JMIR Publications
The Leading eHealth Publisher

Large Language Model for Mental Health: A Systematic Review

Zhijun Guo, Alvina Lai, Johan Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li

Submitted to: JMIR Mental Health
on: February 18, 2024

Table of Contents

Large Language Model for Mental Health: A Systematic Review

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

[1]Institute of Health Informatics University College, London London GB
[2]GOS Institute of Child Health, University College London London GB

Corresponding Author:
Kezhi Li
Institute of Health Informatics University College, London
Institute of Health Informatics University College London 222 Euston Road London United Kingdom
London
GB

Abstract

Background: Large language models (LLMs) have received much attention and show their potential in digital health, while their application in mental health is subject to ongoing debate. This systematic review aims to summarize and characterize the use of LLMs in mental health by investigating the strengths and limitations of the latest work in LLMs and discusses the challenges and opportunities for early screening, digital interventions, and other clinical applications in mental health.

Objective: This systematic review aims to summarize how LLMs are used in mental health. We focus on the models, data sources, methodologies, and main outcomes in existing work, in order to assess the applicability of LLMs to early screening, digital interventions, and other clinical applications.

Methods: Adhering to the PRISMA guidelines, this review searched three open-access databases: PubMed, DBLP Computer Science Bibliography (DBLP), and IEEE Xplore (IEEE). Keywords used were: (mental health OR mental illness OR mental disorder OR psychology OR depression OR anxiety) AND (large language models OR LLMs OR GPT OR ChatGPT OR BERT OR Transformer OR LaMDA OR PaLM OR Claude). We included articles published between January 1, 2017, and September 1, 2023, and excluded non-English articles.

Results: In total, 32 articles were evaluated, including mental health analysis using social media datasets (n=13), LLMs usage for mental health chatbots (n=10), and other applications of LLMs in mental health (n=9). LLMs exhibit substantial effectiveness in classifying and detecting mental health issues and offer more efficient and personalized healthcare to improve telepsychological services. However, assessments also indicate that the current risks associated with the clinical use might surpass their benefits. These risks include inconsistencies in generated text, the production of hallucinatory content, and the absence of a comprehensive ethical framework.

Conclusions: This systematic review examines the clinical applications of LLMs in mental health, highlighting their potential and their inherent risks. The study identifies significant concerns, including inherent biases in training data, ethical dilemmas, challenges in interpreting the 'black box' nature of LLMs, and concerns about the accuracy and reliability of the content they produce. Consequently, LLMs should not be considered substitutes for professional mental health services. Despite these challenges, the rapid advancement of LLMs may highlight their potential as new clinical tools, emphasizing the need for continued research and development in this field.

(JMIR Preprints 18/02/2024:57400)
DOI: https://doi.org/10.2196/preprints.57400

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

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

Original Manuscript

Large Language Model for Mental Health: A Systematic Review

Zhijun Guo[1]; Alvina Lai[1]; Johan Hilge Thygesen[1]; Joseph Farrington[1]; Thomas Keen[1,2]; Kezhi Li[1]

[1]Institute of Health Informatics, University College London
[2]GOS Institute of Child Health, University College London

Corresponding author:
Kezhi Li
Institute of Health Informatics
University College London
222 Euston Road
London
United Kingdom
Phone: +44 7859 995590
Email: ken.li@ucl.ac.uk

Abstract

Background

Large l angmuage (h odedls m LLMas) a avse t eceived potential i d hn igwital t a ealth, i m hile h hieirs pplication to ongoing Tdheibsa ste ystematic review aims to characterize the use of LLMs in mental health byinvestigating the strengths and limitations of the latest work in LLMs and discusses the challenges and opportunitifeosr early screening, digital interventions, and other clin applications in mental health.

Objective:

This systematic review aims to summarize how LLMs are used in m health. We focus on the models, data sources, methodologie outcomes in existing work, in order to assess the applicability of LLMs early screening, digital interventions, and other clinical applications.

Methods:
Adhering to the PRISMA guidelines, this review searched three open-access databases: PubMed, DBLP Computer Science Bibliography (DBLP), and IEEE Xplore (IEEE). Keywords used were: (mental health OR mental illness OR mental disorder OR psychology OR depression OR anxiety) AND (large language models OR LLMs OR GPT OR ChatGPT OR BERT OR Transformer OR LaMDA OR PaLM OR Claude). We included articles published between January 1, 2017, and September 1, 2023, and excluded non-English articles.

Results:
In total, 32 articles were evaluated, including mental health analysis using social media datasets (n=13), LLMs usage for mental health chatbots (n=10), and other applications of LLMs in mental health (n=9). LLMs exhibit substantial effectiveness in classifying and detecting mental health issues and offer more efficient and personalized healthcare to improve telepsychological services. However, assessments also indicate that the current risks associated with the clinical use might surpass their benefits. These risks include inconsistencies in generated text, the production of hallucinatory content, and the absence of a comprehensive ethical framework.

Conclusions:
This systematic review examines the clinical applications of LLMs in mental health, highlighting their potential and their inherent risks. The study identifies significant concerns, including inherent biases in training data, ethical dilemmas, challenges in interpreting the 'black box' nature of LLMs, and concerns about the accuracy and reliability of the content they produce. Consequently, LLMs should not be considered substitutes for professional mental health services. Despite these challenges, the rapid advancement of LLMs may highlight their potential as new clinical tools, emphasizing the need for continued research and development in this field.

Keywords
Large language models; Mental health; Digital healthcare; ChatGPT; BERT

---

1. Introduction and Background

1.1 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 suffered from 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. However, they often go undetected or untreated, and the resources allocated to the diagnosis and treatment of mental illness are far less than the negative impact it has on society 3. Focusing specifically on the UK, the scale of the mental health crisis is significant. Figures show that one in six individuals in England reported experiencing a common mental health problem in a given week 4. The COVID-19 pandemic has further intensified existing mental health challenges globally. Specifically, the World Health Organization (WHO) reported a 26% rise in anxiety disorders and a 28% increase in major depression disorders within just one year of the pandemic 5. This escalating crisis underscores the urgent need for innovative approaches to mental health. The negative effects of poor mental health are far-reaching, with over 90% of people who die by suicide annually being diagnosed with a mental illness 6. These statistics point to the need to raise awareness of mental health in society and to take proactive early intervention and preventive measures.

Mental illness treatment encompasses a range of modalities including medication, psychotherapy, support groups, hospitalization, and complementary & alternative medicine 7. However, societal stigma attached to mental illnesses often deters people from seeking appropriate care 8. Many individuals with mental illness are afraid to discuss their condition with others or seek help from a professional psychologist 9. The COVID-19 crisis and global pandemics have highlighted the importance of digital tools such as telemedicine and apps to deliver care in times of need 10, accelerating a paradigm shift in healthcare. In this evolving landscape, LLMs offer new possibilities for mental health care delivery and support.

Recent technological advancements have revealed some unique advantages of LLMs in the mental health area. These models, capable of processing and generating text similar to human communication, offer a non-judgmental space for individuals to share concerns and receive support 11. LLMs also contribute to psychoeducation, enhance therapeutic methods, and may provide timely interventions, particularly when traditional mental health resources are limited or unavailable 12. Recent data indicates that 23% of people with mental illness report having to wait more than 12 weeks to begin treatment, and 43% of adults with mental illness report that the long wait for treatment has exacerbated their conditions 13. In this landscape of limited healthcare availability, LLMs present a feasible solution for providing timely access to mental health services. For instance, a range of mental health chatbots, developed by using language models, have been gaining recognition, such as Woebot 14 and Wysa 15. Both chatbots follow the principles of Cognitive Behavioural Therapy. These platforms are designed to provide users with self-help tools to help them cope with mental health issues such as stress, anxiety, and depression 16.

Meanwhile, the application of LLMs in the mental health sector presents several risks, especially concerning vulnerable groups. Challenges such as inconsistencies in the content generated and the production of 'hallucinatory' content which may mislead or harm users 17. Given these concerns, a thorough and rigorous assessment of LLMs' responsible and effective use in healthcare is essential. The following section will further examine the workings of LLMs, and their potential mental health applications, and critically evaluate the opportunities and challenges they introduce.

1.2 Large Language Models

LLMs represent significant advancements in machine learning (ML), characterized by their ability to understand and generate human-like text with high accuracy. Distinguished from traditional language models by their scale, LLMs often contain billions of parameters 18. This breakthrough is largely due to the Transformer architecture, a deep neural network structure that employs a 'self-attention' mechanism, developed by Vaswani et al. in 2017. This allows LLMs to process information in parallel rather than sequentially, greatly enhancing speed and contextual understanding 19. Notable examples of such state-of-the-art LLMs include Generative Pre-trained Transformers (GPT), and Bidirectional Encoder Representations from Transformers (BERT), among others (Fig.1).

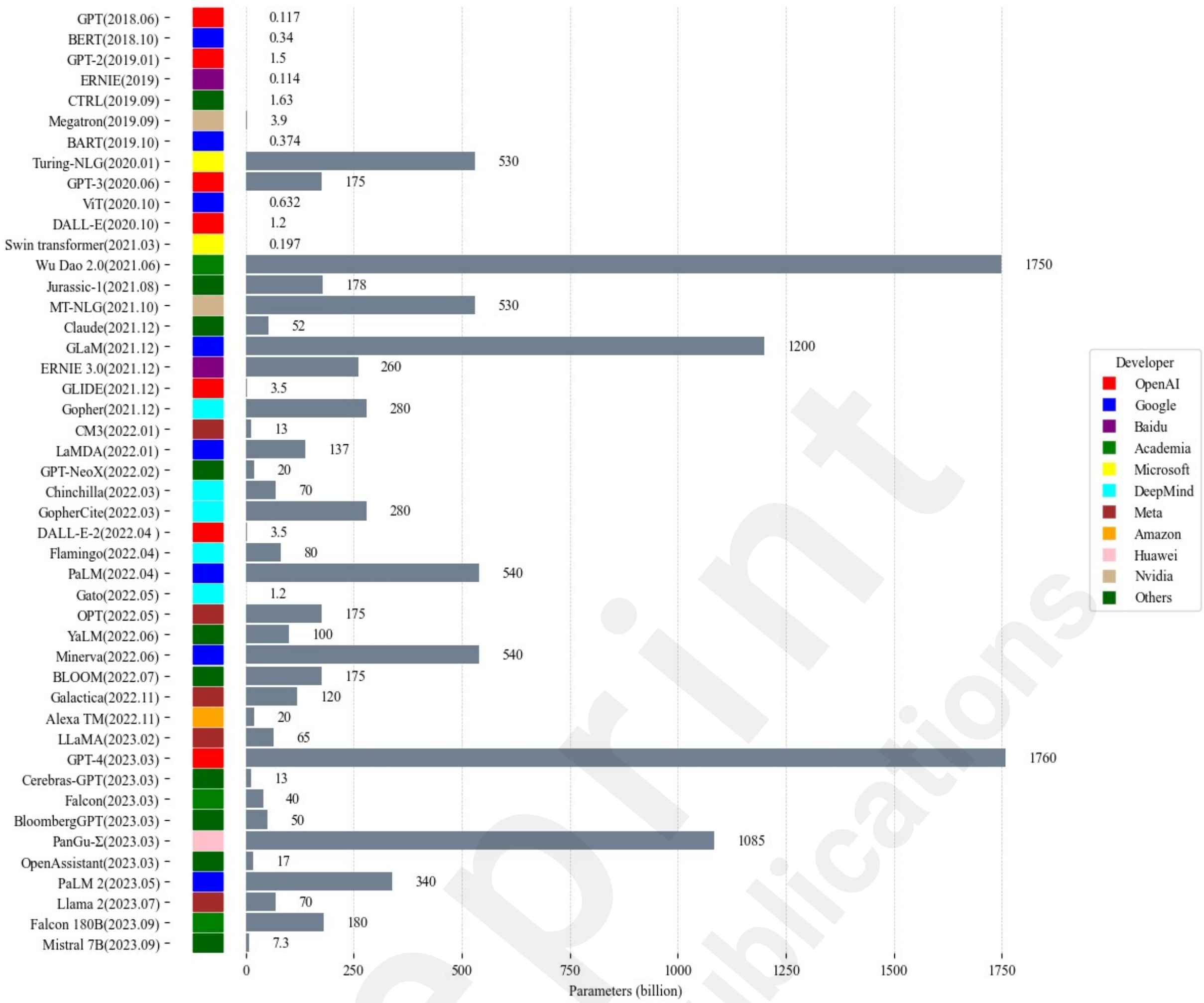


Fig1. Comparative analysis of large language models by parameter size and developer entity. The bar chart represents the number of parameters in billions for various language models by date of publication, with the oldest models at the top. The legend is color-coded by the development entity. Data was summarized with the latest models up to September 2023, with data for parameters and developers from GPT to LLaMA adapted from the work of Thirunavukarasu AJ et al 20.

