Review

The Applications of Large Language Models in Mental Health: Scoping Review

Yu Jin[1*], PhD; Jiayi Liu[1*]; Pan Li[1*]; Baosen Wang[1*]; Yangxinyu Yan[2]; Huilin Zhang[3], BSc; Chenhao Ni[3], BEng; Jing Wang[4], PhD; Yi Li[5], MM; Yajun Bu[5], PhD; Yuanyuan Wang[2], PhD

[1]Department of Statistics, Faculty of Arts and Sciences, Beijing Normal University, Beijing, China

[2]School of Psychology, Center for Studies of Psychological Application, and Guangdong Key Laboratory of Mental Health and Cognitive Science, Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education, South China Normal University, Guangzhou, Guangdong, China

[3]School of Statistics, Beijing Normal University, Beijing, China

[4]Faculty of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, Guangdong, China

[5]The People's Hospital of Pingbian County, Honghe, Yunnan, China

[*]these authors contributed equally

Corresponding Author:
Yuanyuan Wang, PhD
School of Psychology, Center for Studies of Psychological Application, and Guangdong Key Laboratory of Mental Health and Cognitive Science
Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education
South China Normal University
Room 219, Floor 2
School of Psychology
Guangzhou, Guangdong, 510660
China
Phone: 86 13076729124
Fax: 86 13076729124
Email: angelayuanyuanwang@gmail.com

Abstract

Background: Mental health is emerging as an increasingly prevalent public issue globally. There is an urgent need in mental health for efficient detection methods, effective treatments, affordable privacy-focused health care solutions, and increased access to specialized psychiatrists. The emergence and rapid development of large language models (LLMs) have shown the potential to address these mental health demands. However, a comprehensive review summarizing the application areas, processes, and performance comparisons of LLMs in mental health has been lacking until now.

Objective: This review aimed to summarize the applications of LLMs in mental health, including trends, application areas, performance comparisons, challenges, and prospective future directions.

Methods: A scoping review was conducted to map the landscape of LLMs' applications in mental health, including trends, application areas, comparative performance, and future trajectories. We searched 7 electronic databases, including Web of Science, PubMed, Cochrane Library, IEEE Xplore, Weipu, CNKI, and Wanfang, from January 1, 2019, to August 31, 2024. Studies eligible for inclusion were peer-reviewed articles focused on LLMs' applications in mental health. Studies were excluded if they (1) were not peer-reviewed or did not focus on mental health or mental disorders or (2) did not use LLMs; studies that used only natural language processing or long short-term memory models were also excluded. Relevant information on application details and performance metrics was extracted during the data charting of eligible articles.

Results: A total of 95 articles were drawn from 4859 studies using LLMs for mental health tasks. The applications were categorized into 3 key areas: screening or detection of mental disorders (67/95, 71%), supporting clinical treatments and interventions (31/95, 33%), and assisting in mental health counseling and education (11/95, 12%). Most studies used LLMs for depression detection and classification (33/95, 35%), clinical treatment support and intervention (14/95, 15%), and suicide risk prediction (12/95, 13%). Compared with nontransformer models and humans, LLMs demonstrate higher capabilities in information acquisition and analysis and efficiently generating natural language responses. Various series of LLMs also have different advantages and disadvantages in addressing mental health tasks.

Conclusions: This scoping review synthesizes the applications, processes, performance, and challenges of LLMs in the mental health field. These findings highlight the substantial potential of LLMs to augment mental health research, diagnostics, and intervention strategies, underscoring the imperative for ongoing development and ethical deliberation in clinical settings.

(J Med Internet Res 2025;27:e69284) doi: 10.2196/69284

KEYWORDS

mental health; large language models; application; process; performance; comparison

Introduction

Background

Mental health is a growing global public issue, impacting nearly a billion people worldwide, with an estimated 1 in 7 adolescents affected [1,2]. Nevertheless, >70% of individuals with mental health disorders lack access to essential support and services [3]. Furthermore, >720,000 people commit suicide annually, with nearly three-quarters of these suicides occurring in low- and middle-income countries [4]. Consequently, there is an urgent need in mental health to facilitate efficient detection from large-scale data; deliver effective treatments and interventions to large populations; and ensure private, affordable health care and increased access to specialized psychiatrists. To address inadequate access to effective and equitable mental health care, large-scale, innovative solutions are imperative.

Large language models (LLMs), emerging in 2019, are advanced natural language processing (NLP) models capable of analyzing vast textual data and generating humanlike language [5]. Notable LLMs such as GPT-3/4 [6], Pathways Language Model (PaLM) [7], and LLaMA [8], constitute a category of foundational models, each with billions of parameters, trained on extensive textual data [9]. Using the transformer architecture and self-supervised pretraining, LLMs are adept at tackling a variety of NLP tasks, including information extraction, interaction, content generation, and logical reasoning [10]. In comparison to prior NLP models [11,12], LLMs exhibit superior performance in computational efficiency, large-scale data analyses, interaction, and external validity and applicability [9]. Furthermore, LLMs can be fine-tuned to cater to specific domains, including mental health, thereby empowering them to engage in natural language interactions and accomplish mental health–related tasks. LLMs would help address insufficient mental health care system capacity and provide efficient or personalized treatments. Therefore, the application of LLMs in mental health is expanding across diverse domains [13-16].

Objective

Researchers have explored the applications of LLMs in mental health in various areas, encompassing screening or detecting mental disorders [17-19], supporting clinical treatments and interventions [20-22], and assisting in mental health counseling and education [17,20,23,24]. Nonetheless, few comprehensive reviews have yet synthesized these applications, assessed the performance of LLMs, or elucidated their advantages within the mental health domain [25,26]. Therefore, we conducted this scoping review to address 4 questions. First, we identified the challenges in mental health and compared the processes adopted by humans, nontransformer models, and LLMs. Second, we summarized the main areas of LLMs' applications and presented specific processes of these applications. Third, we examined comparative performance studies between LLMs and humans, as well as among different LLMs. Finally, we presented the fine-tuning LLMs for mental health and compared their advantages and disadvantages, which could be directly used by researchers and psychiatrists. This review aimed to provide a foundational understanding of LLM applications in mental health by examining current trends, comparing performance, identifying challenges, and outlining a road map for future research and clinical practice.

Methods

Protocol Registration

We drafted the study protocol based on the relevant items from the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; Multimedia Appendix 1). Compared with the original protocol, the previous query strings were capable of searching for articles related to the applications of LLMs in mental health. However, to enhance the reproducibility of the search process, we have added more detailed query strings. As for databases used, we excluded the Google Scholar database from our search strategy for two reasons: (1) the other 4 English language databases we selected already included the relevant articles and (2) it is not feasible to accurately count the specific number of articles retrieved from the Google Scholar database. Instead, we used official academic databases for the article search. These databases have strict selection criteria and typically include only peer-reviewed, high-quality publications from reputable publishers. In addition, we have updated the literature search period, revising the range from January 1, 2017, to August 31, 2024. This change is important for the following reasons. First, at the beginning of the study, we searched for articles published between January 1, 2017, and August 31, 2024. However, we found that all articles meeting the inclusion criteria were published after 2019. Second, previous studies [9,25] suggested that the term "LLM" was first proposed and widely used starting in October 2019. Therefore, we have revised our inclusion criteria to include only studies published from January 1, 2019, to August 31, 2024. The final protocol was registered prospectively in the Open Science Framework [27].

Search Strategy and Selection Criteria

A scoping review is a preliminary systematic review that aims to map the existing evidence on a specific topic or field of research [28]. It provides a broad overview of the literature by identifying the nature and extent of existing research, including

types of studies, variables, and gaps in the evidence base [29]. This approach is particularly useful when the body of evidence is large or diverse, or when there is a need to understand the scope of a research area before conducting a more focused systematic review.

This scoping review followed the five-stage framework: (1) identifying the research question; (2) identifying relevant studies; (3) study selection; (4) charting the data; and (5) collating, summarizing, and reporting the results. The search terms for mental health include the following: "psychiatr*," "mental health," "depress*," "anxiety," "posttraumatic stress disorder," "PTSD," "bipolar disorder," "schizophrenia," "obsessive-compulsive disorder," "personality disorder," "insomnia," and "suicid*." The keywords and search terms for LLMs in mental health include the following: "large language model," "OpenAI language model," "generative AI," "generative artificial intelligence," "BERT," and "GPT."

According to previous studies [9,25], the term "LLM" is used to distinguish language models based on their parameter scale (eg, containing tens or hundreds of billions of parameters). The term "LLM" has been proposed and widely used since October 2019. In addition, we planned to conduct this scoping review on August 31, 2024. Thus, we searched 4 English language databases (Web of Science, PubMed, IEEE Xplore, and Cochrane Library) and 3 Chinese language databases (Weipu, CNKI, and Wanfang) for peer-reviewed articles published between January 1, 2019, and August 31, 2024. We only included papers in Chinese published in high-quality journals. To find other possibly relevant studies and reports that were missed by the automated searches, the reference lists of the included articles and reports were examined.

The inclusion and exclusion criteria for studies are shown in Textbox 1.

Textbox 1. Study inclusion and exclusion criteria.

Inclusion criteria

- Studies focused on the applications of large language models (LLMs) in the mental health field.
- The LLMs included, but were not limited to, GPT-3/4, ChatGPT, Pathways Language Model (PaLM), LLaMA, and the improved and fine-tuned LLMs.
- Published in a peer-reviewed journal or conference.

Exclusion criteria

- Articles that were not peer-reviewed or did not focus on mental health or mental disorders.
- Studies that did not use large language models (those using natural language processing or long short-term memory were also excluded).

The first round of screening was based on titles, keywords, and abstracts (4392/4859, 90.39%). These articles were divided into 6 subsets (each subset includes 732 papers). A total of 6 researchers (JL, PL, BW, YY, HZ, and CN) independently reviewed each part of the papers, including the titles, keywords, and abstracts. During the initial review, each article was categorized into one of the following groups: (1) fully met the inclusion criteria, (2) did not focus on mental health or mental disorders, (3) did not use LLMs, and (4) had unclear eligibility for inclusion. In the second round of screening, we conducted a cross-check process to ensure the classification of these articles (eg, CN reviewed the paper part handled by HZ, and HZ reviewed the paper part handled by JL). Articles from the fourth group were discussed by YJ and YW one by one for their eligibility. Several questions were discussed:

- NLP use uncertainty—some articles used NLP, but it was unclear whether LLMs were used. A full-text review was necessary to make a definitive determination. Thus, we reserved these articles for further review.
- Exclusion of cognitive disorders—it remains uncertain whether aphasia should be classified as part of the mental health domain. After the discussion, we excluded the articles about aphasia as they were not considered within the scope of mental health.
- Electronic medical records (EMRs) in the context—some articles mentioned EMRs, which could potentially be relevant to the mental health domain. However, these articles were not specifically focused on mental health. After the discussion, we decided to reserve these articles for further review.

The third round of review was based on the full texts to assess their eligibility (308/4392, 7.01%). These potentially eligible articles were divided into 6 parts (each part included 51 or 52 papers). A total of 6 researchers (JL, PL, BW, YY, HZ, and CN) independently reviewed each part of the papers, including the titles, abstracts, and full texts of each paper. To ensure the accuracy of paper screening, we conducted a double-check process. The researchers cross-checked each other's work, reviewing the paper selection process (eg, JL reviewed the paper part handled by PL, and PL reviewed the paper part handled by BW). Disagreements were discussed with third reviewers (YJ and YW) until a consensus was reached. We excluded preprints, reviews, books, studies that did not use LLMs, and those not published in journals or conferences. Ultimately, 95 articles were retained for the final analysis. The search terms for English language databases and Chinese language databases are shown in Multimedia Appendices 2 and 3.

Data Extraction, Categorization, and Labeling

The final data collection form used for peer-reviewed articles is shown in Table 1 (N=95). The information of each study included categories, regions, application tasks, mental conditions, data sources, sample information, and applied models. These articles were divided into 6 parts (each part

included 15 or 16 papers). A total of 6 researchers (JL, PL, BW, YY, HZ, and CN) independently extracted data from each part of the papers. To ensure the accuracy of data extraction, we conducted a double-check process. The researchers cross-checked each other's work, reviewing the data extraction process. The excluded studies, along with the reasons for their exclusion (eg, "It's a preprint"), are listed in Multimedia Appendix 4.

For this scoping review, we developed a categorization framework based on the applications of LLMs in mental health: (1) screening or detection of mental disorders, (2) supporting clinical treatments and interventions, and (3) assisting in mental health counseling and education. At least 1 reviewer categorized each study manually by examining the title and abstract to assign categories. When the study categories could not be clearly determined, the methods and results sections were reviewed to assist with classification.

Table 1. The basic information of the included studies (N=95).

