Interactive Guided ImagerySM on Salivary Cortisol in Overweight Latino Adolescents*.
29. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *J Biomed Inform*. 2024;151(104620):104620. doi:10.1016/j.jbi.2024.104620
30. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the chat-GPT model. *Res Sq*. Published online February 28, 2023. doi:10.21203/rs.3.rs-2566942/v1
31. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. 2023;29(3):721-732. doi:10.3350/cmh.2023.0089
32. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv [csCL]*. Published online January 27, 2022. Accessed September 10, 2024. http://arxiv.org/abs/2201.11903
33. Zhang Z, Zhang A, Li M, Smola A. Automatic chain of thought prompting in large language models. *arXiv [csCL]*. Published online October 7, 2022. Accessed September 10, 2024. http://arxiv.org/abs/2210.03493
34. Min S, Lyu X, Holtzman A, et al. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv [csCL]*. Published online February 25, 2022. Accessed September 10, 2024. http://arxiv.org/abs/2202.12837
35. Ahsan MM, Luna SA, Siddique Z. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare (Basel)*. 2022;10(3):541. doi:10.3390/healthcare10030541
36. Zhang Y, Yuan Y, Yao ACC. Meta Prompting for AI systems. *arXiv [csAI]*. Published online November 19, 2023. http://arxiv.org/abs/2311.11482
37. Prochaska JO, DiClemente CC. Transtheoretical therapy: Toward a more integrative model of change. *Psychotherapy (Chic)*. 1982;19(3):276-288. doi:10.1037/h0088437
38. Ferrara E. Should ChatGPT be biased? Challenges and risks of bias in large language models. *arXiv [csCY]*. Published online April 7, 2023. doi:10.5210/fm.v28i11.13346
39. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? . In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM; 2021. doi:10.1145/3442188.3445922
40. Timmons AC, Duong JB, Simo Fiallo N, et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspect Psychol Sci*. 2023;18(5):1062-1096. doi:10.1177/17456916221134490
41. Ryder AG, Yang J, Zhu X, et al. The cultural shaping of depression: somatic symptoms in China, psychological symptoms in North America? *J Abnorm Psychol*. 2008;117(2):300-313. doi:10.1037/0021-843X.117.2.300
42. Sue S. Cultural competency: From philosophy to research and practice. *J Community Psychol*.

2006;34(2):237-245. doi:10.1002/jcop.20095

Supplementary Files

Figures

PRISMA Flowchart.


Identification
Screening
Eligibility
Included
Articles identified through database searching (n=1046)
PubMed (n=205)
Embase (n=518)
Engineering Village (n=11)
Web of Science (n=36)
PsycInfo (n=276)
Articles after duplicates removed (n=875)
Titles and abstracts screened (n=875)
Records excluded (n=829)
Full-text screening (n=46)
Reasons for exclusion
n=36 (not generative AI)
n=1 (not original research)
n=1 (not a chatbot)
Studies included in the review (n=8)

Multimedia Appendixes

Study Characteristics Extraction Table.
URL: http://asset.jmir.pub/assets/c867b3d44df0b9d2bb75ecda26bef80f.docx