LLMs are designed to learn the fundamental statistical patterns of language 21. Initially, these models are the basis for fine-tuning task-specific models rather than training those models from scratch. This fine-tuning process involves adjusting a pre-trained model to a specific task by further training it on a smaller, task-specific dataset 22. However, with the advent of even larger and more complex models, this fine-tuning step is often unnecessary for a wide range of tasks. These advanced LLMs are capable of understanding and executing tasks specified in natural language prompts. A 'prompt' is a natural language text that describes a task that the AI should perform 23. While highly effective, one must be cautious of 'hallucinations' – a phenomenon where these models confidently generate incorrect or irrelevant outputs 24. This can be particularly challenging in scenarios requiring high accuracy, such as healthcare and medical applications 25-262728.

The existing literature includes a review of the application of ML and natural learning processing (NLP) in mental health 29, as well as an analysis of LLMs in medicine 20. Studies have demonstrated NLP's efficacy in performing statistical tasks, such as text categorization and sentiment analysis 29. Despite these findings, a systematic review of the use of state-of-the-art LLMs specifically for mental health has yet to be conducted. Furthermore, there is a lack of in-depth discussion on the ethical challenges unique to the application of LLMs in various mental health contexts. This study aims to fill these gaps by providing a comprehensive review of the application of LLMs in mental health, examining the relevant ethical considerations, and assessing their ability as tools for early screening of mental health conditions and support in therapeutic interventions.

2. Methods

This systematic review followed the Preferred Reporting Items for Systematic Review and Meta-analysis (PRISMA) guidelines 30. The protocol was registered on PROSPERO under the ID: CRD42024508617.

2.1 Inclusion and Exclusion Criteria

Three major databases were investigated in this systematic review: PubMed, DBLP, and IEEE. These databases were chosen because of their open search capabilities. 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 September 1, 2023. The search was conducted from July 1 to September 1, 2023, primarily targeting the titles, abstracts, models used, data sources, methodology, and main outcomes of the articles. This timeframe was chosen considering the significant 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. Due to the limited literature specifically addressing the mental health applications of LLMs, we included review articles to ensure a comprehensive perspective. Our 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.

2.2 Search Strategies

The mental health-related terms were combined with LLM descriptors using Boolean operators. The search query was: ((mental health OR mental illness OR mental disorder OR psychology OR depression OR anxiety) AND (large language models OR LLMs OR GPT OR ChatGPT OR Bert OR Transformer OR LaMDA OR PaLM OR Claude)). For the articles that matched our search criteria, a meticulous and iterative assessment has been conducted by two independent reviewers (ZG, KL) to ensure each article fell within the scope of LLMs in mental health (Multimedia Appendix 1). This 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.

2.3 Information Extraction

We examined the application scenarios, model architecture, data sources, methodologies used, and main outcomes from selected studies on LLMs in mental health. Firstly, we categorized each study to clarify their primary goals and applications, providing an overview of how LLMs are being utilized in the mental health field. This categorization helps in understanding the diverse ways in which these models are applied. Following that the main model architecture of

LLMs used was summarized. Secondly, we conducted a thorough examination of data sources. Our analysis covered both public and private datasets used in these studies. Noted that some review articles lacked detail on dataset content, we focused on providing comprehensive information on public datasets, including their origins and sample sizes. After that, various methods employed across different scenarios are investigated. They include data collection strategies and analytical methodologies. We examined their comparative structures and statistical techniques to provide a clear understanding of how these methods are applied in practice. Finally, the main outcome of each study was addressed. We documented significant results, aligning them with relevant performance metrics and evaluation criteria, and providing quantitative data where applicable to underscore these findings. This allowed us to highlight the efficacy and impact of LLMs in mental health, providing quantitative data where applicable to underscore these findings.

3. Results

3.1 Strategy and Screening Process

The PRISMA diagram of the systematic screening process can be seen in Figure. 2. Our initial search across three academic databases: PubMed, DBLP, and IEEE yielded 205 papers: 107 from PubMed, 17 from DBLP, and 81 from IEEE. After duplication, 199 unique papers were retained. Subsequent screening is based on predefined inclusion and exclusion criteria, narrowing down the selection to 32 papers included in this review. The reasons for the full-text exclusion of 56 papers can be found in Multimedia Appendix 2.

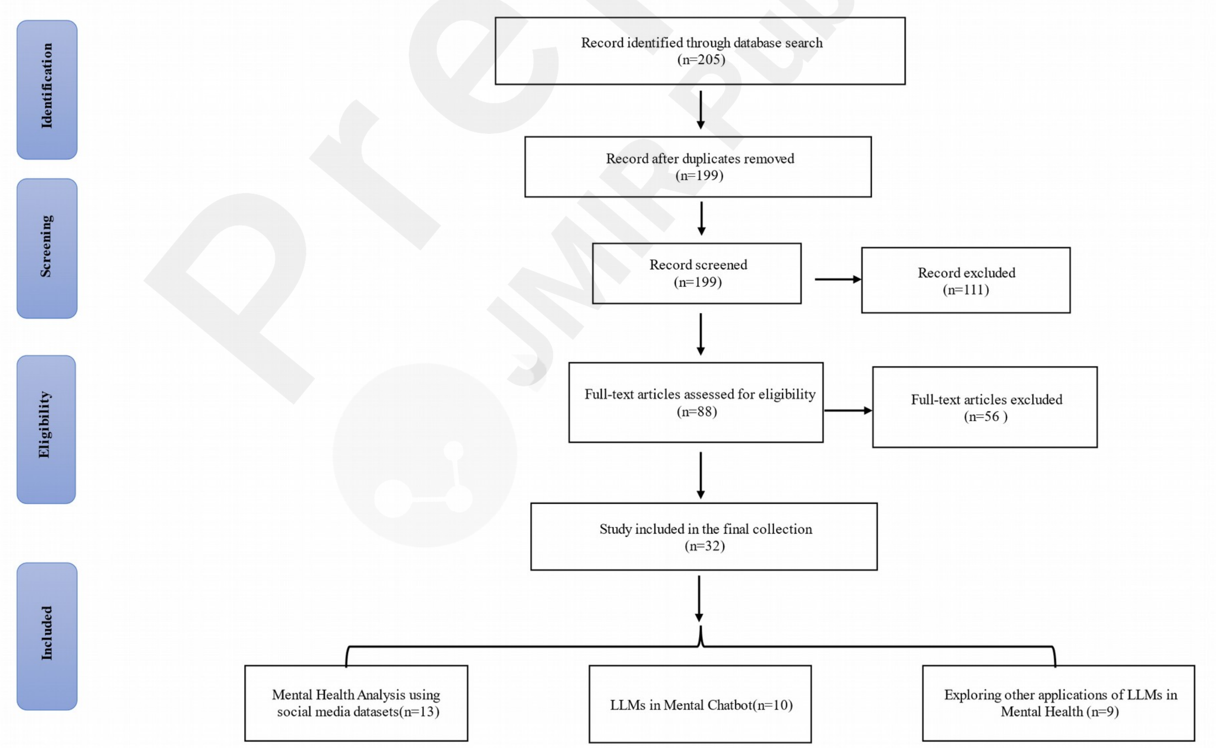


Fig 2. PRISMA flow of selection process.

Papers were classified into three main categories: mental health analysis using social media (n=13), LLMs in mental health chatbots (n=10), and the other applications of LLMs in mental health (n=9). Figure 3 highlights the significant increase in related publications over the last two years, indicating the emerging nature of LLMs in mental health.

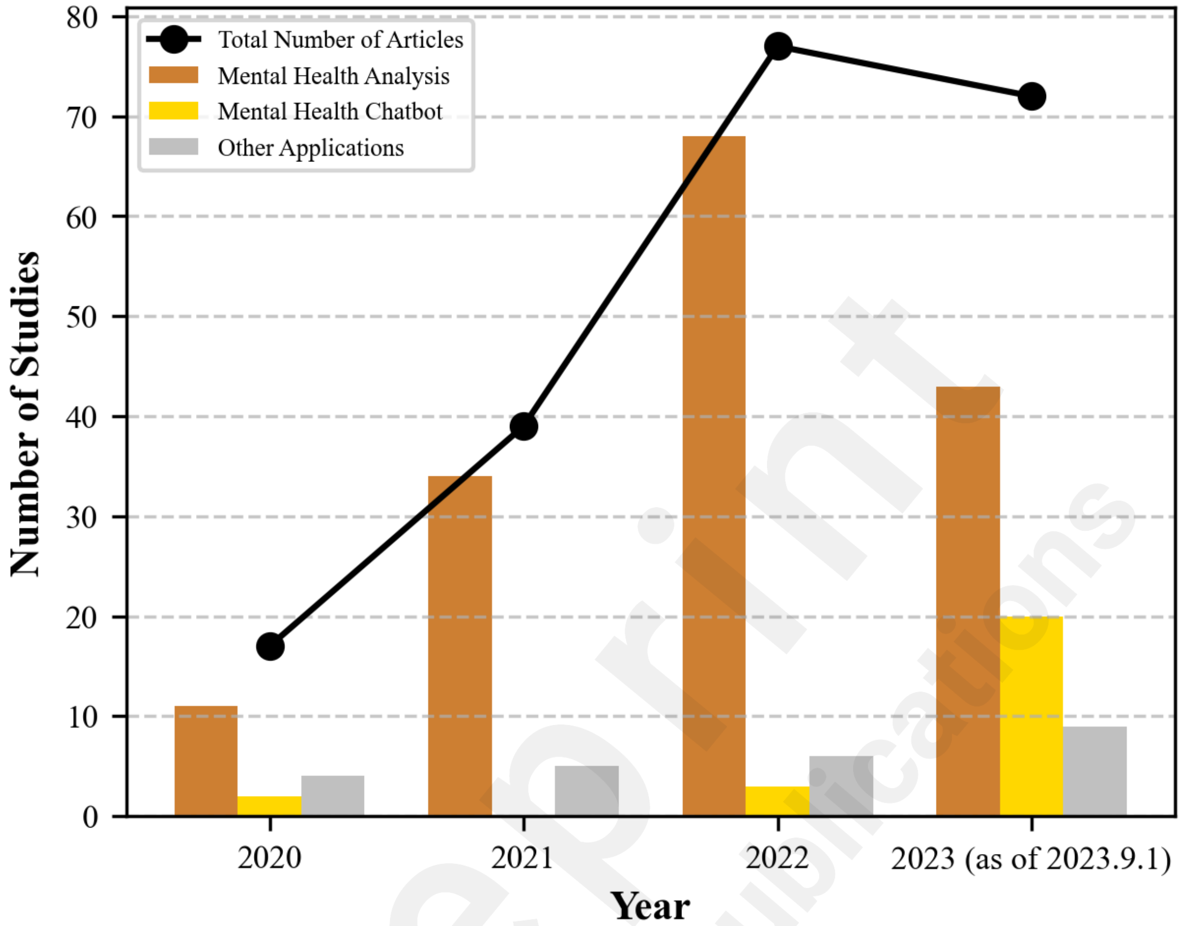


Fig 3. Number of articles after keyword search grouped by the year of publication and application field. The black line indicates the total number of articles in each year.

Table 1. Summary of the 13 selected articles from the literature on LLMs in mental health analysis using social media datasets.