Categories and study	Basic information			Data information		Models
	Region	Application	Mental conditions	Data sources	Sample information	
Depression detection and classification						
[20]	Israel	Prognosis	Depression	Case vignettes and previous studies	1074 experts	GPT-3.5, GPT-4, and Claude and Bard
[24]	United States	Prediction	Stress and depression	Dreaddit, CSSRS-Suicide	3553 posts	Mental-Alpaca and Mental-FLAN-T5
[30]	Singapore	Analysis	Depression, PTSD[a], anxiety	Reddit, SMS text messaging, Twitter, and MHI dataset	105,000 data samples	MentaLLaMA
[31]	United States	Detection	Depression, anxiety, SI[b]	Reddit and Twitter	Conversations	PsychBERT
[32]	United States	Detection	Depression	Twitter	2575 tweets from Twitter users with depression	BERT[c], RoBERTa[d], and XLNet
[33]	India	Detection	Depression	Twitter	189 interviews	BERT with multimodal frameworks
[34]	China	Detection	Depression	DAIC-WOZ[e]	Respondents with depression labels	BERT
[35]	United Arab Emirates	Detection	Depression	E-DAIC[f]	7650 unique entries	BERT-based custom classifier
[36]	Canada	Prediction	Depression, ADHD[g], anxiety	Reddit	2514 users' posts and 167,444 clinical posts, 2,987,780	BERT, ROBERTa, open AI GPT, and GPT 2
[37]	China	Detection	Depression and SI	Dialogues from real-life scenarios	Depression (64 samples) and anxiety (75 samples)	GPT-3.5
[38]	United States	Detection	Depression	DAIC[h], E-DAIC, and EATD	Nondepression and depression, DAIC, E-DAIC, and EATD	BERT and RoBERTa
[39]	China	Detection	Depression	DAIC-WOZ	189 participants	BERT
[40]	United States	Detection	Depression	Twitter (sentiment dataset)	632,000 tweets	RoBERTa, DeBERTa, DistilBERT, and SqueezeBERT
[41]	Malaysia	Detection	Depression	Interviews, Facebook, Reddit, and Twitter	53 participants (11 of them were with depression)	GPT-3 (ADA model)
[42]	United States	Detection	Depression	DAIC-WOZ, extended-DAIC, and simulated data	DAIC-WOZ and E-DAIC (data from 122 participants with depression)	BERT, GPT-3.5, and GPT-4
[43]	China	Prediction	Depression	Weibo	13,993 microblogs with depression labels	BERT, RoBERTa, and XLNET
[44]	United Kingdom	Detection	Depression	RSDD and RSDD-Time	Posts (9210 users with depression)	ALBERT, BioBERT, Longformer, MentalBERT, and MentalRoBERTa
[45]	Switzerland	Classification	Depression	DAIC-WOZ	Respondents with depression labels	BERT, RoBERTa, and DistilBERT
[46]	United States	Responses	PPD[i]	ACOG[j] and PPD	14 questions	GPT-4 and LaMDA
[47]	Germany	Detection	Depression	E-DAIC	275 participants	DepRoBERTa

Categories and study	Basic information			Data information		Models
	Region	Application	Mental conditions	Data sources	Sample information	
[48]	Canada	Detection	Depression	DTR dataset	42,691 tweets from Depression-Candidate-Tweets, 6077 tweets from Depression Tweets Repository and 1500 tweets from DSD-Clinician-Tweets	Mental-BERT
[49]	Malaysia	Detection	Depression	Scraped and survey PHQ-9[k]	250 users	BERT
[50]	Netherlands	Detection	Depression	Clinical data (16,159 patients)	Survey PHQ-9 and scraped	DistilBERT
[51]	Greece	Detection	Stress and depression	Dreaddit dataset	16,159 patients	M-BERT and M-MentalBERT
[52]	Israel	Evaluations	Depression	Clinical vignettes	14 questions	GPT-3.5 and GPT-4
[53]	United States	Screening	Depression	DAIC-WOZ	Diagnosed with depression or 8 vignettes	AudiBERT (I, II, and III)
[54]	Canada	Assessment	Depression	DAIC-WOZ	15 thematic datasets	Prefix-tuned RoBERTa
[55]	India	Detection	Depression	Reddit (mental health corpus and depression)	189 participants	RoBERTa
[56]	Iran	Prediction	Depression	Autodep dataset (Twitter)	219 samples and 20 real participants	DBUFS2E, BBU, MBBU, and DRB
[57]	United States	Classification	Depression	GLOBEM dataset	Collection of passive sensing or tweets and bioscriptions	GPT-3.5, GPT-4 and pathways language model 2
[58]	Russia	Detection	Depression	DAIC-WOZ	Respondents with depression labels	BERT, MentalBERT, MentalRoBERTa, PsycBERT, and Clinical-BERT
[59]	China	Diagnosis	Depression	Labeled text data	NR[l]	DepGPT
[60]	South Korea	Detection	Depression	Mind station app data	428 diaries	GPT-3.5 and GPT-4
Suicide						
[17]	United States	Prediction	SI	Brightside Telehealth platform	460 (SI at intake, SI later, and without SI)	GPT-4
[23]	Austria	Detection	Suicide	Twitter	3202 English tweets	BERT and XLNet
[24]	United States	Prediction	Stress and depression	Dreaddit, CSSRS-Suicide	3553 posts	Mental-Alpaca and Mental-FLAN-T5
[30]	Singapore	Analysis	Depression, PTSD[a], anxiety	Reddit, SMS, Twitter, and MHI dataset	105,000 data samples	MentaLLaMA
[31]	United States	Detection	Depression and anxiety	Reddit and Twitter	148,700 conversations	PsychBERT
[61]	Morocco	Detection	Suicide	Reddit	Suicide and nonsuicide content or 232,074 posts	GPT and BERT
[62]	Canada	Classification	Suicide	Reachout.com forum posts and UMD Reddit dataset	Posts in/r/SuicideWatch on Reddit or 1588 labeled posts	GPT-1
[63]	Israel	Assessment	Suicide	Professional mental health assessments	4 vignettes	GPT-3.5 and GPT-4
[64]	Canada	Detection	SI	UMD dataset and LLM[m] synthetic datasets	>100,000 posts and comments	BERT

Categories and study	Basic information			Data information		Models
	Region	Application	Mental conditions	Data sources	Sample information	
[65]	Brazil	Detection	SI[b]	Twitter	Suicide-related text or 5699 tweets	Boamente
[66]	China	Classification	Suicide	Microblog user data (ZouFan comments)	4500 pieces	Knowledge-perception BERT model
[67]	United States	Detection	Suicide	Reddit (SuicideWatch section)	2.9 million posts and 30 sub-Reddits (including mental health and control sub-Reddits)	BERT
Other mental disorders						
[24]	United States	Prediction	Stress and depression	Dreaddit, CSSRS-Suicide	3553 posts	Mental-Alpaca and Mental-FLAN-T5
[30]	Singapore	Analysis	Depression, PTSD[a], anxiety	Reddit, SMS, Twitter, and MHI dataset	105,000 data samples	MentaLLaMA
[31]	United States	Detection	Depression and anxiety	Reddit and Twitter	148,700 conversations	PsychBERT
[68]	India	Detection	Stress and anxiety	Reddit	3553 labeled posts	RoBERTa and XLNet
[69]	India	Detection	Stress, depression, and suicide	Reddit and Twitter	Stress, depression, and suicide	GPT-2 and GPT-Neo-125M
[70]	Canada	Screening	GAD[n]	Prolific platform data	2000 participants	Longformer
[71]	United States	Diagnosis	OCD[o]	Clinical vignettes	OCD vignettes	GPT-4, Gemini pro, and LLaMA 3
[72]	Iran	Diagnosis	Mental health disease	Clinical cases	Selected by a medical expert or 20 cases	GPT-3.5, GPT-4, Nemotron, and Aya
[73]	China	Prediction	Mental disorder	Kaggle	16,950 categories or 41,851 reviews	MentalBERT
[51]	Greece	Detection	Stress and depression	Dreaddit dataset	16,159 patients	M-BERT and M-MentalBERT
[74]	Israel	Diagnosis	BPD[p] and SPD[q]	Emotional scenarios (20 cases)	20 scenarios	GPT-3.5
[75]	United States	Rating	Emotion	Psychotherapy transcripts and interrater dataset	97,497 ratings	BERT
[76]	China	Extraction	Psychiatric disorder	Clinical notes (12,006 records)	Human or 12,006 anonymous clinical notes	BERT
[77]	China	Screening	Schizophrenia, BPD, major depressive disorder, and DD	EHRs[r]	500 EHRs	BERT, RoBERTa, DistilBERT, and ALBERT
[78]	Iran	Diagnosis	Depression, OCD, GAD, BPD, and schizophrenia	*Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*–based case scenarios	13 case scenarios	GPT-3.5, GPT-4, AYA, and Nemotron-3-8B
[79]	United States	Text analysis	Sentiment	Tweets and news	Annotated by human or 47,925 tweets and news headlines	GPT3.5 turbo, GPT4, and GPT4 turbo
[80]	United States	Sentiment	Sentiment	Multiple sources	Tokens	OPT, GPT-3.5, and BERT
[81]	United States	Classification	Emotion and sentiment	Social media	417,423 and 1,176,509 samples	EmoBERTTiny
[82]	United States	Emotion	Depression	Stress and coping process questionnaire	100 nonstudent adults	Text-davinci-003, GPT-3.5, and GPT-4

Categories and study	Basic information			Data information		Models
	Region	Application	Mental conditions	Data sources	Sample information	
[83]	India	Identification	Emotion	GoEmotions dataset and Twitter dataset	27 different emotion categories or comments and tweets	MobileBERT
[84]	Israel	Emotion	Emotion	RMET and LEAS	36 photos and 20 questions	GPT-4 and Bard
[85]	United States	Classification	Psychotherapy	Smart home images	7 different environments or 10,767 images	GPT-4
Supporting clinical treatments						
[15]	Canada	Diagnosis	PTSD	E-DAIC dataset and ChatGPT-generated transcripts	Severe depression and PTSD or 219 participants	GPT-3.5-turbo
[16]	United States	Detection	CB-PTSD[s]	Participant narratives	1295 narratives	GPT-3.5-turbo
[18]	South Korea	Therapy	ADHD and dementia	USPTO patent data	8656 patients and 205 DTx patents	BERTopic and PatentSBERTa
[20]	Israel	Assessment	Depression	Case vignettes and previous studies	1074 experts	GPT-3.5, GPT-4, Claude, and Bard
[21]	United States	Disorder	Bipolar depression	EHR-based generated data	50 sets of clinical vignettes	GPT-4
[86]	United States	Screening	General mental health issues	EHRs and clinical notes	2,476,628 patients or 290,482,002 clinical notes	GatorTron
[87]	China	Counseling	Stress, LGBTQ issues	Consultation web sites, Weibo, and Zhihu	31 unique questions	ChatGLM, ERNIE Bota, and Qianwen
[88]	United States	Counseling	NR	MentalCLOUDS dataset	11,543 utterances	BART, T5, GPT series, Phi-2, MentalBART, Flan-T5, Mistral, LLaMA-2, and MentalLLaMA
[89]	United Kingdom	Therapy	Anxiety	Therapist-written thoughts	20 tasks at each of 3 stages	GPT-4 and Bard
[90]	United Kingdom	Questionnaires validation	Depression, anxiety, and PTSD	C19PRC study data	2058 adults	Sentence-BERT
[91]	Australia	Diagnosis	Various psychiatric conditions	Clinical case vignettes	100 cases	GPT-3.5
[92]	China	Diagnosis	Depression	MedDialog, Metal Real	NR	LLaMA, ChatGLM, and Alpaca
[93]	United States	Analysis	Psychoactive experiences	Erowid and PDSP-Ki dataset	11,816 testimonials	BERTowid, BERTiment, and CCA
[94]	India	Prediction	Stress and anxiety	Survey dataset	41,000 entries	Gemini
Chatbots						
[20]	Israel	Assessment	Depression	Case vignettes and previous studies	2 vignettes differed in gender	GPT-3.5, GPT-4, Claude, and Bard
[22]	China	Chatbots	Autistic	DXY platform (medical consultation data)	100 consultation samples	GPT-4 and ERNIE bot
[35]	United Arab Emirates	Detection	Depression	E-DAIC dataset	219 samples and 20 real participants or 219 E-DAIC samples	BERT-based custom classifier
[95]	United States	Evaluation	Depression and SI	Human made	25 conversational agents	GPT-3.5

