Review

Large Language Models for Mental Health Applications: Systematic Review

Zhijun Guo[1], MSc; Alvina Lai[1], DPhil; Johan H Thygesen[1], DPhil; Joseph Farrington[1], MSc; Thomas Keen[1,2], MMathPhil; Kezhi Li[1], DPhil

[1]Institute of Health Informatics University College, London, London, United Kingdom

[2]Great Ormond Street Institute of Child Health, University College London, London, United Kingdom

Corresponding Author:
Kezhi Li, DPhil
Institute of Health Informatics University College, London
222 Euston Road
London, NW1 2DA
United Kingdom
Phone: 44 7859 995590
Email: ken.li@ucl.ac.uk

Abstract

Background: Large language models (LLMs) are advanced artificial neural networks trained on extensive datasets to accurately understand and generate natural language. While they have received much attention and demonstrated potential in digital health, their application in mental health, particularly in clinical settings, has generated considerable debate.

Objective: This systematic review aims to critically assess the use of LLMs in mental health, specifically focusing on their applicability and efficacy in early screening, digital interventions, and clinical settings. By systematically collating and assessing the evidence from current studies, our work analyzes models, methodologies, data sources, and outcomes, thereby highlighting the potential of LLMs in mental health, the challenges they present, and the prospects for their clinical use.

Methods: Adhering to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, this review searched 5 open-access databases: MEDLINE (accessed by PubMed), IEEE Xplore, Scopus, JMIR, and ACM Digital Library. Keywords used were (*mental health* OR *mental illness* OR *mental disorder* OR *psychiatry*) AND (*large language models*). This study included articles published between January 1, 2017, and April 30, 2024, and excluded articles published in languages other than English.

Results: In total, 40 articles were evaluated, including 15 (38%) articles on mental health conditions and suicidal ideation detection through text analysis, 7 (18%) on the use of LLMs as mental health conversational agents, and 18 (45%) on other applications and evaluations of LLMs in mental health. LLMs show good effectiveness in detecting mental health issues and providing accessible, destigmatized eHealth services. However, assessments also indicate that the current risks associated with clinical use might surpass their benefits. These risks include inconsistencies in generated text; the production of hallucinations; and the absence of a comprehensive, benchmarked ethical framework.

Conclusions: This systematic review examines the clinical applications of LLMs in mental health, highlighting their potential and inherent risks. The study identifies several issues: the lack of multilingual datasets annotated by experts, concerns regarding the accuracy and reliability of generated content, challenges in interpretability due to the "black box" nature of LLMs, and ongoing ethical dilemmas. These ethical concerns include the absence of a clear, benchmarked ethical framework; data privacy issues; and the potential for overreliance on LLMs by both physicians and patients, which could compromise traditional medical practices. As a result, LLMs should not be considered substitutes for professional mental health services. However, the rapid development of LLMs underscores their potential as valuable clinical aids, emphasizing the need for continued research and development in this area.

Trial Registration: PROSPERO CRD42024508617; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=508617

Introduction

Mental Health

Mental health, a critical component of overall well-being, is at the forefront of global health challenges [1]. In 2019, an estimated 970 million individuals worldwide experienced mental illness, accounting for 12.5% of the global population [2]. Anxiety and depression are among the most prevalent psychological conditions, affecting 301 million and 280 million individuals, respectively [2]. In addition, 40 million people experienced bipolar disorder, 24 million experienced schizophrenia, and 14 million experienced eating disorders [3]. These mental disorders collectively contribute to an estimated US $5 trillion in global economic losses annually [4]. Despite the staggering prevalence, many cases remain undetected or untreated, with the resources allocated to the diagnosis and treatment of mental illness far less than the negative impact it has on society [5]. Globally, untreated mental illnesses affect 5% of the population in high-income countries and 19% of the population in low- and middle-income countries [3]. The COVID-19 pandemic has further exacerbated the challenges faced by mental health services worldwide [6], as the demand for these services increased while access was decreased [7]. This escalating crisis underscores the urgent need for more innovative and accessible mental health care approaches.

Mental illness treatment encompasses a range of modalities, including medication, psychotherapy, support groups, hospitalization, and complementary and alternative medicine [8]. However, the societal stigma attached to mental illnesses often deters people from seeking appropriate care [9]. Influenced by the fear of judgment and concerns about costly, ineffective treatments [10], many people with mental illness avoid or delay psychotherapy [11]. The COVID-19 crisis and other global pandemics have underscored the importance of digital tools, such as telemedicine and mobile apps, in delivering care during critical times [12]. In this evolving context, large language models (LLMs) present new possibilities for enhancing the delivery and effectiveness of mental health care.

Recent technological advancements have revealed some unique advantages of LLMs in mental health. These models, capable of processing and generating text akin to human communication, provide accessible support directly to users [13]. A study analyzing 2917 Reddit (Reddit, Inc) user reviews found that conversational agents (CAs) powered by LLMs are valued for their nonjudgmental listening and effective problem-solving advice. This aspect is particularly beneficial for individuals considered socially marginalized, as it enables them to be heard and understood without the need for direct social interaction [14]. Moreover, LLMs enhance the accessibility of mental health services, which are notably undersupplied globally [15]. Recent data reveals substantial delays in traditional mental health care delivery; 23% of individuals with mental illnesses report waiting for >12 weeks for face-to-face psychotherapy sessions [16], with 12% waiting for >6 months and 6% waiting for >1 year [16]. In addition, 43% of adults with mental illness indicate that such long waits have exacerbated their conditions [16].

Telemedicine, enhanced by LLMs, offers a practical alternative that expedites service delivery and could flatten traditional health care hierarchies [17]. This includes real-time counseling sessions through CAs that are not only cost-effective but also accessible anytime and from any location. By reducing the reliance on physical visits to traditional health care settings, telemedicine has the potential to decentralize access to medical expertise and diminish the hierarchical structures within the health care system [17]. Mental health chatbots developed using language models, such as Woebot [18] and Wysa [19], have been gaining recognition. Both chatbots follow the principles of cognitive behavioral therapy and are designed to equip users with self-help tools for managing their mental health issues [20]. In clinical practice, LLMs hold the potential to support the automatic assessment of therapists' adherence to evidence-based practices and the development of systems that offer real-time feedback and support for patient homework between sessions [21]. These models also have the potential to provide feedback on psychotherapy or peer support sessions, which is especially beneficial for clinicians with less training and experience [21]. Currently, these applications are still in the proposal stage. Although promising, they are not yet widely used in routine clinical settings, and further evaluation of their feasibility and effectiveness is necessary.

The deployment of LLMs in mental health also poses several risks, particularly for groups considered vulnerable. Challenges such as inconsistencies in the content generated and the production of "hallucinatory" content may mislead or harm users [22], raising serious ethical concerns. In response, authorities such as the World Health Organization have developed ethical guidelines for artificial intelligence (AI) research in health care, emphasizing the importance of data privacy; human oversight; and the principle that AI tools should augment, rather than replace, human practitioners [23]. These potential problems with LLMs in health care have gained considerable industry attention, underscoring the need for a comprehensive and responsible evaluation of LLMs' applications in mental health. The following section will further explore the workings of LLMs and their potential applications in mental health and critically evaluate the opportunities and challenges they introduce.

LLMs in Mental Health

LLMs represent advancements in machine learning, characterized by their ability to understand and generate human-like text with high accuracy [24]. The efficacy of these models is typically evaluated using benchmarks designed to assess their linguistic fidelity and contextual relevance. Common metrics include Bilingual Evaluation Understudy for translation accuracy and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) for summarization tasks [25]. LLMs are characterized

by their scale, often encompassing billions of parameters, setting them apart from traditional language models [26]. This breakthrough is largely due to the transformer architecture, a deep neural network structure that uses a "self-attention" mechanism developed by Vaswani et al [27]. This allows LLMs to process information in parallel rather than sequentially, greatly enhancing speed and contextual understanding [27]. To clearly define the scope of this study concerning LLMs, we specify that an LLM must use the transformer architecture and contain a high number of parameters, traditionally at least 1 billion, to qualify as "large" [28]. This criterion encompasses models such as GPT (OpenAI) and Bidirectional Encoder Representations from Transformers (BERT; Google AI). Although the standard BERT model, with only 0.34 billion parameters [29], does not meet the traditional criteria for "large," its sophisticated bidirectional design and pivotal role in establishing new natural language processing (NLP) benchmarks justify its inclusion among notable LLMs [30]. The introduction of ChatGPT (OpenAI) in 2022 generated substantial public and academic interest in LLMs, underlining their transformative potential within the field of AI [31]. Other state-of-the-art LLMs include Large Language Model Meta AI (LLaMA; Meta AI) and Pathways Language Model (PaLM; Google AI), as illustrated in Table 1 [32-35].

Table 1. Comparative analysis of large language models (LLMs) by parameter size and developer entity. Data were summarized with the latest models up to June 2024, with data for parameters and developers from GPT (OpenAI) to Large Language Model Meta AI (LLaMA; Meta AI) adapted from the study by Thirunavukarasu et al [32].