Ref.	Cases	Models	Data Sources	Methodology Used	Main Outcomes
(Baird et al., 2022) 31	Analysis of telehealth topics using LLMs	BERT	Twitter (10,689 tweets)	Data from Twitter at the intersection of telehealth and mental health was processed with BERT, optimized for both pre and during-pandemic periods, followed by human evaluation and refinement.	A fourfold increase in tweets about telehealth for mental health or substance abuse during the pandemic was revealed, with a notable shift towards mental health topics in discussions about funding and support.
(Senn et al., 2022) 32	Depression classification by using LLMs	RoBERTa 33; DistilBERT 34; their ensemble combinations	Distress Analysis Interview Corpus - Wizard of Oz 35	The paper compared three BERT variants (BERT, RoBERTa, and DistilBERT) and four ensembles of BERT variants for depression classification using transcripts of 12 clinical interview questions.	Ensemble models of BERT variants outperformed individual BERT models in depression classification, with the ensemble models Ens 2 and Ens achieving the highest mean F1 score of 0.62.
(Xu et al., 2023) 36	Mental health prediction by using LLMs	Alpaca; Alpaca-LoRA; FLAN-T5; GPT-3.5/4.	Dreaddit dataset (posts from Reddit in domains: abuse, social, anxiety, PTSD, and financial)	This paper evaluated several LLMs using various setups such as zero-shot prompting, few-shot prompting, and instruction fine-tuning for mental health prediction tasks based on online text data.	Instruction fine-tuning significantly enhanced LLM performance for all tasks, with Mental-Alpaca and Mental-FLAN-T5 models surpassing GPT-3.5 and GPT-4 designs; ethical concerns and limitations were highlighted.
(Lamichhane et al., 2023) 37	Mental health classification by using LLMs	ChatGPT 3.5-turbo	Stress Detection Dataset 38; Depression Detection Dataset 39; Suicidality Detection 40	ChatGPT was employed to perform mental health classification, specifically stress detection, depression detection, and suicidality detection using labeled datasets from social media posts.	ChatGPT showed potential for mental health classifications, achieving F1 scores of 0.73, 0.86, and 0.37 for stress detection, depression detection, and suicidality detection respectively, outperforming a baseline model.
(Hassan et al., 2023) 41	Evaluation of ChatGPT in mental health	ChatGPT; GPT-3	11 benchmark datasets, including T-SID 42 and CAMS 43	The paper employed various prompting strategies for ChatGPT, enhanced with chain-of-thought and emotional cues, coupled with human evaluations on explanation quality.	ChatGPT had shown promise in mental health care with certain limitations, benefited from emotional cue prompts, and surpassed GPT-3 in explanation quality.
(Vajre et al., 2021) 44	Mental health classification by using LLMs	PsychBERT	Twitter hashtags and Subreddit (6 domains: anxiety, mental health, suicide, etc)	The paper developed a taxonomy based on HiTOP, implemented a two-stage framework for mental health text identification and behavior detection, and incorporated interpretability	The introduced framework, including the PsychBERT model, surpassed existing methods in mental health behavior detection from social media, proving both

				components.	effective and interpretable.
(El-Ramly et al., 2021) 45	Mental health prediction by using LLMs	ARABERT; MARBERT 46	CairoDep v1.0 (7,000 depressed and non-depressed posts in Arabic)	BERT models, specifically ARABERT and MARBERT, were trained and tested. Data collection encompassed crowdsourcing, Arabic forums, and translated English datasets, using Python.	Both ARABERT and MARBERT models achieved high accuracy, precision, recall, and F1-score values in detecting depression, surpassing traditional lexicon-based and ML approaches.
(Bajaj et al., 2021) 47	Mental health classification by using LLMs	Transformer-based classification models	Subreddits (50,242 samples)	A framework consisting of four parts was proposed to analyze Reddit users potentially affected by the pandemic. Transformer-based classification models were applied to this data, focusing particularly on the March-May period.	6.4% of the user base was mentally healthy before the pandemic. An observable relationship was found between the onset of depression and COVID-19, with a significant number of users starting to post about their struggles during the initial stages of the pandemic.
(Vishwakarma et al., 2021) 48	Speech emotion recognition by using LLMs	BERT; Multilayer perceptron; Convolutional neural network (CNN); Generative adversarial networks (GANs)	ISEAR 49; Emotion dataset 50; RAVDESS 51; TESSl 52; Emo-DB 53	The paper developed a personalized multi-modal architecture integrating text, speech, and facial expressions, using GANs for human-like interaction and lip-synced post-emotion analysis.	The multi-modal system surpassed existing single-mode emotion detection models in predicting cumulative emotional status and offering timely support through enhanced human-like GAN-generated responses.
(William et al., 2022) 54	Mental health prediction by using LLMs	BERT	Reddit's "rsuicidewatch" and "rcasualconversation" sub-forum (3,412 data); Twitter (from 1st Jan 2015 to 20th Sep 2020)	Using BERT combined with extractive summarization, data was pre-processed and benchmarked against other text classification techniques for depression detection, considering metrics like accuracy and F1-score.	Using BERT with extractive summarization enhanced depression detection on social media, outperforming base BERT and BiLSTM models, though XLNet remained superior in detection effectiveness.
(Kaseb et al., 2022) 55	Mental health prediction by using LLMs	Transformer-based pre-trained language models	Twitter (1,200 non-depressed tweets and 800 depressed tweets) 56; Kaggle (1,350k	Various pre-trained language models, including BERT and RoBERTa 33, were trained on a depression detection dataset, with RoBERTa then applied to pseudo-label datasets related to COVID-19 and vaccinations for	The RoBERTa model achieved a 78.85% F1-score and revealed, through pseudo-labeling, an increase in depression levels on tweets during the pandemic, highlighting the impact of

			tweets) 57; Kaggle (3,000k tweets) 58	depression insights.	COVID-19 and vaccinations.
(Zeberga et al., 2022) 59	Mental health prediction by using LLMs	BERT; Bi-LSTM	Reddit (95,000 posts); Twitter (100,000 tweets)	A framework was designed integrating BERT, Bi-LSTM, word2vec, and knowledge distillation for depression and anxiety detection from social media.	The developed model successfully detected depression and anxiety-related posts with a 98% accuracy rate.
(Heinz et al., 2023) 60	Evaluation of AI models in mental health assessment	GPT-3	59 distinct clinical vignettes	The paper employed a generative AI model, tested it with clinical vignettes, and utilized balanced accuracy (BAC), generalized linear mixed-effects models, and odds ratios (ORs) to analyze domain knowledge and demographic bias.	The AI model displayed variable diagnostic performance with high BAC for certain psychiatric disorders and low BAC for others, underscoring the need for caution and further development before deployment in critical healthcare settings.

Table 2. Summary of the 10 selected articles from the literature on LLMs in mental health chatbot.

Ref.	Cases	Models	Data Sources	Methodology Used	Main Outcomes
(Wei et al., 2023) 61	Designed mental chatbot using LLMs to investigate the design factors of prompts	GPT-3	User Self-Reported (48 participants)	Using an online study with 48 participants, four chatbot prompt designs were assessed for dialogue flow and user perception. to correct these harmful behaviors in AI chatbots.	Chatbots captured 79% of desired information, with prompt design significantly affecting conversation quality and data collection.
(Lin et al., 2023) 62	Designed models using LLMs to correct harmful behaviors in AI chatbots	GPT-3.5	Conservations between chatbots and a hypothetical user	The SafeguardGPT framework was introduced, incorporating four AI agents, and its effectiveness was demonstrated via a social conversation simulation.	SafeguardGPT effectively detected and corrected harmful chatbot behaviors, but challenges in evaluation and alignment with human values persisted.
(Kumar et al., 2022) 63	Designed mental chatbot for managing mood using LLMs	GPT-3	Survey responses from Amazon Mechanical Turk (945 participants)	The paper centered on the GPT-3 chatbot's prompt design dimensions, specifically identity, intent, and behavior, and applied both quantitative and qualitative analyses of user interactions and perceptions.	Users found the chatbot was helpful in many scenarios, but raised concerns about repetitiveness and privacy, with certain prompt designs showing promise around problem-solving and cognitive behavioral therapy.
(Lai et al., 2023) 64	Designed mental chatbot	PanGu; WenZhong	PsyQA dataset (5000 samples) 65	The paper introduced the creation and assessment of the Psy-LLM	The Psy-LLM framework effectively generated coherent and relevant

	in psychological consultation settings using LLMs	Model		framework. It merged pre-trained LLMs with expert inputs and articles, analyzed data via word and sentence metrics, and measured outcomes using both intrinsic metrics and human feedback on effectiveness and applicability.	answers to psychological questions, held the potential to enhance mental health support using AI, and improved overall societal well-being.
(Crasto et al., 2021) 66	Designed mental chatbot with LLMs for student mental health support on an online platform	DialoGPT	Counselchat (includes tags of illness); question answers from 100 college students	Recognized mental health questionnaires (Patient Health Questionnaire-9 & WHO-5) were completed. The DialoGPT fine-tuned with Counselchat data, was employed for chatbot interaction. Micro-interventions were suggested based on identified issues, and a student survey was administered.	The DialoGPT model, demonstrating higher perplexity and preferred by 63% of college participants for its human-like and empathetic responses, was chosen as the most suitable system for addressing student mental health issues.
(Chen et al., 2023) 67	Designed mental health chatbots using LLMs and assessed their viability in psychiatric outpatient scenarios	ChatGPT	Human evaluation results (14 patients and 11 psychiatrists)	Using a human-centered design, the paper collaborated with psychiatrists to harness ChatGPT for simulating psychiatrist-patient interactions, influenced by real-world scenarios, and evaluated through interactions with professionals and patients.	The paper revealed a ChatGPT-backed evaluation framework that validated the effectiveness of chatbots in psychiatric contexts based on evaluations with real psychiatrists and patients.
(Cabrera et al., 2023) 68	A review of the mental chatbot for bioethical dilemmas	ChatGPT 4.0	33 scientific abstracts; 13 media	Scientific literature and media news were systematically reviewed using predefined criteria on the Web of Science and Microsoft Bing search engines, focusing on the relationship between chatbots and mental health.	Bioethical dilemmas about chatbots in mental health were systematically identified and classified into four major areas, with a call for tailored development and ethical regulation.
(Fournier-Tombs et al., 2023) 69	A review of mental chatbots to inform safe and appropriate future developments in the use of chatbots in	GPT-3	Literature, policies, and recommendations from the European Union, UNESCO and WHO	The paper employed a conceptual analysis of conversational chatbots in medicine, likely informed by a literature review, focusing on their ethical implications, and proposed an integrated framework connecting AI and medical ethics.	A framework was proposed to understand the impacts of conversational chatbots on patients and the broader medical community, with an emphasis on aligning AI ethics with traditional medical ethics, in hopes of guiding future development and regulations in a safe and relevant manner.

	healthcare				
(Jo et al., 2023) 70	Designed mental health chatbot using LLMs to provide support for socially isolated individuals in public health interventions	HyperCLOVA	34 audio-recorded interviews; observational notes; codebook (10 parent and 24 child codes)	Focus group workshop sessions with 14 CareCall users were observed, and interviews with 20 people spanning three stakeholder groups (CareCall users, teleoperators, and developers) were conducted, to understand the benefits, challenges, and perspectives of using LLM-based chatbots in public health.	Insights into the holistic support provided by CareCall to mitigate loneliness, offload public health workloads, and challenges LLM-driven chatbots pose for public and personal health were highlighted, and considerations for the design and deployment of such chatbots in public health interventions were addressed.
(Webster et al., 2023) 71	A review of mental chatbot: examining the capabilities, concerns, and ethical considerations	ChatGPT; Bard; Med-PaLM	MultiMedQA; HealthSearchQA Database; PaLM Training Corpu; research papers	The paper investigated the capabilities, limitations, and potential risks of AI chatbots like Med-PaLM in the medical field by exploring their training, accuracy in medical exams, representation of doubt, and potential for misdiagnosis.	Concerns were raised about the potential inaccuracy of AI chatbot diagnoses and the importance of medical governance, informed consent, and the collaboration of AI scientists with clinicians. OpenAI, the maker of ChatGPT, warned against using their model for critical medical decisions.

Table 3. Summary of the 9 selected articles from the literature on other applications of LLMs in mental health.

Ref.	Cases	Model	Data Sources	Methodology Used	Main Outcomes
(Salah et al., 2023) 72	Evaluation of ChatGPT in mental health	ChatGPT	Literature review; research papers; online textual data	The study offered a thorough overview of ChatGPT's application in social psychology research, discussed ethical, theoretical, and methodological challenges, and underscored the importance of a theoretical framework that integrates Generative AI with current social psychology theories.	The study found that ChatGPT could transform social psychology research through data analysis and modeling of social interactions, but researchers must address associated challenges and follow guidelines for ethical and responsible use, including bias management, data validation, and adherence to privacy standards.
(Farhat et al., 2023) 73	Evaluation of ChatGPT in mental health	ChatGPT	Responses generated by ChatGPT	The study evaluated ChatGPT's effectiveness in mental health support by analyzing its responses and cross-questioning, particularly focusing on issues related to anxiety and depression and its suggestions regarding medications.	ChatGPT displayed significant inconsistencies and low reliability when providing mental health support for anxiety and depression, underlining the necessity of validation by medical professionals and cautious use in mental health contexts.
(Woodnutt et al.,	Evaluation of	ChatGPT	Responses	The study input basic text commands	The study found that OpenAI's ChatGPT