Categories and study	Basic information			Data information		Models
	Region	Application	Mental conditions	Data sources	Sample information	
[96]	Spain	Emotion	General emotional states	Internet-based human conversations	64 participants	GPT-3
[97]	Germany	Therapy	ADHD	NR	NR	GPT3.5, GPT-4 turbo, and Claude-3 opus
[98]	Poland	Sentiment	Mental health	Corpus of Translated Emotional Texts and Polish common crawl	Sentences and web pages	GPT-3.5
[99]	United States	Detection	Loneliness and SI	Survey data	1006 users of Replika	Replika
Data augmentation						
[15]	Canada	Diagnosis	PTSD	E-DAIC dataset and ChatGPT-generated transcripts	Severe depression and PTSD or augmented data	GPT-3.5
[16]	United States	Detection	CB-PTSD	Participant narratives	1295 narratives	GPT-3.5-turbo
[47]	Germany	Detection	Depression	E-DAIC dataset	275 participants	DepRoBERTa
[64]	Canada	Detection	SI	UMD dataset and LLM synthetic datasets	>100,000 posts and comments	BERT
[92]	China	Diagnosis	Depression	EATD-Corpus, Med-Dialog dataset (Chinese)	NR	LLaMA-7B, ChatGLM-6B, and Alpaca
[100]	Canada	Augmentation	PTSD	E-DAIC dataset, generated data	219 interview records	CALLM[t], GPT-4, DistilBERT, and BERT
[101]	China	Augmentation	Mental health	ChatGPT-generated narratives	3017 instances; 80/20 train-test split	BERT, BLOOM-560M, BLOOMZ-3B, ChatGLM2-6B, and Baichuan-7B
[102]	Turkey	Generation	Various disorders	DAIC-WOZ, ChatGPT-dataset, and Bard-dataset	Real patients and synthetic patients	GPT-3.5 and Bard
[103]	China	Generation	Psychiatry	DSM-5 diagnostic criteria	2000 records	Mistral7B-DSM-5 model
Assisting in counseling						
[104]	India	Chatbot	Depression and anxiety	Reddit	Questions related to the illness or NR	CareBot
[105]	United States	Chatbot	General mental health	Reddit	120 posts (2917 user comments)	Replika
[106]	Poland	Chatbot	General mental health	Empathetic dialogues and DailyDialog datasets	DailyDialog dataset or NR	BERT
[14]	Philippines	Chatbot	Not suitable	Well-being conversations, PERMA Lexica	24,850 conversations	VHope
[107]	Canada	Chatbot	General mental well-being	Prompts made by author	With mindfulness experience or NR	GPT-3 based chatbots
[108]	United States	Health care	General mental well-being	NR	NR	GPT-4o
[99]	United States	Detection	Loneliness and suicide	Survey data	1006 users of Replika	Replika
[109]	United States	Generation	Depression	Psychiatric questions	4 questions	GPT-3.5 and GPT4

Categories and study	Basic information			Data information		Models
	Region	Application	Mental conditions	Data sources	Sample information	
[110]	United Kingdom	Measurement	General mental health	Qwell platform therapy transcripts	254 conversations	RoBERTa and CTM[u]
Assisting in mental health education						
[19]	United States	Education	ADHD and ED[v]	Interview data	With signs of a disorder or 102 students	GPT-3
[111]	Australia	Education	Substance use	Mental health portals	"Cracks in the Ice" website	GPT-4

[a]PTSD: posttraumatic stress disorder.

[b]SI: suicidal ideation.

[c]BERT: bidirectional encoder representations from transformers.

[d]RoBERTa: a robustly optimized bidirectional encoder representations from transformers pretraining approach.

[e]DAIC-WOZ: Distress Analysis Interview Corpus-Wizard of Oz.

[f]E-DAIC: Extended Distress Analysis Interview Corpus.

[g]ADHD: attention-deficit hyperactivity disorder.

[h]DAIC: Distress Analysis Interview Corpus.

[i]PPD: postpartum depression.

[j]ACOG: American College of Obstetricians and Gynecologists.

[k]PHQ: Patient Health Questionnaire

[l]NR: not reported.

[m]LLM: large language model.

[n]GAD: generalized anxiety disorder.

[o]OCD: obsessive-compulsive disorder.

[p]BPD: borderline personality disorder.

[q]SPD: schizoid personality disorder.

[r]EHR: electronic health record.

[s]CB-PTSD: childbirth related posttraumatic stress disorder.

[t]CALLM: Clinical Interview Data Augmentation via Large Language Models.

[u]CTM: contextualized topic model.

[v]ED: erectile dysfunction.

Statistical Analysis

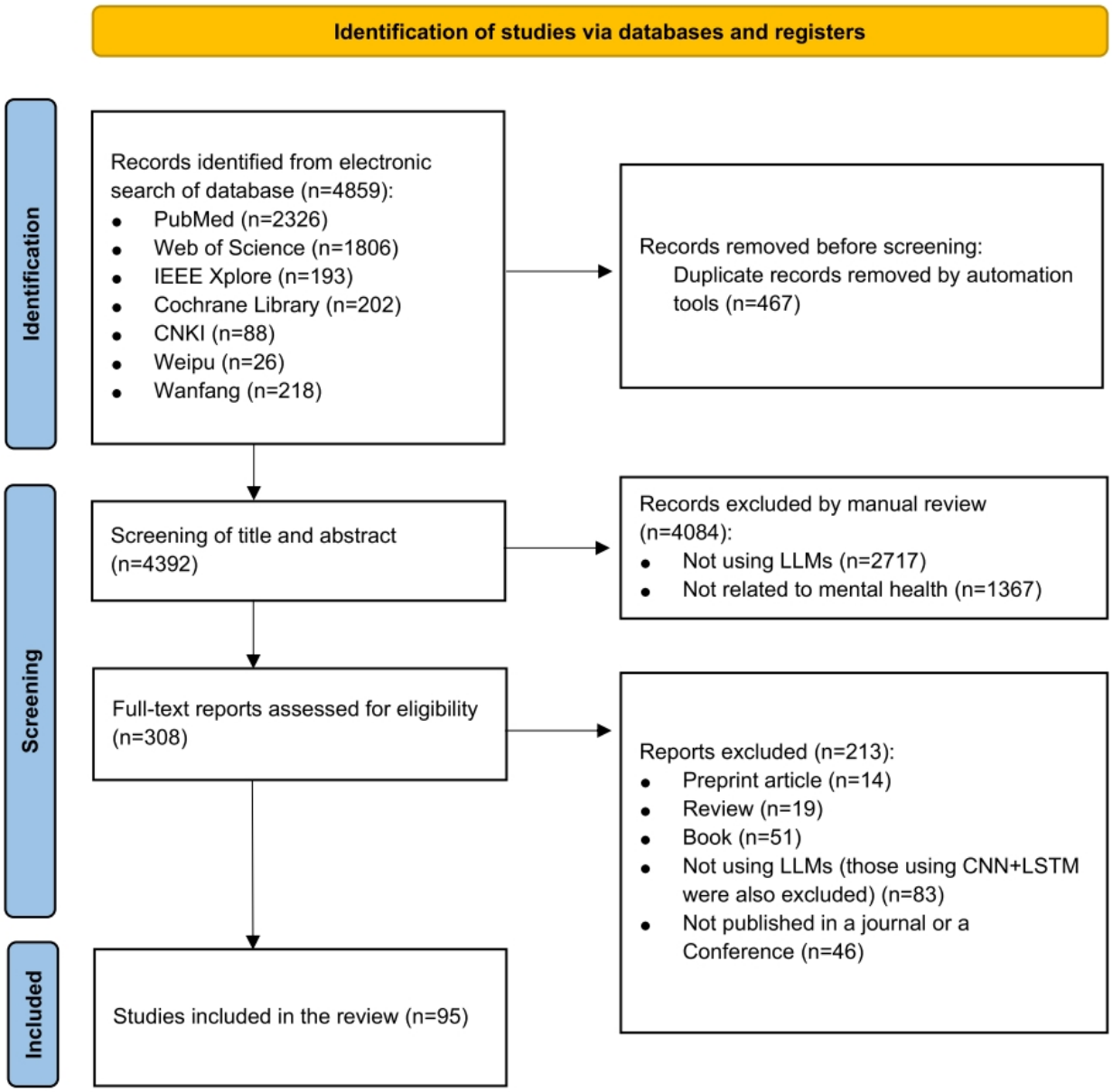
Descriptive statistics were used to summarize the distribution of studies across different application areas. For each application task, we calculated frequencies and percentages. For the performance comparisons between humans and LLMs, between LLMs and nontransformer models, also between various LLMs, we collected the metrics results and plotted the results. Calculations and data charting were performed using R software (version 4.4.2) developed at Bell Laboratories by John Chambers and colleagues, and PyCharm software developed by JetBrains.

Results

Overview

The initial search yielded 4859 records, of which 467 duplicates were removed. Of the remaining 4392 records, 4084 records were removed due to the contents irrelevant to mental health or not using LLMs. After the full-text screening, 95 articles fulfilled the inclusion criteria (Figure 1). Table 1 demonstrates the basic information of each study, including categories, regions, application tasks, mental conditions, data sources, sample information, and applied models.

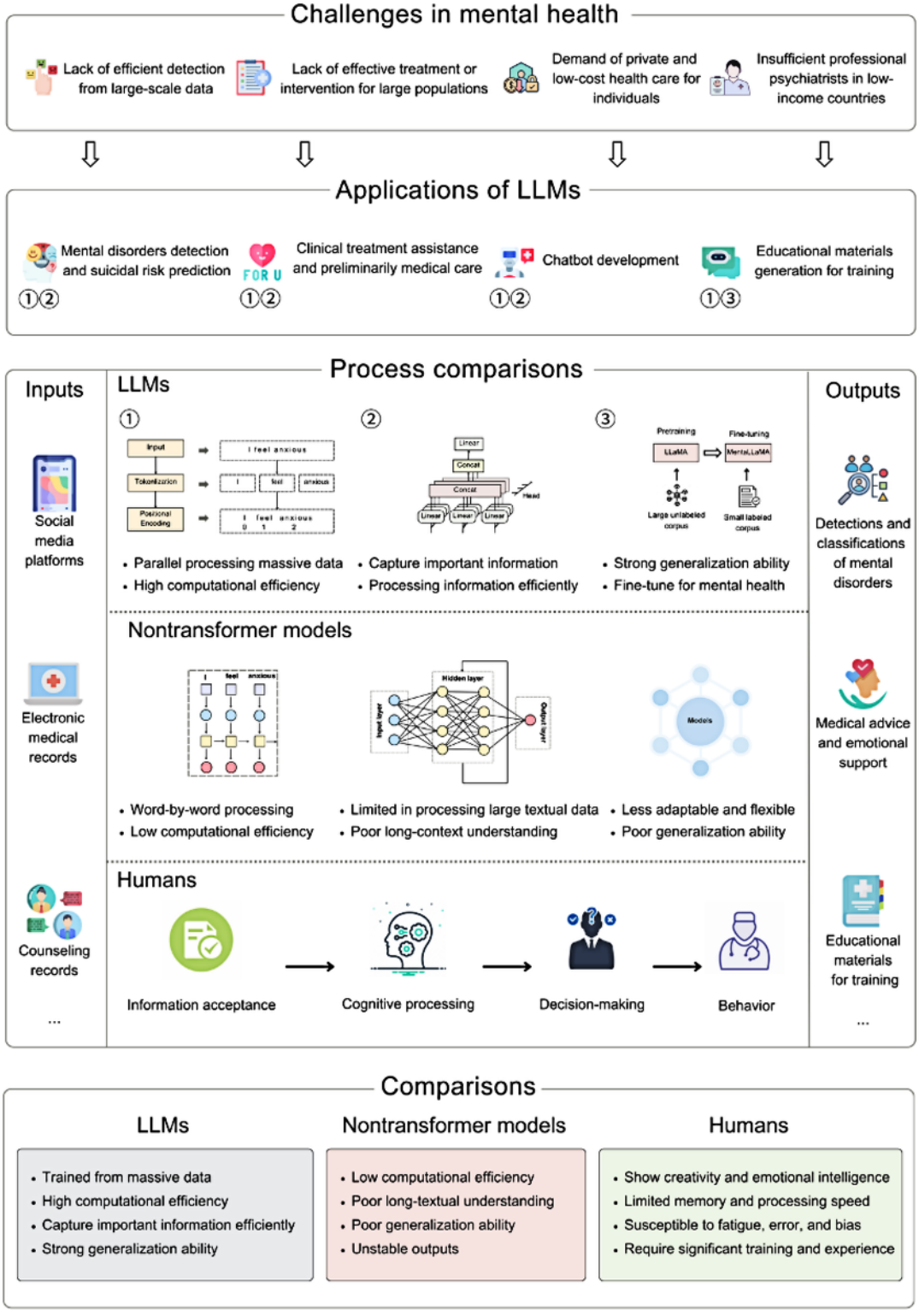
Figure 1. PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flowchart of the studies identified for inclusion in the scoping review about applications of large language models (LLMs) in mental health. CNN: convolutional neural networks; LSTM: long short-term memory.



Comparisons Between LLMs, Nontransformer Models, and Humans

The current challenges in the mental health field include difficulty in efficient mental disorders detection from large-scale data, effective treatment or intervention for large populations, private and low-cost health care, demand for professional psychiatrists, and so on. Compared with nontransformer models and humans, LLMs present higher capabilities in efficient parallel computing of massive data, textual generation, information capture, strong generalization ability, and fine-tuned mental health tasks. The process comparisons between LLMs, nontransformer models, and humans were presented with the framework (Figure 2, a higher resolution version of figure is also available in Multimedia Appendix 5). Therefore, LLMs have been applied to tasks such as detecting or predicting mental disorders, supporting clinical treatment, providing preliminary medical care, and generating educational materials—using datasets from sources such as social media platforms, EMRs, and counseling transcripts.