Model name	Publication date	Parameters (billion)	Developer
Generative Pretrained Transformer (GPT)	June 2018	0.117	OpenAI
Bidirectional Encoder Representations from Transformers (BERT)	October 2018	0.34	Google
GPT-2	January 2019	1.5	OpenAI
Enhanced Representation through Knowledge Integration (ERNIE)	September 2019	0.114	Baidu
Conditional Transformer Language Model (CTRL)	September 2019	1.63	OpenAI
Megatron	September 2019	3.9	NVIDIA
Bidirectional and Auto-Regressive Transformers (BART)	October 2019	0.374	Meta
Turing Natural Language Generation (Turing-NLG)	January 2020	530	Microsoft
GPT-3	June 2020	175	OpenAI
Vision Transformer (ViT)	October 2020	0.632	Google
Inspired by artist Salvador Dalí and Pixar's WALL·E (DALL-E)	October 2020	1.2	OpenAI
Swin Transformer	March 2021	0.197	Microsoft
Wu Dao 2.0	June 2021	1750	Huawei
Jurassic-1	August 2021	178	AI21 Labs
Megatron-Turing Natural Language Generation (MT-NLG)	October 2021	530	Microsoft & Nvidia
Claude	December 2021	52	Anthropic
Generalist Language Model (GLAM)	December 2021	1200	Google
ERNIE 3.0	December 2021	260	Baidu
Guided Language-to-Image Diffusion for Generation and Editing (GLIDE)	December 2021	3.5	OpenAI
Gopher	December 2021	280	DeepMind
Causal Masked Modeling 3 (CM3)	January 2022	13	Meta
Language Model for Dialogue Applications (LaMDA)	January 2022	137	Google
GPT-NeoX	February 2022	20	EleutherAI
Chinchilla	March 2022	70	DeepMind
GopherCite	March 2022	280	DeepMind
DALL-E 2	April 2022	3.5	OpenAI
Flamingo	April 2022	80	DeepMind
Pathways Language Model (PaLM)	April 2022	540	Google
Gato	May 2022	1.2	DeepMind
Open Pretrained Transformer (OPT)	May 2022	175	Meta
Yet Another Language Model (YaLM)	June 2022	100	Yandex
Minerva	June 2022	540	Google
BigScience Large Open-science Open-access Multilingual Language Model (BLOOM)	July 2022	175	Hugging Face
Galactica	November 2022	120	Meta
Alexa Teacher Model (Alexa TM)	November 2022	20	Amazon
Large Language Model Meta AI (LLAMA)	February 2023	65	Meta
GPT-4	March 2023	1760	OpenAI
Cerebras-GPT	March 2023	13	Cerebras
Falcon	March 2023	40	Technology Innovation Institute

Model name	Publication date	Parameters (billion)	Developer
Bloomberg Generative Pretrained Transformer (BloombergGPT)	March 2023	50	Bloomberg
PanGu-2	March 2023	1085	Huawei
OpenAssistant	March 2023	17	LAION
PaLM 2	May 2023	340	Google
Llama 2	July 2023	70	Meta
Falcon 180B	September 2023	180	Technology Innovation Institute
Mistral 7B	September 2023	7.3	Mistral
Claude 2.1	November 2023	200	Anthropic
Grok-1	November 2023	314	xAI
Mixtral 8x7B	December 2023	46.7	Mistral
Phi-2	December 2023	2.7	EleutherAI
Gemma	February 2024	7	Google
DBRX	March 2024	136	Databricks
Llama 3	April 2024	70	Meta AI
Fugaku-LLM	May 2024	13	Fujitsu, Tokyo Institute of Technology, etc
Nemotron-4	June 2024	340	Nvidia

LLMs are primarily designed to learn fundamental statistical patterns of language [36]. Initially, these models were used as the basis for fine-tuning task-specific models rather than training those models from scratch, offering a more resource-efficient approach [37]. This fine-tuning process involves adjusting a pretrained model to a specific task by further training it on a smaller, task-specific dataset [38]. However, developments in larger and more sophisticated models have reduced the need for extensive fine-tuning in some cases. Notably, some advanced LLMs can now effectively understand and execute tasks specified through natural language prompts without extensive task-specific fine-tuning [39]. Instruction fine-tuned models undergo additional training on pairs of user requests and appropriate responses. This training allows them to generalize across various complex tasks, such as sentiment analysis, which previously required explicit fine-tuning by researchers or developers [40]. A key part of the input to these models, such as ChatGPT and Gemini (Google AI), includes a system prompt, often hidden from the user, which guides the model on how to interpret and respond to user prompts. For example, it might direct the model to act as a helpful mental health assistant. In addition, "prompt engineering" has emerged as a crucial technique in optimizing model performance. Prompt engineering involves crafting input texts that guide the model to produce the desired output without additional training. For example, refining a prompt from "Tell me about current events in health care" to "Summarize today's top news stories about technology in health care" provides the model with more specific guidance, which can enhance the relevance and accuracy of its responses [41]. While prompt engineering can be highly effective and reduce the need to retrain the model, it is important to be wary of "hallucinations," a phenomenon where models confidently generate incorrect or irrelevant outputs [42]. This can be particularly challenging in high-accuracy scenarios, such as

health care and medical applications [43-46]. Thus, while prompt engineering reduces the reliance on extensive fine-tuning, it underscores the need for thorough evaluation and testing to ensure the reliability of model outputs in sensitive applications.

The existing literature includes a review of the application of machine learning and NLP in mental health [47], analyses of LLMs in medicine [32], and a scoping review of LLMs in mental health. These studies have demonstrated the effectiveness of NLP for tasks such as text categorization and sentiment analysis [47] and provided a broad overview of LLM applications in mental health [48]. However, a gap remains in systematically reviewing state-of-the-art LLMs in mental health, particularly in the comprehensive assessment of literature published since the introduction of the transformer architecture in 2017.

This systematic review addresses these gaps by providing a more in-depth analysis; evaluating the quality and applicability of studies; and exploring ethical challenges specific to LLMs, such as data privacy, interpretability, and clinical integration. Unlike previous reviews, this study excludes preprints, follows a rigorous search strategy with clear inclusion and exclusion criteria (using Cohen κ to assess the interreviewer agreement), and uses a detailed assessment of study quality and bias (using the Risk of Bias 2 tool) to ensure the reliability and reproducibility of the findings.

Guided by specific research questions, this systematic review critically assesses the use of LLMs in mental health, focusing on their applicability and efficacy in early screening, digital interventions, and clinical settings, as well as the methodologies and data sources used. The findings of this study highlight the potential of LLMs in enhancing mental health diagnostics and

interventions while also identifying key challenges such as inconsistencies in model outputs and the lack of robust ethical guidelines. These insights suggest that, while LLMs hold promise, their use should be supervised by physicians, and they are not yet ready for widespread clinical implementation.

Methods

This systematic review followed the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) guidelines [49]. The protocol was registered on PROSPERO (CRD42024508617). A PRISMA checklist is available in Multimedia Appendix 1.

Search Strategies

The search was initiated on August 3, 2024, and completed on August 6, 2024, by 1 author (ZG). ZG systematically searched 5 databases: MEDLINE, IEEE Xplore, Scopus, JMIR, and ACM Digital Library using the following search keywords: (*mental health* OR *mental illness* OR *mental disorder* OR *psychiatry*) and (*large language models*). These keywords were consistently applied across each database to ensure a uniform search strategy. To conduct a comprehensive and precise search for relevant literature, strategies were tailored for different databases. All metadata were searched in MEDLINE and IEEE Xplore, whereas the search in Scopus was confined to titles, abstracts, and keywords. The JMIR database used the criteria *exact match* feature to refine search results and enhance precision. In the ACM Digital Library database, the search focused on full text. The screening of all citations involved four steps:

1. Initial search. All relevant citations were imported into a Zotero (Corporation for Digital Scholarship) citation manager library.
2. Preliminary inclusion. Citations were initially screened based on predefined inclusion criteria.
3. Duplicate removal. Citations were consolidated into a single group, from which duplicates were eliminated.
4. Final inclusion. The remaining references were carefully evaluated against the inclusion criteria to determine their suitability.

Study Selection and Eligibility Criteria

All the articles that matched the search criteria were double screened by 2 independent reviewers (ZG and KL) to ensure that each article fell within the scope of LLMs in mental health. This process involved the removal of duplicates followed by a detailed manual evaluation of each article to confirm adherence to our predefined inclusion criteria, ensuring a comprehensive and focused review. To quantify the agreement level between the reviewers and ensure objectivity, interrater reliability was calculated using Cohen κ, with a score of 0.84 indicating a good level of agreement. In instances of disagreement, a third reviewer (AL) was consulted to achieve consensus.

To assess the risk of bias, we used the Risk of Bias 2 tool, as recommended for Cochrane Reviews. The results have been visualized in Multimedia Appendix 2. We thoroughly examined each study for potential biases that could impact the validity of the results. These included biases from the randomization process, deviations from intended interventions, missing outcome data, inaccuracies in outcome measurement, and selective reporting of results. This comprehensive assessment ensures the credibility of each study.

The criteria for selecting articles were as follows: we limited our search to English-language publications, focusing on articles published between January 1, 2017, and April 30, 2024. This timeframe was chosen considering the substantial developments in the field of LLMs in 2017, marked notably by the introduction of the transformer architecture, which has greatly influenced academic and public interest in this area.

In this review, the original research articles and available full-text papers have been carefully selected, aiming to focus on the application of LLMs in mental health. To comply with the PRISMA guidelines, articles that have not been published in a peer-reviewed venue, including those only available on a preprint server, were excluded. Owing to the limited literature specifically addressing the mental health applications of LLMs, we included review articles to ensure a comprehensive perspective. The selection criteria focused on direct applications, expert evaluations, and ethical considerations related to the use of LLMs in mental health contexts, with the goal of providing a thorough analysis of this rapidly developing field.

Information Extraction

The data extraction process was jointly conducted by 2 reviewers (ZG and KL), focusing on examining the application scenarios, model architecture, data sources, methodologies used, and main outcomes from selected studies on LLMs in mental health.

Initially, we categorized each study to determine its main objectives and applications. The categorization process was conducted in 2 steps. First, after reviewing all the included articles, we grouped them into 3 primary categories: detection of mental health conditions and suicidal ideation through text, LLM use for mental health CAs, and other applications and evaluation of the LLMs in mental health. In the second step, we performed a more detailed categorization. After a thorough, in-depth reading of each article within these broad categories, we refined the classifications based on the specific goals of the studies. Following this, we summarized the main model architectures of the LLMs used and conducted a thorough examination of data sources, covering both public and private datasets. We noted that some review articles lacked detail on dataset content; therefore, we focused on providing comprehensive information on public datasets, including their origins and sample sizes. We also investigated the various methods used across different studies, including data collection strategies and analytic methodologies. We examined their comparative structures and statistical techniques to offer a clear understanding of how these methods are applied in practice.

Finally, we documented the main outcomes of each study, recording significant results and aligning them with relevant performance metrics and evaluation criteria. This included providing quantitative data where applicable to underscore these findings. We used a narrative approach to synthesize the information, integrating and comparing results from various studies to emphasize the efficacy and impact of LLMs on mental health. This narrative synthesis allowed us to highlight the

efficacy and impact of LLMs in mental health, providing quantitative data where applicable to underscore these findings. The results of this analysis are presented in Tables S1-S3 in Multimedia Appendix 3 [14,50-131], each corresponding to 1 of the primary categories.