2023) 74	ChatGPT in mental health		generated by ChatGPT	into ChatGPT regarding a fictitious person with self-harming tendencies, and the output was assessed for quality, accuracy, errors, ethical concerns, and potential harms using the authors' clinical expertise and current care guidelines.	generated outputs with significant errors and ethical issues, presenting a risk of potential harm if used in mental health care without expert oversight; AI's use could decrease the quality of care provided by nurses and affect aspects of recovery tied to personal relationships and social interactions.
(Elyoseph et al., 2023) 75	Evaluation of ChatGPT in mental health	ChatGPT	Responses generated by ChatGPT	The Levels of Emotional Awareness Scale (LEAS) was used to measure ChatGPT's emotional awareness (EA) by analyzing its responses to twenty scenarios. This performance was then compared with two evaluations conducted on ChatGPT in January and February 2023 using different versions of the model.	ChatGPT showcased a superior performance on the LEAS scales compared to the general population. The AI's scores were particularly high in the second evaluation, indicating potential improvement over time, which suggests its capability to generate appropriate EA responses and its utility in clinical applications.
(Bhattacharyya et al., 2023) 76	Evaluation of ChatGPT in mental health	ChatGPT	N/A	The paper focused on ChatGPT's application in mental health, its ethical considerations, its interaction dynamics, and its synergy with other digital health tools.	ChatGPT was highlighted for its potential in mental health, its ability to work alongside other tools, and the need for ethical caution due to potential inaccuracies.
(Qiu et al., 2023) 77	Solutions to mental health data problems	SMILE	ESCon 78; AugESC 79; PsyQA 65	Utilizing the SMILE approach, the paper expanded single-turn dialogues into multi-turn ones using ChatGPT and validated this through an exploratory study and contrastive analysis.	The SMILE method effectively produced a comprehensive, real-life-like multi-turn mental health support conversation dataset, with utterance lengths aligning with genuine counseling sessions.
(Perlis et al., 2023) 80	Evaluation of ChatGPT in mental health	ChatGP; GPT-4	10 antidepressant prescribing vignettes	ChatGPT-4 was given 10 antidepressant prescribing vignettes in a randomized order, results were regenerated five times for consistency, and model outcomes were then juxtaposed with expert clinician consensus.	GPT-4 included at least one optimal medication choice in 76% of vignettes but also presented less optimal or contraindicated choices in 48% of them.
(Elyoseph et al., 2023) 81	Evaluation of ChatGPT in mental health	ChatGPT	A hypothetical text vignette	ChatGPT's evaluations of the vignette were contrasted with mental health professional norms reported by Levi-Belz and Gamliel using two-sample t-tests.	ChatGPT consistently underestimated the risk of suicide attempts compared to mental health professionals, suggesting it may not provide accurate suicide risk assessments.
(Egli et al., 2023)	Evaluation of	ChatGP;	N/A	A descriptive analysis of how chatbot	The paper delved into the potential

82	ChatGPT in mental health	GPT-4		technologies, like ChatGPT and GPT-4, operate and explores their potential applications, strengths, and weaknesses in clinical microbiology.	applications of LLMs in clinical microbiology, focusing on their functionalities, quality-control measures, and potential biases in their training data.

3.2 Mental Health Analysis Using Social Media Datasets

Early intervention and screening are crucial in mitigating the global burden of mental health issues 33. We examined LLMs' performance in predicting mental h c ealth a onditions m h nd i ategorizing t er analysis. For instance, by integrating extractive summarization improve depression detection on social media platforms like Twitter by outperforming the base BERT model 54. ARABERT (F1-Score: 96.92%) and MARBERT (F1-Score: 96.07%) also excelled in Arabic contexts, surpassing traditional ML techniques 45. Additionally, the RoBERTa model, a fine-tuned transformer-b effectively tracked the rise in depression levels on T COVID-19 pandemic, outperforming LSTM (F1-Score: 78.85% vs. 55. Mental-Alpaca and Mental-FLAN-T5, improved LLMs' performance with fewer prompts for mental health tasks, meanwhile highlighting limitations such as racial and gender bias issues 36.

ChatGPT can also detect mental health issues tha baseline m t odels t a hrough F s ext 0 nalysis, a 0 ith in s tress d etection, d epression d etection, nd 37. Moreover, Hassan et al.'s research on ChatGPT contextual learning capabilities in mental health analysis but also revealed shortcomings in comparison to advanced task-specific methods 41. LLMs, particularly BERT, have been instrumental in analyzing health text dialogues. A study analyzing Twitter data indicated a fou increase in telemedicine discussions related to mental health and substance abuse d uring t he p andemic, h ighlighting n 31. Ensemble models combining various BERT variants RoBERTa, a D nd d istilBERT, e p emonstrated i c nhance depression 32. Continuous advancements are notable in th instance, P sychBERT b as i stablished m h enchmarks behaviors on social media, providing high interpretability for both bi and multi-class classification 44.

3.3 LLMs in mental health chatbot

The use of a mental health chatbot employing LLMs shows promise for early intervention i m i n ental illnesses 71. These r emote t ntervention can encourage people who were reluctant to seek health-r 62 t o m ake i nteractions, c a h s ut hey an h lso elp systems 69. Multiple studies have shown that LLMs are effective in creating chatbots for mental health use. Crasto et al. observed that 63 out o college students surveyed showed a prefe DialoGPT, a variant of the GPT architecture, over those from LSTM and RNN models 66. The participants noted that DialoGPT-gen appeared more human-like and resonant. However, it was noted t response q f uaGlity m womv PTW e aodiels t ast ariable. format and personality modifiers in prompt design significantly influ the chatbots' capability in slot fi 61. This finding was further supported by Kumar et al.'s research with 945 participants, which revealed that p perceptions of chatbot risk, trustworthiness, expertise, and willingnes interact 63.

The COVID-19 pandemic has significantly in psychological counseling. The Psy-LLM framework by integrates pre-trained LLMs with professional Q&A from psychologists and psychological articles, has been evaluated for coherent, relevant responses to psychological inq 64. uiriThe criteria for these evaluations included usefulness, fluency, relevance, and logic, proving the Psy-LLM to be an effective front-end tool for healthcare professionals to enhance healthcare effic 64. ieHncoywever, the deployment of this single case does not alleviate the safety concerns of LLMs i Webster h ighlighted o m he L imitatioMs i d f edical patient care due to limited healthcare data despite high scores in medic examin 71. ationsConsequently, establishing framework focused on informed consent is crucial.

To a ddress c thical C hallenges, b abdera i t l. chatbot use in mental health, including issues of privacy accountability, conflicts of interest, cultural sensit 68. Fournier-Tombs et al. emphasized the expanded role of v during and following the COVID-19 pandemic for a range of health services, including providing COVID-19 information, personalized health advice, and prescription a 69. ssistance However, p his e henomenon t i xace of b oundaries A a etween a t tI pplications h nd h professionals. To mitigate these concerns, inclu gender, race, and data from vulnerable populations, an ethical frame was discussed to guide chatbot deployment based on the core principles of medical e i thics, b ncluding eneficence, a j onmalefice 69.

3.4 Exploring other applications of LLMs in mental health

ChatGPT has gained attention for its advar capabilities, attracting the interest of many researchers and practit 82 due to its unparalleled ability to generate human-like text and analyz large amounts of textual data 72. ChatGPT has been used to address the scarcity of comprehensive data on problems in mental health. The SMILE approach uses ChatGPT to expand single rounds of dialo extensive, multi-turn conversations and enables the creation of large-scale, diverse, and close-to-real-life mental health support conversation cor 77. However, this approach has not been thoroughly and reliably evaluated yet, and the virtual dialogue system exhibits frequent an behavior 77. In the context of social psychology, ChatGPT has demonstrated potential in analyzing complex human behaviors and interactions, as discussed by Salah et al. These investigations into ChatGPT have shown mixed results. Elyoseph et al. highlighted ChatGPT's proficiency in EA, noting that ChatGPT's performance in generating ap responses was impressive, achieving nearly the highest possible score (9 out of 100) on a subsequent assessment using the LEAS 75.

Despite this, the model demonstrated limitations in accurately eval suicide risk, falling short of the assessments ma professionals 81. In psychopharmacology, C c i hile hatGPT at least one of the optimal 7 medications n 6% contraindicated drugs in 48% of vignettes 80, giving the caveat that it does not operate independently in medication management. Furthermore, Farhat et al. observed inconsistencies in ChatGPT's mental health supp varying prompts leading to different outputs 73. Similarly, Woodnutt et al. highlighted ethical concerns regarding the use of ChatGPT in mental health care strategies, particularly pointing out potential inaccuracies and ethical dilemmas in handling sensitive topics 74. These findings underscore the vital importance of meticulous supervision by healthcare professionals and institutions in the application of LLMs like ChatGPT environments.

3.5 Strengths and limitations of using LLMs in mental health

Based on the works of literature the strengths and weaknesses of applying the LLMs in mental health are summarized in Table 4.

LLMs h d ave emtonstpated t c aheirp otentialh o la problems through text a 23 1 73, a 9 4 a,4 1 5y 4 s 7,5 5 5 9 LLMs have shown the potential for the early detection of mental illness using media37, thereby possibly enhancing treatment outcomes and alleviati healthcare burdens. In the context of the global mental health crisis, LLMs present a promising digital intervention strategy that could m effects of the current shortage of healthcare resources, particularly during crises such as COVID-1649 The LLM is adept at predicting the emotional trends of crowds from social media texts, which can be used to increase the social awareness of specific groups at particular times and encourage early intervention in areas with limited medical resources. Additi excel i u ni ser o nteractaion, a r fferings w nonymity encourage more people suffering from mental illness to actively participate in their treatment66. The ability of LLMs to personalize interactions caters to the diverse interests of different user groups thus enhancing engagement 70. Meanwhile, it is crucial to recognize that mental health chatbots are not substitutes for professional psychologi revealing t p heo L oteinthial n y b f r LM, f t c as u ot et due to technological risks and ethical issues that remain to be thorough addressed.

The application of LLMs in mental health, particularly those fine-tuned for specific instructions like ChatGPT, re effectiveness of instruction fine-tuned LLMs specificity of user-generated prompts. In these models, inadequate prompts can disrupt dialogue flow and red 68, leading to inconsistent dialogue quality due to users' unfamiliarity with prompt engineering Deviations in training data for LLMs inevitably lead to issues like hallucinations and instability, which are challenging to 76. Another critical concern is the 'black box' nature of LLMs, which raises interpretability issues, particularly in clinical and mental heal 73 This lack of interpretability complicates the applicati mental health, where trustworthiness and clarity are important. When talk about neural networks as black boxes, we know what they were trained with, how they were trained, what the weights are, etc. However, with many new LLMs like GPT 3.5/4, researchers and practitioners often acc models via web interfaces or APIs without complete kn 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 wit LLMs in mental health. Debates are emerging around issues personhood, informed consent, the ri appropriateness of AI in mimicking human interactions 81. Additionally, human rights risks such as discrimination are pivotal concerns. There is a possibility that LLMs could generate inconsistent display increased error rates influenced by patient demographics su gender, ethnicity, race, and religion, stereotypes or biased perspectives 69. Accompanying these issues worries about data security and violations of user privacy and auto 62. Therefore, there is a need to strengthen the LLM as an adjunctive tool in the mental health field, which is a key focus for future exploration and action.

Table 4: Summary of main strengths, limitations, and suggestions of LLMs in mental health from the selected articles.

CATEGORY	STRENGTH	LIMITATION	SUGGESTION
MENTAL HEALTH ANALYSIS	• LLMs have shown great potential in mental health categorization tasks and mental health analysis, especially in stress and depression detection 37,41,47,48. • LLMs can show better performance and accuracy by integrating with other models 32,36,44,55,59. • LLM can measure emerging trends related to mental health at scale. This can inform public health strategies and resource allocation 47.	• LLMs display diminished efficacy in non-English settings and struggle with identifying complex psychological conditions 37,47,60. • Assessment relies heavily on dataset annotations, and there are limitations to the use of sentiment and emotion lexicons due to annotation bias, limited vocabulary, and the evolving nature of online language 37,60. • Most of the studies used limited prompt settings and relied on the first response of the LLMs as a prediction. Different prompts or variations may produce more optimized or varied results 37.	• Pay more attention to social media in other countries where other languages are dominant and build databases specifically for depression screening and training of LLMs 31. • The current dataset for stress and depression testing can be re-annotated by multiple psychologists to minimize bias 37. • Future experiments should broaden datasets, models, and prompt designs, and refine evaluations to improve categorization accuracy through varied prompt settings and response analysis across iterations 36,37.
MENTAL HEALTH CHATBOT	• The LLMs show potential as a mental health intervention tool and the ability to generate coherent and relevant answers to psychological questions 64. • LLMs-driven mental chatbots can reduce the burden of healthcare to some extent, and the online dialogue approach offers the possibility of telemedicine 70. • LLMs-driven mental chatbots can help users reduce loneliness and emotional burden and reduce the stigma associated with mental health as well as prejudice 62,70.	• Over-reliance on prompts results in the quality and relevance of the generated response being directly related to the accuracy of the input prompts, leading to inaccurate responses if the prompts are inadequate 68. • Inappropriate responses persist in the public health context, in part because LLMs rely on inherently biased training data 62,70. • LLM-driven chatbot output is influenced by user biases and conversational styles, which can impact its message slot-filling performance 61,68.	• Refine prompts by integrating sample dialogues for incremental learning to sharpen questioning abilities and broaden the scope of prompt dimensions 61,62. • Customising LLMs-driven chatbots to target specific groups of people and expanding the dataset used to train LLMs through a wide range of sources representing more diverse demographics, perspectives, and scenarios 70. • Advanced techniques like multi-agent reinforcement learning enable chatbots to adapt slot-filling strategies to conversation context, moving beyond static training data 62.
	• The LLMs (ChatGPT) outperform humans in the assessment of EA, recognizing and describing emotions	• LLMs may generate 'hallucinatory' content, presenting inconsistencies and difficulties in grasping the nuances of	• Detailed documentation of training datasets, shared model architectures, and third-party

OTHER APPLICATIONS OF LLMS IN MENTAL HEALTH	in specific scenarios 70. • The LLMs can be invoked to build a corpus of mental health support dialogues using its inclusive linguistic scalability from single-turn to multi-turn 77. • LLMs improve the accuracy of data analysis and the fluency of interactions 72.	social language 73,7680. • LLMs that mimic human communication prompt critical ethical discussions on digital identity, consent, manipulation risks, and the potential for deceptive simulations when users are unaware, they're engaging with AI 73,82. • LLMs are not yet precise enough for standalone clinical application in mental health diagnostics and therapy and lack interpretability 76.	audits; outlining logical relationships and facts with knowledge graphs 84. • Rigorous risk assessments and enhanced transparency are essential for user interactions with AI; moreover, establishing channels for user feedback on AI and developing stringent ethical codes and industry standards is imperative 22,36,62,70,73,74,76,81. • To build and refine a large database of LLMs specifically for mental health training 77, to improve the performance of LLMs in empathy 74, psychopharmacology 80, etc., and to develop their potential as supportive tools 73.