Figure 2. The current challenges of the mental health field; comparisons between large language models (LLMs), nontransformer models, and humans. LLaMA: large language model; Meta AI; Mental-LLM: large language model for mental health.



Categorization of Studies Based on Mental Health Applications

We categorized studies in terms of the applications of LLMs in the mental health field. These applications were divided into 3 categories: screening or detection of mental disorders (67/95, 71%), supporting the clinical treatments and intervention (31/95, 33%), and assisting in mental health counseling and education (11/95, 12%; Multimedia Appendix 6). Each study was categorized with ≥1 applications; thus, the percentages sum to >100%. Most studies applied LLMs for depression detection and classification (33/95, 34.7%), supporting clinical treatments and interventions (14/95, 15%), and suicide risk prediction (12/95, 13%; Multimedia Appendix 6). These studies used data from social media platforms such as Reddit, Twitter, Facebook, and Weibo, or clinical datasets (Distress Analysis Interview

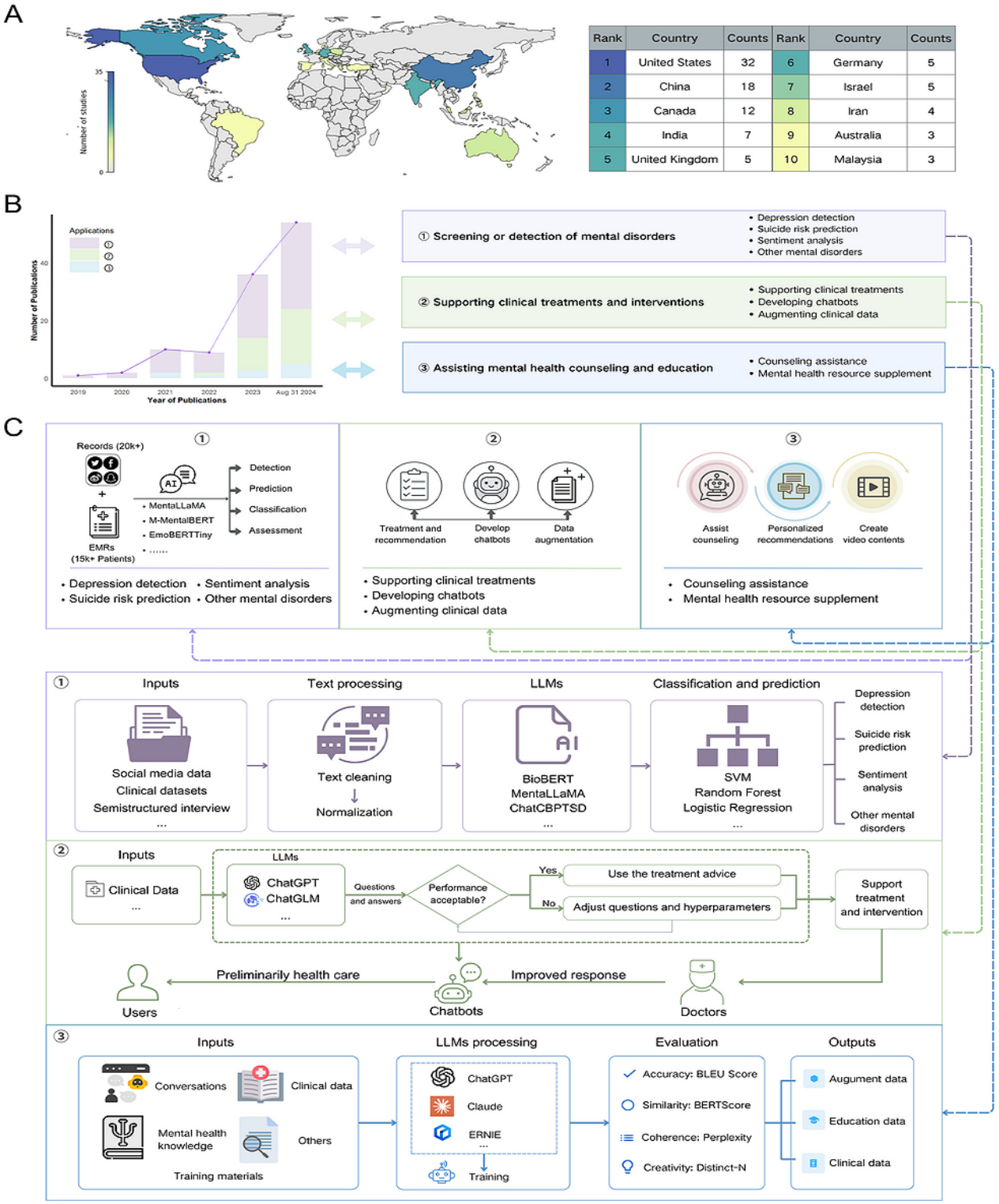
Corpus-Wizard of Oz and Extended Distress Analysis Interview Corpus), as well as semistructured interviews from hospitals. When evaluating the performance of these LLMs, most studies measured the performance of LLMs with various metrics, such as F_1-score (54/95, 57%), precision (34/95, 36%), accuracy (45/95, 47%), and recall (32/95, 34%).

The number of studies mapped by country is presented in Figure 3A (a higher resolution version of figure is also available in Multimedia Appendix 5). The United States is the country that has explored the applications of LLMs in the field of mental health the most, followed by China and Canada. The number of included studies increased year on year, and this trend is shown in Figure 3B. We can find that since 2019, an increasing number of researchers have been exploring the applications of LLMs in the field of mental health. As for the application areas, the first area mainly focused on screening or detection of mental disorders, including depression detection, suicide risk prediction, sentiment analysis, and other mental disorders. The second area focused on supporting clinical treatments and interventions, including supporting clinical treatments, developing chatbots, and augmenting clinical data. The third area focused on assisting mental health counseling and education, including counseling assistance and mental health resource supplement (Figure 3C). These studies applied basic LLMs or fine-tuned LLMs (eg, MentaLLaMA [30], PsychBERT [31], and RoBERTa [68]) to detect or predict depression [30,32-41,49,61,69]; suicide risk [23,24,62]; and other mental disorders, such as anxiety [37,70], obsessive-compulsive disorder [71], and posttraumatic stress disorder [15]. The detailed process for these applications of LLMs is presented in Figure 3C.

In the second application area, most studies explored the capability of LLMs in supporting clinical treatments and interventions [17,21,23,62,63,86], developing chatbots, and augmenting unbalanced clinical data [95,104-106]. These studies applied LLMs to provide treatment advice, assist diagnostic services, and assess prognosis through a question-answering approach. The performance of LLMs was evaluated by professional clinicians and compared with the related performance of humans. A total of 3 studies also applied LLMs to elicit emotion [20-22]. Furthermore, to address the imbalance of clinical data and enhance diagnosis and treatment, LLMs could augment clinical data and targeted dialogues in safe and stable ways.

Moreover, LLMs have been applied for counseling assistance [14,104-106] and educational resource supplements [19,111]. These results showed that the introduction of any interaction (video or chatbot) improved intent to practice and overall experience compared to the baseline [107]. Furthermore, LLMs showed the potential to generate educational materials for training.

Figure 3. Trends in and applications of large language models (LLMs) in mental health: (A) number of studies mapped by country, (B) trends of included studies published per year, and (C) the framework of 3 application categories of LLMs. BERT: bidirectional encoder representations from transformers; BERTScore: bidirectional encoder representations from transformers–based evaluation score; BioBERT: biomedical bidirectional encoder representations from transformers; BLEU: bilingual evaluation understudy; ChatCBPTSD: chat-based cognitive behavioral therapy for posttraumatic stress disorder; ChatGLM: chat general language model; Distnct-N: automatic metric for evaluating diversity in language generation tasks; EMR: electronic medical record; EmoBERTTiny: emotion-aware bidirectional encoder representations from transformers tiny; ERNIE: enhanced representation through knowledge integration; MentalLLaMA: large language model for mental health; M-MentalBERT: multilingual mental health bidirectional encoder representations from transformers; SVM: support vector machine.

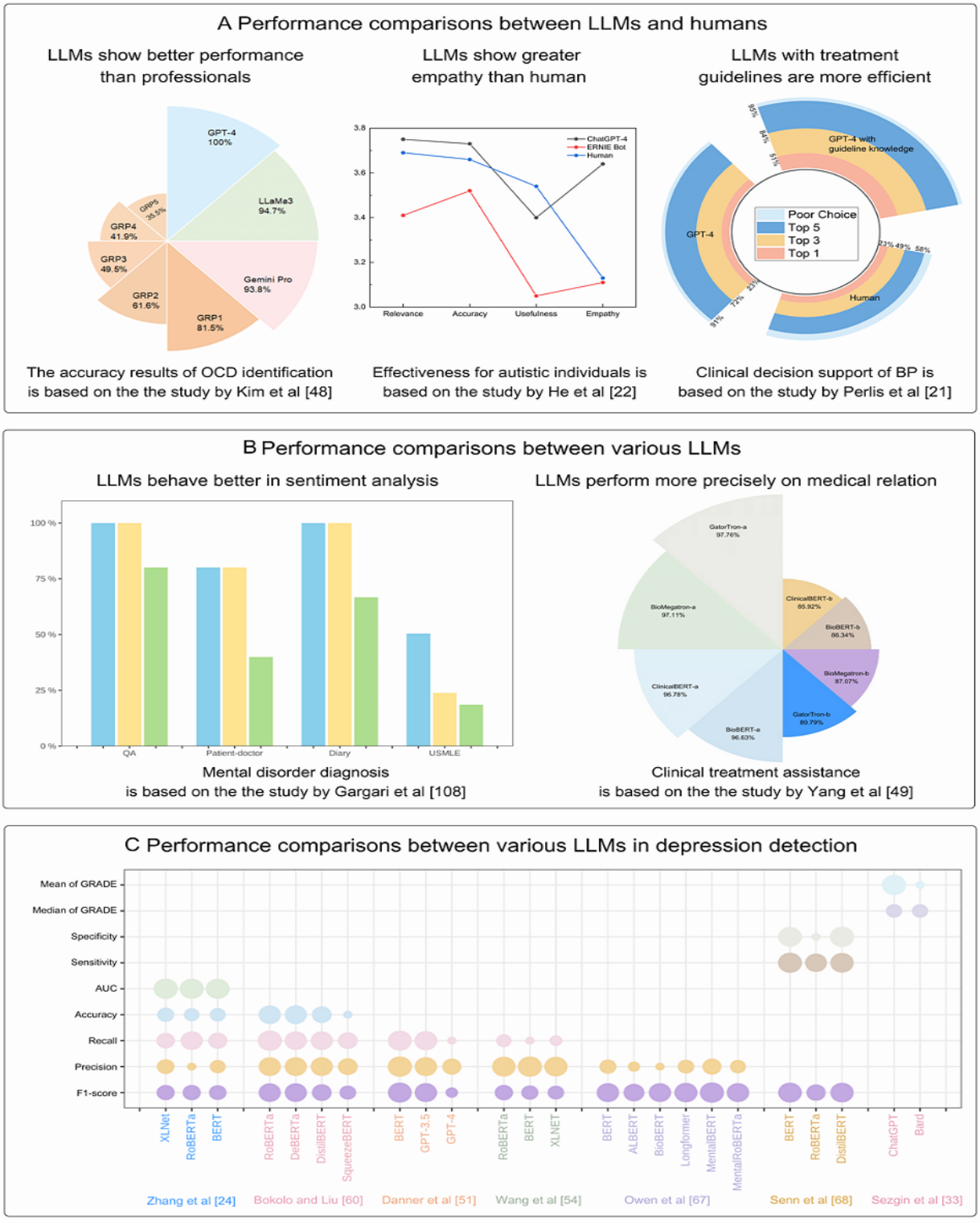


Performance Comparisons Between LLMs and Humans and Between Various LLMs

Several studies compared the performance between humans and LLMs. The metrics' results of performance were displayed in Multimedia Appendix 7. Figure 4A (a higher resolution version of figure is also available in Multimedia Appendix 5) presents the results of 3 studies on obsessive-compulsive disorder identification [71], chatbots efficiency [22], and treatment support [21]. According to these results, most LLMs showed efficient and promising performance for mental health tasks. Several LLMs, such as GPT-4, Claude, and Bard, aligned closely with mental health professionals' perspectives.

Figures 4B and 4C show the comparisons of model performance between different LLMs in mental disorder diagnosis [72], clinical treatment assistance [86], and depression detection [32,40,42-46]. These results found that the latest LLMs (eg, ChatGPT) perform better than traditional and previous models. The complete results are presented in Multimedia Appendix 8.