Results

Strategy and Screening Process

The PRISMA diagram of the systematic screening process can be seen in Figure 1. Our initial search across 5 academic databases, namely, MEDLINE, IEEE Xplore, Scopus, JMIR, and ACM Digital Library, yielded 14,265 papers: 907 (6.36%) from MEDLINE, 102 (0.72%) from IEEE Xplore, 204 (1.43%) from Scopus, 211 (1.48%) from JMIR, and 12,841 (90.02%) from ACM Digital Library. After duplication, 97.91% (13,967/14,265) of the unique papers were retained. Subsequent screening was based on predefined inclusion and exclusion criteria, narrowing down the selection to 0.29% (40/13,967) of the papers included in this review. The reasons for the full-text exclusion of 61 papers can be found in Multimedia Appendix 4.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow of the selection process. LLM: large language model.
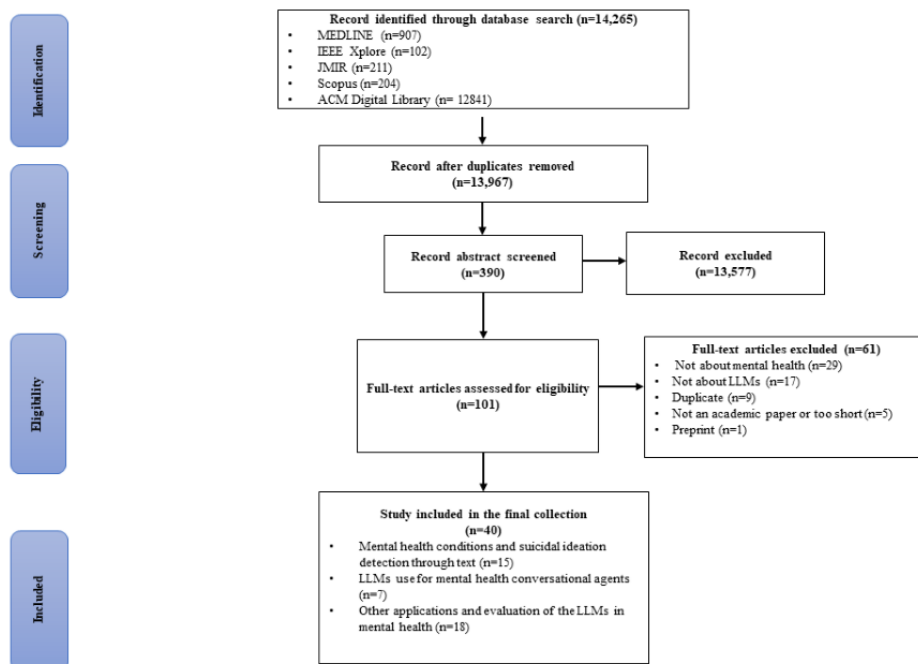
In our review of the literature, we classified the included articles into 3 broad categories: detection of mental health conditions and suicidal ideation through text (15/40, 38%), LLMs' use for mental health CAs (7/40, 18%), and the other applications and evaluation of the LLMs in mental health (18/40, 45%). The first category investigates the potential of LLMs for the early detection of mental illness and suicidal ideation via social media and other textual sources. Early screening is highlighted as essential for preventing the progression of mental disorders and mitigating more severe outcomes. The second category assesses LLM-supported CAs used as teletherapeutic interventions for mental health issues, such as loneliness, with a focus on evaluating their effectiveness and validity. The third category covers a broader range of LLM applications in mental health, including clinical uses such as decision support and therapy enhancement. It aims to assess the overall effectiveness, utility, and ethical considerations associated with LLMs in these settings. All selected articles are summarized in Tables S1-S3 in Multimedia Appendix 3 according to the 3 categories.

Mental Health Conditions and Suicidal Ideation Detection Through Text

Early intervention and screening are crucial in mitigating the global burden of mental health issues [132]. We examined the performance of LLMs in detecting mental health conditions and suicidal ideation through textual analysis. Of 40 articles, 6 (15%) assessed the efficacy of early screening for depression using LLMs [50,57,60,61,66,68], while another (1/40, 2%) simultaneously addressed both depression and anxiety [60]. One comprehensive study examined various psychiatric conditions, including depression, social anxiety, loneliness, anxiety, and other prevalent mental health issues [69]. Two (5%) of the 40 articles assessed and compared the ability of LLMs to perform sentiment and emotion analysis [75,81], and 5 (12%) articles focused on the capability of LLMs to analyze textual content for detecting suicidal ideation [54,65,70,72,78]. Most studies (10/40, 25%) used BERT and its variants as one of the primary models [50,54,57,62,65,66,68,69,75,78], while GPT models were also commonly used (8/40, 20%) [57,60,61,66,70,72,78,81]. Most training data (10/40, 25%) comprised social media posts [50,54,62,65,68,69,72,75,78,81] from platforms such as Twitter (Twitter, Inc), Reddit, and Sina Weibo (Sina corporation), covering languages such as English, Malay dialects, Chinese, and Portuguese. In addition, 5 (12%) of the 40 studies used datasets consisting of clinical transcripts and patient interviews [50,57,60,61,66], providing deeper insights into LLM applications in clinical mental health settings.

In studies focusing on early screening for depression, comparing results horizontally is challenging due to variations in datasets, training methods, and models across different investigations. Nonetheless, substantial evidence supports the significant potential of LLMs in detecting depression from text-based data. For example, Danner et al [57] conducted a comparative analysis using a convolutional neural network on the Distress Analysis Interview Corpus-Wizard of Oz dataset, achieving F_1-scores of 0.53 and 0.59; however, their use of GPT-3.5 demonstrated superior performance, with an F_1-score of 0.78. Another study involving the E-Distress Analysis Interview Corpus dataset (an extension of Distress Analysis Interview Corpus-Wizard of Oz) used the Robustly Optimized BERT Approach for Depression Detection to predict the Patient Health Questionnaire-8 scores from textual data. This approach identified 3 levels of depression and achieved the lowest mean absolute error of 3.65 in Patient Health Questionnaire–8 scores [66].

LLMs play an important role in sentiment analysis [75,81], which categorizes text into overall polarity classes, such as positive, neutral, negative, and occasionally mixed, and emotion classification, which assigns labels such as "joy," "sadness," "anger," and "fear" [75]. These analyses enable the detection of emotional states and potential mental health issues from textual data, facilitating early intervention [133]. Stigall et al [75] demonstrated the efficacy of these models, with their study showing that Emotion-aware BERT Tiny, a fine-tuned variant of BERT, achieved an accuracy of 93.14% in sentiment analysis and 85.46% in emotion analysis. This performance surpasses that of baseline models, including BERT-Base Cased and BERTTiny-Pretrained [75], underscoring the advantages and validity of fine-tuning in enhancing model performance. LLMs have also demonstrated robust accuracy in detecting and classifying a range of mental health syndromes, including social anxiety, loneliness, and generalized anxiety. Vajre et al [69] introduced PsychBERT, developed using a diverse training dataset from both social media texts and academic literature, which achieved an F_1-score of 0.63, outperforming traditional deep learning approaches such as convolutional neural networks and long short-term memory networks, which recorded F_1-scores of 0.57 and 0.51, respectively [69]. In research on detecting suicidal ideation using LLMs, Diniz et al [54] showcased the high efficacy of the BERTimbau large model within a non-English (Portuguese) context, achieving an accuracy of 0.955, precision of 0.961, and an F_1-score of 0.954. The assessment of the BERT model by Metzler et al [65] found that it correctly identified 88.5% of tweets as suicidal or off-topic, performing comparably to human analysts and other leading models. However, Levkovich et al [70] noted that while GPT-4 assessments of suicide risk closely aligned with those by mental health professionals, it overestimated suicidal ideation. These results underscore that while LLMs have the potential to identify tweets reflecting suicidal ideation with accuracy comparable to psychological professionals, extensive follow-up studies are required to establish their practical application in clinical settings.

LLMs in Mental Health CAs

In the growing field of mental health digital support, the implementation of LLMs as CAs has exhibited both promising advantages [14,84,91,96] and significant challenges [92,96]. The studies by Ma et al [14] and Heston [96] demonstrate the effectiveness of CAs powered by LLMs in providing timely, nonjudgmental mental health support. This intervention is particularly important for those who lack ready access to a therapist due to constraints such as time, distance, and work, as well as for certain populations considered socially marginalized, such as older adults who experience chronic loneliness and a lack of companionship [14,97]. The qualitative analysis of user interactions on Reddit by Ma et al [14] highlights that LLMs encourage users to speak up and boost their confidence by providing personalized and responsive interactions. In addition, VHope, a DialoGPT-enabled mental health CA, was evaluated by 3 experts who rated its responses as 67% relevant, 78% human-like, and 79% empathetic [84]. Another study found that after observing 717 evaluations by 100 participants on 239 autism-specific questions, 46.86% of evaluators preferred responses of the chief physicians, whereas 34.87% preferred the responses of GPT-4, and 18.27% favored the responses of Enhanced Representation through Knowledge Integration Bot (ERNIE Bot; version 2.2.3; Baidu, Inc). Moreover, ChatGPT (mean 3.64, 95% CI 3.57-3.71) outperformed physicians (mean 3.13, 95% CI 3.04-3.21) in terms of empathy [98], indicating that LLM-powered CAs are not only effective but also acceptable by users. These findings highlight the potential for LLMs to complement mental health intervention systems and provide valuable medical guidance.

The development and implementation of a non-English CA for emotion capture and categorization was explored in a study by Zygadlo et al [92]. Faced with a scarcity of Polish datasets, the study adapted by translating an existing database of personal conversations from English into Polish, which decreased the accuracy in tasks from 90% in English to 80% in Polish [92]. While the performance remained commendable, it highlighted the challenges posed by the lack of robust datasets in languages other than English, impacting the effectiveness of CAs across different linguistic environments. However, findings by He et al [98] suggest that the availability of language-specific datasets is not the sole determinant of CA performance. In their study, although ERNIE Bot was trained in Chinese and ChatGPT in English, ChatGPT demonstrated greater empathy for Chinese users [98]. This implies that factors beyond the training language and dataset availability, such as model architecture or training methodology, can also affect the empathetic responsiveness of LLMs, underscoring the complexity of human-AI interaction.

Meanwhile, the reliability of LLM-driven CAs in high-risk scenarios remains a concern [14,96]. An evaluation of 25 CAs found that in tests involving suicide scenarios, only 2 included suicide hotline referrals during the conversation [96]. This suggests that while these CAs can detect extreme emotions, few are equipped to take effective preventive measures. Furthermore, CAs often struggle with maintaining consistent communication due to limited memory capacity, leading to disruptions in conversation flow and negatively affecting user experience [14].

Other Applications and Evaluation of the LLMs in Mental Health

ChatGPT has gained attention for its unparalleled ability to generate human-like text and analyze large amounts of textual data, attracting the interest of many researchers and practitioners [100]. Numerous evaluations of LLMs in mental health have focused on ChatGPT, exploring its utility across various scenarios such as clinical diagnosis [100,106,111], treatment planning [106,128,131], medication guidance [105,109,129], patient management [106], psychiatry examinations [118], and psychology education [102], among others [107,110,127,130].