4. Discussion

4.1 Principal findings

In the context of the wider prominence of LLMs in the literature 72,76,82, our research supports the assertion that interest in LLMs is growing in the field of mental health. Figure 3 shows the increasing trends in mental health studies using LLMs, while we note that our search for 2023 ended on 1st September so it only includes part of the year's data. Key areas of interest identified by our work are mental health chatbots and the use of LLMs on social media datasets for primary mental health screenings. This likely reflects the promise of these two contexts of use as opportunities for LLM-driven strategies that can be scaled at a low cost to improve mental healthcare provision. This may be particularly relevant in scenarios where the existing capacity to provide care is limited. We also note that none of the existing work discovered classes as the strong standard of clinical evaluation evidence required to support live clinical use.

4.2 Limitations of the Selected Articles

Beyond a lack of high-quality clinical evidence, much of the work discovered falls outside the peer-reviewed literature. As detailed in Tables 1, 2, and 3 listing the relevant articles, 11 of the 32 articles were from arXiv without peer review. These include 4 articles on mental health analysis using social media datasets, 6 on mental health chatbots, and 1 on other applications of LLMs in mental health. Articles from arXiv were marked in the tables. The scarcity of peer-reviewed literature in the public domain presents a significant challenge for new researchers in the field.

Throughout the literature review, several further research limitations were identified. A key concern is the age bias in social media data used for depression and mental health screening. Users of social media skew towards younger demographics, leading to the underrepresentation of older age groups. This skew is further compounded by the frequent absence of crucial metadata such as age and gender, adding to the problem of representation in datasets 47. Language barriers also shape the current development of the field. While tools like ChatGPT show proficiency in English, their accuracy in other languages can still be lacking 76. The resultant focus on English-centric social media platforms can overlook insights from non-English-speaking regions. Future research could gain from incorporating data from platforms prevalent in these areas to provide a more holistic view 31. Another limitation is the diversity of LLMs used was low. Most articles in our review focused on variants of BERT and ChatGPT. Therefore, this review provides only a limited picture of the variability we might expect in applicability between different LLMs. Another limitation is the rapid evolution of LLMs. For example, studies using GPT-3.5-turbo do not incorporate advancements in subsequent models like GPT-4, making it hard to just find the applicability to subsequent models 37. Additionally, the common practice of binary analysis in these studies risks oversimplifying complex psychological conditions by categorizing social media posts merely as 'depressed' or 'non-depressed' 55. In the case of complex mental health conditions, assessment is often more subjective, relying on expert judgment or assessment of people’s behavior (rather than just text or voice fragments).

4.3 Opportunities and Future Work

Implementing technologies involving LLMs within the healthcare provision of real patients demands thorough and multi-faceted 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, the latest version of ChatGPT-4 introduces the ability to process and interpret image inputs within conversational contexts, leading OpenAI to issue an official warning that ChatGPT-4 is not approved for analyzing specialized medical images, such as CT scans 85.

A key challenge to address in LLM research is the tendency to produce incoherent text or hallucinations. This is a reasonable cause of concern in considering LLM’s role in future mental healthcare 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. The creation of specialized datasets could leverage the customizable nature of LLMs, fostering the development of models that cater to the distinct needs of varied demographic groups. For example, in contrast to models designed for healthcare professionals to assist in tasks like data documentation, symptom analysis, medication management, and postoperative care, LLMs for patient interaction might be trained with an emphasis on empathy and comfortable dialogue.

Data privacy is another significant area of concern. Many LLMs, like ChatGPT and Claude, involve sending data to third-party servers opening up the risk of data leakage. One potential solution lies in locally hosting open-source models or deploying LLMs in private clouds, enabling enhanced control over data storage and access 86.

The lack of interpretability of LLM decision-making is another important area for new work on healthcare 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 like 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 84.

Another area for research is reducing LLMs' dependence on user inputs, by optimizing prompt design. Future developers can explore advanced reinforcement learning approaches, such as multi-agent reinforcement learning, which enhance LLMs' capabilities to learn from interactions and advance their understanding of natural language 62. In the realm of mental health, LLMs' abilities for sentiment analysis, personalized responses, and empathy are especially important. Using randomized factorial experiments could deepen understanding of prompt design 63. Additionally, expanding the scope of experimental testing in prompt engineering, including the strategic use of minimal learning with psychologically validated dialogue examples, can refine the models' questioning acumen. Further exploration is required to see how various parameters influence prompt effectiveness, alongside experimental investigations into zero-shot, one-shot, and few-shot prompting 41. These investigations are expected to reveal how different prompt designs might impact LLM output in terms of accuracy, reliability, and ultimately applicability.

Prior to deployment in the mental health sector, it is also imperative for medical professionals to rigorously evaluate these models to prevent any intervention that might cause harm. Developing ethical guidelines is a priority for future research. In clinical settings, combining LLMs with physician oversight could enhance effectiveness. For example, while ChatGPT has demonstrated initial competence in recommending medication, it is not appropriate to be used independent of clinician scrutiny. However, if viewed instead as a decision-making aid, it could save the physician's time and increase efficiency. Future evaluation schemes might include combined impact and expert assessments to investigate criteria such as reliability, security, fairness, resistance to misuse, interpretability, adherence to social norms, robustness, performance, linguistic accuracy, and cognitive competence 87. These measures are fundamental for creating ethical frameworks suitable for mental health care.

4.4 Conclusion

This review thoroughly examines LLMs in mental health, covering social media data analysis, chatbots, and their evaluation and solutions. Despite their potential, challenges like training data bias, model accuracy, and ethical concerns persist. As data quality and ethical guidelines improve, LLMs are expected to become more integral and important as they provide an alternative solution to mental health, this global healthcare issue.

4.5 Contributors

ZG and KL contributed to the conception and design of the study. ZG and KL also contributed to the development of the search strategy. Database search outputs were screened by ZG, and data were extracted by ZG. An assessment of the risk of bias of 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.

4.6 Acknowledgements

This work was funded by the UKRI Centre for Doctoral Training in AI-enabled healthcare systems (grant EP/S021612/1). The funders were not involved in the study design, data collection, analysis, publication decisions, or manuscript writing. The views expressed in the text are those of the authors and not those of the funder.

4.7 Conflicts of Interest

The authors declare no conflict of interest.

4.8 Data sharing statement

The authors ensure that all pertinent data have been incorporated within the article and/or its supplementary materials. For access to the research data, interested parties may contact the corresponding author, Kezhi Li (ken.li@ucl.ac.uk), subject to a reasonable request.

4.9 Abbreviations

BAC: balanced accuracy
BERT: Bidirectional Encoder Representations from Transformers
CNN: Convolutional Neural Network
DBLP: DBLP Computer Science Bibliography
EA: emotional awareness
GANs: General Adversarial Networks
GPT: Generative Pre-trained Transformer
IEEE: IEEE Xplore
LEAS: Levels of Emotional Awareness Scale
LLM: large language model
ML: machine learning
NLP: natural language processing
OR: odds ratio
PRISMA: Preferred Reporting Items for Systematic Review and Meta-analysis
WHO: World Health Organization

5. Multimedia Appendix 1

Supplementary material 1: Risk of bias assessment

	Study	Selection Bias	Performance Bias	Detection Bias	Attrition Bias	Reporting Bias	Overall Risk of Bias
1	Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data	Low	Low	Low	Low	Low	Low
2	LLM-Empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation	Low	Low	Low	Low	Low	Low
3	Ethical Dilemmas, Mental Health, Artificial Intelligence, and LLM-Based Chatbots	Low	Low	Low	Low	Low	Low
4	A Medical Ethics Framework for Conversational Artificial Intelligence	Moderate	Low	Low	Low	Low	Low
5	Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention	Moderate	Low	Low	Low	Low	Low
6	Towards Healthy AI: Large Language Models Need Therapists Too	Low	Low	Low	Low	Low	Low
7	Exploring The Design of Prompts for Applying GPT-3 Based Chatbots: A Mental Wellbeing Case Study on Mechanical Turk	Low	Low	Low	Low	Low	Low
8	Psy-LLM: Scaling Up Global Mental Health Psychological Services with AI-Based Large Language Models	Low	Low	Low	Low	Low	Low

9	Carebot: A Mental Health Chatbot	Moderate	Low	Low	Low	Low	Low
10	May The Force of Text Data Analysis Be with You: Unleashing the Power of Generative AI for Social Psychology Research	Unclear	Unclear	Unclear	Low	Low	Low
11	Chatgpt as a Complementary Mental Health Resource: A Boon or A Bane	Unclear	Unclear	Unclear	Low	Low	Low
12	Testing Domain Knowledge and Risk of Bias of A Large-Scale General Artificial Intelligence Model in Mental Health	Moderate	Low	Low	Low	Low	Low
13	Could Artificial Intelligence Write Mental Health Nursing Care Plans?	Moderate	Low	Low	Low	Low	Low
14	ChatGPT Outperforms Humans in Emotional Awareness Evaluations	Low	Low	Low	Low	Low	Low
15	ChatGPT and Its Application in The Field of Mental Health	Moderate	Low	Low	Low	Low	Low
16	SMILE: Single-Turn to Multi-Turn Inclusive Language Expansion Via ChatGPT for Mental Health Support	Moderate	Low	Low	Low	Low	Low
17	Research Letter: Application of GPT-4 to Select Next-Step Antidepressant Treatment in Major Depression	Moderate	Low	Low	Low	Low	Low
18	Beyond Human Expertise: The Promise and Limitations of ChatGPT in Suicide Risk Assessment	Low	Low	Low	Low	Low	Low
19	ChatGPT, GPT-4, and Other Large Language Models – The Next Revolution for Clinical Microbiology?	Unclear	Unclear	Unclear	Low	Low	Low
20	Medical AI Chatbots: Are They Safe to Talk to Patients?	Unclear	Unclear	Unclear	Low	Low	Low
21	Consumer Perceptions of Telehealth for Mental Health or Substance Abuse: A Twitter-Based Topic Modeling Analysis	Moderate	Low	Low	Low	Low	Low
22	Ensembles of BERT for Depression Classification	Moderate	Low	Low	Low	Low	Low
23	Mental-LLM: Leveraging Large Language Models for Mental Health Prediction Via Online Text Data	Moderate	Low	Low	Low	Low	Low
2	Evaluation of ChatGPT for NLP-Based	Moderate	Low	Low	Low	Low	Low

4	Mental Health Applications						
25	Towards Interpretable Mental Health Analysis with ChatGPT	Moderate	Low	Low	Low	Low	Low
26	Psychbert: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis	Low	Low	Low	Low	Low	Low
27	Cairodep: Detecting Depression in Arabic Posts Using BERT Transformers	Low	Low	Low	Low	Low	Low
28	Mental Health Analysis During COVID-19: A Comparison Before and During the Pandemic	Moderate	Low	Low	Low	Low	Low
29	An Emotionally Aware Friend: Moving Towards Artificial General Intelligence	Low	Low	Low	Low	Low	Low
30	Leveraging BERT With Extractive Summarization for Depression Detection on Social Media	Moderate	Low	Low	Low	Low	Low
31	Analysis on Tweets Towards COVID-19 Pandemic: An Application of Text-Based Depression Detection	Moderate	Low	Low	Low	Low	Low
32	A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model	Low	Low	Low	Low	Low	Low