Figure 4. Performance comparisons between large language models (LLMs) and humans and between various LLMs: (A) performance comparisons between LLMs and humans, (B) performance comparisons between various LLMs, (C) performance comparisons between various LLMs in depression detection. AUC: area under the curve; BioBERT: biomedical bidirectional encoder representations from transformers; BP: bipolar depression; ClinicalBERT: clinical bidirectional encoder representation from transformers; ERNIE: enhanced representation through knowledge integration; GPT: generative pretrained transformer; GRADE: Grading of Recommendations Assessment, Development and Evaluation; GRP: group; GRP1: doctoral trainees; GRP2: APA members; GRP3: primary care physicians; GRP4: medical providers; GRP5: clergy members; OCD: obsessive-compulsive disorder; poor choice: poor or contraindicated medications; QA: question and answers; Top 1/3/5: the first 1/3/5 plans with optimal decisions; USMLE: United States Medical Licensing Exam.



Existing Fine-Tuned LLMs for Mental Health

Table 2 provides the fine-tuned LLMs for mental health, including availability, base models, the number of parameters, training strategy, and published year. These fine-tuned LLMs could be applied specifically to mental health tasks.

Table 2. Fine-tuned large language models (LLM) for mental health.

Availability of the fine-tuned models and base models	Fine-tuned models	Parameters	Training strategy	Year
Yes				
BERT[a]	MBBU[b]	Unreported	Fine-tuning	2024
BERT	BioBERT	Unreported	Unreported	2023
BERT	MentalRoBERTa	Unreported	Unreported	2023
BERT	PsychBERT	Unreported	Domain adaptation (domain-adaptive pretraining)	2021
LLaMA	MentaLLaMA	7 billion-13 billion	IFT[c]	2024
LLaMA	ChatCounselor	7 billion	IFT	2023
FLAN-T5[d]	Mental-FLAN-T5	7 billion-1700 billion	IFT	2024
FLAN-T5	Mental-LLM	7 billion or 11 billion	IFT	2023
GPT	LLM-counselors	Unreported	TFP[e]	2024
GPT	ChatCBPTSD	Unreported	TFP	2023
Alpaca	Mental-alpaca	7 billion-1700 billion	IFT	2024
Not specified				
BERT	CALLM[f]	Unreported	IFT	2024
BERT	EmoBERTTiny	4.4 million	IFT	2024
BERT	M-MentalBERT	Unreported	IFT	2024
BERT	Boamente	Unreported	IFT	2022
BERT	AudiBERT (I, II, and III)	Unreported	IFT	2021
GPT	Psy-LLM	Unreported	TFP	2023
GPT	CareBot	Unreported	IFT	2021

[a]BERT: bidirectional encoder representation from transformers.

[b]MBBU: mentalBERT-base-uncased.

[c]IFT: instruction fine-tuning.

[d]FLAN-T5: fine-tuned language net-t5.

[e]TFP: tuning-free prompting.

[f]CALLM: a framework for systematic contrastive analysis of large language models.

Common Advantages and Disadvantages

Figure 5 summarizes the common advantages and disadvantages of LLMs in mental health. These LLMs could be divided into the BERT series, GPT series, LLaMA series, and others. For example, the shared strengths of BERT-based models make them well-suited for fine-tuning on specific mental health issues. However, the BERT series models require large computational resources. The ChatGPT series models can conduct multiround dialogue, even based on small-sample learning. Nevertheless, the accuracy of ChatGPT models should be improved. Furthermore, GPT-4 could receive multimodal data and show more powerful performance in comprehension and generation. As for the LLaMA series models, they are open source for the public and beneficial for interactive applications, although their complex task performance is inferior to large-scale proprietary models.

Figure 5. The common advantages and disadvantages of large language models (LLMs). ALBERT: a lite bidirectional encoder representations from transformers; BERT: bidirectional encoder representations from transformers; BioBERT: biomedical bidirectional encoder representations from transformers; DistilBERT: distilled version of bidirectional encoder representations from transformers; ERNIE: enhanced representation through knowledge integration; GPT: generative pretrained transformer; LLaMA-7B: LLaMAwith 7 billion parameters; MentalBERT: bidirectional encoder representations from transformers for mental health; RoBERTa: a robustly optimized bidirectional encoder representations from transformers pretraining approach; XLNet: a unsupervised language representation learning method.

Discussion

Principal Findings

This scoping review explored the applications of LLMs in the mental health field and summarized trends, application areas, performance comparisons, challenges, and prospective future directions. The applications of LLMs were categorized into 3 key areas: screening or detection of mental disorders, supporting clinical treatments and interventions, and assisting in mental health counseling and education. Most studies used LLMs for depression detection and classification (33/95, 35%), clinical treatment support and intervention (14/95, 15%), and suicide risk prediction (12/95, 13%). Compared with nontransformer models and humans, LLMs demonstrate higher capabilities in information acquisition and analysis and efficiently generating natural language responses. Furthermore, we summarized the fine-tuning LLMs for mental health and compared their advantages and disadvantages, offering insights for future researchers and psychiatrists.

Advantages of LLMs' Applications in Mental Health

Compared to nontransformer models and humans, LLMs demonstrate higher capabilities in information acquisition, analysis, and generating professional responses. These enhanced capabilities posit LLMs as potential tools for the detection and prediction of mental disorders through the analysis of extensive datasets, including social media content [30,32,33], EMRs [21,86], and counseling notes [87,88]. Their applications in treatment and intervention are noteworthy. LLMs could assimilate patient clinical records, summarize treatment sessions, and support diagnoses for mental disorders. This potential streamlines the workflow for patients and mental health care systems efficiently. On the basis of the information interaction and generation ability, LLMs can be instrumental in the development of chatbots designed for initial medical consultation and emotional support. Such applications help to offer discreet and affordable health care solutions to individuals who, due to stigma or financial constraints, are reluctant to seek assistance for mental health issues.

In this scoping review, most studies applied LLMs in detecting depression and suicidal risk [34-41,49,61,69]. These studies have demonstrated the potential of LLMs such as MentaLLaMA [30], PsychBERT [31], RoBERTa [68], and GPT-4 [89], in detecting and identifying mental disorders from social media platforms and clinical datasets. These models have been trained to recognize depression and recommend evidence-based treatments, with some, such as GPT-4 and Claude, closely aligning with the perspectives of mental health professionals [20]. Furthermore, the integration of linguistic features and the appending of mental disorders' background information have been shown to enhance classification performance and calibration of these LLMs [17,23,62,63]. LLMs perform comparably to experienced clinicians in identifying the risk of suicide ideation, with the addition of suicide attempt history enhancing sensitivity. GPT-4 has demonstrated superior performance over GPT-3.5 in recognizing SI, despite a tendency to underestimate resilience [24]. Other studies showed that the use of LLMs produced effective strategies for predicting suicidal risk with sparsely labeled data [23,62]. This efficiency in labeling and analysis of large datasets is a better advancement over previous methods, which were often hindered by the requirement of manual annotation and consequently limited by small sample sizes or finite datasets.

Due to their efficient ability to acquire information and generate humanlike language, LLMs show potential in preliminary care [14], providing treatment guidelines, augmenting unbalanced clinical datasets [95,104-106], and assisting in training professionals. LLMs also have shown potential to assist with cognitive behavioral therapy tasks, such as reframing unhelpful thoughts and identifying cognitive biases [89]. Several studies also investigated the potential of LLMs in addressing health care regional disparities and enhancing diagnostic and therapeutic services, although these LLMs have only shown initial results so far [20-22]. Furthermore, in the diagnosis of posttraumatic stress disorder [15], LLM-augmented datasets have shown improved performance over original datasets, with both zero-shot and few-shot approaches outperforming the original dataset, highlighting the effectiveness of LLMs in enhancing diagnostic accuracy with minimal additional data. According to the results of performance comparisons, LLMs present similar performance as professionals and may even surpass physicians [22]. These results demonstrate the potential of LLMs in clinical assistance and support. However, when evaluating the quality of LLM-generated responses to queries and educational health materials, several studies have indicated that, despite GPT-4's strong face validity, it remains insufficiently reliable for direct-to-consumer use [111]. Therefore, the outputs by LLMs for educational health require cautious human editing and oversight. These findings underscore the growing importance of LLMs in mental health research and practice, offering new avenues for early detection, risk assessment, and intervention strategies.

Prompt Engineering Techniques of LLMs

Prompt engineering techniques are becoming increasingly essential across various applications and mental health conditions [109]. By optimizing prompt strategies in a targeted manner, models can more effectively perform tasks such as the preliminary detection of mental disorders, intervention recommendations, and emotional assessments, all while adhering to professional and ethical standards. Previous studies have reported that the use of diverse prompts when applying LLMs can lead to substantial variations in the stability and accuracy of the models' outputs [112,113]. Notably, without modifying the underlying architecture of the LLMs, adjusting the input prompts alone can significantly influence the quality and relevance of the outputs. This underscores the critical role of prompt design in optimizing model performance. Thus, it is vital to standardize prompts and share them openly in academic and clinical contexts to enhance the robustness and reproducibility of research findings. In mental health dialogue scenarios, prompts must balance professional standards with ethical considerations to prevent the generation of misleading or inappropriate content [114]. Clinicians should be aware that poorly designed prompts may pose potential risks, such as inaccurate treatment suggestions or unsuitable discussion topics. The ongoing refinement and standardization of prompt engineering not only enhance the performance and explainability of LLMs but also enable health care professionals and researchers to provide more efficient and safer preliminary support and services to patients [115].

Challenges of LLMs' Applications in Mental Health

While LLMs have demonstrated high performance on certain benchmarks, their practical applications in clinical settings are still limited. Recent studies have shown that even advanced LLMs struggle with complex clinical tasks, such as interpreting medical records and providing accurate treatment recommendations [30,32,33]. The efficiency gains from LLMs in clinical settings must be balanced against their limitations and potential risks [108]. As these models continue to evolve, their role in mental health support is likely to expand, with a focus on enhancing data privacy, biases, and ethical considerations in clinical implementation. LLMs are trained on vast amounts of data, which may include sensitive personal and health information. Ensuring compliance with privacy regulations such as the general data protection regulation is

essential to protect patient confidentiality [116,117]. The inadvertent inclusion of personally identifiable information in pretraining datasets can compromise patient privacy, and LLMs can make privacy-invading inferences from seemingly innocuous data. Implementing measures such as data anonymization and safe data storage procedures is crucial to address these issues. Biases in LLMs are another critical issue that must be considered. These models would perpetuate and magnify biases in their training data, leading to differential treatment and outcomes, particularly in populations considered vulnerable [118]. For instance, biases in gender, race, and socioeconomic status can result in inaccurate or misleading information, which may exacerbate existing health disparities [119]. To mitigate these risks, it is important to develop techniques for identifying, alleviating, and preventing biases in LLMs. Ethical concerns are paramount when using LLMs in mental health. The potential for LLMs to generate false or misleading information is a particular concern in health care. Ensuring that LLMs are rigorously tested and monitored is crucial to maintaining reliability in patient care. In addition, transparency is vital for health care professionals to trust the results of LLMs. These LLMs should provide clear, understandable insights for operation with sufficient explanation.

Prospective Future Directions of LLMs in Mental Health

In the future, the application trend of LLMs in mental health is expected to continue to rise, and the aspects of their applications will be broader. Initially, most studies based on the LLMs with textual data and multimodal data, such as pictures, videos, and sounds, could be integrated with multimodal LLMs. Various data types might further improve the performance of mental disorders' detection and identification. Several studies have explored multimodal LLMs in mental health research [72,108]. Moreover, existing studies mainly focus on depression and suicidality; more mental disorders should be investigated with LLMs, especially for rare mental disorders, such as borderline personality disorder and bipolar disorder. The applications of LLMs in treatments, interventions, and preliminary care would be beneficial for these patients. Furthermore, although several studies have developed chatbots for early detection or intervention in mental disorders, further research is needed to enhance the accuracy and robustness of LLMs. In addition, it is important to provide open-resource and fine-tuned LLMs for mental health, especially for low- and middle-income countries. Although LLMs show great performance in various applications in mental health, several areas of LLMs should be improved. First, there is a need for more high-quality, diverse, and representative datasets to train LLMs for mental health applications, ensuring that the models can understand and respond to a wide range of mental health–related queries and scenarios. Second, while LLMs can generate coherent and contextually appropriate responses, they still lag behind human performance in terms of empathy and emotional intelligence, which are crucial in mental health support. Third, LLMs need to improve their ability to reason and understand the context and nuances of mental health dialogues, which often require a deep understanding of human emotions and psychological states. Finally, it is important to establish standards for privacy, safety, and ethical considerations when LLMs process sensitive personal and health information [120,121]. These standards are essential to mitigate the potential risks and ensure the responsible use of LLMs in health care settings. On the one hand, the model training process for LLMs inevitably involves the collection of stigmatizing or biased data [118], which can generate hallucinatory content that may mislead or harm patients. For instance, biased data can result in inaccurate or inappropriate medical advice, which could have serious consequences for individuals seeking health information [105]. Several studies underscore the risk of sensitive data exposure and emphasize the prevention of harmful content generation [71,73]. On the other hand, the convenience and low cost of LLMs may lead teenagers to become overly dependent on them for mental support. This excessive reliance could result in several negative outcomes. It may lead to addiction to the internet-based world, negatively impacting their daily lives and social interactions. Moreover, it could delay the optimal time for teenagers to seek professional help, potentially exacerbating their mental health issues. While LLMs can provide initial support and guidance, they cannot replace the nuanced understanding and empathy that human professionals can offer. Furthermore, LLMs cannot address crises effectively [122]. Although they can identify extreme emotions during conversations, they typically only suggest seeking professional assistance without providing direct and effective measures. LLMs lack the clinical judgment required to handle emergencies, which means they cannot offer the immediate support that may be necessary in critical situations. Therefore, in the practical application of LLMs, professional intervention is essential [123]. Experts should develop a dedicated system of ethical standards to guide the use of LLMs in health care. This system should include regular supervision and evaluations to ensure that LLMs are used responsibly and ethically. LLMs should be used as supplementary tools rather than complete replacements for human roles. They can provide initial support and guidance, but ultimately, the responsibility for clinical judgment and patient care should remain with trained health care professionals. Future advancements depend on collaborative efforts to refine technology, develop standardized evaluations, and ensure ethical applications, aiming to realize LLMs' full potential in supporting mental health care.