Research has highlighted ChatGPT's accuracy in diagnosing various psychiatric conditions [106,110,111,126]. For example, Franco D'Souza et al [100] evaluated ChatGPT's responses to 100 clinical psychiatric cases, awarding it an "A" rating in 61 cases, with no errors in the diagnoses of different psychiatric disorders and no unacceptable responses, underscoring ChatGPT's expertise and interpretative capacity in psychiatry. Further supporting this, Schubert et al [118] assessed the performance of ChatGPT 4.0 using neurology board-style examination questions, finding that it answered 85% of the questions correctly, surpassing the average human performance of 73.8%. Meanwhile, in a study of LLMs regarding the prognosis and long-term outcomes of depression, GPT-4, Claude (Anthropic), and Bard (Google AI) showed strong agreement with mental health professionals. They all recommended a combination of psychotherapy and antidepressant medication in every case [130]. This not only proves the reliability of LLMs for mental health assessment but also highlights their usefulness in providing valuable support and guidance for individuals seeking information or coping with mental illness.

However, the direct deployment of LLMs, such as ChatGPT, in clinical settings carries inherent risks. The outputs of LLMs are heavily influenced by prompt engineering, which can lead to inconsistencies that undermine clinical reliability [102,105-107,109]. For example, Farhat et al [105] conducted a critical evaluation of ChatGPT's ability to generate medication guidelines through detailed cross-questioning and noted that altering prompts substantially changed the responses. While ChatGPT typically provided helpful advice and recommended seeking expert consultation, it occasionally produced inappropriate medication suggestions. Perlis et al [129] verified this, showing that GPT-4 Turbo suggested medications that were considered less efficient choices or contraindicated by experts in 12% of the cases. Moreover, LLMs often lack the necessary clinical judgment capabilities. This issue was highlighted in the study by Grabb [109], which revealed that despite built-in safeguards, ChatGPT remains susceptible to generating extreme and potentially hazardous recommendations. A particularly alarming example was ChatGPT advising a patient with depression to engage in high-risk activities such as bungee jumping as a means of seeking pleasure [109]. These LLMs depend on prompt engineering [102,105,109], which means their responses can vary widely depending on the wording and context of the prompts given. The system prompts, which are predefined instructions given to the model, and the prompts used by the experimental team, such as those in the study by Farhat et al [105], guide the behavior of ChatGPT and similar LLMs. These prompts are designed to accommodate a variety of user requests within legal and ethical boundaries. However, while these boundaries are intended to ensure safe and appropriate responses, they often fail to align with the nuanced sensitivities required in psychological contexts. This mismatch underscores a significant deficiency in the clinical judgment and control of LLMs within sensitive mental health settings.

Further research into other LLMs in the mental health sector has shown a range of capabilities and limitations. For example, a study by Sezgin et al [111] highlighted Language Model for Dialogue Applications' (LaMDA's) proficiency in managing complex inquiries about postpartum depression that require medical insight or nuanced understanding; however, they pointed out its challenges with straightforward, factual questions, such as "What are antidepressants?" [111]. Assessments of LLMs such as LLaMA-7B, ChatGLM-6B, and Alpaca, involving 50 interns specializing in mental illness, received favorable feedback regarding the fluency of these models in a clinical context, with scores >9.5 out of 10. However, the results also indicated that the responses of these LLMs often failed to address mental health issues adequately, demonstrated limited professionalism, and resulted in decreased usability [116]. Similarly, a study on psychiatrists' perceptions of using LLMs such as Bard and Bing AI (Microsoft Corp) in mental health care revealed mixed feelings. While 40% of physicians indicated that they would use such LLMs to assist in answering clinical questions, some expressed serious concerns about their reliability, confidentiality, and potential to damage the patient-physician relationship [130].

Discussion

Principal Findings

In the context of the wider prominence of LLMs in the literature [14,50,57,60,61,69,96,130], this study supports the assertion that interest in LLMs is growing in the field of mental health. Figure 2 indicates an increase in the number of mental health studies using LLMs, with a notable surge observed in 2023 following the introduction of ChatGPT in late 2022. Although we included articles only up to the end of April 2024, it is evident that the number of articles related to LLMs in the field of mental health continues to show a steady increase in 2024. This marks a substantial change in the discourse around LLMs, reflecting their broader acceptance and integration into various aspects of mental health research and practice. The progression from text analysis to a diverse range of applications highlights the academic community's recognition of the multifaceted uses of LLMs. LLMs are increasingly used for complex psychological assessments, including early screening, diagnosis, and therapeutic interventions.

Figure 2. Number of articles included in this literature review, grouped by year of publication and application field. The black line indicates the total number of articles in each year. CA: conversational agent.
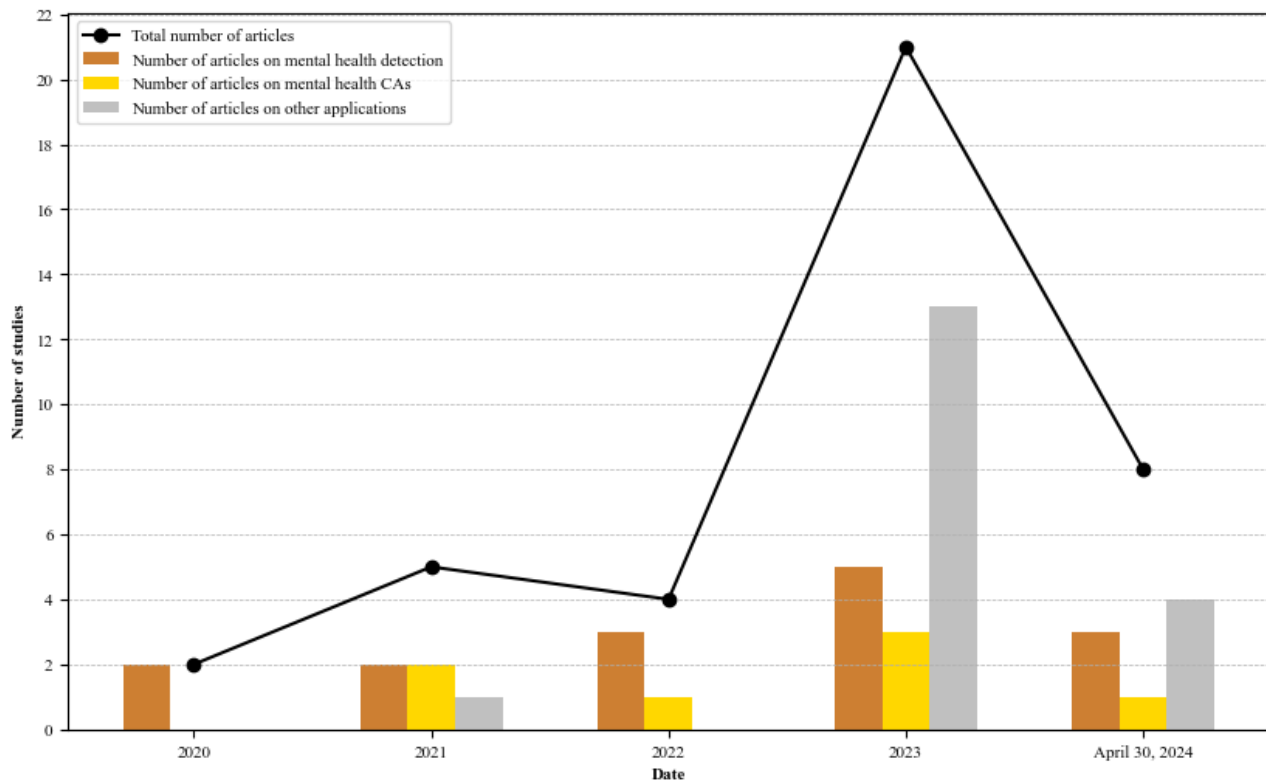
The findings of this study demonstrate that LLMs are highly effective in analyzing textual data to assess mental states and identify suicidal ideation [50,54,57,60,61,65,66,68,69,72,78], although their categorization often tends to be binary [50,54,65,68,69,72,78]. These LLMs possess extensive knowledge in the field of mental health and are capable of generating empathic responses that closely resemble human interactions [97,98,107]. They show great potential for providing mental health interventions with improved prognoses [50,96,110,127,128,131], with the majority being recognized by psychologists for their appropriateness and accuracy [98,100,129]. The careful and rational application of LLMs can enhance mental health care efficiently and at a lower cost, which is crucial in areas with limited health care capacity. However, there are currently no studies available that provide evaluative evidence to support the clinical use of LLMs.

Limitations

Limitations of Using LLMs in Mental Health

On the basis of the works of literature, the strengths and weaknesses of applying the LLMs in mental health are summarized in Multimedia Appendix 5.

LLMs have a broad range of applications in the mental health field. These models excel in user interaction, provide empathy and anonymity, and help reduce the stigma associated with mental illness [14,107], potentially encouraging more patients to participate in treatment. They also offer a convenient, personalized, and cost-effective way for individuals to access mental health services at any time and from any location, which can be particularly helpful for populations considered socially isolated, especially older adults [60,84,97]. In addition, LLMs can help reduce the burden of care during times of severe health care resource shortages and patient overload, such as during the COVID-19 pandemic [68]. Although previous research has highlighted the potential of LLMs in mental health, it is evident that they are not yet ready for clinical use due to unresolved technical risks and ethical issues.

The use of LLMs in mental health, particularly those fine-tuned for specific tasks such as ChatGPT, reveals clear limitations. The effectiveness of these models heavily depends on the specificity of user-generated prompts. Inappropriate or imprecise prompts can disrupt the conversation's flow and diminish the model's effectiveness [75,96,105,107,109]. Even small changes in the content or tone of prompts can sometimes lead to significant variations in responses, which can be particularly problematic in health care settings where interpretability and consistency are critical [14,105,107]. Furthermore, LLMs lack clinical judgment and are not equipped to handle emergencies [95,108]. While they can generally capture extreme emotions and recognize scenarios requiring urgent action, such as suicide ideation [54,65,70,72,78], they often fail to provide direct, practical measures, typically only advising users to seek professional help [96]. In addition, the inherent bias in LLM training data [66,106] can lead to the propagation of stereotypical, discriminatory, or biased viewpoints. This bias can also give rise to hallucinations, that is, LLMs producing erroneous or misleading information [105,131]. Furthermore, hallucinations may stem from overfitting the training data or a lack of context understanding [134]. Such inaccuracies can have serious consequences, such as providing incorrect medical information, reinforcing harmful stereotypes, or failing to

recognize and appropriately respond to mental health crises [131]. For example, an LLM might reinforce a harmful belief held by a user, potentially exacerbating their mental health issues. It could also generate nonfactual, overly optimistic, or pessimistic medical advice, delaying appropriate professional intervention. These issues could undermine the integrity and fairness of social psychology [102,105,106,110].