Table S1: Risk of bias assessment

6. Multimedia Appendix 2

Supplementary material 2: List of studies excluded at full-text screening stage

	Title	Year	Author	Exclusion reason
1	Global Mental Health Services and the Impact of Artificial Intelligence-Powered Large Language Models	2023	Alastair C. van Heerden	Review paper, too short
2	Safety Profile of Methylphenidate Under Long-Term Treatment in Adult ADHD Patients - Results of the COMPAS Study	2020	Bernhard Kis	Not about LLMs
3	Mental Health and Discrimination among Migrants from Africa: An Italian Cross-Sectional Study	2022	Gianluca Voglino	Not about LLMs
4	Chat-GPT: Opportunities and Challenges in Child Mental Healthcare	2023	Nazish Imran	Review article, too short
5	The Emergent Role of Artificial Intelligence, Natural Learning Processing, and Large Language Models in Higher Education and Research	2023	Tariq Alqahtani	Not about mental health
6	Mental Health and Adherence to Mediterranean Diet among University Students: An Italian Cross-Sectional Study	2021	Giuseppina Lo Moro	Not about mental health

7	Exploring Cyberaggression and Mental Health Consequences among Adults: An Italian Nationwide Cross-Sectional Study	2023	Giuseppina Lo Moro	Not about mental health
8	Can Natural Language Processing Models Extract and Classify Instances of Interpersonal Violence in Mental Healthcare Electronic Records: An Applied Evaluative Study	2022	Riley Botelle	Not about LLMs
9	Effects of Covid-19 Lockdown on Mental Health and Sleep Disturbances in Italy	2020	Maria Rosaria Gualano	Not about LLMs
10	The Culture of Health in Early Care and Education: Workers' Wages, Health, And Job Characteristics	2019	Jennifer J. Otten	Not about LLMs
11	ChatGPT on ECT: Can Large Language Models Support Psychoeducation?	2023	Robert M Lundin	Review article, too short
12	Effectiveness of Guided and Unguided Online Alcohol Help: A Real-Life Study	2022	Ans Vangrunderbeek	Not about LLMs
13	Listening to Mental Health Crisis Needs at Scale: Using Natural Language Processing to Understand and Evaluate a Mental Health Crisis Text Messaging Service	2021	Zhaolu Liu	Not about LLMs
14	Emotional Eating and Depression During the Pandemic: QuarantEat, an Italian Nationwide Survey	2022	Giuseppina Lo Moro M.D.	Not about LLMs
15	Technology Enhanced Health and Social Care for Vulnerable People During the COVID-19 Outbreak	2021	Evangelia D Romanopoulou	Not about LLMs
16	A CNN-Transformer Hybrid Approach for Decoding Visual Neural Activity into Text	2022	Jiang Zhang	Not about mental health
17	Multimodal Automatic Coding of Client Behavior in Motivational Interviewing	2020	Leili Tavabi	Not about LLMs
18	The Effectiveness of Psychological Interventions Alone, or in Combination with Phosphodiesterase-5 Inhibitors, for the Treatment of Erectile Dysfunction: A Systematic Review	2021	Sandrine Atallah	Not about LLMs
19	Treatment of Pain in Cancer: Towards Personalised Medicine	2018	Marieke H J van den Beuken-van Everdingen	Not about LLMs
20	Automatic Depression Severity Assessment with Deep Learning Using Parameter-Efficient Tuning	2023	Clinton Lau	Not about LLMs
21	Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums using Language Models: Longitudinal Analysis and Evaluation	2022	David Owen	Duplicate
22	Authentic Engagement: A Conceptual Model for Welcoming Diverse and Challenging Consumer and Survivor Views in Mental Health Research, Policy, and Practice	2019	Indigo Daya BBus	Not about LLMs
23	The Impact of COVID-19 on Mental Health in Medical Students: A Cross-Sectional Survey Study in Italy	2022	Sara Carletto	Not about LLMs
2	How do you feel? Using natural language processing to automatically	2021	Michael J. Tanana	Not about LLMs

4	rate emotion in psychotherapy			
25	Depression Risk Prediction for Chinese Microblogs via Deep-Learning Methods: Content Analysis	2020	Xiaofeng Wang	Duplicate
26	Depression, Suicidal Ideation and Perceived Stress in Italian Humanities Students: A Cross-Sectional Study	2020	Fabrizio Bert	Not about LLMs
27	Transfer Learning for Risk Classification of Social Media Posts: Model Evaluation Study	2019	Derek Howard	Duplicate
28	AI Assisted Attention Mechanism for Hybrid Neural Model to Assess Online Attitudes About COVID-19	2022	Harnain Kour	Not about mental health
29	Behavioural, Emotional and Rhythm-Related Disturbances in Toddlers: Preliminary Findings from a Community-Based Study in Kerala, India	2021	Preeti Jacob	Not about LLMs
30	MMASleepNet: A Multimodal Attention Network Based on Electrophysiological Signals for Automatic Sleep Staging	2022	Yubo Zheng	Not about LLMs
31	Development of Internet Suicide Message Identification and the Monitoring-Tracking-Rescuing Model in Taiwan	2023	En-Liang Wu	Not about LLMs
32	LGCCT: A Light Gated and Crossed Complementation Transformer for Multimodal Speech Emotion Recognition	2022	Feng Liu	Not about mental health
33	Nursing Education in the Age of Artificial Intelligence Powered Chatbots (AI-Chatbots): Are We Ready Yet?	2023	Wilson Tam	Not about mental health
34	Neural Mediation of Greed Personality Trait on Economic Risk-Taking	2019	Weiwei Li	Not about LLMs
35	Data-driven Depression Detection System for Textual Data on Twitter Using Deep Learning	2022	Mushrifah Hasan	Duplicate
36	Stress Identification in Online Social Networks	2022	Ashok Kumar	Duplicate
37	Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study	2022	Aryan Rastogi	Duplicate
38	A Radical Approach to Depression Detection	2022	Xue Lei	Not about LLMs
39	Doing Well-Being: Self-Reported Activities Are Related to Subjective Well-Being	2022	August Håkan Nilsson	Not about LLMs
40	Increased Online Aggression During COVID-19 Lockdowns: Two-Stage Study of Deep Text Mining and Difference-in-Differences Analysis	2022	Jerome Tze-Hou Hsu	Duplicate
41	Surveilling COVID-19 Emotional Contagion on Twitter by Sentiment Analysis	2020	Cristina Crocamo	Duplicate
42	Efficacy Of A Group Psychoeducation Treatment in Binge Eating Disorder: An Open-Label Study	2022	Silvia Liquori	Not about LLMs
43	Network Sentiment Analysis of College Students in Different Epidemic Stages Based on Text Clustering	2022	Zhenghuai Song	Duplicate

44	Rvm-Gsm: Classification of Oct Images of Genitourinary Syndrome of Menopause Based on Integrated Model of Local-Global Information Pattern	2023	Kaiwen Song	Not about mental health
45	Monitoring The Impact of Covid-19 Pandemic on Mental Health: A Public Health Challenge? Reflection On Italian Data	2021	Maria Rosaria Gualano	Review article, too short
46	Analysis of sentiment changes in online messages of depression patients before and during the COVID-19 epidemic based on BERT+BiLSTM	2022	Chaohui Guo	Duplicate
47	Effectiveness of A Brief Dialectical Behavior Therapy Intensive-Outpatient Community Health Program	2022	Craig A Warlick	Not about mental health
48	The Auto Segmentation for Cardiac Structures Using a Dual-Input Deep Learning Network Based on Vision Saliency and Transformer	2022	Jing Wang	Not about mental health
49	Exploring the Possible Health Consequences of Job Insecurity: A Pilot Study Among Young Workers	202	Fabrizio Bert	Not about LLMs
50	Sentiment Analysis of Insomnia-Related Tweets via A Combination of Transformers Using Dempster-Shafer Theory: Pre- and Peri-COVID-19 Pandemic Retrospective Study	2022	Arash Maghsoudi	Duplicate
51	Opioid Death Projections with Ai-Based Forecasts Using Social Media Language	2023	Matthew Matero	Not about mental health
52	Predicting Generalized Anxiety Disorder from Impromptu Speech Transcripts Using Context-Aware Transformer-Based Neural Networks: Model Evaluation Study	2023	Bazen Gashaw Teferra	Duplicate
53	Multimodal Treatment Efficacy Differs in Dependence of Core Symptom Profiles in Adult Attention-Deficit/Hyperactivity Disorder: An Analysis of the Randomized Controlled Compas Trial	2022	Benjamin Selaskowski	Not about LLMs
54	Examining the Psychometric Properties of the Integrative Hope Scale's English Translation in A Mixed-Diagnostic Community Health Sample	2022	Craig A Warlick	Not about LLMs
55	Social Media for Psychological Support of Patients with Chronic Non-Infectious Diseases: A Systematic Review	2023	Fabrizio Bert	Not about LLMs
56	Systematic Review and Meta-Analysis of the Effects of Group Painting Therapy on the Negative Emotions of Depressed Adolescent Patients	2021	Zhaoxia Yuan	Not about LLMs

Table S2: Studies excluded after full text screening. LLMs=large language models

7. Multimedia Appendix 3

Supplementary material 3: PRISMA Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			

Section and Topic	Item #	Checklist item	Location where item is reported
Title	1	Identify the report as a systematic review.	Title Page- Pg 1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract- Pg 1-2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Introduction- Pg 2-5
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Introduction- Pg 5
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Methods- Pg 5
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Methods- Pg 6
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Methods- Pg 5
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Methods- Pg 5
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Methods- Pg 5-6
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Methods- Pg 5-6
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Methods- Pg 6
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Methods- Pg 5-6; Multimedia Appendix 1
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Methods- Pg 6
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Methods- Pg 5-6
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of	Methods- Pg 5-6

Section and Topic	Item #	Checklist item	Location where item is reported
		missing summary statistics, or data conversions.	
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Methods- Pg 5-6
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Methods- Pg 5-6
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Methods- Pg 5-6
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Methods- Pg 5-6
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Methods- Pg 5-6
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Results- Pg 6
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Results- Pg 6; Multimedia Appendix 2
Study characteristics	17	Cite each included study and present its characteristics.	Results- Pg 9-12 (Table1-3)
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Multimedia Appendix 1
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Results- Pg 9-12 (Table1-3)
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Results – Pg 9-12, 15-16
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Results - Pg 13-16
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Results - Pg 13-16
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized	

Section and Topic	Item #	Checklist item	Location where item is reported
		results.	
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Methods - Pg 15-16; Multimedia Appendix 1
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Results - Pg 9-16
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion- Pg 17-20
	23b	Discuss any limitations of the evidence included in the review.	Methods- Pg 15-16; Discussion- Pg 17
	23c	Discuss any limitations of the review processes used.	Discussion- Pg 18-19
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion- Pg 19-20
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Methods- Pg 5
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Methods- Pg 5
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Methods- Pg 5
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Acknowledgment- Pg 21
Competing interests	26	Declare any competing interests of review authors.	Conflicts of Interest- Pg 21
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data sharing statement- Pg 21

Table S3: PRISMA Checklist

8. Reference

1. World Health Organization. Mental health: strengthening our response [Internet]. 2022. Available from: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response.

2. World Health Organization. Mental disorders [Internet]. 2022. Available from: https://www.who.int/news-room/fact-sheets/detail/mental-disorders/?gclid=CjwKCAiApaarBhB7EiwA YiMwqi4yDXDAAvPftDkV_3GkuAV2lJxAyFdFWvHbEomzPBKgVpCpqupx_RoC 2_IQAvD_BwE.

3. 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. 2023 Apr; 90. Available from: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00106-8/fulltext

4. McManus S, Bebbington P, Jenkins R, Brugha T. Mental health and wellbeing in England: Adult psychiatric morbidity survey 2014. Leeds: NHS Digital; 2016.

5. World Health Organization. Mental health and COVID-19: Early evidence of the pandemic's impact: Scientific brief, 2 March 2022 [Internet]. 2022. Available from: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1.

6. Bertolote JM, Fleischmann A. Suicide and psychiatric diagnosis: a worldwide perspective. World Psychiatry. 2002 Oct;1(3):181-185.

7. Mental Health America. Mental health treatments [Internet]. 2023. Available from: https://mhanational.org/mental-health-treatments.

8. Rüsch N, Angermeyer MC, Corrigan PW. Mental illness stigma: concepts, consequences, and initiatives to reduce stigma. Eur Psychiatry. 2005 Dec;20(8):529-539. DOI: 10.1016/j.eurpsy.2005.04.004.

9. Corrigan PW, Watson AC. Understanding the impact of stigma on people with mental illness. World Psychiatry. 2002 Feb;1(1):16-20.

10. Torous J, Myrick KJ, 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. 2020;7(3):e18848.

11. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023;103:102274. ISSN 1041-6080. DOI: https://doi.org/10.1016/j.lindif.2023.102274.

12. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. arXiv. 2023. https://doi.org/10.48550/arXiv.2304.03347. [Preprint]

13. The Guardian. NHS mental health patients wait times [Internet]. 2022. Available from: https://www.theguardian.com/society/2022/oct/10/nhs-mental-health-patients-wait-times.

14. 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. 2017;4(2):e19. Available from: https://mental.jmir.org/2017/2/e19/.

15. Wysa—Everyday Mental Health. n.d. Wysa - Everyday mental health [Internet]. Available from: https://www.wysa.com/.

16. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative AI to interpret human emotions from visual and textual data: Pilot evaluation study. JMIR Ment Health. 2024;11:e54369. DOI: 10.2196/54369. PMID: 38319707.

17. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. eBioMedicine. 2023 Apr;90. Available from: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00077-4/fulltext.

18. OpenAI. Better language models [Internet]. 2023. Available from: https://openai.com/research/better-language-models.

19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. 2017. Available from: https://doi.org/10.48550/arXiv.1706.03762. [Preprint].

20. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):Article 8. DOI: 10.1038/s41591-023-02448-8.

21. Priest M. Large Language Models explained [Internet]. Boost.ai. 2023. Available from: https://boost.ai/blog/llms-large-language-models.

22. Kumar M. Understanding large language models and fine-tuning for business scenarios: a simple guide [Internet]. Medium. 2023. Available from: https://medium.com/@careerInAI/understanding-large-language-models-and-fine-tuning-for-business-scenarios-a-simple-guide-42f44cb687f0.

23. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners [Internet]. 2019. Available from: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

24. 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;5185-5198. https://doi.org/10.18653/v1/2020.acl-main.463.

25. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019;36(4):1234-1240. DOI: 10.1093/bioinformatics/btz682.

26. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv. 2020. Available from: https://arxiv.org/abs/1904.05342. [Preprint].

27. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, Zha Y, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell. 2020;182(5):1360. DOI: 10.1016/j.cell.2020.08.029.

28. 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. 2023 Jul;93. Available from: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00236-0/fulltext

29. Le Glaz A, Haralambous Y, Kim-Dufor D, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res.

2021;23(5):e15708. DOI: 10.2196/15708. PMID: 33944788. PMCID: PMC8132982.

30. Moher D. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Ann Intern Med. 2009;151(4):264. DOI: 10.7326/0003-4819-151-4-200908180-00135.

31. Baird A, Xia Y, Cheng Y. Consumer perceptions of telehealth for mental health or substance abuse: A Twitter-based topic modeling analysis. JAMIA Open. 2022;5(2). DOI: 10.1093/jamiaopen/ooac028.

32. Senn S, Tlachac ML, Flores R, Rundensteiner E. Ensembles of BERT for depression classification. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2022;4691-4694. DOI: https://doi.org/10.1109/EMBC48229.2022.9871120.

33. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. 2019. Available from: https://doi.org/10.48550/arXiv.1907.11692. [Preprint].

34. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. 2020. Available from: https://doi.org/10.48550/arXiv.1910.01108. [Preprint].

35. Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. Proceedings of the 9th International Conference on Language Resources and Evaluation; 2014.

36. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-LLM: leveraging large language models for mental health [Internet]. arXiv.org. 2023. Available from: https://arxiv.org/pdf/2307.14385.pdf. [Preprint].

37. Lamichhane B. Evaluation of ChatGPT for NLP-based mental health applications [Internet]. arXiv.org. 2023. Available from: https://arxiv.org/abs/2303.15727. [Preprint].

38. Chiang G, Stepanyan A. Insight stress analysis [Internet]. 2020. Available from: https://github.com/gillian850413/Insight-Stress-Analysis.

39. Inna P. Identifying depression [Internet]. 2018. Available from: https://github.com/Inusette/Identifying-depression.

40. Alambo A. Suicide risk assessment using reddit [Internet]. 2022. Available from: https://github.com/AmanuelF/Suicide-Risk-Assessment-using-Reddit.

41. Hassan T. Towards robust and interpretable practical applications of automatic mental state analysis using a dynamic and hybrid facial action estimation approach. [Year not provided];[Preprint]. DOI: 10.20378/irb-48641.

42. Ji S, Li X, Huang Z, Cambria E. Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput Appl. 2022;34:10309-10319.

43. Garg M, Saxena C, Saha S, Krishnan V, Joshi R, Mago V. CAMS: an annotated corpus for causal analysis of mental health issues in social media posts. Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022. pp. 6387–6396.

44. Vajre V, Naylor M, Kamath U, Shehu A. Psychbert: A mental health language model for social media mental health behavioral analysis. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2021. DOI: 10.1109/BIBM52615.2021.9669469.

45. El-Ramly M, Abu-Elyazid H, Mo'men Y, Alshaer G, et al. CairoDep: Detecting depression in Arabic posts using BERT transformers. 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS). 2021; Cairo, Egypt. pp. 207-212. DOI: 10.1109/ICICIS52592.2021.9694178.

46. Abdul-Mageed M, Elmadany A, Nagoudi EMB. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021. pp. 7088-7105. DOI: https://doi.org/10.18653/v1/2021.acl-long.551.

47. Bajaj GS, Yadav H, Sahdev HS, Sah S, Kaur P. Mental health analysis during COVID-19: A comparison before and during the pandemic. 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). 2021; Kuala Lumpur, Malaysia. pp. 1-7. DOI: 10.1109/GUCON50781.2021.9573763.

48. Vishwakarma A, Sawant S, Sawant P, Shankarmani R. An emotionally aware friend: moving towards artificial general intelligence. 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). 2021; Coimbatore, India. pp. 1094–1100. DOI: 10.1109/ICIRCA51532.2021.9544616.

49. Abdel Razek M, Frasson C. Text-based intelligent learning emotion system. J Intell Learn Syst Appl. 2017;9:17-20.

50. Saravia E, Liu HT, Huang YH, Wu J, Chen YS. CARER: Contextualized affect representations for emotion recognition. EMNLP. 2018.

51. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic multimodal set of facial and vocal expressions in North American English. PLoS ONE. 2018;13(5):e0196391.

52. Pichora-Fuller MK, Dupuis K. Toronto emotional speech set (TESS). Scholars Portal Dataverse. 2020. Available from: https://doi.org/10.5683/SP2/E8H2MF.

53. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. Ninth European Conference on Speech Communication and Technology. 2005.

54. William D, Achmad S, Suhartono D, Gema AP. Leveraging BERT with extractive summarization for depression detection on social media. 2022 International Seminar on Intelligent Technology and Its Applications (ISITIA). 2022; Surabaya, Indonesia. pp. 63-68. DOI: 10.1109/ISITIA56226.2022.9855370.

55. Kaseb A, Galal O, Elreedy D. Analysis on tweets towards COVID-19 pandemic: An application of text-based depression detection. 2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES). 2022. DOI: 10.1109/NILES56402.2022.9942363.

56. Wang S, Rupty LK, Mohona MH, Alagammai A, Omar M, Qabeel M. Depression detection using twitter data. 2019.

57. Shannak Y, Shurrab S. US COVID tweets. 2020.

58. Database HD. Vaccine tweets. 2021.

59. Zeberga K, Attique M, Shah B, Ali F, Jembre YZ, Chung TS. A novel text mining approach for mental health prediction using Bi-LSTM and BERT model. Comput Intell Neurosci. 2022;2022:7893775. DOI: https://doi.org/10.1155/2022/7893775.

60. Heinz MV, Bhattacharya S, Trudeau B, Quist R, Song SH, Lee CM, Jacobson NC. Testing domain knowledge and risk of bias of a large-scale general artificial intelligence model in mental health. Digit Health. 2023;9:20552076231170499. DOI: https://doi.org/10.1177/20552076231170499.

61. Wei J, Kim S, Jung H, Kim YH. Leveraging large language models to power chatbots for collecting user self-reported data [Internet]. arXiv.org. 2023. Available from: https://arxiv.org/abs/2301.05843. [Preprint].

62. Lin B, Bouneffouf D, Cecchi G, Varshney KR. Towards healthy AI: Large language models need therapists too [Internet]. arXiv.org. 2023. Available from: https://arxiv.org/abs/2304.00416. [Preprint].

63. Kumar H, Musabirov I, Shi J, Lauzon A, Choy KK, Gross O, Kulzhabayeva D, Williams JJ. Exploring the design of prompts for applying GPT-3 based chatbots: A mental wellbeing case study on mechanical turk [Internet]. arXiv.org. 2022. Available from: https://arxiv.org/abs/2209.11344. [Preprint].

64. Lai T, Shi Y, Du Z, Wu J, Fu K, Dou Y, Wang Z. Psy-LLM: Scaling up global mental health psychological services with AI-based large language models [Internet]. arXiv.org. 2023. Available from: https://arxiv.org/abs/2307.11991. [Preprint].

65. Sun H, Lin Z, Zheng C, Liu S, Huang M. PsyQA: A Chinese dataset for generating long counseling text for mental health support. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics. 2021. pp. 1489-1500.

66. Crasto R, Dias L, Miranda D, Kayande D. CareBot: A mental health chatbot. 2021 2nd International Conference for Emerging Technology (INCET). 2021. DOI: 10.1109/incet51464.2021.9456326.

67. Chen S, Wu M, Zhu KQ, Lan K, Zhang Z, Cui L. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. arXiv.org. 2023. Available from: https://arxiv.org/abs/2305.13614. [Preprint].

68. Cabrera J, Loyola MS, Magaña I, Rojas R. Ethical dilemmas, mental health, artificial intelligence, and LLM-based chatbots. Bioinformatics and Biomedical Engineering. IWBBIO 2023. Lecture Notes in Computer Science. Vol 13920. Cham: Springer; 2023. DOI: https://doi.org/10.1007/978-3-031-34960-7_22.

69. Fournier-Tombs E, McHardy J. A medical ethics framework for conversational artificial intelligence. J Med Internet Res. 2023;25. DOI: 10.2196/43068.

70. Jo E, Epstein DA, Jung H, Kim YH. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). New York, NY, USA: Association for Computing Machinery; 2023. Article 18, pages 1–16. DOI: https://doi.org/10.1145/3544548.3581503.

71. Webster P. Medical AI chatbots: Are they safe to talk to patients? Nature Med. 2023;[Preprint]. DOI: 10.1038/s41591-023-02535-w.

72. Salah M, Al Halbusi H, Abdelfattah F. May the force of text data analysis be with you: Unleashing the power of generative AI for social psychology research. Comput Hum Behav: Artif Humans. 2023;1(2):100006. DOI: 10.1016/j.chbah.2023.100006.

73. Farhat F. ChatGPT as a complementary mental health resource: a boon or a bane. Ann Biomed Eng. 2023 Jul 21. DOI: 10.1007/s10439-023-03326-7. Epub ahead of print. PMID: 37477707.

74. Woodnutt S, Allen C, Snowden J, Flynn M, Hall S, Libberton P, Purvis F. Could artificial intelligence write mental health nursing care plans? J Psychiatr Ment Health Nurs. 2023 Aug 4. DOI: 10.1111/jpm.12965. Epub ahead of print. PMID: 37538021.

75. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol. 2023;14. DOI: 10.3389/fpsyg.2023.1199058.

76. Bhattacharyya R, Chakraborty K, Neogi R. ChatGPT and its application in the field of mental health. J SAARC Psychiatr Fed. 2023;1(1):6. DOI: 10.4103/jspf.jspf_9_23.

77. Qiu H, He H, Zhang S, Li A, Lan Z. Smile: single-turn to multi-turn inclusive language expansion via CHATGPT for mental health support. arXiv.org. 2023. Available from: https://arxiv.org/abs/2305.00450. [Preprint].

78. Liu S, Zheng C, Demasi O, Sabour S, Li Y, Yu Z, Jiang Y, Huang M. Towards emotional support dialog systems. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021;3469-3483. https://doi.org/10.18653/v1/2021.acl-long.269.

79. Zheng C, Sabour S, Wen J, Zhang Z, Huang M. AugESC: dialogue augmentation with large language models for emotional support conversation. arXiv. 2023. Available from: https://doi.org/10.48550/arXiv.2202.13047. [Preprint].

80. Perlis RH. Research letter: application of GPT-4 to select next-step antidepressant treatment in major depression. [Year not provided]. DOI: 10.1101/2023.04.14.23288595. [Preprint].

81. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry. 2023;14:1213141. DOI: https://doi.org/10.3389/fpsyt.2023.1213141.

82. Egli A. ChatGPT, GPT-4, and other large language models: The next revolution for clinical microbiology? Clin Infect Dis. 2023;77(9):1322-1328. DOI: https://doi.org/10.1093/cid/ciad407.

83. 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. 2020;14:23. https://doi.org/10.1186/s13033-020-00356-9

84. The Black Box Problem: opaque inner workings of large language models. Prompt Engineering. 2023 Oct 23. Available from: https://promptengineering.org/the-black-box-problem-opaque-inner-workings-of-large-language-models/.

85. OpenAI Help Center. Image inputs for ChatGPT - FAQ. [No publication date available]. Available at: https://help.openai.com/en/articles/8400551-image-inputs-for-chatgpt-faq.

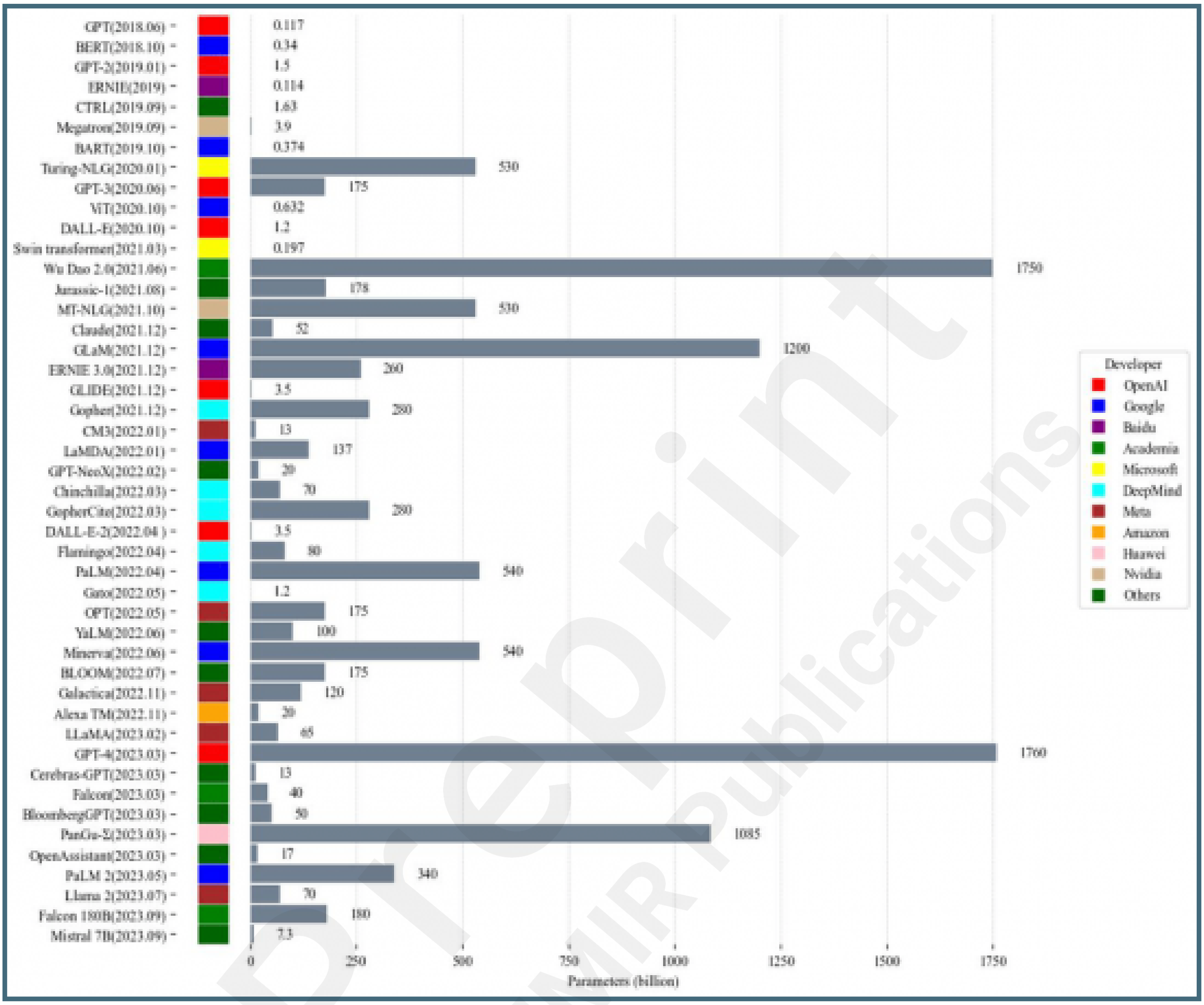
86. Hinkle M. LLMs and data privacy: navigating the new frontiers of AI. The New Stack. 2023. Available from: https://thenewstack.io/llms-and-data-privacy-navigating-the-new-frontiers-of-ai.

87. Emaminejad N, Akhavian R. Trustworthy AI and robotics: implications for the AEC industry. Automation in Construction. 2022;139:104298. Available from: https://doi.org/10.1016/j.autcon.2022.104298.

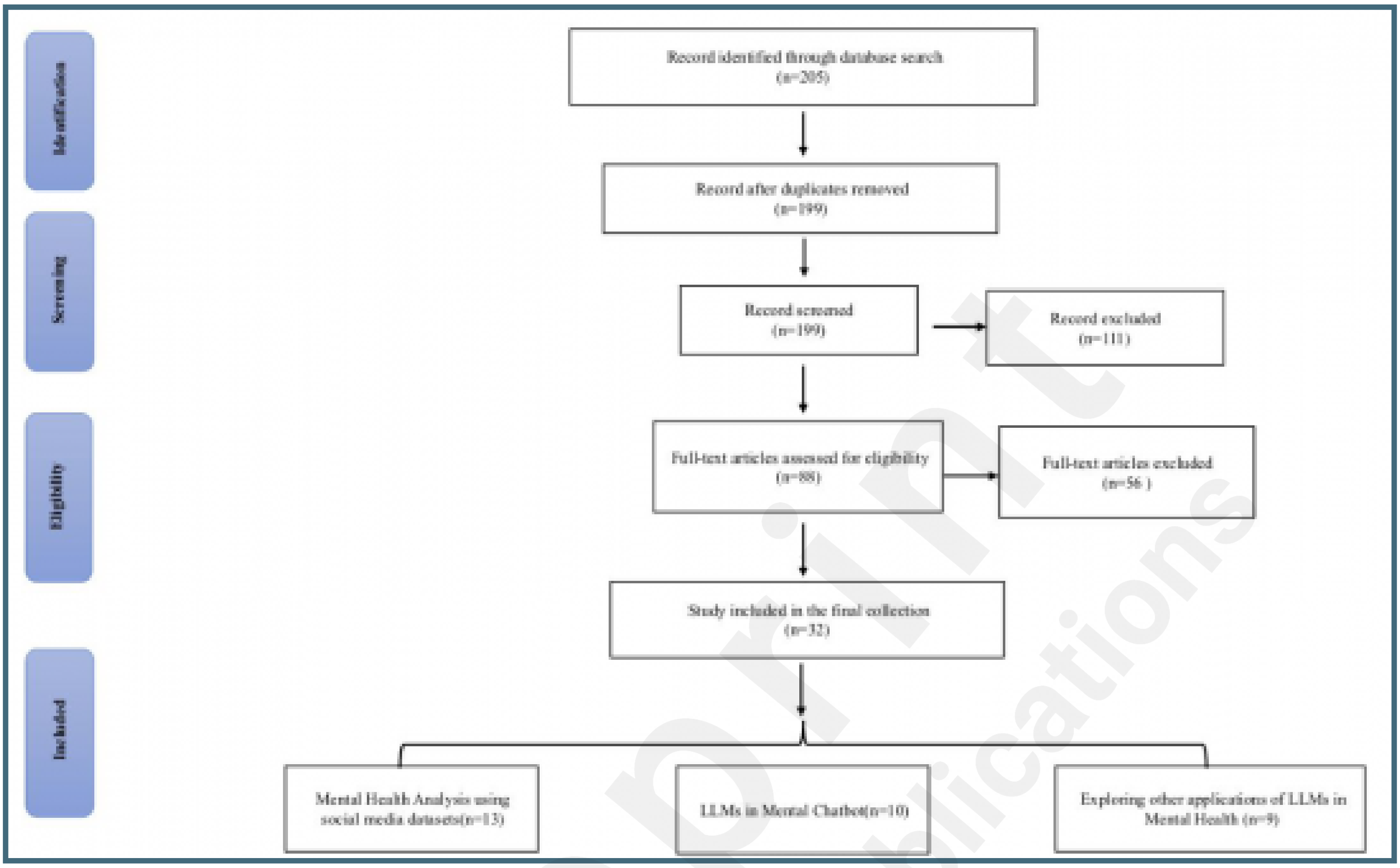
Supplementary Files

Figures

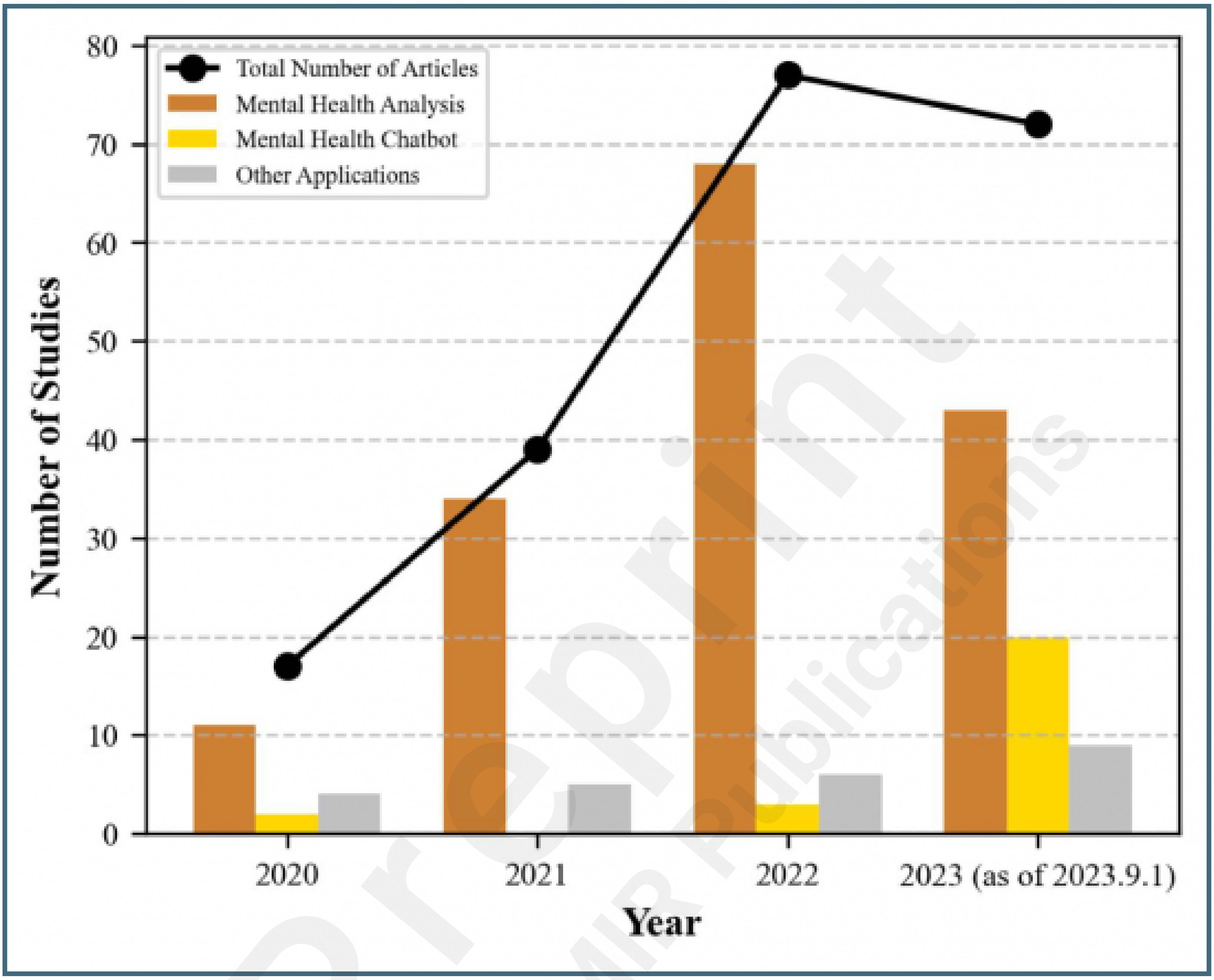
Comparative analysis of large language models by parameter size and developer entity.



PRISMA flow of selection process.



Number of articles after keyword search grouped by the year of publication and application field.



Multimedia Appendixes

Risk of bias assessment.
URL: http://asset.jmir.pub/assets/9f883b502d85e8e296e2013a852c0cf5.pdf

List of studies excluded at full-text screening stage.
URL: http://asset.jmir.pub/assets/b4ecdde2b52b5343e75a1b540de7c6b6.pdf

PRISMA Checklist.
URL: http://asset.jmir.pub/assets/68a692b050ce59bb258b2f5e879c89a7.pdf