Comparison With Previous Studies

Several reviews have explored the applications of LLMs in mental health care from various perspectives. A viewpoint article and a preprint summarized the opportunities and risks associated with using LLMs in mental health [115,124]. The viewpoint article focuses on application scenarios, including education, assessment, and intervention, while the preprint clarifies the potential opportunities and risks of LLMs from an ecological conceptualization perspective, encompassing the care seeker, caregiver, institution, and society [124]. The analysis methods and application categories in these reviews differ from those in this scoping review. Another preprint scoping review also identified diverse applications of LLMs in mental health care, including preprint articles [25]. On the basis of 34 articles screened from 313 articles published between October 1, 2019,

and December 2, 2023, the researchers categorized the application areas into 3 domains: the deployment of conversational agents, resource enrichment, and detailed diagnosis classification. However, this scoping review delved deeper into LLMs and training techniques, dataset characteristics, validation measures and metrics, and challenges. In contrast, our study offers a more targeted and comprehensive reference guide for LLM applications in mental health. Specifically, we compare the application effects of LLMs with human evaluations and among different models, aiming to provide a detailed and structured overview of LLMs' performance and potential in mental health scenarios. Another systematic review with 40 articles assessed the clinical applications of LLMs in mental health [26], focusing on their potential and inherent risks. Although this review summarized the results of each article in tabular form, it did not extract key information. Our scoping review searched 4 English language databases and 3 Chinese language databases with more specific search terms and a wider time range. We categorized these applications into 3 key areas: screening or detection of mental disorders, supporting clinical treatments and interventions, and assisting in mental health counseling and education based on 95 articles screened from 4859 articles. We also extracted key information about LLMs' applications and developed frameworks for each application area. These frameworks are designed to help researchers and clinicians understand the use of LLMs in mental health, even without a background in artificial intelligence. Furthermore, we provided the fine-tuned LLMs for mental health and compared the advantages and disadvantages of LLMs. This comparative approach offers a more nuanced understanding of how different models and human expertise can be leveraged in mental health applications, aiding researchers and clinicians in selecting suitable LLMs for their specific mental health tasks.

Although this scoping review explored the applications of LLMs in mental health, several limitations should be considered. First, there is an absence of assessment of the risk of bias due to the unique nature of these included studies. Moreover, we cannot perform a meta-analysis due to the diversity of methods and tasks in the included studies. Second, with the rapid development of LLMs, the results of comparative studies would be cautious. The performance of these LLMs may have significantly improved. Third, the preprint studies (eg, from arXiv and medRxiv platforms) were not included in this review due to the lack of peer review, though they were published recently. Finally, this review is limited to studies published in English and Chinese, which may compromise the comprehensiveness and representativeness of the findings. This approach may overlook the applications and contributions of LLMs in other regions that speak different languages, thereby failing to provide a comprehensive global overview of LLM applications in mental health. In addition, different languages present unique challenges and opportunities for LLM applications in mental health. Although this scoping review identified a limited number of studies applying LLMs for depression detection in Malay dialects [41] and developing therapeutic dialogue agents in the Polish language [106], the performance of LLMs across various languages can vary based on linguistic characteristics, data availability, and cultural contexts. Limiting the review to English and Chinese studies may underestimate these differences. Moreover, low- and middle-income regions, which often rely on technologies such as LLMs to enhance mental health service accessibility and quality, may use languages other than English or Chinese. Limiting the review to English and Chinese studies could neglect the actual needs and potential contributions of these regions, thereby affecting the development and application of LLMs in those areas. To provide a more holistic view of LLM applications in mental health, future scoping reviews should consider a broader range of languages to provide a more comprehensive understanding.

Conclusions

This scoping review summarized the applications of LLMs in the mental health field, including trends, application areas, performance comparisons, challenges, and future directions. The applications mainly focused on 3 areas: screening or detection of mental disorders, supporting clinical treatments and interventions, and assisting in mental health counseling and education. Compared with nontransformer models and human experts, LLMs demonstrate higher capabilities in information acquisition and analysis, generating natural language responses, and addressing complex reasoning problems. As these models continue to evolve, their role in mental health support is likely to expand, with a focus on enhancing accuracy, sensitivity, and ethical considerations in clinical implementation.

Acknowledgments

The study was funded by the National Natural Science Foundation of China (82201708 and 72304040), the South China Normal University's Striving for the First-Class, Improving Weak Links and Highlighting Features (SIH) Key Discipline for Psychology, and the National Center for Mental Health and Prevention (MHSK2024B24). Additional support was provided by the Guangdong Planning Office of Philosophy and Social Science Program (GD23YXL01) and the Guangdong Education Science Planning Project (2023GXJK668).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. For detailed data, readers may refer to the original studies cited in our review. The complete list of the included studies can be found in the References section.

Authors' Contributions

YW and YB contributed equally to this study as cocorresponding authors. YJ, YW, and YB conceptualized the study, developed the methodology, and led the project administration. YJ also drafted the original manuscript. JL supervised the data analysis process and contributed to data curation, visualization, and validation. PL, BW, YY, HZ, and CN participated in data curation and visualization. JW, YL, and YB provided supervision throughout the study. All authors reviewed, revised, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-ScR checklist.
[DOCX File , 31 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search terms used in the main review for English-language databases.
[DOCX File , 16 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Search terms used in the main review for Chinese-language databases.
[DOCX File , 16 KB-Multimedia Appendix 3]

Multimedia Appendix 4

List of studies excluded at the full-text screening stage.
[DOCX File , 49 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Higher resolution versions of Figures 2-4.
[RAR File , 34363 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Categorization of articles based on mental health applications across 95 studies.
[DOCX File , 19 KB-Multimedia Appendix 6]

Multimedia Appendix 7

The metrics results of performance between large language models and humans.
[DOCX File , 44 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Performance comparisons between various large language models.
[DOCX File , 110 KB-Multimedia Appendix 8]

References

1. Carswell K, Cuijpers P, Gray B, Kestel D, Malik A, Weissbecker I, et al. WHO recommendations on psychological interventions for mental disorders. Lancet Psychiatry. Sep 2024;11(9):678-679. [FREE Full text] [doi: 10.1016/S2215-0366(24)00220-7] [Medline: 39067470]
2. Sacco R, Camilleri N, Eberhardt J, Umla-Runge K, Newbury-Birch D. A systematic review and meta-analysis on the prevalence of mental disorders among children and adolescents in Europe. Eur Child Adolesc Psychiatry. Sep 2024;33(9):2877-2894. [FREE Full text] [doi: 10.1007/s00787-022-02131-2] [Medline: 36581685]
3. Brohan E, Chowdhary N, Dua T, Barbui C, Thornicroft G, Kestel D. The WHO Mental Health Gap Action Programme for mental, neurological, and substance use conditions: the new and updated guideline recommendations. Lancet Psychiatry. Feb 2024;11(2):155-158. [FREE Full text] [doi: 10.1016/S2215-0366(23)00370-X] [Medline: 37980915]

4. Flett GL, Hewitt PL. The need to focus on perfectionism in suicide assessment, treatment and prevention. World Psychiatry. Feb 12, 2024;23(1):152-154. [FREE Full text] [doi: 10.1002/wps.21157] [Medline: 38214627]
5. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A Survey on Evaluation of Large Language Models. ACM Trans Intell Syst Technol. Mar 29, 2024;15(3):1-45. [FREE Full text] [doi: 10.1145/3641289]
6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on May 28, 2020. [FREE Full text]
7. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. J Mach Learn Res. 2023;24(1):11324-11436. [FREE Full text]
8. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on February 27, 2023. [FREE Full text]
9. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv. Preprint posted online on March 31, 2023. [FREE Full text]
10. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond). Oct 10, 2023;3(1):141. [FREE Full text] [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
11. Gao J, Lin CY. Introduction to the special issue on statistical language modeling. ACM Trans Asian Lang Inf Process. Jun 2004;3(2):87-93. [FREE Full text] [doi: 10.1145/1034780.1034781]
12. Kombrink S, Mikolov T, Karafiát M, Burget L. Recurrent neural network based language modeling in meeting recognition. In: Proceedings of the Interspeech 2011. 2011. Presented at: Interspeech 2011; August 27-31, 2011; Florence, Italy. [doi: 10.21437/interspeech.2011-720]
13. Torous J, Blease C. Generative artificial intelligence in mental health care: potential benefits and current challenges. World Psychiatry. Feb 2024;23(1):1-2. [FREE Full text] [doi: 10.1002/wps.21148] [Medline: 38214643]
14. Beredo JL, Ong EC. A hybrid response generation model for an empathetic conversational agent. In: Proceedings of the 2022 International Conference on Asian Language Processing. 2022. Presented at: IALP 2022; October 27-28, 2022; Singapore, Singapore. [doi: 10.1109/ialp57159.2022.9961311]
15. Wu Y, Chen J, Mao K, Zhang Y. Automatic post-traumatic stress disorder diagnosis via clinical transcripts: a novel text augmentation with large language models. In: Proceedings of the 2023 IEEE Biomedical Circuits and Systems Conference. 2023. Presented at: BioCAS 2023; October 19-21, 2023; Toronto, ON. URL: https://ieeexplore.ieee.org/abstract/document/10388714 [doi: 10.1109/biocas58349.2023.10388714]
16. Bartal A, Jagodnik KM, Chan SJ, Dekel S. AI and narrative embeddings detect PTSD following childbirth via birth stories. Sci Rep. Apr 11, 2024;14(1):8336. [FREE Full text] [doi: 10.1038/s41598-024-54242-2] [Medline: 38605073]
17. Lee C, Mohebbi M, O'Callaghan E, Winsberg M. Large language models versus expert clinicians in crisis prediction among telemental health patients: comparative study. JMIR Ment Health. Aug 02, 2024;11:e58129. [FREE Full text] [doi: 10.2196/58129] [Medline: 38876484]
18. Jeon E, Yoon N, Sohn SY. Exploring new digital therapeutics technologies for psychiatric disorders using BERTopic and PatentSBERTa. Technol Forecast Soc Change. Jan 2023;186:122130. [FREE Full text] [doi: 10.1016/j.techfore.2022.122130]
19. Mármol-Romero AM, García-Vega M, García-Cumbreras MÁ, Montejo-Ráez A. An empathic GPT-based chatbot to talk about mental disorders with Spanish teenagers. Int J Hum Comput Interact. May 08, 2024;41(7):1-17. [FREE Full text] [doi: 10.1080/10447318.2024.2344355]
20. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health. Jan 09, 2024;12(Suppl 1):e002583. [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
21. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology. Aug 13, 2024;49(9):1412-1416. [doi: 10.1038/s41386-024-01841-2] [Medline: 38480911]
22. He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician versus large language model chatbot responses to web-based questions from autistic patients in Chinese: cross-sectional comparative analysis. J Med Internet Res. Apr 30, 2024;26:e54706. [FREE Full text] [doi: 10.2196/54706] [Medline: 38687566]
23. Metzler H, Baginski H, Niederkrotenthaler T, Garcia D. Detecting potentially harmful and protective suicide-related content on Twitter: machine learning approach. J Med Internet Res. Aug 17, 2022;24(8):e34705. [FREE Full text] [doi: 10.2196/34705] [Medline: 35976193]
24. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. Proc ACM Interact Mob Wearable Ubiquitous Technol. Mar 06, 2024;8(1):1-32. [doi: 10.1145/3643540] [Medline: 39925940]
25. Hua Y, Liu F, Yang K, Li Z, Na H, Sheu YH, et al. Large language models in mental health care: a scoping review. arXiv. Preprint posted online on January 1, 2024. [FREE Full text]
26. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. Oct 18, 2024;11:e57400. [doi: 10.2196/57400] [Medline: 39423368]