Another critical concern is the "black box" nature of LLMs [105,107,131]. This lack of interpretability complicates the application of LLMs in mental health, where trustworthiness and clarity are important. When we talk about neural networks as black boxes, we know details such as what they were trained with, how they were trained, and what the weights are. However, with many new LLMs, such as GPT-3.5 and 4, researchers and practitioners often access the models via web interfaces or application programming interfaces without complete knowledge of the training data, methods, and model updates. This situation not only presents the traditional challenges associated with neural networks but also has all these additional problems that come from the "hidden" model.

Ethical concern is another significant challenge associated with applying LLMs in mental health. Debates are emerging around issues such as digital personhood, informed consent, the risk of manipulation, and the appropriateness of AI in mimicking human interactions [60,102,105,106,135]. A primary ethical concern is the potential alteration of the traditional therapist-patient relationship. Individuals may struggle to fully grasp the advantages and disadvantages of LLM derivatives, often choosing these options for their lower cost or greater convenience. This shift could lead to an increased reliance on the emotional support provided by AI [14], inadvertently positioning AI as the primary diagnostician and decision maker for mental health issues, thereby undermining trust in conventional health care settings. Moreover, therapists may become overly reliant on LLM-generated answers and use them in clinical decision-making, overlooking the complexities involved in clinical assessment. This reliance could compromise their professional judgment and reduce opportunities for in-depth engagement with patients [17,129,130]. Furthermore, the dehumanization and technocratic nature of mental health care has the potential to depersonalize and dehumanize patients [136], where decisions are more driven by algorithms than by human insight and empathy. This can lead to decisions becoming mechanized, lacking empathy, and detached from ethics [137]. AI systems may fail to recognize or adequately interpret the subtle and often nonverbal cues, such as the tone of voice, facial expressions, and the emotional weightage behind words, which are critical in traditional therapeutic settings [136]. These cues are essential for comprehensively understanding a patient's condition and providing empathetic care.

In addition, the current roles and accuracy of LLMs in mental health are limited. For instance, while LLMs can categorize a patient's mood or symptoms, most of these categorizations are binary, such as *depressed* or *not depressed* [50,65]. This oversimplification can lead to misdiagnoses. Data security and user privacy in clinical settings are also of utmost concern [14,54,60,96,130]. Although approximately 70% of psychiatrists believe that managing medical documents will be more efficient

using LLMs, many still have concerns about their reliability and privacy [97,130,131]. These concerns could have a devastating impact on patient privacy and undermine the trust between physicians and patients if confidential treatment records stored in LLM databases are compromised. Beyond the technical limitations of AI, the current lack of an industry-benchmarked ethical framework and accountability system hinders the true application of LLMs in clinical practice [131].

Limitations of the Selected Articles

Several limitations were identified in the literature review. A significant issue is the age bias present in the social media data used for depression and mental health screening. Social media platforms tend to attract younger demographics, leading to an underrepresentation of older age groups [65]. Furthermore, most studies have focused on social media platforms, such as Twitter, primarily used by English-speaking populations, which may result in a lack of insight into mental health patterns in non–English-speaking regions. Our review included studies in Polish, Chinese, Portuguese, and Malay, all of which highlighted the significant limitations of LLMs caused by the availability and size of databases [54,61,92,98,116]. For instance, due to the absence of a dedicated Polish-language mental health database, a Polish study had to rely on machine-translated English databases [92]. While the LLMs achieve 80% accuracy in categorizing emotions and moods in Polish, this is still lower than the 90% accuracy observed in the original English dataset. This discrepancy highlights that the accuracy of LLMs can be affected by the quality of the database.

Another limitation of this study is the low diversity of LLMs studied. Although we used "large language models" as keywords in our search phase, the vast majority of identified studies (39/40, 98%) focused on BERT and its variants, as well as the GPT model, as one of the models studied. Therefore, this review provides only a limited picture of the variability expected in applicability between different LLMs. In addition, the rapid development of LLM technologies presents a limitation; this study can only reflect current advancements and may not encompass future advances or the full potential of LLMs. For instance, in tests involving psychologically relevant questions and answers, GPT-3.5 achieved an accuracy of 66.8%, while GPT-4.0 reached an accuracy of 85%, compared to the average human score of 73.8% [118]. Evaluating ChatGPT at different stages separately and comparing its performance to that of humans can lead to varied conclusions. In the assessment of prognosis and treatment planning for depression using LLMs, GPT 3.5 demonstrated a distinctly pessimistic prognosis that differed significantly from those of GPT-4, Claude, Bard, and mental health professionals [128]. Therefore, continuous monitoring and evaluation are essential to fully understand and effectively use the advancements in LLM technologies.

Opportunities and Future Work

Implementing technologies involving LLMs within the health care provision of real patients demands thorough and multifaceted evaluations. It is imperative for both industry and researchers to not let rollout exceed proportional requirements for evidence on safety and efficacy. At the level of the service provider, this includes providing explicit warnings to the public

to discourage mistaking LLM functionality for clinical reliability. For example, GPT-4 introduced the ability to process and interpret image inputs within conversational contexts, leading OpenAI to issue an official warning that GPT-4 is not approved for analyzing specialized medical images such as computed tomography scans [138].

A key challenge to address in LLM research is the tendency to produce incoherent text or hallucinations. Future efforts could focus on training LLMs specifically for mental health applications, using datasets with expert labeling to reduce bias and create specialized mental health lexicons [84,102,116]. The creation of specialized datasets could take advantage of the customizable nature of LLMs, fostering the development of models that cater to the distinct needs of varied demographic groups. For instance, unlike models designed for health care professionals that assist in tasks such as data documentation, symptom analysis, medication management, and postoperative care, LLMs intended for patient interaction might be trained with an emphasis on empathy and comfortable dialogue.

Another critical concern is the problem of outdated training data in LLMs. Traditional LLMs, such as GPT-4 (with a cutoff date up to October 2023), rely on potentially outdated training data, limiting their ability to incorporate recent events or information. This can compromise the accuracy and relevance of their responses, leading to the generation of uninformative or incorrect answers, known as "hallucinations" [139]. Retrieval-augmented generation (RAG) technology offers a solution by retrieving facts from external knowledge bases, ensuring that LLMs use the most accurate and up-to-date information [140]. By searching for relevant information from numerous documents, RAG enhances the generation process with the most recent and contextually relevant content [141]. In addition, RAG includes evidence-based information, increasing the reliability and credibility of LLM responses [139].

To further enhance the reliability of LLM content and minimize hallucinations, recent studies suggest adjusting model parameters, such as the "temperature" setting [142-144]. The temperature parameter influences the randomness and predictability of outputs [145]. Lowering the temperature typically results in more deterministic outputs, enhancing coherence and reducing irrelevant content [146]. However, this adjustment can also limit the model's creativity and adaptability, potentially making it less effective in scenarios requiring diverse or nuanced responses. In mental therapy, where nuanced and sensitive responses are essential, maintaining an optimal balance is crucial. While a lower temperature can ensure accuracy, which is important for tasks such as clinical documentation, it may not suit therapeutic dialogues where personalized engagement is key. Low temperatures can lead to repetitive and impersonal responses, reducing patient engagement and therapeutic effectiveness. To mitigate these risks, regular updates of the model incorporating the latest therapeutic practices and clinical feedback are essential. Such updates could refine the model's understanding and response mechanisms, ensuring it remains a safe and effective tool for mental health care. Nevertheless, determining the "optimal" temperature setting is challenging, primarily due to the variability in tasks and interaction contexts, which require different levels of creativity and precision.

Data privacy is another important area of concern. Many LLMs, such as ChatGPT and Claude, involve sending data to third-party servers, which poses the risk of data leakage. Current studies have found that LLMs can be enhanced by privacy-enhancing techniques, such as zero-knowledge proofs, differential privacy, and federated learning [147]. In addition, privacy can be preserved by replacing identifying information in textual data with generic tokens. For example, when recording sensitive information (eg, names, addresses, or credit card numbers), using alternatives to mask tokens can help protect user data from unauthorized access [148]. This obfuscation technique ensures that sensitive user information is not stored directly, thereby enhancing data security.

The lack of interpretability in LLM decision-making is another crucial area for future research on health care applications. Future research should examine the models' architecture, training, and inferential processes for clearer understanding. Detailed documentation of training datasets, sharing of model architectures, and third-party audits would ideally form part of this undertaking. Investigating techniques such as attention mechanisms and modular architectures could illuminate aspects of neural network processing. The implementation of knowledge graphs might help in outlining logical relationships and facts [149]. In addition, another promising approach involves creating a dedicated embedding space during training, guided by an LLM. This space aligns with a causal graph and aids in identifying matches that approximate counterfactuals [146].

Before deploying LLMs in mental health settings, a comprehensive assessment of their reliability, safety, fairness, abuse resistance, interpretability, compliance with social norms, robustness, performance, linguistic accuracy, and cognitive ability is essential. It is also crucial to foster collaborative relationships among mental health professionals, patients, AI researchers, and policy makers. LLMs, for instance, have demonstrated initial competence in providing medication advice; however, their responses can sometimes be inconsistent or include inappropriate suggestions. As such, LLMs require professional oversight and should not be used independently. Nevertheless, when used as decision aids, LLMs have the potential to enhance health care efficiency. This study calls on developers of LLMs to collaborate with authoritative regulators in actively developing ethical guidelines for AI research in health care. These guidelines should aim to adopt a balanced approach that considers the multifaceted nature of LLMs and ensures their responsible integration into medical practice. They are expected to become industry benchmarks, facilitating the future development of LLMs in mental health.

Conclusions

This review examines the use of LLMs in mental health applications, including text-based screening for mental health conditions, detection of suicidal ideation, CAs, clinical use, and other related applications. Despite the potential of LLMs, challenges such as the production of hallucinatory or harmful information, output inconsistency, and ethical concerns remain. Nevertheless, as technology advances and ethical guidelines improve, LLMs are expected to become increasingly integral

and valuable in mental health services, providing alternative   solutions to this global health care issue.

Acknowledgments

Data Availability

The authors ensure that all pertinent data have been incorporated in the manuscript and the multimedia appendices. For access to the research data, interested parties may contact the corresponding author (KL) subject to a reasonable request.

Authors' Contributions

ZG and KL contributed to the conception and design of the study. ZG, KL, and AL contributed to the development of the search strategy. Database search outputs were screened by ZG, and data were extracted by ZG and KL. An assessment of the risk of bias in the included studies was performed by ZG and KL. ZG completed the literature review, collated the data, performed the data analysis, interpreted the results, and wrote the first draft of the manuscript. KL, AL, JHT, JF, and TK reviewed the manuscript and provided multiple rounds of guidance in the writing of the manuscript. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 27 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Risk of bias assessment.
[DOCX File , 559 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Summary of the 40 selected articles from the literature on large language models in mental applications, categorized into each group.
[DOCX File , 78 KB-Multimedia Appendix 3]

Multimedia Appendix 4

List of the studies excluded at the full-text screening stage.
[DOCX File , 30 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Summary of the strengths and weaknesses of applying the large language models in mental health.
[DOCX File , 24 KB-Multimedia Appendix 5]

References

1. Mental health. World Health Organization. Jun 17, 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response [accessed 2024-04-15]
2. Mental disorders. World Health Organization. Jun 8, 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [accessed 2024-04-15]
3. MHPSS worldwide: facts and figures. Government of the Netherlands. URL: https://www.government.nl/topics/mhpss/mhpss-worldwide-facts-and-figures [accessed 2024-04-15]
4. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine. Dec 2022;54:101675. [FREE Full text] [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]

5.   Zhang W, Yang C, Cao Z, Li Z, Zhuo L, Tan Y, et al. Detecting individuals with severe mental illness using artificial intelligence applied to magnetic resonance imaging. EBioMedicine. Apr 2023;90:104541. [FREE Full text] [doi: 10.1016/j.ebiom.2023.104541] [Medline: 36996601]

6.   Mental health and COVID-19: early evidence of the pandemic's impact: scientific brief, 2 March 2022. World Health Organization. Mar 2, 2022. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1 [accessed 2024-04-15]

7.   Duden GS, Gersdorf S, Stengler K. Global impact of the COVID-19 pandemic on mental health services: a systematic review. J Psychiatr Res. Oct 2022;154:354-377. [FREE Full text] [doi: 10.1016/j.jpsychires.2022.08.013] [Medline: 36055116]

8.   Mental health treatments. Mental Health America. URL: https://mhanational.org/mental-health-treatments [accessed 2024-04-15]

9.   Stigma, prejudice and discrimination against people with mental illness. American Psychiatric Association. URL: https://www.psychiatry.org/patients-families/stigma-and-discrimination [accessed 2024-04-15]

10.   Nietzel MT. Almost half of Americans don't seek professional help for mental disorders. Forbes. May 24, 2021. URL: https://www.forbes.com/sites/michaeltnietzel/2021/05/24/why-so-many-americans-do-not-seek-professional-help-for-mental-disorders/?sh=55b4ec4b3de7 [accessed 2024-04-15]

11.   Why do people avoid mental health treatment? Thriveworks. Aug 9, 2022. URL: https://thriveworks.com/blog/why-people-avoid-mental-health-treatment/ [accessed 2024-04-15]

12.   Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health. Mar 26, 2020;7(3):e18848. [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]

13.   Kumar V, Srivastava P, Dwivedi A, Budhiraja I, Ghosh D, Goyal V, et al. Large-language-models (LLM)-based AI chatbots: architecture, in-depth analysis and their performance evaluation. In: Santosh KC, Makkar A, Conway M, Singh AK, Vacavant A, Abou el Kalam A, et al, editors. Recent Trends in Image Processing and Pattern Recognition. Cham, Switzerland. Springer; 2024.

14.   Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. AMIA Annu Symp Proc. 2023;2023:1105-1114. [FREE Full text] [Medline: 38222348]

15.   Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. arXiv. Preprint posted online on April 6, 2023. [FREE Full text] [doi: 10.18653/v1/2023.emnlp-main.370]

16.   Patients turning to A and E as wait times for NHS mental health treatment spiral. The Guardian. Oct 10, 2022. URL: https://www.theguardian.com/society/2022/oct/10/nhs-mental-health-patients-wait-times [accessed 2024-04-16]

17.   Elyoseph Z, Gur T, Haber Y, Simon T, Angert T, Navon Y, et al. An ethical perspective on the democratization of mental health with generative artificial intelligence. JMIR Preprints. Preprint posted online on March 2, 2024. [FREE Full text] [doi: 10.2196/preprints.58011]

18.   Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health. Jun 06, 2017;4(2):e19. [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]

19.   Wysa - everyday mental health. Wysa. URL: https://www.wysa.com/ [accessed 2024-04-16]

20.   Haque MD, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. JMIR Mhealth Uhealth. May 22, 2023;11:e44838. [FREE Full text] [doi: 10.2196/44838] [Medline: 37213181]

21.   Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res. Apr 02, 2024;3(1):12. [FREE Full text] [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]

22.   Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine. Apr 2023;90:104512. [FREE Full text] [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]

23.   Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2024. URL: https://iris.who.int/bitstream/handle/10665/375579/9789240084759-eng.pdf? [accessed 2024-04-16]

24.   What are large language models (LLMs)? IBM. URL: https://www.ibm.com/topics/large-language-models [accessed 2024-04-16]

25.   LLM evaluation: key metrics and best practices. AISERA. URL: https://aisera.com/blog/llm-evaluation/ [accessed 2024-04-16]

26.   Better language models and their implications. OpenAI. Feb 14, 2019. URL: https://openai.com/research/better-language-models [accessed 2024-04-16]

27.   Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. Preprint posted online on June 12, 2017. [FREE Full text]

28.   Kerner SM. What are large language models (LLMs)? TechTarget. URL: https://www.techtarget.com/whatis/definition/large-language-model-LLM [accessed 2024-05-16]

29.   Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on October 11, 2018. [FREE Full text] [doi: 10.5260/chara.21.2.8]

30.	Rogers A, Kovaleva O, Rumshisky A. A primer in BERTology: what we know about how BERT works. arXiv. Preprint posted online on February 27, 2020. [FREE Full text] [doi: 10.1162/tacl_a_00349]

31.	ChatGPT a year on: 3 ways the AI chatbot has completely changed the world in 12 months. Euronews. Nov 30, 2023. URL: https://www.euronews.com/next/2023/11/30/chatgpt-a-year-on-3-ways-the-ai-chatbot-has-completely-changed-the-world-in-12-months [accessed 2024-04-16]

32.	Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 17, 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]

33.	Hickey G. The best Large Language Models (LLMs) of 2024. TechRadar. 2024. URL: https://www.techradar.com/computing/artificial-intelligence/best-llms [accessed 2024-08-08]

34.	Dilmegani C. 10+ large language model examples – benchmark and use cases in '24. AIMultiple. URL: https://research.aimultiple.com/large-language-models-examples/ [accessed 2024-04-16]

35.	Timeline of AI and language models. Life Architect. URL: https://lifearchitect.ai/timeline/ [accessed 2024-04-16]

36.	Priest M. Large language models explained. boost.ai. Feb 20, 2024. URL: https://boost.ai/blog/llms-large-language-models [accessed 2024-04-16]

37.	Vucetic D, Tayaranian M, Ziaeefard M, Clark JJ, Meyer BH, Gross WJ. Efficient fine-tuning of BERT models on the edge. In: Proceedings of the IEEE International Symposium on Circuits and Systems. 2022. Presented at: ISCAS 2022; May 27-June 1, 2022; Austin, TX. [doi: 10.1109/iscas48785.2022.9937567]

38.	Kumar M. Understanding large language models and fine-tuning for business scenarios: a simple guide. Medium. Oct 27, 2023. URL: https://medium.com/@careerInAI/understanding-large-language-models-and-fine-tuning-for-business-scenarios-a-simple-guide-42f44cb687f0 [accessed 2024-04-16]

39.	Mishra S, Khashabi D, Baral C, Hajishirzi H. Cross-task generalization via natural language crowdsourcing instructions. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022. Presented at: ACL 2022; May 22-27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.acl-long.244]

40.	Zhang S, Dong L, Li X, Zhang S, Sun X, Wang S, et al. Instruction tuning for large language models: a survey. arXiv. Preprint posted online on August 21, 2023. [doi: 10.48550/arXiv.2308.10792]

41.	Berryman J, Ziegler A. A developer's guide to prompt engineering and LLMs. GitHub. 2024. URL: https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/ [accessed 2024-04-15]

42.	Bender EM, Koller A. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. Presented at: ACL 2020; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-main.463]

43.	Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15, 2020;36(4):1234-1240. [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]

44.	Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on April 10, 2019. [doi: 10.48550/arXiv.1904.05342]

45.	Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell. Sep 03, 2020;182(5):1360. [FREE Full text] [doi: 10.1016/j.cell.2020.08.029] [Medline: 32888496]

46.	Trengove M, Vandersluis R, Goetz L. Response to "attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine". EBioMedicine. Jul 2023;93:104671. [FREE Full text] [doi: 10.1016/j.ebiom.2023.104671] [Medline: 37327676]

47.	Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res. May 04, 2021;23(5):e15708. [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]

48.	Hua Y, Liu F, Yang K, Li Z, Sheu YH, Zhou P, et al. Large language models in mental health care: a scoping review. arXiv. Preprint posted online on January 1, 2024. [doi: 10.48550/arXiv.2401.02984]

49.	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. Aug 18, 2009;151(4):264-9, W64. [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]

50.	Verma S, Vishal, Joshi RC, Dutta MK, Jezek S, Burget R. AI-enhanced mental health diagnosis: leveraging transformers for early detection of depression tendency in textual data. In: Proceedings of the 15th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. 2023. Presented at: ICUMT 2023; October 30-November 1, 2023; Ghent, Belgium. [doi: 10.1109/icumt61075.2023.10333301]

51.	Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on July 26, 2019. [doi: 10.48550/arXiv.1907.11692]

52.	Namdari R. Mental health corpus. Kaggle. URL: https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus [accessed 2024-04-17]

53.   Depression: Reddit dataset (cleaned). Kaggle. URL: https://www.kaggle.com/datasets/infamouscoder/ depression-reddit-cleaned [accessed 2024-04-17]

54.   Diniz EJ, Fontenele JE, de Oliveira AC, Bastos VH, Teixeira S, Rabêlo RL, et al. Boamente: a natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. Healthcare (Basel). Apr 08, 2022;10(4):698. [FREE Full text] [doi: 10.3390/healthcare10040698] [Medline: 35455874]

55.   Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020. Presented at: EMNLP 2020 - Demos; November 16-20, 2020; Online. [doi: 10.18653/v1/2020.emnlp-demos.6]

56.   Souza F, Nogueira R, Lotufo R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Proceedings of the 9th Brazilian Conference on Intelligent Systems. 2020. Presented at: BRACIS 2020; October 20-23, 2020; Rio Grande, Brazil. [doi: 10.1007/978-3-030-61377-8_28]

57.   Danner M, Hadzic B, Gerhardt S, Ludwig S, Uslu I, Shao P, et al. Advancing mental health diagnostics: GPT-based method for depression detection. In: Proceedings of the 62nd Annual Conference of the Society of Instrument and Control Engineers. 2023. Presented at: SICE 2023; September 6-9, 2023; Tsu, Japan. [doi: 10.23919/sice59929.2023.10354236]

58.   Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. 2014. Presented at: LREC 2014; May 26-31, 2014; Reykjavik, Iceland.

59.   Extended DAIC database. University of Southern California. URL: https://dcapswoz.ict.usc.edu/ extended-daic-database-download/ [accessed 2024-04-17]

60.   Tao Y, Yang M, Shen H, Yang Z, Weng Z, Hu B. Classifying anxiety and depression through LLMs virtual interactions: a case study with ChatGPT. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. 2023. Presented at: BIBM 2023; December 5-8, 2023; Istanbul, Turkiye. [doi: 10.1109/bibm58861.2023.10385305]

61.   Hayati MF, Md. Ali MA, Md. Rosli AN. Depression detection on Malay dialects using GPT-3. In: Proceedings of the IEEE-EMBS Conference on Biomedical Engineering and Sciences. 2022. Presented at: IECBES 2022; December 7-9, 2022; Kuala Lumpur, Malaysia. [doi: 10.1109/iecbes54088.2022.10079554]

62.   Wang X, Chen S, Li T, Li W, Zhou Y, Zheng J, et al. Depression risk prediction for Chinese microblogs via deep-learning methods: content analysis. JMIR Med Inform. Jul 29, 2020;8(7):e17958. [FREE Full text] [doi: 10.2196/17958] [Medline: 32723719]

63.   Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. arXiv. Preprint posted online on June 19, 2019. [doi: 10.48550/arXiv.1906.08237]

64.   Wang X, Chen S, Li T, Li W, Zhou Y, Zheng J, et al. Assessing depression risk in Chinese microblogs: a corpus and machine learning methods. In: Proceedings of the IEEE International Conference on Healthcare Informatics. 2019. Presented at: ICHI 2019; June 10-13, 2019; Xi'an, China. [doi: 10.1109/ichi.2019.8904506]

65.   Metzler H, Baginski H, Niederkrotenthaler T, Garcia D. Detecting potentially harmful and protective suicide-related content on Twitter: machine learning approach. J Med Internet Res. Aug 17, 2022;24(8):e34705. [FREE Full text] [doi: 10.2196/34705] [Medline: 35976193]

66.   Sadeghi M, Egger B, Agahi R, Richer R, Capito K, Rupp LH, et al. Exploring the capabilities of a language model-only approach for depression detection in text data. In: Proceedings of the IEEE EMBS International Conference on Biomedical and Health Informatics. 2023. Presented at: BHI 2023; October 15-18, 2023; Pittsburgh, PA. [doi: 10.1109/bhi58575.2023.10313367]

67.   DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A, et al. SimSensei kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems. 2014. Presented at: AAMAS '14; May 5-9, 2014; Paris, France. [doi: 10.1609/aaai.v29i1.9777]

68.   Zhang Y, Lyu H, Liu Y, Zhang X, Wang Y, Luo J. Monitoring depression trends on Twitter during the COVID-19 pandemic: observational study. JMIR Infodemiology. Jul 18, 2021;1(1):e26769. [FREE Full text] [doi: 10.2196/26769] [Medline: 34458682]

69.   Vajre V, Naylor M, Kamath U, Shehu A. PsychBERT: a mental health language model for social media mental health behavioral analysis. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. 2021. Presented at: BIBM 2021; December 9-12, 2021; Houston, TX. [doi: 10.1109/BIBM52615.2021.9669469]

70.   Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health. Sep 20, 2023;10:e51232. [FREE Full text] [doi: 10.2196/51232] [Medline: 37728984]

71.   Levi-Belz Y, Gamliel E. The effect of perceived burdensomeness and thwarted belongingness on therapists' assessment of patients' suicide risk. Psychother Res. Jul 2016;26(4):436-445. [doi: 10.1080/10503307.2015.1013161] [Medline: 25751580]

72.   Howard D, Maslej MM, Lee J, Ritchie J, Woollard G, French L. Transfer learning for risk classification of social media posts: model evaluation study. J Med Internet Res. May 13, 2020;22(5):e15371. [FREE Full text] [doi: 10.2196/15371] [Medline: 32401222]

73.   Overview DeepMoji. Massachusetts Institute of Technology Media Lab. URL: https://www.media.mit.edu/projects/deepmoji/ overview/ [accessed 2024-04-17]

74.   Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, St. John R, et al. Universal sentence encoder. arXiv. Preprint posted online on March 29, 2018. [FREE Full text] [doi: 10.18653/v1/d18-2029]

75.   Stigall W, Khan MA, Attota D, Nweke F, Pei Y. Large language models performance comparison of emotion and sentiment classification. In: Proceedings of the 2024 ACM Southeast Conference. 2024. Presented at: ACMSE '24; April 18-20, 2024; Marietta, GA. [doi: 10.1145/3603287.3651183]

76.   dair-ai / emotion. Hugging Face. URL: https://huggingface.co/datasets/dair-ai/emotion/tree/main [accessed 2024-08-06]

77.   Twitter tweets sentiment dataset. Kaggle. URL: https://www.kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset [accessed 2024-08-06]

78.   Ghanadian H, Nejadgholi I, Osman HA. Socially aware synthetic data generation for suicidal ideation detection using large language models. IEEE Access. 2024;12:14350-14363. [doi: 10.1109/access.2024.3358206]

79.   FLAN-T5. Hugging Face. URL: https://huggingface.co/docs/transformers/model_doc/flan-t5 [accessed 2024-08-06]

80.   Ghanadian H, Nejadgholi I, Osman HA. ChatGPT for suicide risk assessment on social media: quantitative evaluation of model performance, potentials and limitations. In: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. 2023. Presented at: WASSA@ACL 2023; July 14, 2023; Toronto, ON. [doi: 10.18653/v1/2023.wassa-1.16]

81.   Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, et al. A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data. JMIR Ment Health. Jan 25, 2024;11:e50150. [FREE Full text] [doi: 10.2196/50150] [Medline: 38271138]

82.   Chung JY, Gibbons A, Atlas L, Ballard E, Ernst M, Japee S, et al. COVID-19 and mental health: predicted mental health status is associated with clinical symptoms and pandemic-related psychological and behavioral responses. medRxiv. Preprint posted online on October 14, 2021. [FREE Full text] [doi: 10.1101/2021.10.12.21264902] [Medline: 34671781]

83.   Nelson LM, Simard JF, Oluyomi A, Nava V, Rosas LG, Bondy M, et al. US public concerns about the COVID-19 pandemic from results of a survey given via social media. JAMA Intern Med. Jul 01, 2020;180(7):1020-1022. [FREE Full text] [doi: 10.1001/jamainternmed.2020.1369] [Medline: 32259192]

84.   Beredo JL, Ong EC. A hybrid response generation model for an empathetic conversational agent. In: Proceedings of the International Conference on Asian Language Processing (IALP). 2022. Presented at: IALP 2022; October 27-28, 2022; Singapore, Singapore. [doi: 10.1109/ialp57159.2022.9961311]

85.   Santos KA, Ong E, Resurreccion R. Therapist vibe: children's expressions of their emotions through storytelling with a chatbot. In: Proceedings of the Interaction Design and Children Conference. 2020. Presented at: IDC '20; June 21-24, 2020; London, UK. [doi: 10.1145/3392063.3394405]

86.   Ong E, Go MJ, Lao R, Pastor J, To LB. Towards building mental health resilience through storytelling with a chatbot. In: Proceedings of the 29th International Conference on Computers in Education. 2021. Presented at: ICCE 2021; November 22-26, 2021; Online.

87.   PERMA model. Corporate Finance Institute. URL: https://corporatefinanceinstitute.com/resources/management/perma-model/ [accessed 2024-04-17]

88.   Rashkin H, Smith EM, Li M, Boureau YL. Towards empathetic open-domain conversation models: a new benchmark and dataset. arXiv. Preprint posted online on November 1, 2018. [FREE Full text] [doi: 10.18653/v1/p19-1534]

89.   Sia DE, Yu MJ, Daliva JL, Montenegro J, Ong E. Investigating the acceptability and perceived effectiveness of a chatbot in helping students assess their well-being. In: Proceedings of the Asian CHI Symposium 2021. 2021. Presented at: Asian CHI '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3429360.3468177]

90.   Schwartz HA, Sap M, Kern ML, Eichstaedt JC, Kapelner A, Agrawal M, et al. Predicting individual well-being through the language of social media. Pac Symp Biocomput. 2016;21:516-527. [FREE Full text] [Medline: 26776214]

91.   Crasto R, Dias L, Miranda D, Kayande D. CareBot: a mental health ChatBot. In: Proceedings of the 2nd International Conference for Emerging Technology. 2021. Presented at: INCET 2021; May 21-23, 2021; Belagavi, India. [doi: 10.1109/incet51464.2021.9456326]

92.   Zygadlo A. A therapeutic dialogue agent for Polish language. In: Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2021. Presented at: ACIIW 2021; September 28-October 1, 2021; Nara, Japan. [doi: 10.1109/aciiw52867.2021.9666281]

93.   Conversational AI platform. Rasa Technologies Inc. URL: https://rasa.com/ [accessed 2024-04-18]

94.   Industrial-strength natural language processing in Python. spaCy. URL: https://spacy.io/ [accessed 2024-04-18]

95.   Li Y, Su H, Shen X, Li W, Cao Z, Niu S. DailyDialog: a manually labelled multi-turn dialogue dataset. arXiv. Preprint posted online on October 11, 2017. [doi: 10.48550/arXiv.1710.03957]

96.   Heston TF. Safety of large language models in addressing depression. Cureus. Dec 18, 2023;15(12):e50729. [FREE Full text] [doi: 10.7759/cureus.50729] [Medline: 38111813]

97.   Alessa A, Al-Khalifa H. Towards designing a ChatGPT conversational companion for elderly people. In: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments. 2023. Presented at: PETRA '23; July 5-7, 2023; Corfu, Greece. [doi: 10.1145/3594806.3596572]

98. He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician versus large language model chatbot responses to web-based questions from autistic patients in Chinese: cross-sectional comparative analysis. J Med Internet Res. Apr 30, 2024;26:e54706. [FREE Full text] [doi: 10.2196/54706] [Medline: 38687566]

99. Deng Z, Deng Z, Liu S, Evans R. Knowledge transfer between physicians from different geographical regions in China's online health communities. Inf Technol Manag. May 19, 2023:1-18. [FREE Full text] [doi: 10.1007/s10799-023-00400-3] [Medline: 37359990]

100. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr. Nov 2023;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]

101. Wright B, Dave S, Dogra N. 100 Cases in Psychiatry, Second Edition. Boca Raton, FL. CRC Press; 2017.

102. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use ChatGPT for mental health and substance use education? Examining its quality and potential harms. JMIR Med Educ. Nov 30, 2023;9:e51243. [FREE Full text] [doi: 10.2196/51243] [Medline: 38032714]

103. Evidence-based information for the community. Cracks in the Ice. URL: https://cracksintheice.org.au/ [accessed 2024-04-18]

104. Positive choices: drug and alcohol education - get informed, stay smart, stay safe. Positive Choices. URL: https://positivechoices.org.au/ [accessed 2024-04-18]

105. Farhat F. ChatGPT as a complementary mental health resource: a boon or a bane. Ann Biomed Eng. May 2024;52(5):1111-1114. [doi: 10.1007/s10439-023-03326-7] [Medline: 37477707]

106. Wei Y, Guo L, Lian C, Chen J. ChatGPT: opportunities, risks and priorities for psychiatry. Asian J Psychiatr. Dec 2023;90:103808. [doi: 10.1016/j.ajp.2023.103808] [Medline: 37898100]

107. Yongsatianchot N, Torshizi PG, Marsella S. Investigating large language models' perception of emotion using appraisal theory. In: Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2023. Presented at: ACIIW 2023; September 10-13, 2023; Cambridge, MA. [doi: 10.1109/aciiw59127.2023.10388194]

108. Xenova / text-davinci-003. Hugging Face. URL: https://huggingface.co/Xenova/text-davinci-003 [accessed 2024-04-18]

109. Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. J Med Artif Intell. Oct 2023;6. [doi: 10.21037/jmai-23-71]

110. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry. Sep 01, 2023;14:1234397. [FREE Full text] [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]

111. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res. Sep 11, 2023;25:e49240. [FREE Full text] [doi: 10.2196/49240] [Medline: 37695668]

112. Collins E, Ghahramani Z. LaMDA: our breakthrough conversation technology. Google. May 18, 2021. URL: https://blog.google/technology/ai/lamda/ [accessed 2024-04-18]

113. Tanana MJ, Soma CS, Kuo PB, Bertagnolli NM, Dembe A, Pace BT, et al. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. Behav Res Methods. Oct 2021;53(5):2069-2082. [FREE Full text] [doi: 10.3758/s13428-020-01531-z] [Medline: 33754322]

114. Welcome to LIWC-22. LIWC. URL: https://www.liwc.app/ [accessed 2024-04-18]

115. Publisher of streaming video, audio, and text library databases that promote research, teaching, and learning across disciplines, including music, counseling, history, business and more. Alexander Street. URL: https://alexanderstreet.com/ [accessed 2024-04-18]

116. Wang X, Liu K, Wang C. Knowledge-enhanced pre-training large language model for depression diagnosis and treatment. In: Proceedings of the IEEE 9th International Conference on Cloud Computing and Intelligent Systems. 2023. Presented at: CCIS 2023; August 12-13, 2023; Dali, China. [doi: 10.1109/ccis59572.2023.10263217]

117. ChineseNLP/docs /language_modeling.md. GitHub. URL: https://github.com/didi/ChineseNLP/blob/master/docs/language_modeling.md [accessed 2024-04-18]

118. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. JAMA Netw Open. Dec 01, 2023;6(12):e2346721. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.46721] [Medline: 38060223]

119. Neurology board review questions and practice tests. BoardVitals. URL: https://www.boardvitals.com/neurology-board-review [accessed 2024-04-18]

120. Friedman SF, Ballentine G. Trajectories of sentiment in 11,816 psychoactive narratives. Hum Psychopharmacol. Jan 2024;39(1):e2889. [doi: 10.1002/hup.2889] [Medline: 38117133]

121. Taylor WL. "Cloze" readability scores as indices of individual differences in comprehension and aptitude. J Appl Psychol. 1957;41(1):19-26. [doi: 10.1037/h0040591]

122. Erowid homepage. Erowid. URL: https://www.erowid.org/ [accessed 2024-04-18]

123. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang XP, et al. Automated design of ligands to polypharmacological profiles. Nature. Dec 13, 2012;492(7428):215-220. [FREE Full text] [doi: 10.1038/nature11691] [Medline: 23235874]

124. Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res. Jan 2013;41(Database issue):D996-1008. [FREE Full text] [doi: 10.1093/nar/gks1042] [Medline: 23193282]

125. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. Presented at: ACL 2020; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-main.372]

126. Wu Y, Chen J, Mao K, Zhang Y. Automatic post-traumatic stress disorder diagnosis via clinical transcripts: a novel text augmentation with large language models. In: Proceedings of the IEEE Biomedical Circuits and Systems Conference. 2023. Presented at: BioCAS 2023; October 19-21, 2023; Toronto, ON. [doi: 10.1109/biocas58349.2023.10388714]

127. Kumar H, Wang Y, Shi J, Musabirov I, Farb NA, Williams JJ. Exploring the use of large language models for improving the awareness of mindfulness. In: Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. 2023. Presented at: CHI EA '23; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544549.3585614]

128. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health. Jan 09, 2024;12(Suppl 1):e002583. [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]

129. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology. Aug 2024;49(9):1412-1416. [doi: 10.1038/s41386-024-01841-2] [Medline: 38480911]

130. Blease C, Worthen A, Torous J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey. Psychiatry Res. Mar 2024;333:115724. [FREE Full text] [doi: 10.1016/j.psychres.2024.115724] [Medline: 38244285]

131. Berrezueta-Guzman S, Kandil M, Martín-Ruiz ML, Pau de la Cruz I, Krusche S. Future of ADHD care: evaluating the efficacy of ChatGPT in therapy enhancement. Healthcare (Basel). Mar 19, 2024;12(6):683. [FREE Full text] [doi: 10.3390/healthcare12060683] [Medline: 38540647]

132. Colizzi M, Lasalvia A, Ruggeri M. Prevention and early intervention in youth mental health: is it time for a multidisciplinary and trans-diagnostic model for care? Int J Ment Health Syst. Mar 24, 2020;14:23. [FREE Full text] [doi: 10.1186/s13033-020-00356-9] [Medline: 32226481]

133. Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min. 2021;11(1):81. [FREE Full text] [doi: 10.1007/s13278-021-00776-6] [Medline: 34484462]

134. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv. Preprint posted online on November 9, 2023. [doi: 10.48550/arXiv.2311.05232]

135. Egli A. ChatGPT, GPT-4, and other large language models: the next revolution for clinical microbiology? Clin Infect Dis. Nov 11, 2023;77(9):1322-1328. [FREE Full text] [doi: 10.1093/cid/ciad407] [Medline: 37399030]

136. Palmer A, Schwan D. Beneficent dehumanization: employing artificial intelligence and carebots to mitigate shame-induced barriers to medical care. Bioethics. Feb 2022;36(2):187-193. [doi: 10.1111/bioe.12986] [Medline: 34942057]

137. Haque OS, Waytz A. Dehumanization in medicine: causes, solutions, and functions. Perspect Psychol Sci. Mar 2012;7(2):176-186. [doi: 10.1177/1745691611429706] [Medline: 26168442]

138. Image inputs for ChatGPT - FAQ. OpenAI. URL: https://help.openai.com/en/articles/8400551-image-inputs-for-chatgpt-faq [accessed 2024-04-18]

139. What is RAG (Retrieval Enhanced Generation)? Amazon Web Services. URL: https://aws.amazon.com/cn/what-is/retrieval-augmented-generation/ [accessed 2024-08-08]

140. Models. OpenAI Platform. URL: https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4 [accessed 2024-08-08]

141. What is retrieval-augmented generation (RAG)? Google Cloud. URL: https://cloud.google.com/use-cases/retrieval-augmented-generation [accessed 2024-08-08]

142. Mündler N, He J, Jenko S, Vechev M. Self-contradictory hallucinations of large language models: evaluation, detection and mitigation. arXiv. Preprint posted online on May 25, 2023. [doi: 10.48550/arXiv.2305.15852]

143. Forbes GC, Katlana P, Ortiz Z. Metric ensembles for hallucination detection. arXiv. Preprint posted online on October 16, 2023. [doi: 10.48550/arXiv.2310.10495]

144. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology. Jan 2024;310(1):e232756. [doi: 10.1148/radiol.232756] [Medline: 38226883]

145. What is LLM temperature? Iguazio. URL: https://www.iguazio.com/glossary/llm-temperature/ [accessed 2024-04-27]

146. LLM optimization parameters. Attri. URL: https://attri.ai/generative-ai-wiki/llm-optimization-parameters [accessed 2024-04-27]

147. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High Confid Comput. Jun 2024;4(2):100211. [doi: 10.1016/j.hcc.2024.100211]

148. Vats A, Liu Z, Su P, Paul D, Ma Y, Pang Y, et al. Recovering from privacy-preserving masking with large language models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 2024. Presented at: ICASSP 2024; April 14-19, 2024; Seoul, Republic of Korea. [doi: 10.1109/icassp48485.2024.10448234]

149. Ramlochan S. The black box problem: opaque inner workings of large language models. Prompt Engineering. URL: https:/ /promptengineering.org/the-black-box-problem-opaque-inner-workings-of-large-language-models/ [accessed 2024-04-18]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
CA: conversational agent
ERNIE Bot: Enhanced Representation through Knowledge Integration Bot
LaMDA: Language Model for Dialogue Application
LLM: large language model
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analyses
RAG: retrieval-augmented generation
ROUGE: Recall-Oriented Understudy for Gisting Evaluation