27. The applications of large language models in mental health: a scoping review. OSF Registries. 2024. URL: https://osf.io/yg9xj [accessed 2025-04-22]
28. Armstrong R, Hall BJ, Doyle J, Waters E. Cochrane update. 'Scoping the scope' of a Cochrane review. J Public Health (Oxf). Mar 23, 2011;33(1):147-150. [doi: 10.1093/pubmed/fdr015] [Medline: 21345890]
29. Munn Z, Peters MD, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. Nov 19, 2018;18(1):1-7. [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
30. Yang K, Zhang T, Kuang Z, Xie Q, Huang J, Ananiadou S. MentaLLaMA: interpretable mental health analysis on social media with large language models. In: Proceedings of the ACM Web Conference 2024. 2024. Presented at: WWW '24; May 13-17, 2024; Singapore, Singapore. URL: https://dl.acm.org/doi/abs/10.1145/3589334.3648137 [doi: 10.1145/3589334.3648137]
31. Vajre V, Naylor M, Kamath U, Shehu A. PsychBERT: a mental health language model for social media mental health behavioral analysis. In: Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine. 2021. Presented at: BIBM 2021; December 09-12, 2021; Houston, TX. [doi: 10.1109/bibm52615.2021.9669469]
32. Zhang Y, Lyu H, Liu Y, Zhang X, Wang Y, Luo J. Monitoring depression trends on Twitter during the COVID-19 pandemic: observational study. JMIR Infodemiology. 2021;1(1):e26769. [FREE Full text] [doi: 10.2196/26769] [Medline: 34458682]
33. Suri M, Semwal N, Chaudhary D, Gorton I, Kumar B. I don’t feel so good! Detecting depressive tendencies using transformer-based multimodal frameworks. In: Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing. 2022. Presented at: MLNLP '22; December 23-25, 2022; Sanya, China. [doi: 10.1145/3578741.3578817]
34. Guo Y, Liu J, Wang L, Qin W, Hao S, Hong R. A prompt-based topic-modeling method for depression detection on low-resource data. IEEE Trans Comput Soc Syst. Feb 2024;11(1):1430-1439. [doi: 10.1109/tcss.2023.3260080]
35. Abilkaiyrkyzy A, Laamarti F, Hamdi M, Saddik AE. Dialogue system for early mental illness detection: toward a digital twin solution. IEEE Access. 2024;12:2007-2024. [doi: 10.1109/access.2023.3348783]
36. Abdullah M, Negied N. Detection and prediction of future mental disorder from social media data using machine learning, ensemble learning, and large language models. IEEE Access. 2024;12:120553-120569. [doi: 10.1109/access.2024.3406469]
37. Tao Y, Yang M, Shen H, Yang Z, Weng Z, Hu B. Classifying anxiety and depression through LLMs virtual interactions: a case study with ChatGPT. In: Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine. 2023. Presented at: BIBM 2023; December 5-8, 2023; Istanbul, Turkiye. [doi: 10.1109/bibm58861.2023.10385305]
38. Sood P, Yang X, Wang P. Enhancing depression detection from narrative interviews using language models. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. 2023. Presented at: BIBM 2023; December 5-8, 2023; Istanbul, Turkiye. [doi: 10.1109/bibm58861.2023.10385480]
39. Lu KC, Thamrin SA, Chen AL. Depression detection via conversation turn classification. Multimed Tools Appl. Apr 01, 2023;82(25):39393-39413. [doi: 10.1007/s11042-023-15103-8]
40. Bokolo BG, Liu Q. Deep learning-based depression detection from social media: comparative evaluation of ML and transformer techniques. Electronics. Oct 24, 2023;12(21):4396. [doi: 10.3390/electronics12214396]
41. Hayati MF, Ali MA, Rosli AN. Depression detection on Malay dialects using GPT-3. In: Proceedings of the 2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences. 2022. Presented at: IECBES 2022; December 7-9, 2022; Kuala Lumpur, Malaysia. [doi: 10.1109/iecbes54088.2022.10079554]
42. Danner M, Hadzic B, Gerhardt S, Ludwig S, Uslu I, Shao P. Advancing mental health diagnostics: GPT-based method for depression detection. In: Proceedings of the 62nd Annual Conference of the Society of Instrument and Control Engineers. 2023. Presented at: SICE 2023; September 06-09, 2023; Tsu, Japan. [doi: 10.23919/sice59929.2023.10354236]
43. Wang X, Chen S, Li T, Li W, Zhou Y, Zheng J, et al. Depression risk prediction for Chinese microblogs via deep-learning methods: content analysis. JMIR Med Inform. Jul 29, 2020;8(7):e17958. [FREE Full text] [doi: 10.2196/17958] [Medline: 32723719]
44. Owen D, Antypas D, Hassoulas A, Pardiñas AF, Espinosa-Anke L, Collados JC. Enabling early health care intervention by detecting depression in users of web-based forums using language models: longitudinal analysis and evaluation. JMIR AI. Mar 24, 2023;2:e41205. [FREE Full text] [doi: 10.2196/41205] [Medline: 37525646]
45. Senn S, Tlachac ML, Flores R, Rundensteiner E. Ensembles of BERT for depression classification. In: Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society. 2022. Presented at: EMBC 2022; July 11-15, 2022; Glasgow, UK. [doi: 10.1109/embc48229.2022.9871120]
46. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res. Sep 11, 2023;25:e49240. [FREE Full text] [doi: 10.2196/49240] [Medline: 37695668]
47. Sadeghi M, Egger B, Agahi R, Richer R, Capito K, Rupp LH. Exploring the capabilities of a language model-only approach for depression detection in text data. In: Proceedings of the IEEE EMBS International Conference on Biomedical and Health Informatics. 2023. Presented at: BHI 2023; October 15-18, 2023; Pittsburgh, PA. [doi: 10.1109/bhi58575.2023.10313367]
48. Farruque N, Goebel R, Sivapalan S, Zaïane OR. Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach. Lang Resour Eval. Apr 04, 2024;58(3):1013-1041. [doi: 10.1007/s10579-024-09720-4]

49. Tey W, Goh H, Lim AH, Phang C. Pre- and post-depressive detection using deep learning and textual-based features. Int J Technol. Oct 31, 2023;14(6):1334-1343. [doi: 10.14716/ijtech.v14i6.6648]
50. de Hond A, van Buchem M, Fanconi C, Roy M, Blayney D, Kant I, et al. Predicting depression risk in patients with cancer using multimodal data: algorithm development study. JMIR Med Inform. Jan 18, 2024;12:e51925. [FREE Full text] [doi: 10.2196/51925] [Medline: 38236635]
51. Ilias L, Mouzakitis S, Askounis D. Calibration of transformer-based models for identifying stress and depression in social media. IEEE Trans Comput Soc Syst. Apr 2024;11(2):1979-1990. [doi: 10.1109/tcss.2023.3283009]
52. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health. Sep 16, 2023;11(4):e002391. [FREE Full text] [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
53. Toto E, Tlachac ML, Rundensteiner EA. AudiBERT: a deep transfer learning multimodal classification framework for depression screening. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021. Presented at: CIKM '21; November 1-5, 2021; Virtual Event. [doi: 10.1145/3459637.3481895]
54. Lau C, Zhu X, Chan WY. Automatic depression severity assessment with deep learning using parameter-efficient tuning. Front Psychiatry. Jun 15, 2023;14:1160291. [FREE Full text] [doi: 10.3389/fpsyt.2023.1160291] [Medline: 37398577]
55. Verma S, Vishal, Joshi RC, Dutta MK, Jezek S, Burget R. AI-enhanced mental health diagnosis: leveraging transformers for early detection of depression tendency in textual data. In: Proceedings of the 15th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. 2023. Presented at: ICUMT 2023; October 30-November 1, 2023; Ghent, Belgium. [doi: 10.1109/icumt61075.2023.10333301]
56. Pourkeyvan A, Safa R, Sorourkhah A. Harnessing the power of hugging face transformers for predicting mental health disorders in social networks. IEEE Access. 2024;12:28025-28035. [doi: 10.1109/access.2024.3366653]
57. Englhardt Z, Ma C, Morris ME, Chang CC, Xu X, Qin L, et al. From classification to clinical insights: towards analyzing and reasoning about mobile and behavioral health data with large language models. Proc ACM Interact Mob Wearable Ubiquitous Technol. May 15, 2024;8(2):1-25. [doi: 10.1145/3659604]
58. Firoz N, Berestneva O, Aksyonov SV. Dual layer Cogni - insight deep-mood encoder: a two- tiered approach for depression detection. In: Proceedings of the International Russian Smart Industry Conference. 2024. Presented at: SmartIndustryCon 2024; March 25-29, 2024; Sochi, Russian Federation. [doi: 10.1109/smartindustrycon61328.2024.10516113]
59. Huang L. A study on the design of a depression diagnostic framework based on DepGPT and neural networks. In: Proceedings of the 2nd International Conference on Mechatronics, IoT and Industrial Informatics. 2024. Presented at: ICMIII 2024; June 12-14, 2024; Melbourne, Australia. [doi: 10.1109/icmiii62623.2024.00137]
60. Shin D, Kim H, Lee S, Cho Y, Jung W. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. J Med Internet Res. Sep 18, 2024;26:e54617. [FREE Full text] [doi: 10.2196/54617] [Medline: 39292502]
61. Qorich M, El Ouazzani R. Advanced deep learning and large language models for suicide ideation detection on social media. Prog Artif Intell. Jun 21, 2024;13(2):135-147. [doi: 10.1007/s13748-024-00326-z]
62. Howard D, Maslej MM, Lee J, Ritchie J, Woollard G, French L. Transfer learning for risk classification of social media posts: model evaluation study. J Med Internet Res. May 13, 2020;22(5):e15371. [FREE Full text] [doi: 10.2196/15371] [Medline: 32401222]
63. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health. Sep 20, 2023;10:e51232. [FREE Full text] [doi: 10.2196/51232] [Medline: 37728984]
64. Ghanadian H, Nejadgholi I, Osman HA. Socially aware synthetic data generation for suicidal ideation detection using large language models. IEEE Access. 2024;12:14350-14363. [doi: 10.1109/access.2024.3358206]
65. Diniz EJ, Fontenele JE, de Oliveira AC, Bastos VH, Teixeira S, Rabêlo RL, et al. Boamente: a natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. Healthcare (Basel). Apr 08, 2022;10(4):698. [doi: 10.3390/healthcare10040698] [Medline: 35455874]
66. Wang R, Yang BX, Ma Y, Wang P, Yu Q, Zong X, et al. Medical-level suicide risk analysis: a novel standard and evaluation model. IEEE Internet Things J. Dec 1, 2021;8(23):16825-16834. [doi: 10.1109/jiot.2021.3052363]
67. Bauer B, Norel R, Leow A, Rached ZA, Wen B, Cecchi G. Using large language models to understand suicidality in a social media-based taxonomy of mental health disorders: linguistic analysis of Reddit posts. JMIR Ment Health. May 16, 2024;11:e57234. [FREE Full text] [doi: 10.2196/57234] [Medline: 38771256]
68. Kumar A, Trueman TE, Cambria E. Stress identification in online social networks. In: Proceedings of the IEEE International Conference on Data Mining Workshops. 2022. Presented at: ICDMW 2022; November 28-December 1, 2022; Orlando, FL. [doi: 10.1109/icdmw58026.2022.00063]
69. Jain B, Goyal G, Sharma M. Evaluating emotional detection and classification capabilities of GPT-2 and GPT-neo using textual data. In: Proceedings of the 14th International Conference on Cloud Computing, Data Science & Engineering. 2024. Presented at: Confluence 2024; January 18-19, 2024; Noida, India. [doi: 10.1109/confluence60223.2024.10463396]
70. Teferra BG, Rose J. Predicting generalized anxiety disorder from impromptu speech transcripts using context-aware transformer-based neural networks: model evaluation study. JMIR Ment Health. Mar 28, 2023;10:e44325. [FREE Full text] [doi: 10.2196/44325] [Medline: 36976636]

71. Kim J, Leonte KG, Chen ML, Torous JB, Linos E, Pinto A, et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. NPJ Digit Med. Jul 19, 2024;7(1):193. [FREE Full text] [doi: 10.1038/s41746-024-01181-x] [Medline: 39030292]
72. Gargari OK, Fatehi F, Mohammadi I, Firouzabadi SR, Shafiee A, Habibi G. Diagnostic accuracy of large language models in psychiatry. Asian J Psychiatr. Oct 2024;100:104168. [doi: 10.1016/j.ajp.2024.104168] [Medline: 39111087]
73. Wang Y, Yu Y, Liu Y, Ma Y, Pang PC. Predicting patients' satisfaction with mental health drug treatment using their reviews: unified interchangeable model fusion approach. JMIR Ment Health. Dec 05, 2023;10:e49894. [FREE Full text] [doi: 10.2196/49894] [Medline: 38051580]
74. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry. Sep 1, 2023;14:1234397. [FREE Full text] [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
75. Tanana MJ, Soma CS, Kuo PB, Bertagnolli NM, Dembe A, Pace BT, et al. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. Behav Res Methods. Oct 22, 2021;53(5):2069-2082. [FREE Full text] [doi: 10.3758/s13428-020-01531-z] [Medline: 33754322]
76. Wan C, Ge X, Wang J, Zhang X, Yu Y, Hu J, et al. Identification and impact analysis of family history of psychiatric disorder in mood disorder patients with pretrained language model. Front Psychiatry. May 20, 2022;13:861930. [FREE Full text] [doi: 10.3389/fpsyt.2022.861930] [Medline: 35669265]
77. Dai HJ, Su CH, Lee YQ, Zhang YC, Wang CK, Kuo CJ, et al. Deep learning-based natural language processing for screening psychiatric patients. Front Psychiatry. Jan 15, 2020;11:533949. [FREE Full text] [doi: 10.3389/fpsyt.2020.533949] [Medline: 33584354]
78. Gargari OK, Habibi G, Nilchian N, Shafiee A. Comparative analysis of large language models in psychiatry and mental health: a focus on GPT, AYA, and Nemotron-3-8B. Asian J Psychiatr. Sep 2024;99:104148. [doi: 10.1016/j.ajp.2024.104148] [Medline: 39047354]
79. Rathje S, Mirea DM, Sucholutsky I, Marjieh R, Robertson CE, Van Bavel JJ. GPT is an effective tool for multilingual psychological text analysis. Proc Natl Acad Sci U S A. Aug 20, 2024;121(34):e2308950121. [FREE Full text] [doi: 10.1073/pnas.2308950121] [Medline: 39133853]
80. Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, et al. A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data. JMIR Ment Health. Jan 25, 2024;11:e50150. [FREE Full text] [doi: 10.2196/50150] [Medline: 38271138]
81. Stigall W, Khan MA, Attota D, Nweke F, Pei Y. Large language models performance comparison of emotion and sentiment classification. In: Proceedings of the 2024 ACM Southeast Conference. 2024. Presented at: ACMSE '24; April 18-20, 2024; Marietta, GA. [doi: 10.1145/3603287.3651183]
82. Yongsatianchot N, Torshizi PG, Marsella S. Investigating large language models' perception of emotion using appraisal theory. In: Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2023. Presented at: ACIIW 2023; September 10-13, 2023; Cambridge, MA. [doi: 10.1109/aciiw59127.2023.10388194]
83. Goyal T, Rajeshbai DH, Gopalkrishna N, M T, HR M. Mobile machine learning models for emotion and sarcasm detection in text: a solution for alexithymic individuals. In: Proceedings of the 3rd International Conference for Innovation in Technology. 2024. Presented at: INOCON 2024; March 01-03, 2024; Bangalore, India. [doi: 10.1109/inocon60754.2024.10511772]
84. 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. Feb 06, 2024;11:e54369. [FREE Full text] [doi: 10.2196/54369] [Medline: 38319707]
85. Fan Y, Nie J, Sun X, Jiang X. Exploring foundation models in detecting concerning daily functioning in psychotherapeutic context based on images from smart home devices. In: Proceedings of the IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things. 2024. Presented at: FMSys 2024; May 13-15, 2024; Hong Kong. [doi: 10.1109/fmsys62467.2024.00012]
86. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. Dec 26, 2022;5(1):194. [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
87. Huang S, Fu F, Yang K, Zhang K, Yang F. Empowerment of large language models in psychological counseling through prompt engineering. In: Proceedings of the 2024 IEEE 4th International Conference on Software Engineering and Artificial Intelligence. 2024. Presented at: SEAI 2024; June 21-23, 2024; Xiamen, China. [doi: 10.1109/seai62072.2024.10674052]
88. Adhikary PK, Srivastava A, Kumar S, Singh SM, Manuja P, Gopinath JK, et al. Exploring the efficacy of large language models in summarizing mental health counseling sessions: benchmark study. JMIR Ment Health. Jul 23, 2024;11:e57306. [FREE Full text] [doi: 10.2196/57306] [Medline: 39042893]
89. Hodson N, Williamson S. Can large language models replace therapists? Evaluating performance at simple cognitive behavioral therapy tasks. JMIR AI. Jul 30, 2024;3:e52500. [FREE Full text] [doi: 10.2196/52500] [Medline: 39078696]

90. McElroy E, Wood T, Bond R, Mulvenna M, Shevlin M, Ploubidis GB, et al. Using natural language processing to facilitate the harmonisation of mental health questionnaires: a validation study using real-world data. BMC Psychiatry. Jul 24, 2024;24(1):530. [FREE Full text] [doi: 10.1186/s12888-024-05954-2] [Medline: 39049010]
91. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr. Nov 2023;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
92. Wang X, Liu K, Wang C. Knowledge-enhanced pre-training large language model for depression diagnosis and treatment. In: Proceedings of the IEEE 9th International Conference on Cloud Computing and Intelligent Systems. 2023. Presented at: CCIS 2023; August 12-13, 2023; Dali, China. [doi: 10.1109/ccis59572.2023.10263217]
93. Friedman SF, Ballentine G. Trajectories of sentiment in 11,816 psychoactive narratives. Hum Psychopharmacol. Jan 20, 2024;39(1):e2889. [doi: 10.1002/hup.2889] [Medline: 38117133]
94. Kamoji S, Rozario S, Almeida S, Patil S, Patankar S, Pendhari H. Mental health prediction using machine learning models and large language model. In: Proceedings of the 2024 Second International Conference on Inventive Computing and Informatics. 2024. Presented at: ICICI 2024; June 11-12, 2024; Bangalore, India. [doi: 10.1109/icici62254.2024.00040]
95. Heston T. Safety of large language models in addressing depression. Cureus. Dec 2023;15(12):e50729. [FREE Full text] [doi: 10.7759/cureus.50729] [Medline: 38111813]
96. Llanes-Jurado J, Gómez-Zaragozá L, Minissi ME, Alcañiz M, Marín-Morales J. Developing conversational virtual Humans for social emotion elicitation based on large language models. Expert Syst Appl. Jul 2024;246:123261. [doi: 10.1016/j.eswa.2024.123261]
97. Berrezueta-Guzman S, Kandil M, Martin-Ruiz ML, de la Cruz IP, Krusche S. Exploring the efficacy of robotic assistants with ChatGPT and Claude in enhancing ADHD therapy: innovating treatment paradigms. In: Proceedings of the International Conference on Intelligent Environments. 2024. Presented at: IE 2024; June 17-20, 2024; Ljubljana, Slovenia. [doi: 10.1109/ie61493.2024.10599903]
98. Gabor-Siatkowska K, Sowański M, Rzatkiewicz R, Stefaniak I, Kozłowski M, Janicki A. AI to train AI: using ChatGPT to improve the accuracy of a therapeutic dialogue system. Electronics. Nov 18, 2023;12(22):4694. [doi: 10.3390/electronics12224694]
99. Maples B, Cerit M, Vishwanath A, Pea R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. Npj Ment Health Res. Jan 22, 2024;3(1):4. [FREE Full text] [doi: 10.1038/s44184-023-00047-6] [Medline: 38609517]
100. Wu Y, Mao K, Zhang Y, Chen J. CALLM: enhancing clinical interview analysis through data augmentation with large language models. IEEE J Biomed Health Inform. Dec 2024;28(12):7531-7542. [doi: 10.1109/JBHI.2024.3435085] [Medline: 39074002]
101. Cai Z, Fang H, Liu J, Xu G, Long Y. Instruction tuning of LLM for unified information extraction in mental health domain. J Chin Inf Process. 2024;38(8):112-127. [FREE Full text]
102. Aygün İ, Kaya M. Use of large language models for medical synthetic data generation in mental illness. IET Conf Proc. Feb 27, 2024;2023(44):652-656. [doi: 10.1049/icp.2024.1033]
103. Sung CW, Lee YK, Tsai YT. A new pipeline for generating instruction dataset via RAG and self fine-tuning. In: Proceedings of the 2024 IEEE 48th Annual Computers, Software, and Applications Conference. 2024. Presented at: COMPSAC 2024; July 02-04, 2024; Osaka, Japan. [doi: 10.1109/compsac61105.2024.00371]
104. Crasto R, Dias L, Miranda D, Kayande D. CareBot: a mental health ChatBot. In: Proceedings of the 2nd International Conference for Emerging Technology. 2021. Presented at: INCET 2021; May 21-23, 2021; Belagavi, India. [doi: 10.1109/incet51464.2021.9456326]
105. Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. AMIA Annu Symp Proc. 2023;2023:1105. [FREE Full text] [Medline: 38222348]
106. Zygadlo A. A therapeutic dialogue agent for Polish language. In: Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2021. Presented at: ACIIW 2021; September 28-October 1, 2021; Nara, Japan. [doi: 10.1109/aciiw52867.2021.9666281]
107. Kumar H, Wang Y, Shi J, Musabirov I, Farb NA, Williams JJ. Exploring the use of large language models for improving the awareness of mindfulness. In: Proceedings o the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. 2023. Presented at: CHI EA '23; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544549.3585614]
108. Thapa S, Adhikari S. GPT-4o and multimodal large language models as companions for mental wellbeing. Asian J Psychiatr. Sep 2024;99:104157. [doi: 10.1016/j.ajp.2024.104157] [Medline: 39053243]
109. Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. J Med Artif Intell. Oct 2023;6:20. [doi: 10.21037/jmai-23-71]
110. Milligan G, Bernard A, Dowthwaite L, Vallejos EP, Davis J, Salhi L, et al. Developing a single‐session outcome measure using natural language processing on digital mental health transcripts. Couns Psychother Res. May 30, 2024;24(3):1057-1068. [doi: 10.1002/capr.12766]
111. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use ChatGPT for mental health and substance use education? Examining its quality and potential harms. JMIR Med Educ. Nov 30, 2023;9:e51243. [FREE Full text] [doi: 10.2196/51243] [Medline: 38032714]

112. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med. Feb 20, 2024;7(1):41. [FREE Full text] [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
113. Lainwright N, Pemberton M. Assessing the response strategies of large language models under uncertainty: a comparative study using prompt engineering. OSF Preprints. Preprint posted online on August 01, 2024. [FREE Full text] [doi: 10.31219/osf.io/34yqj]
114. Hadi MU, Tashi QA, Qureshi R, Shah A, Irfan M, Zafar A, et al. A survey on large language models: applications, challenges, limitations, and practical usage. TechRxiv. Preprint posted online on July 10, 2023. [FREE Full text] [doi: 10.36227/techrxiv.23589741.v1]
115. Lawrence HR, Schneider RA, Rubin SB, Matarić MJ, McDuff DJ, Jones Bell M. The opportunities and risks of large language models in mental health. JMIR Ment Health. Jul 29, 2024;11:e59479. [FREE Full text] [doi: 10.2196/59479] [Medline: 39105570]
116. Neame R. Privacy protection for personal health information and shared care records. Inform Prim Care. 2014;21(2):84-91. [FREE Full text] [doi: 10.14236/jhi.v21i2.55] [Medline: 24841409]
117. Paavola J, Ekqvist J. Privacy preserving and resilient cloudified IoT architecture to support eHealth systems. In: Proceedings of the Third International Conference on Interoperability, Safety and Security in IoT. 2017. Presented at: InterIoT 2017; November 6-7, 2017; Valencia, Spain. [doi: 10.1007/978-3-319-93797-7_15]
118. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med. Feb 08, 2025;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
119. Potter L, Zawadzki MJ, Eccleston CP, Cook JE, Snipes SA, Sliwinski MJ, et al. The intersections of race, gender, age, and socioeconomic status: implications for reporting discrimination and attributions to discrimination. Stigma Health. Aug 2019;4(3):264-281. [FREE Full text] [doi: 10.1037/sah0000099] [Medline: 31517056]
120. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA. Apr 25, 2023;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
121. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. Lancet Digit Health. Sep 2024;6(9):e662-e672. [doi: 10.1016/s2589-7500(24)00124-9]
122. Han T, Kumar A, Agarwal C, Lakkaraju H. Towards safe large language models for medicine. ArXiv. Preprint posted online on May 1, 2024. [FREE Full text]
123. Fu G, Zhao Q, Dan L, Li J, Song J, Wei Z, et al. Enhancing psychological counseling with large language model: a multifaceted decision-support system for non-professionals. arXiv. Preprint posted online on Septeber 11, 2023. [doi: 10.2196/preprints.52656]
124. De Choudhury M, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health. PsyArXiv Preprints. Preprint posted online on November 15, 2023. [FREE Full text] [doi: 10.31234/osf.io/y8ax9]

Abbreviations

EMR: electronic medical record
LLM: large language model
NLP: natural language processing
PaLM: Pathways Language Model
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by T de Azevedo Cardoso; submitted 26.11.24; peer-reviewed by M Naguib, B Davis, G Shu, S Chowdhury, S Olalere; comments to author 11.01.25; revised version received 09.02.25; accepted 14.03.25; published 05.05.25

Please cite as:
Jin Y, Liu J, Li P, Wang B, Yan Y, Zhang H, Ni C, Wang J, Li Y, Bu Y, Wang Y
The Applications of Large Language Models in Mental Health: Scoping Review
J Med Internet Res 2025;27:e69284
URL: https://www.jmir.org/2025/1/e69284
doi: 10.2196/69284
PMID: