

Large Language Models in Mental Health Care

A Scoping Review

**Yining Hua, MSc^{1,2,*}, Fenglin Liu, MSc³, Kailai Yang, BEng⁴, Zehan Li, MSc⁵, Hongbin Na, BEng⁶,
Yi-han Sheu, MD, MPH, ScD^{7,8}, Peilin Zhou, MSc⁹, Lauren V. Moran, MD¹⁰, Sophia Ananiadou,
PhD^{4,11}, David A. Clifton³, Andrew Beam, PhD^{1,*}, and John Torous, MD, MBI^{2,8,*}**

¹Department of Epidemiology, Harvard T.H. Chan School of Public Health, MA, USA

² Department of Psychiatry, Beth Israel Deaconess Medical Center, MA, USA

³Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK

⁴Department of Computer Science, The University of Manchester, Manchester, UK

⁵School of Biomedical Informatics, University of Texas Health Science at Houston, Houston, USA

⁶Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia

⁷Center for Precision Psychiatry, Massachusetts General Hospital, MA, USA

⁸Department of Psychiatry, Harvard Medical School, MA, USA

⁹Data Science and Analytics Thrust, Hong Kong University of Science and Technology, Guangzhou, China

¹⁰Division of Psychotic Disorders, McLean Hospital, MA, USA

¹¹The Alan Turing Institute, London, UK

*Corresponding authors

ABSTRACT

The integration of large language models (LLMs) in mental health care is an emerging field. There is a need to systematically review the application outcomes and delineate the advantages and limitations in clinical settings.

This review aims to provide a comprehensive overview of the use of LLMs in mental health care, assessing their efficacy, challenges, and potential for future applications.

A systematic search was conducted across multiple databases including PubMed, Web of Science, Google Scholar, arXiv, medRxiv, and PsyArXiv in November 2023. All forms of original research, peer-reviewed or not, published or disseminated between October 1, 2019, and December 2, 2023, are included without language restrictions if they used LLMs developed after T5 and directly addressed research questions in mental health care settings.

From an initial pool of 313 articles, 34 met the inclusion criteria based on their relevance to LLM application in mental health care and the robustness of reported outcomes. Diverse applications of LLMs in mental health care are identified, including diagnosis, therapy, patient engagement enhancement, etc. Key challenges include data availability and reliability, nuanced handling of mental states, and effective evaluation methods. Despite successes in accuracy and accessibility improvement, gaps in clinical applicability and ethical considerations were evident, pointing to the need for robust data, standardized evaluations, and interdisciplinary collaboration.

LLMs hold substantial promise for enhancing mental health care. For their full potential to be realized, emphasis must be placed on developing robust datasets, development and evaluation frameworks, ethical guidelines, and interdisciplinary collaborations to address current limitations.

1 Introduction

Mental health is an important component of modern public health. Statistics from the US National Institute of Mental Health (NIMH) show that 22.8% of U.S. adults experienced mental illness of any type in 2021.^{1,2} Globally, mental health disorders account for 30% of the non-fatal disease burden, highlighting their prevalence as a major cause of disability as published by the World Health Organization (WHO).³ Moreover, the WHO estimates that depression and anxiety disorders cost the global economy \$1 trillion per year in lost productivity.⁴ These data emphasize the critical need for innovations in the prevention and management of mental health conditions, both to alleviate individual suffering and to reduce the substantial societal and economic burdens associated with mental health issues.

A large portion of mental health issues and their management take place within the realm of natural language.^{5,6} This encompasses a range of elements, including the evaluation of symptoms and signs associated with mental health disorders and various forms of interventions, such as talk therapies.^{5,6} In this context, the wealth of information embedded in recorded textual expressions and interactions serves as a crucial resource, enhancing research and practical applications in mental health.

Accordingly, Natural Language Processing (NLP), a branch of computer science that enables computers to process unstructured text (natural language) information in meaningful ways, has shown promise as a tool to aid mental health-related tasks. This includes identifying specific mental conditions,⁷ building emotion-support chatbots,⁸ and supporting interventions.⁹ These tasks utilize data from various sources, such as clinical data^{10,11} and social media data.¹²⁻¹⁵

Large Language Models (LLMs), one of the most recent advances in NLP, have further expanded the potential for innovative mental health care.^{16,17} As a type of Artificial Intelligence (AI) that understands and generates human-like and fluent texts, LLMs offer many promising applications for mental health. For example, their ability to efficiently retrieve and summarize information from large volumes of unstructured text contained in diverse sources (e.g., electronic health records, mobile device interactions, social media platforms,¹⁸ etc.) can provide insights into patient behaviors and experiences for clinicians. This may help with early intervention strategies and tailored treatment plans. In addition, through conversational interactions, they have the potential to offer a relatable and accessible means for individuals to articulate their emotional states and personal experiences.¹⁹ This aids in both user self-expression and also potentially enhances clinical therapeutic processes if they can be successfully validated and integrated into the clinical workflows.

While the interest in deploying LLMs to mental health care is increasing, to our knowledge, there are no comprehensive reviews on this topic. This study aims to bridge this gap, presenting the first comprehensive review of LLM applications in mental health care. To differentiate LLMs from smaller-sized language models, we follow an established review for general domain LLMs²⁰ to focus exclusively on model architectures developed since the introduction of T5²¹ in 2019. This period, spanning the last four years, marks a significant phase in the evolution of LLMs, offering a rich context to explore their influence and advancements in the field of mental health care. Through this study, we aim to explore:

1. Datasets, models, and training techniques.
2. Mental health applications, conditions, and validation measures.
3. Ethical, privacy, safety, and other challenges.
4. Gaps between tools currently available and clinical practicality.

1.1 Background on Natural Language Processing and Large Language Models

Natural Language Processing (NLP) emerged in the 1950s with machine translation systems.²² Initially, the focus was on rule-based systems that required explicit coding of language rules,^{23,24} which limited their scope and adaptability to handle the complexity and variability of natural languages. Statistical Language Models arose in the 1990s and they modeled word sequences using statistical methods.²⁵ These models were typically hindered by issues such as data sparsity. Neural language models, proposed in the 2000s, utilized neural network-inspired architectures (e.g., multi-layer perceptrons and recurrent neural networks) to refine the prediction of word sequences.²⁶ This era witnessed the emergence of distributed word representations, notably through the word2vec algorithm,²⁷ which shifted the focus towards more effective feature learning for varied NLP tasks. Pre-trained language models like ELMo,²⁸ BERT,²⁹ T5,²¹ and GPT³⁰ were introduced in the late 2010s. These models leveraged advanced neural architectures, including

bidirectional LSTM³¹ and Transformers,³² revolutionizing NLP with their ‘pre-training and fine-tuning’ paradigm. Starting from around 2020, NLP gradually moved toward the ‘large language model (LLM)’ era. LLMs can be analogized to sophisticated, pre-built statistical models but are massive in their scales and can offer much more in terms of versatility and depth.

Typical LLMs include GPT-3,³³ GPT-3.5,³⁴ GPT-4³⁵ (ChatGPT is an application developed based on GPT-3.5 and GPT-4), LLaMA,³⁶ LLaMA-2³⁷ and PaLM.³⁸ Thanks to their impressive performance in understanding and generating fluent human language, natural language “prompts” are proposed to interact with and instruct these LLMs with their internal weights (i.e. model parameters) fixed.³⁹ This strategy is generally referred to as “prompt-tuning,” a paradigm distinct from “fine-tuning,” where adjusting model weights requires training with additional data.

Today, due to differing policies from developers, LLMs are divided into two categories: open-source and non-open-source. Open-source LLMs, e.g., LLaMA, can be deployed in local environments, allowing for customization with specific datasets. This local deployment ensures data privacy is maintained, as it obviates the necessity for data to be uploaded to cloud-based services, which is crucial in sensitive fields like mental health. In contrast, non-open-source LLMs like GPT-4 and PaLM do not share their model architectures or training details. These models are typically hosted by third-party organizations and can be accessed on their websites or through Application Programming Interfaces (APIs). While offering robust language processing capabilities, they require different data privacy and control considerations.

1.2 Potential Applications in Mental Health Care

LLMs, effective in processing large volumes of natural language texts and simulating human-like interactions, have the potential to assist with a wide range of tasks in mental health care. These include interpreting and predicting behavioral patterns, identifying psychological stressors, and providing emotional support. With adequate regulatory, ethical, and privacy safeguards, they may also serve as supplementary roles to support clinically-oriented tasks. This could involve aiding in preliminary assessments for diagnostic processes, facilitating the management of psychiatric disorders, and enhancing therapeutic interventions, such as by helping to improve treatment adherence.

2 Methods

2.1 Search Strategy and Selection Criteria

In conducting this review, our approach adhered to the Preferred Reporting Items for Systematic Reviews and MetaAnalysis (PRISMA) guidelines (2020 version)⁴⁰ to ensure a transparent and rigorous process. The PRISMA diagram is shown in Figure 2.

We established specific criteria for paper inclusion: (1) The study must employ at least one LLM published after T5, as defined in,²⁰ and (2) it should directly address research questions of these models in mental health care settings. This model scope was chosen to ensure that our review concentrates on the latest advancements in the field, deliberately excluding early pre-trained language models such as BERT and GPT-2, which have already received extensive coverage in prior literature.

Our initial search on PubMed highlighted a limited number of published studies in this domain. Recognizing the fast-evolving nature of LLMs, we extended our scope beyond traditional peer-reviewed literature. This inclusion of both peer-reviewed and non-peer-reviewed studies, such as preprints, is essential to capture the latest advancements in LLMs, which are rapidly advancing and often reported in non-traditional formats. We considered all forms of original research, including long and short articles and case reports, published or disseminated between October 1, 2019, and Dec 2, 2023, without any language restrictions.

We conducted comprehensive searches across several databases and registries, including ArXiv, MedRxiv, the ACM Digital Library, PubMed, Web of Science, and Google Scholar. The search strategy used the logical combination of keywords “Large Language Model” AND “mental OR psychiatry OR psychology.” Whenever possible, we restricted the search scope to titles and abstracts. In databases without this functionality, we extended the search scope to the full text of the papers.

Further details regarding the search process and specific search queries can be found in Appendix 7.1. For the Google Scholar registry, we employed the same set of keywords and reviewed the first 100 studies, ordered by relevance, that met our inclusion criteria.

2.2 Screening with GPT-4

After removing duplicated articles and those without abstracts, we had 281 papers left for initial screening. Recognizing that GPT-4 has been shown to be adept in assisting with article screening and exhibits performance comparable to humans,^{41,42} we used GPT-4 as a secondary reviewer for this process. Prior to its use, we tested different prompts to optimize GPT-4's screening capability (7.1).

Both YH and GPT-4 independently screened the article titles and abstracts to determine their eligibility for inclusion in the study. Each had three response options: 1 for inclusion, 0 for exclusion, and 2 for uncertainty. Any disagreements identified were subsequently resolved through discussion with the rest of the review team (YH, KY, ZL, and FL). To quantitatively evaluate the agreement level between the human reviewer (YH) and the AI (GPT-4), Cohen's Kappa⁴³ was calculated to be approximately 0.9024. This high score indicates a strong agreement. It was observed that GPT-4 tended to be more inclusive, often categorizing more articles as related to mental health care than the human reviewer. Nonetheless, implementing the option of uncertainty, while slightly reducing the kappa score, was crucial in ensuring comprehensive inclusion of relevant articles, thereby balancing thoroughness with precision.

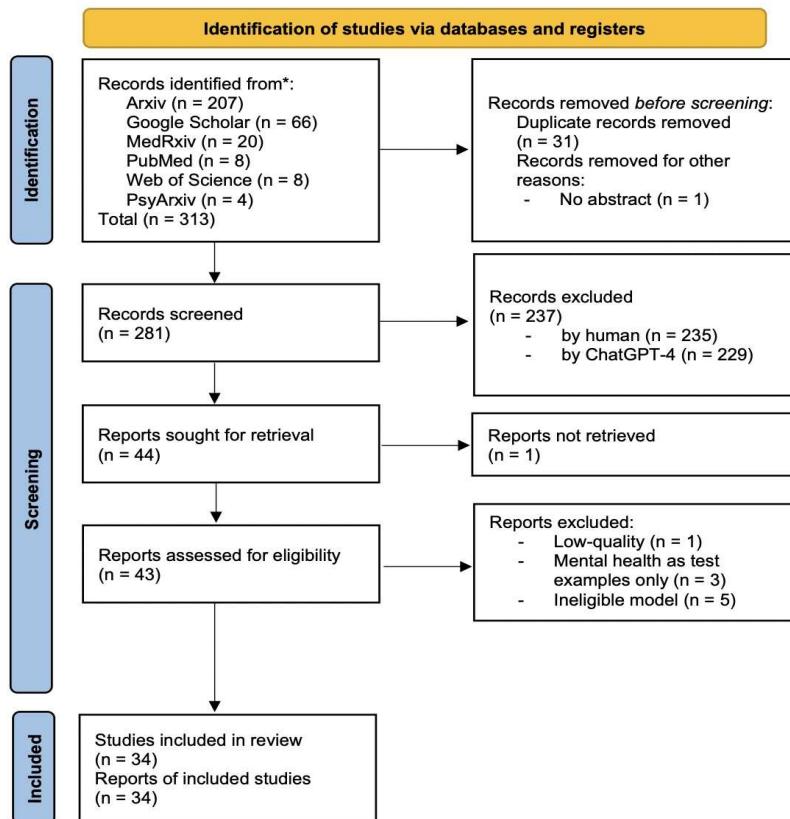


Figure 1. A PRISMA figure for the systematic search.

3 Findings

Forty-three articles remained for full-text review. YH, KY, ZL, and FL reviewed all articles and removed nine articles due to low quality ($n = 1$), only having mental health as test cases ($n = 3$), or ineligible model size ($n = 5$). During the review, studies are categorized based on their research questions and intentions:

1. Dataset and Benchmark: These studies use standardized tests or benchmark datasets to evaluate and compare the performance of different methods, systems, or models under controlled conditions. Some of them curate or compile new datasets.
2. Methods Development: These studies propose new LLMs or introduce methods, such as fine-tuning and prompting, to adapt and improve existing LLMs for mental health care.
3. Specific Applications: These studies employ LLMs to evaluate their performance in mental health-related tasks in real-world applications. They may include the evaluation of LLMs on certain tasks (inference-only).
4. Ethics, Privacy, and Safety consideration: These studies examine the potential risks, ethical dilemmas, and privacy concerns associated with deploying LLMs in sensitive mental health contexts and propose frameworks or guidelines to mitigate these issues.

Thirty-four articles met our criteria and were included in subsequent analyses. To maintain a clear focus and ensure a thorough analysis to answer the application-focused study questions, we summarized "Dataset and Benchmark" studies separately.

Figure 1 shows the submission/publication time¹ and type of the papers included in the final analysis. Research on LLMs in mental health care emerged in Sep 2022 and has shown a progressive increase in publication volume, with a surge in the latter half of the year, particularly noticeable in October. The majority of these papers are concentrated in the "Prompt-tuning and Application" category, which saw a rise in July. In contrast, "Model Development and Fine-tuning" studies were absent in the initial months, only to experience a marked rise in October. Notably, there have been only two papers concerning "Dataset and Benchmark", both emerging later in the year. Additionally, there is only one study¹⁹ that addresses ethics, privacy, and other challenges, which was published mid-year.

¹ Papers with pre-prints have submission time entries.

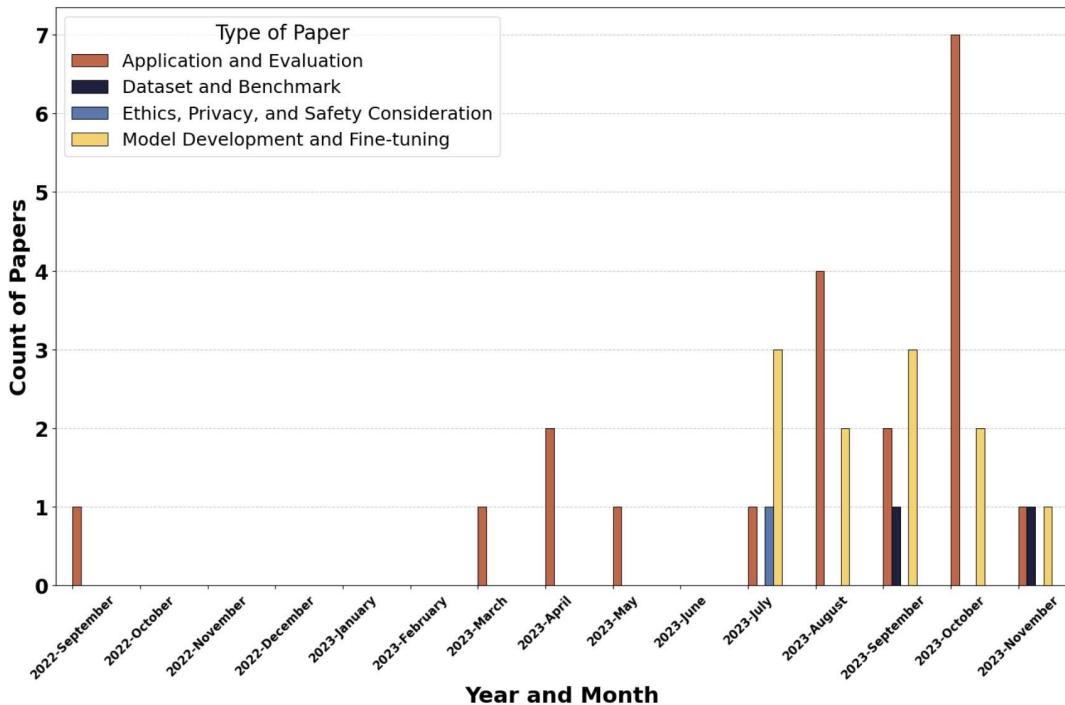


Figure 2. Submission/Publication time and type distribution of qualified studies.

3.1 Areas of Application and Mental Health Conditions Involved

During our review, we observed a strong correlation between the scope of each study and the dataset employed. This section offers a high-level overview of the application areas across these studies, along with the mental health conditions they intended to address. In Table 2, we present a comprehensive summary of all the datasets utilized by these studies, detailing their intended purposes.

3.1.1 Application Areas

The studies roughly cover three main areas. The first area is the deployment of conversational agents,^{19,44–58} with 16 studies aiming to improve the models' capacity to produce empathetic and context-aware responses. These agents are typically not tailored to specific mental disorders, instead, they cater to a broad spectrum of mental health needs. Among these studies, 12 are designed to interact directly with individuals seeking mental health support through various platforms such as personal digital companions,¹⁹ on-demand online counseling,^{45–47,49–51,53,54,56–58} and emotional support.⁵² Some studies extend to specialized applications like couple therapy.^{56,57} The remaining four studies are developed to assist care providers by providing specific recommendations and analyses.^{46,48} This helps to mitigate the issue of provider shortage while enhancing the delivery of care. It's noteworthy that some studies have taken a different approach by developing screening tools. These tools are designed to identify urgent cases⁴⁶ and distinguish between emotional experiences in individuals with different personality disorder subtypes.⁵⁵

The second area focuses on resource enrichment. Studies in this category^{59–64} have explored LLM-generated explanations for multi-task analysis^{59,60} and the development of educational content.^{63,64} The latter includes creating hypothetical case vignettes relevant to social psychiatry⁶³ and personalized psychoeducation material.⁶⁴ The potential of synthetic data provided by LLMs is also being tapped into as a method for augmenting data and fine-tuning downstream, particularly in augmenting clinical questionnaires to enrich the symptomatology of depression cases.⁶²

The third area involves the use of LLMs as classification models for detailed diagnosis.^{59,60,65–69} This often involves binary classification to detect the presence or absence of a single condition providing given contexts,^{60,66,68,69} and multiclass classification for more granular information about a condition, including severity levels and subtypes.^{59,60,65–67} Section 3.1.2 provides more information about the mental health conditions explored in these studies. Some examples

of multi-class classification include depression severity prediction (minimal, mild, moderate, and severe according to DSM-5⁷⁰), subtypes of suicide (supportive, indicator, ideation, behavior, and attempt according to the Columbia Suicide Severity Rating Scale (C-SSRS)⁷¹), and identification of sources of stress span (school, finance, family, social relation, etc., based on the classifications established in SAD.⁷²

3.1.2 Mental Health Problems Involved

Out of the 34 articles reviewed, 23 focused on specific mental health problems, while the remaining articles explored general mental health knowledge or dialogues without specific conditions. Of the 23 articles examining specific mental health problems, 14 focused on single mental health problems, and nine involved multiple.

A wide range of mental health conditions has been studied, but most of them are limited by the availability of open-source datasets. The most frequently studied conditions include stress,^{46,47,52,57,59,60,65,66,73} suicide,^{48,59,60,65,66,68,69} and depression.^{49,59,60,62,65,66,74-77} Other conditions include anxiety,^{49,60} bipolar disorder,⁶⁰ Post-traumatic demoralization syndrome (PTDS),^{60,78} Autism spectrum disorder (ASD),⁵³ and personality disorder,⁵⁵ loneliness,^{60,68} cognitive distortion,⁶¹ In addition to these conditions, life circumstances that can profoundly affect mental health are also studied, covering relationship problems,^{56,57,68} sleep problems, sexual violence, custody issues, and bullying.⁶⁸

3.2 Models and Training Techniques

The suitability and effectiveness of a pre-trained model for mental health care are fundamentally influenced by its training data, size, and open-source status. These factors collectively determine the model's potential biases or lack of representativeness for certain tasks and populations. In Table 1, we provide a summary of existing LLMs developed for mental health care. This summary includes details about underlying base models, the scale indicated by the number of parameters, the transparency of the base model training data, the strategies employed during training, and information regarding accessibility if they are open-source.

3.2.1 Distribution of Model Architectures and Prompting Algorithms

A significant amount of existing work has been dedicated to directly prompting GPT-3.5 or GPT-4 for mental health applications without training. Some examples include depression detection,^{75,76} suicide detection,⁶⁸ cognitive distortion detection,⁶¹ and relationship counseling.^{56,57} In these studies, LLMs function as intelligent chatbots, engaging with users to provide a range of mental health services, including analysis,^{19,59,66} prediction⁶⁵ and support.^{46,48,74}

To enhance their effectiveness, methods like few-shot prompting, which presents the LLMs with a small number of task demonstrations before requiring them to perform a task, and chain-of-thought (CoT) prompting,⁸⁴ which prompt the model to generate intermediate steps or paths of reasoning when dealing with problems, are employed. Drawing from the success of CoT in NLP, novel approaches such as chain-of-empathy prompting,⁵¹ which incorporates the insights from psychotherapy (i.e., the therapists' reasoning process) to prompt the LLMs to generate the cognitive reasoning of human emotion, and diagnosis-of-thought prompting,⁶¹ which prompt the LLMs to make a decision using three diagnosis stages (i.e., subjectivity assessment, contrastive reasoning, and schema analysis), have been proposed to further improve LLM performance in the mental health domain.

Table 1. Summary of existing LLMs for mental health, in terms of their model development, the number of parameters, the availability of training data, the training strategy, and the available public links. "B" denotes billion. **TFP** and **IFT** stand for "tuning-free prompting" and "instruction fine-tuning", respectively.

Table 1. Summary of existing LLMs for mental health, in terms of their model development, the number of parameters, the availability of training data, the training strategy, and the available public links. "B" denotes billion. **TFP** and **IFT** stand for "tuning-free prompting" and "instruction fine-tuning", respectively.

Models	Base LLM	# Params	Base Model Training Data	Training Strategy	Availability
DSC ⁴⁴	GPT-3.5 ³⁴	Unknown	Unknown	TFP	✗
Replika ¹⁹	GPT-3 ³³	175B	Partially known	TFP	✓
Mental-LLM ⁶⁵	Alpaca ⁷⁹ /FLAN-T5 ²¹	7B/11B	Partially known/Fully known	IFT	✓
MentaLLaMA ^{59,60}	LLaMA-2 ³⁷	33B	Partially known	IFT	✓
MindShift ⁷³	GPT-3.5	Unknown	Unknown	TFP	✗
Psy-LLM ⁴⁶	Pangu ⁸⁰	350M ¹	Partially known ²	TFP	✗
MindfulDiary ⁷⁴	GPT-4 ³⁵	Unknown	Unknown	TFP	✗
ChatCounselor ⁴⁷	LLaMA ³⁶	7B	Partially known	IFT	✓
LLM-Counselors ⁴⁸	GPT-3.5	Unknown	Unknown	TFP	✗
Task-Adaptive Tokenization ⁵⁰	LLaMA	7B	Partially known	TFP	✓
Chain of Empathy Prompting ⁵¹	GPT-3.5	Unknown	Unknown	TFP	✗
Diagnosis of Thought Prompting ⁶¹	GPT-3.5/GPT-4	Unknown/Unknown	Unknown	TFP	✓
ExTES-LLaMA ⁵²	LLaMA	7B	Public	IFT	✓
Explainable Depression Detection ⁷⁵	Vicuna ⁸¹ /GPT-3.5	13B/Unknown	Unknown	TFP	✗
BBMHR ⁵⁴	BlenderBot-BST ⁸²	2.7B	Public	IFT	✓
ChatCBPTSD ⁷⁸	GPT-3.5	Unknown	Unknown	TFP	✓
Female Firearm Suicides ⁶⁸	FLAN-UL2 ⁸³	20B	Partially known	TFP	✗
Plasticity ⁵⁵	GPT-3.5	Unknown	Unknown	TFP	✗
Depression Detection ⁷⁶	GPT-3	Unknown	Unknown	TFP	✗
Psychiatric ⁷⁷	GPT-4	Unknown	Unknown	TFP	✗
MindWatch ⁶⁹	GPT-3.5	Unknown	Unknown	TFP	✗
Social psychiatry ⁶³	GPT-3.5	Unknown	Unknown	TFP	✗
Relationship counseling ⁵⁶	GPT-3.5	Unknown	Unknown	TFP	✗
Relationship expert ⁵⁷	GPT-4	Unknown	Unknown	TFP	✗
Mindfulness ⁶⁴	GPT-3	Unknown	Unknown	TFP	✗
Mental Wellbeing ⁵⁸	GPT-3	Unknown	Unknown	TFP	✗

¹ This study uses the 350M parameter version of PanGu, which also has versions with 2.6B, 13.1B, and 207B parameters.

² PanGu claims that its training data is open source, but we have not been able to find it through searches.

Another stream of research focuses on the further training or fine-tuning of general LLMs using mental health specific texts. This approach aims to inject mental health knowledge into existing base LLMs, leading to more relevant and accurate analyses and support. Notable examples include MentaLLaMA^{59,60} and Mental-LLM,⁶⁵ which fine-tuned the LLaMA-2 model and the Alpaca⁷⁹/FLAN-T5⁸⁵ model, respectively, using social media data for enhanced mental health predictions. Similarly, ChatCounselor⁴⁷ leveraged the Psych8k dataset,⁴⁷ comprising real interactions between clients and psychologists, to fine-tune the LLaMA model. Additionally, ExTES-LLaMA⁵² employed the LLaMA model, fine-tuned on emotional support dialogues.

While some studies used open-source LLMs that provide more transparency about their sizes and training data sources, the complete specifics of any base model's training data remain undisclosed. For example, although overarching details about the training data for LLaMA and T5 are available, the detailed documents or sources used in these datasets are not publicly shared.

3.2.2 Fine-Tuning Techniques

Acknowledging the high cost and extensive time required to train LLMs from scratch,^{86–89} existing studies have uniformly adopted fine-tuning existing models on mental health data. This approach allows the models to acquire expert domain knowledge and evolve into dedicated mental health-focused LLMs.

All the studies that involved fine-tuning adopted the instruction fine-tuning (IFT) technique. IFT is a type of fine-tuning where a model is fine-tuned for an instruction-following task that involves instructing a model to perform another task. In contrast, in classical types of supervised fine-tuning, the model does not have access to instructions but is directly tuned to perform a single downstream task. IFT allows domain knowledge to be injected into LLMs while improving LLM's capability to follow human instructions. For example, giving a conversation between clients and psychologist, ChatCounselor⁴⁷ used the prompts: "*Your task is to identify the patient and counsellor in the conversation. Summarize the conversation into only one round of conversation, one query or description by the patient and one feedback by the counsellor*", to prompt GPT-4³⁵ to generate instruction-input-output triples, in which the instruction

could be "If you are a counsellor, please answer the questions based on the description of the patient." The input could be the query/question from the patient and output could be the feedback/answer from the model.

Table 2. Overview of Mental Health-Related Datasets. ⁺Dataset curated or compiled in used in "Dataset and Benchmark" studies. "K" denotes a thousand. Information not provided is marked by N/P. N/A stands for not applicable.

Name	Intended Task	Mental Condition	Data Source	Sample size	Constructed by	Reviewer experience	License
Published Dataset							
PsyQA ⁹⁰	Counseling	N/A	Mental health service platforms	22,346 questions with 56,063 answers	Human	Trained laymen	N/P
SMHD ⁹¹	Mental health condition detection	ADHD ¹ , anxiety disorders, ASD ² , bipolar disorders, BPD ³ , depression, eating disorders, OCD ⁴ , PTSD ⁵ , schizophrenia, SAD ⁶	Reddit	4,399K posts	Weak supervision	N/P	CC-BY-4.0
D4 ⁹²	Depression diagnosis-oriented dialogues	Depression	Human-machine dialogue	1,339 dialogues	Human	Experts	Research only
Escon ⁹³	Emotional support conversation	N/A	Crowdworker chatting simulation	1,053 dialogues (31,410 utterances)	Human	Trained laymen	CC-BY-NC-4.0
DialogueSafety ⁹⁴	Dialogue safety	N/A	Online counseling platforms	7,935 multi-turn dialogues	Human	Experts	MIT
Dreaddit ⁹⁵	Stress detection	N/A	Reddit	3,551 post segments	Human	Trained laymen	Publicly Available
DepSeverity ⁹⁶	Depression severity prediction	Depression	Reddit	3,551 posts	Human	N/P	Publicly Available
SDCNL ⁹⁷	Suicide ideation detection	N/A	Reddit	1,895 posts	Weak supervision	N/P	N/P
CSSRS-Suicide ⁹⁸	Suicide risk prediction	N/A	Reddit	500 users	Human	Experts	CC-BY 4.0
Depression_Reddit ⁹⁹	Mental health condition detection	Depression	Reddit	800 posts	Weak supervision	N/P	Publicly available
CLPsych15 ¹⁰⁰	Mental health condition detection	Depression & PTSD	Reddit	1,746 users	Human	N/P	Research only
Ji et al. ¹⁰¹	Suicide ideation & mental disorder detection	Depression & PTSD	Twitter	866 users & 59,212 posts	Weak supervision	N/P	Research only
SAD ⁷²	Stress cause detection	N/A	SMS messages	6,847 SMS-like sentences	Human	Laymen	CC-BY-4.0
CAMS ¹⁰²	Depression & suicide cause detection	Depression	Reddit	5,051 posts	Human	Trained students	Publicly available
EMPATHETICDIALOGUES ¹⁰³	Empathetic dialogue generation	N/A	Crowdworker chatting simulation	~25K multi-turn dialogues	Human	Laymen	CC-BY-NC-4.0
MultIW ¹⁰⁴	Wellness dimensions detection	N/A	Reddit	3,227 posts	Human	Trained students & experts	Research only
IRF ¹⁰⁵	Interpersonal risk factors detection	N/A	Reddit	3,523 cases	Human	Trained students	Research only
GoEmotions ¹⁰⁶	Emotion classification	N/A	Reddit	58,009 cases	Human	Laymen	Apache 2.0
MHP Reddit ¹⁰⁷	Counseling	N/A	Reddit, TalkLife ¹⁰⁹	9,501 question-answer pairs	Human	N/P	MIT
EPITOME ¹⁰⁸	Empathetic response generation	N/A	Reddit, TalkLife ¹⁰⁹	10K interactions on empathy	Human	Trained laymen	Publicly available
Therapist Q&A ¹¹⁰	Cognitive distortions detection	N/A	Kaggle	2,531 patient speech samples	Human	N/P	N/P
ExTES ¹²	Emotional support dialogue	N/A	ChatGPT	11,177 dialogues	LLM generation	Trained laymen	Publicly available
BDI-Sen ¹¹¹	Mental health condition detection	Depression	Reddit	4,973 sentences	Human	Experts	Research only
PsySym ¹¹²	Mental health condition detection	Depression, anxiety, ADHD, bipolar disorder, OCD, PTSD, eating disorder	Reddit	83,779 sentences	Human	Trained laymen & experts	N/P
NVDRS ¹¹³	Suicide behavior detection	N/A	Mortality & incident reports	1,462 reports	Human	Experts	N/P
EATD-Corpus ¹¹⁴	Mental health condition detection	Depression	Interview transcripts	162 interview audio & transcripts	Weak supervision	N/P	Publicly available
ANEW ¹¹⁵	Lexicon sentiment detection	N/A	English vocabulary	1K words	Human	N/P	N/P
Wariner et al. ¹¹⁶	Lexicon sentiment detection	N/A	English vocabulary	13,915 lemmas	Human	Trained laymen	N/P
NRC-VAD ¹¹⁷	Lexicon sentiment detection	N/A	English vocabulary	20K words	Human	Laymen	N/P
Self-constructed Dataset							
Ma et al. ¹⁹	User comment analysis	N/A	Reddit	2,917 posts	Human	N/P	N/P
Psych8k ⁴⁷	Mental health support	N/A	Real-life counseling	260 interviews with 8,187 instruction pairs	N/P	N/P	Publicly available
Yao et al. ⁴⁹	Counseling	Postpartum Mood & Anxiety Disorders	Helpline	7,014 conversations with 65,062 messages	Weak supervision	N/P	N/P
Qi et al. ^{67 +}	Suicide risk prediction & cognitive distortion detection	N/A	Weibo	2,159 comments	Human	Experts	Publicly available
Bhaumik et al. ⁶⁹	Suicide ideation detection & personalized psychoeducation	N/A	Reddit	232K posts	N/P	N/P	N/P
Kumar et al. ⁵⁸	Counseling	N/A	Crowdworker chatting simulation	945 survey responses	Human	Laymen	N/P

¹ Attention-deficit/hyperactivity disorder.

² Borderline personality disorder.

³ Obsessive-compulsive disorder.

⁴ Post-traumatic stress disorder.

⁵ Seasonal affective disorder.

3.3 Dataset Characteristics

The integrity of the data profoundly influences outcomes in mental health care studies, particularly in terms of representativeness, quality, and potential bias. Recognizing the sources and characteristics of datasets is fundamental in ensuring accurate and fair research findings. We reviewed datasets used in the studies and reported their associated tasks, data sources, sample sizes, annotation methods if the dataset is annotated, human reviewer experience, and licenses in Table 2.

Among the 34 studies we reviewed, 36 datasets were identified. These datasets contain diverse data applicable to a wide range of intended mental health care tasks. Most datasets are dedicated to detection or classification tasks. This category includes tasks such as mental health condition detection, mainly focusing on depression and post-traumatic stress disorder (PTSD),^{95,96,99,100,111} cause of stress detection,⁷² and interpersonal risk factor prediction.¹⁰⁴ Another major group focuses on text-generation tasks, such as simulating counseling sessions,^{90,107} responding to medical queries and

generating empathetic dialogue.^{103,108} The remainder concerns more specialized applications, such as analysis of user discussions of emotion-support LLMs¹¹⁸ and dialogue safety exploration.⁹⁴

Sources of these datasets are imbalanced, with 20 datasets (55.6%) collected from social media platforms, such as Reddit, Twitter, and Weibo. Eleven datasets relied on more controlled venues, such as mental health counseling or interview dialogues simulated by clinicians⁹⁴ and collected from real-life scenarios.^{47,90,92} suicide behavior detection corpus collected from mortality and incident reports,¹¹³ and helpline conversations.⁴⁹ Less frequently seen sources include LLM-synthesized data,⁵² existing sentiment lexicons,^{115–117}, and crowdworker simulated conversation.^{58,93,103}

Dataset sizes and units vary with sources and annotation methods. Those drawn from social media and created via web crawling with the need for annotation typically have larger sizes. In contrast, those composed of expert content contain fewer samples. For example, SMHD⁹¹ contains 4,399K posts collected from Reddit, while D4⁹² only contains 1,339 dialogues reviewed by experts.¹¹⁸

Most datasets were crafted by manual work with either collection or annotation. Six used weak supervision, which uses heuristics or helper models to collect or annotate data and can thus achieve much larger scales. For example, SWMH⁹¹ obtained annotations through subreddit titles and was able to include over 50K samples. One study⁵² utilized LLM - ChatGPT in their case, to generate data, and had trained laymen to review them.

Most datasets involved human review of the content regardless of their construction methods. Eight datasets explicitly mentioned the involvement of experts in the review process. Fourteen datasets did not specify the experience level of the annotators. The remaining datasets were reviewed by laymen or trained laymen, including those who might be crowdworkers such as Amazon Mechanical Turk (MTurk) workers. Most papers relied on publicly available datasets. Six datasets^{19,47,49,67,69} were self-constructed. Only one of them,⁴⁹ curated in a "Dataset and Benchmark" study, has been released for public use. Most of these datasets are released under licenses that restrict their use to non-commercial purposes. Details and implications of these restrictions and licensing agreements are provided in Appendix 7.1.

3.4 Validation measures and metrics

To ensure robust and unbiased assessments of LLMs, it's crucial to adopt appropriate validation measures and metrics. Our analysis categorized them into two types: automated evaluation and human assessment. Table 3 details metrics used in automated evaluation and attributes evaluated by human assessment. To reflect the appropriateness of these metrics,

we also categorized them based on their evaluation focus: language proficiency and mental health applicability.

For mental health applicability, different variants of F1 scores are the most predominant metric adopted across studies.^{49,51,59–61,66,75,76,78} Accuracy, as another fundamental metric, is also extensively employed.^{59–61,65,66,69} Recall (also known as sensitivity)^{49,51,59,69,78} and precision^{49,51,62,69,75} are also widely used, and frequently reported with F1 and accuracy. Diagnosis-focused studies, while also heavily relying on F1 and Accuracy, tend to employ additional metrics such as Area Under the Receiver Operating Characteristic (AUROC),^{69,78} and Specificity⁷⁸ for a holistic understanding of LLM diagnostic validity.

In evaluating language proficiency, BLEU,^{44,50,54,75,120} ROUGE,^{46,50,52,59,75} Distinct-N,^{44,46,52} and METEOR⁵² are widely used to assess alignment with human-like language, diversity of expression, and overall quality of the generated text. Advanced metrics such as GPT3-Score,⁵⁹ BARTScore,⁵⁹ and BERT-Score^{49,50,54} are tailored to evaluate semantic coherence and relevance of text in specific contexts, addressing the demand for nuanced understanding. Perplexity^{46,52} quantifies model predictability and text naturalness, while Extrema and Vector Extrema⁵² analyze the diversity and distribution in the embedding space to reflect the model's linguistic creativity and depth. The adoption of these traditional language evaluation metrics appears to be driven by a lack of automated, efficient, and easily understandable metrics designed for evaluating the quality of free-text generation of LLMs for mental health care. Human evaluation is thus frequently used across studies.

Among the 34 articles reviewed, 19 employed a combination of human and automated evaluations, 5 exclusively used human evaluations, and 10 relied solely on automated methods. However, there is no single unified evaluation framework. A few studies^{44,45,51,52,54} adopted or adapted published evaluation standards or attributes discussed in previous studies,^{94,108,121–123} yet none of these frameworks are widely adopted. Some of these studies and most other studies added or designed from scratch their evaluation frameworks and selected attributes from interest from their

own judgment.^{46–48,50,52–54,60} A wide range of properties/attributes are proposed to be evaluated, and the frequently overlapped ones are empathy,^{44,46,49,51,53,54,56,57} relevance,^{44,46,49,52–54,56,57,59} fluency,^{44,46,50,52,56,57,59} understanding,^{47,52,53,56,57} and helpfulness.^{44,46,52,54,56,57} Specifically, in intervention applications, studies often measure user engagement with interventions and technology adoption. The rate of intervention acceptance,^{58,64,73} frequency of smartphone usage,⁷³ adherence to digital journaling protocols,⁷⁴ willingness to use the intervention again,^{57,58,64} and intervention adherence rates^{57,58,64,73,74} are observed to evaluate the effectiveness of LLMs in these studies.

We noticed some attributes, despite sharing names, differ in definition across studies. For instance, 'Informativeness' varies in context – in some studies,^{44,47} it relates to the richness of LLM responses, while in others,⁵² it measures the extent of individual elaboration on emotional distress. Expert-led assessments^{48,53,57,63,77} prominently involve direct analysis of model outputs or expert questionnaire scoring. Reliability metrics like Cohen's Weighted Kappa,^{44,49} Krippendorff's Alpha,⁴⁸ Fleiss' Kappa,^{52,59} and inter-annotator agreement^{54,56} have been essential for validating study instruments. The study designs vary, with the number of reviewers ranging from as few as three⁷⁵ to as many as fifty.⁵²

Table 3. Explanation of Metrics Used to Evaluate LLM-Generated Content for Mental Health Care. Metrics are ordered alphabetically for easy indexing. Only model-oriented metrics are shown for simplicity.

Automated Metrics			
Metric	Associated studies	Metric explanation	Target of Evaluation
Accuracy	59–61,65,66,69 69,78	Proportion of correct predictions in classification tasks.	Mental Health
AUROC	49,50,54,59	Measures the model's performance in distinguishing between classes.	Mental Health
BERT-Score ¹²⁴ /BARTScore ¹²⁵ /GPT3-Score ¹²⁶	44,50,54,75,120 44,46,52	Evaluate the quality of text generation against reference texts using respective models.	Language
BLEU-N ¹²⁷	49,51,59–61,66,75,76,78	Measures the overlap of n-grams between the generated text and reference texts.	Language
Distinct-N	52	Measures the diversity of language by calculating unique n-grams.	Language
F1 Score	49,51,59–61,66,75,76,78	A combination of precision and recall into a single metric, balancing the two.	Mental Health
Extrema/Vector Extrema	52	Analyze the diversity and distribution in the embedding space of the texts.	Language
METEOR ¹²⁸	52	Aligns and compares similar words between the generated text and reference translations.	Language
Perplexity	46,52	Measures how well a probability distribution predicts a sample. Lower is better.	Language
Precision	49,51,62,69,75	The proportion of positive identifications that were actually correct.	Mental Health
Recall (sensitivity)	49,51,59,69,78	The proportion of actual positives that were identified correctly.	Mental Health
ROUGE-N/L ¹²⁹	46,50,52,59,75 78	Measures the overlap of n-grams or longest common subsequence between texts.	Language
Specificity	78	The proportion of true negatives that are correctly identified.	Mental Health
Attributes Evaluated in Human Assessment			
Attribute	Associated studies	Attribute definition	Target of Evaluation
Consistency	46,52	Ensuring the chatbot's responses are aligned with its defined role without contradictions.	Language
Empathy	44,46,49,51,53,54,56,57	The degree to which the chatbot shows warmth, sympathy, and concern towards the user's situation.	Mental Health
Engagement	54	The chatbot's ability to maintain prolonged and engaging conversations with the user.	Mental Health
Exploration	45,51,56	Indicators of the user's attempts to investigate deeper into the topics brought up in the conversation.	Mental Health
Guidance	47,57	The chatbot's effectiveness in giving clear, actionable instructions and guidance.	Mental Health
Helpfulness	46,52,54,56	The extent to which the chatbot contributes to reducing the user's emotional distress and improving mood.	Mental Health
Informativeness	44,47	The extent to which the chatbot provides useful and relevant suggestions for the user's issues.	Mental Health
Interpretation	45,51	The user's effort to restate or reinterpret the problems presented, as indicated by linguistic markers.	Mental Health
Professionalism	50,60	Evaluating the chatbot's responses for their logical and psychological accuracy.	Mental Health
Reflectiveness	47,56,57	The chatbot's skill in rephrasing, mirroring user inputs, and demonstrating active listening.	Mental Health
Relevance	44,46,49,52–54,56,57,59	The chatbot's ability to generate content related to the user's specific problems.	Mental Health
Safety	57,120	Adherence to established safety norms and guidelines in the chatbot's interactions.	Mental Health
Understanding	47,52,53,56,57	The chatbot's level of accurately grasping and reflecting the user's experiences and emotions.	Mental Health
Coherence	50,52,56	The chatbot's ability to keep the conversation focused and transition smoothly between topics.	Language

3.5 Ethics, Privacy, Safety, and Other Concerns

Privacy is a major concern across studies, particularly amplified due to the sensitive nature of data often processed by mental health care applications. Several studies underscore this risk of sensitive data exposure and stress the need for robust data protection and ethical standards.^{44,45,62,66} Safety and reliability are identified as foundational requirements, with an emphasis on preventing harmful content generation^{19,49,61,62} and ensuring accurate and appropriate responses.^{45,67,74,120} The literature also draws attention to the essential balance between leveraging LLM's benefits and ensuring safety, with a focus on risk assessment in mental health support^{44,47,51,59} and a continuous pursuit of reliability and consistency.^{48,50,61,120}

An increasing dependency on AI for emotional support or therapeutic guidance is a notable concern,^{19,46,49,73} raising issues about the neglect of real-life interactions and the potential for over-reliance. This is exacerbated by the implications of inaccurate or biased content, which can seriously affect perceptions and decisions in mental health contexts.^{44,54,65,67} Anthropomorphism is pinpointed as an ethical consideration, influencing user expectations and

reliance on AI systems.¹⁹ The intimate nature of mental health discussions amplifies the risk of misunderstandings and misuse,^{60,62,66,67} underscoring the importance of responsible LLM usage.

Technical and performance challenges are also frequently mentioned, ranging from model limitations and generalizability^{44,47,55,120} to memory and context limitations.^{19,50,62,66} These issues impact the reliability and efficacy of AI applications, particularly in complex, real-world settings. Variability in performance^{46,54,59} and the need for robustness and transparency^{51,59,61,63} are areas that require continuous innovation and scrutiny.

The transition to real-world applications brings additional complexity, particularly in mental health where accuracy and sensitivity are crucial. Discussions on long-term effectiveness,^{47,58,73,76} the distinction between controlled and real-world settings,^{45,48,51,62} and addressing the accessibility and digital divide^{46,52,58,63} illustrate the challenges in bridging the gap between LLMs' potential and its practical implementation.

Finally, a few studies noted the need for diverse and large datasets,^{60,62,65,67} specialized training,^{46,59,63,120} and consideration of data annotations.^{62,66,67,76} The necessity of substantial computational resources^{44,50,51,62} and expert involvement^{47,61,62,66} are also highlighted as critical components in advancing the field responsibly.

3.6 Benchmark studies

Benchmark studies are included in this application-focused review to identify areas needing enhancement for objectively assessing the efficacy of various LLMs in mental health care. So far there have been two benchmark studies.^{67,120} Jin et al.¹²⁰ provides a comprehensive evaluation across models including GPT-4, GPT-3.5, Alpaca,⁷⁹ Vicuna,⁸¹ and LLaMA-2,³⁷ on six tasks using data from social media and therapy sessions: diagnosis prediction, sentiment analysis for inferring emotional states, language modeling for understanding conversations, and question answering for therapeutic guidance, along with evaluating therapy session dialogues. In contrast, Qi et al.⁶⁷ have a much more focused scope on classifying cognitive distortions and predicting suicidal risks on Chinese social media data. They evaluated ChatGLM2-6B,¹³⁰ GLM-130B,¹³¹ GPT-3.5, and GPT-4.

4 Discussion

Through an examination of 34 pertinent studies, we synthesized the characteristics, methodologies, datasets, validation measures, application areas, and specific mental health challenges addressed by LLMs. This synthesis aims to serve as a bridge between the computational and mental health communities, detailing insights in an accessible manner. Despite preliminary evidence suggesting the feasibility of LLMs in enhancing mental health care, our results show that notable gaps remain between their current state and actual clinical applicability. Below, we outline three key areas for improvement that could help realize the full potential of LLMs in clinical practice.

Improvements in data quality for LLM development and validation: The predominant method in mental health LLM applications is prompt-tuning favored for its simplicity and conversational interaction style. GPT-3.5 and GPT-4, widely recognized as ChatGPT, are at the forefront of this research. Despite their impressive performance, they occasionally underperform in the complex context of mental health and may yield biased results. As GPT-4 is not modifiable, it is necessary to investigate fine-tuning techniques on open-source LLMs to mitigate these issues.

Current fine-tuning efforts mainly utilize dialogue and social media data, like Reddit postings. While this data can be useful for screening, it falls short of establishing formal clinical diagnoses due to oversimplifications and the diverse nature of clinical symptoms.^{132,133} In addition, they were sourced from a distribution of users that very likely do not represent the general population at risk of the mental conditions in question, and descriptions of compositions of the included users are lacking. Moreover, labels in existing datasets are typically not curated or validated by mental health experts, with substantial variability in their definitions across studies. For example, "depression" and "anxiety" range from everyday mood fluctuations to clinical disorders. Only one study⁶⁵ made a distinction between clinical diagnoses and self-reported mental states. For LLMs to be reliably integrated into clinical workflows, development and validation with high-quality data are crucial. This data should accurately define mental health conditions and provide a solid basis for rigorous assessment.

Enhanced reasoning, empathy capabilities, and evaluation methods: Effective dialogue-based tasks, especially those on mental health care, require advanced reasoning and empathy abilities, allowing for the analysis of statements

provided by the users, and the provision of appropriate feedback. However, the extent to which LLMs' reasoning is a product of their large training datasets or a genuine emergent property remains a subject of debate (i.e., memorization vs. understanding/reasoning). In cases where "reasoning" is based on memorized training data, the models may lack robustness, particularly in handling diverse user dialogues and language use. In addition, although LLMs have shown preliminary capabilities in demonstrating empathy, these results are still inconclusive.

The absence of a unified evaluation framework amplifies this issue. While discriminative tasks with clear right or wrong answers are easy to assess, evaluating free-text generation is challenging due to the absence of a universal standard. Previous studies have used automated metrics like BLEU, ROUGE-L, and Meteor to evaluate language quality, but these fail to capture context. Even advanced metrics like the BERT score and GPT-3 score fall short in capturing the nuances of mental health care dialogues. Most studies have resorted to human evaluations, but the lack of a standardized framework leads to subjective selection of assessment aspects, resulting in varied and potentially incomparable results. This methodology discrepancy hinders the generalization of findings and underscores the need for a more relevant evaluation framework.

Addressing privacy, safety, and ethical/regulatory concerns: Prioritizing patient privacy, safety, and ethical standards is paramount in mental health applications. For LLMs to serve as reliable clinical aids that can be applied in practice, models are expected to process sensitive personal and health information; therefore, strict adherence to data protection laws like the Health Insurance Portability and Accountability Act (HIPAA)¹³⁴ and the General Data Protection Regulation (GDPR)¹³⁵ is necessary. Standards for bias analysis, model transparency, and informed consent should be established and enforced. To maintain alignment with ethical standards and clinical efficacy, developers, mental health professionals, ethicists, and legal experts must collaborate closely and implement continuous monitoring and evaluation of the LLMs. Additionally, emergency protocols should be in place to allow for expert intervention when necessary.

5 Conclusion

LLMs hold promise for enhancing mental health care with their ability to process and generate large volumes of textual information. Our results suggest that while it is feasible to run these models, concerns about the lack of valid training sets, lack of standardized evaluation metrics, and privacy/ethical concerns require careful consideration before any clinical implementation. Addressing these challenges will be more productive than further pilot studies which show the feasibility but also require new collaborations to refine technology, develop standardized evaluations, and ensure ethical applications, aiming to realize LLMs' full potential in supporting mental health care.

References

1. On Mental Illness (NAMI), N. A. Mental health by the numbers (2023). Accessed on December 17, 2023.
2. Of Mental Health (NIMH), N. I. Mental illness (2023). Last updated March 2023.
3. Organization, W. H. Mental health (2023). Accessed on December 17, 2023.
4. Organization, W. H. Who special initiative for mental health (2023). Accessed on December 17, 2023.
5. Sadock, B. J. *et al.* *Kaplan & Sadock's synopsis of psychiatry: behavioral sciences/clinical psychiatry*, vol. 2015 (Wolters Kluwer Philadelphia, PA, 2015).
6. Hales, R. E., Yudofsky, S. C. & Roberts, L. W. *The American Psychiatric Association Publishing Textbook of Psychiatry* (American Psychiatric Publishing, 2014), 6th edn.
7. Zhang, T., Schoene, A., Ji, S. *et al.* Natural language processing applied to mental illness detection: a narrative review. *npj Digit. Medicine* 5, DOI: [10.1038/s41746-022-00589-7](https://doi.org/10.1038/s41746-022-00589-7) (2022).
8. Malgaroli, M., Hull, T., Zech, J. *et al.* Natural language processing for mental health interventions: a systematic review and research framework. *Transl. Psychiatry* 13, DOI: [10.1038/s41398-023-02592-2](https://doi.org/10.1038/s41398-023-02592-2) (2023).

9. Li, H., Zhang, R., Lee, Y. *et al.* Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *npj Digit. Medicine* 6, DOI: [10.1038/s41746-023-00979-5](https://doi.org/10.1038/s41746-023-00979-5) (2023).
10. Irving, J. *et al.* Using natural language processing on electronic health records to enhance detection and prediction of psychosis risk. *Schizophr. bulletin* 47, 405–414 (2021).
11. Levis, M., Westgate, C. L., Gui, J., Watts, B. V. & Shiner, B. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychol. medicine* 51, 1382–1391 (2021).
12. Skaik, R. & Inkpen, D. Using social media for mental health surveillance: A review. *ACM Comput. Surv.* 53, 1–31, DOI: [10.1145/3422824](https://doi.org/10.1145/3422824) (2020).
13. Coppersmith, G., Dredze, M. & Harman, C. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 51–60 (2014).
14. Li, M. *et al.* Tracking the impact of covid-19 and lockdown policies on public mental health using social media: Infoveillance study. *J. Med. Internet Res.* 24, e39676 (2022).
15. Wu, J. *et al.* Exploring social media for early detection of depression in covid-19 patients. 3968–3977, DOI: [10.1145/3543507.3583867](https://doi.org/10.1145/3543507.3583867) (2023).
16. Heerden, A., Ruiz Pozuelo, J. & Kohrt, B. Global mental health services and the impact of artificial intelligence-powered large language models. *JAMA psychiatry* 80, DOI: [10.1001/jamapsychiatry.2023.1253](https://doi.org/10.1001/jamapsychiatry.2023.1253) (2023).
17. Ji, S., Zhang, T., Yang, K., Ananiadou, S. & Cambria, E. Rethinking large language models in mental health applications (2023). [2311.11267](https://doi.org/10.1126/sciadv.2311126).
18. Choudhury, M. D., Pendse, S. R. & Kumar, N. Benefits and harms of large language models in digital mental health (2023). [2311.14693](https://doi.org/10.1126/sciadv.2311146).
19. Ma, Z., Mei, Y. & Su, Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support (2023). [2307.15810](https://doi.org/10.1126/sciadv.2307158).
20. Zhao, W. X. *et al.* A survey of large language models (2023). [2303.18223](https://doi.org/10.1126/sciadv.2303182).
21. Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *The J. Mach. Learn. Res.* 21, 5485–5551 (2020).
22. Jones, K. S. Natural language processing: a historical review. *Curr. issues computational linguistics: honour Don Walk.* 3–16 (1994).
23. Weizenbaum, J. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 36–45, DOI: [10.1145/365153.365168](https://doi.org/10.1145/365153.365168) (1966).
24. Schank, R. C. Conceptual dependency: A theory of natural language understanding. *Cogn. psychology* 3, 552–631 (1972).
25. Beeferman, D., Berger, A. & Lafferty, J. Statistical models for text segmentation. *Mach. learning* 34, 177–210 (1999).
26. Goldberg, Y. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.* 57, 345–420 (2016).
27. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *Proc. Work. at ICLR* 2013 (2013).
28. Sarzynska-Wawer, J. *et al.* Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.* 304, 114135 (2021).

29. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C. & Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423) (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).
30. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* Improving language understanding by generative pre-training. *OpenAI* (2018).
31. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Process.* 45, 2673–2681 (1997).
32. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* 30 (2017).
33. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* 33, 1877–1901 (2020).
34. OpenAI. Chatgpt [large language model]. <https://chat.openai.com> (2023).
35. OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
36. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
37. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models (2023). [2307.09288](https://arxiv.org/abs/2307.09288).
38. Chowdhery, A. *et al.* Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.* 24, 1–113 (2023).
39. Liu, P. *et al.* Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 1–35 (2023).
40. Page, M. J. *et al.* The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372, DOI: [10.1136/bmj.n71](https://doi.org/10.1136/bmj.n71) (2021). <https://www.bmjjournals.org/content/372/bmj.n71.full.pdf>.
41. Khraisha, Q., Put, S., Kappenberg, J., Warratich, A. & Hadfield, K. Can large language models replace humans in the systematic review process? evaluating gpt-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages (2023). [2310.17526](https://arxiv.org/abs/2310.17526).
42. Gargari, O. K., Mahmoudi, M. H., Hajisafarali, M. & Samiee, R. Enhancing title and abstract screening for systematic reviews with gpt-3.5 turbo. *BMJ Evidence-Based Medicine* DOI: [10.1136/bmjebm-2023-112678](https://doi.org/10.1136/bmjebm-2023-112678) (2023). <https://ebm.bmjjournals.org/content/early/2023/11/21/bmjebm-2023-112678.full.pdf>.
43. McHugh, M. Interrater reliability: the kappa statistic. *Biochem. Medica (Zagreb)* 22, 276–282 (2012).
44. Chen, Q. & Liu, D. Dynamic strategy chain: Dynamic zero-shot cot for long mental health support generation (2023). [2308.10444](https://arxiv.org/abs/2308.10444).
45. Loh, S. B. & Raamkumar, A. S. Harnessing large language models' empathetic response generation capabilities for online mental health counselling support (2023). [2310.08017](https://arxiv.org/abs/2310.08017).
46. Lai, T. *et al.* Psy-llm: Scaling up global mental health psychological services with ai-based large language models (2023). [2307.11991](https://arxiv.org/abs/2307.11991).
47. Liu, J. M. *et al.* Chatcounselor: A large language models for mental health support (2023). [2309.15461](https://arxiv.org/abs/2309.15461).
48. Fu, G. *et al.* Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals (2023). [2308.15192](https://arxiv.org/abs/2308.15192).
49. Yao, X. *et al.* Development and evaluation of three chatbots for postpartum mood and anxiety disorders (2023). [2308.07407](https://arxiv.org/abs/2308.07407).

50. Liu, S. *et al.* Task-adaptive tokenization: Enhancing long-form text generation efficacy in mental health and beyond (2023). [2310.05317](#).
51. Lee, Y. K., Lee, I., Shin, M., Bae, S. & Hahn, S. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models (2023). [2311.04915](#).
52. Zheng, Z., Liao, L., Deng, Y. & Nie, L. Building emotional support chatbots in the era of llms (2023). [2308.11584](#).
53. Cho, Y. *et al.* Evaluating the efficacy of interactive language therapy based on llm for high-functioning autistic adolescent psychological counseling (2023). [2311.09243](#).
54. Zhang, Q., Naradowsky, J. & Miyao, Y. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In Rogers, A., Boyd-Graber, J. & Okazaki, N. (eds.) *Findings of the Association for Computational Linguistics: ACL 2023*, 6665–6694, DOI: [10.18653/v1/2023.findings-acl.417](#) (Association for Computational Linguistics, Toronto, Canada, 2023).
55. Hadar Shoval, D., Elyoseph, Z. & Lvovsky, M. The plasticity of chatgpt's mentalizing abilities: Personalization for personality structures. *Front. Psychiatry* Volume 14, DOI: [10.3389/fpsyg.2023.1234397](#) (2023).
56. Vowels, L. M., Francois-Walcott, R. & Darwiche, J. Ai in relationship counselling: Evaluating chatgpt's therapeutic efficacy in providing relationship advice, DOI: [10.31234/osf.io/3zajt](#) (2023).
57. Vowels, L. M. Are chatbots the new relationship experts? insights from three studies, DOI: [10.31234/osf.io/nh3v9](#) (2023).
58. Kumar, H. *et al.* Exploring the design of prompts for applying gpt-3 based chatbots: A mental wellbeing case study on mechanical turk (2022). [2209.11344](#).
59. Yang, K. *et al.* Towards interpretable mental health analysis with large language models (2023). [2304.03347](#).
60. Yang, K. *et al.* Mentallama: Interpretable mental health analysis on social media with large language models (2023). [2309.13567](#).
61. Chen, Z., Lu, Y. & Wang, W. Y. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting (2023). [2310.07146](#).
62. Bucur, A.-M. Utilizing chatgpt generated data to retrieve depression symptoms from social media (2023). [2307.02313](#).
63. Smith, A. *et al.* Old dog, new tricks? exploring the potential functionalities of chatgpt in supporting educational methods in social psychiatry. *The Int. journal social psychiatry* 69, 207640231178451, DOI: [10.1177/00207640231178451](#) (2023).
64. Kumar, H. *et al.* Exploring the use of large language models for improving the awareness of mindfulness. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, DOI: [10.1145/3544549.3585614](#) (Association for Computing Machinery, New York, NY, USA, 2023).
65. Xu, X. *et al.* Mental-llm: Leveraging large language models for mental health prediction via online text data (2023). [2307.14385](#).
66. Lamichhane, B. Evaluation of chatgpt for nlp-based mental health applications (2023). [2303.15727](#).
67. Qi, H. *et al.* Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in chinese social media (2023). [2309.03564](#).
68. Zhou, W., Prater, L., Goldstein, E. & Mooney, S. Identifying rare circumstances preceding female firearm suicides: Validating a large language model approach. *JMIR Mental Heal.* 10, e49359, DOI: [10.2196/49359](#) (2023).

69. Bhaumik, R., Srivastava, V., Jalali, A., Ghosh, S. & Chandrasekharan, R. Mindwatch: A smart cloud-based ai solution for suicide ideation detection leveraging large language models. *medRxiv* DOI: [10.1101/2023.09.25.23296062](https://doi.org/10.1101/2023.09.25.23296062) (2023). <https://www.medrxiv.org/content/early/2023/09/26/2023.09.25.23296062.full.pdf>.
70. American Psychiatric Association, A., Association, A. P. *et al.* *Diagnostic and statistical manual of mental disorders: DSM-IV*, vol. 4 (American psychiatric association Washington, DC, 1994).
71. Posner, K. *et al.* The columbia–suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am. journal psychiatry* 168, 1266–1277 (2011).
72. Mauriello, M. L. *et al.* Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 1–7 (2021).
73. Wu, R. *et al.* Mindshift: Leveraging large language models for mental-states-based problematic smartphone use intervention (2023). [2309.16639](https://doi.org/10.4236/ojs.2023.13091639).
74. Kim, T. *et al.* Mindfuldiary: Harnessing large language model to support psychiatric patients’ journaling (2023). [2310.05231](https://doi.org/10.4236/ojs.2023.1305231).
75. Souto, E. B., Pérez, A. & Parapar, J. Explainable depression symptom detection in social media (2023). [2310.13664](https://doi.org/10.4236/ojs.2023.13013664).
76. Hayati, M. F. M., Ali, M. A. M. & Rosli, A. N. M. Depression detection on malay dialects using gpt-3. In *2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 360–364, DOI: [10.1109/IECBES54088.2022.10079554](https://doi.org/10.1109/IECBES54088.2022.10079554) (2022).
77. Grabb, D. The impact of prompt engineering in large language model performance: a psychiatric example. *J. Med. Artif. Intell.* 6 (2023).
78. Dekel, S., Bartal, A., Jagodnik, K. & Chan, S. Chatgpt demonstrates potential for identifying psychiatric disorders: Application to childbirth-related post-traumatic stress disorder, DOI: [10.21203/rs.3.rs-3428787/v1](https://doi.org/10.21203/rs.3.rs-3428787/v1) (2023).
79. Taori, R. *et al.* Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023).
80. Zeng, W. *et al.* Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation (2021). [2104.12369](https://doi.org/10.4236/ojs.202121012369).
81. Chiang, W.-L. *et al.* Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (2023).
82. Roller, S. *et al.* Recipes for building an open-domain chatbot. In Merlo, P., Tiedemann, J. & Tsarfaty, R. (eds.) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325, DOI: [10.18653/v1/2021.eacl-main.24](https://doi.org/10.18653/v1/2021.eacl-main.24) (Association for Computational Linguistics, Online, 2021).
83. Tay, Y. *et al.* Ul2: Unifying language learning paradigms (2023). [2205.05131](https://doi.org/10.4236/ojs.20230205131).
84. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* 35, 24824–24837 (2022).
85. Scaling instruction-finetuned language models, DOI: [10.48550/ARXIV.2210.11416](https://doi.org/10.48550/ARXIV.2210.11416) (2022).
86. Zhou, H. *et al.* A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112* (2023).
87. Zeng, Q. *et al.* Greenplm: Cross-lingual transfer of monolingual pre-trained language models at almost no cost. *32nd Int. Jt. Conf. on Artif. Intell.* (2022).

88. Ye, Q. *et al.* Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089* (2023).
89. Guo, Z. & Hua, Y. Continuous training and fine-tuning for domain-specific language models in medical question answering (2023). [2311.00204](#).
90. Sun, H., Lin, Z., Zheng, C., Liu, S. & Huang, M. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In Zong, C., Xia, F., Li, W. & Navigli, R. (eds.) *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1489–1503, DOI: [10.18653/v1/2021.findings-acl.130](#) (Association for Computational Linguistics, Online, 2021).
91. Cohan, A. *et al.* SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In Bender, E. M., Derczynski, L. & Isabelle, P. (eds.) *Proceedings of the 27th International Conference on Computational Linguistics*, 1485–1497 (Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018).
92. Yao, B. *et al.* D4: a chinese dialogue dataset for depression-diagnosis-oriented chat. *arXiv preprint arXiv:2205.11764* (2022).
93. Liu, S. *et al.* Towards emotional support dialog systems. In Zong, C., Xia, F., Li, W. & Navigli, R. (eds.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483, DOI: [10.18653/v1/2021.acl-long.269](#) (Association for Computational Linguistics, Online, 2021).
94. Qiu, H. *et al.* A benchmark for understanding dialogue safety in mental health support. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 1–13 (Springer, 2023).
95. Turcan, E. & McKeown, K. Dreaddit: A Reddit dataset for stress analysis in social media. In Holderness, E. *et al.* (eds.) *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 97–107, DOI: [10.18653/v1/D19-6213](#) (Association for Computational Linguistics, Hong Kong, 2019).
96. Naseem, U., Dunn, A. G., Kim, J. & Khushi, M. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, 2563–2572 (2022).
97. Haque, A., Reddi, V. & Giallanza, T. Deep learning for suicide and depression identification with unsupervised label correction. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*, 436–447 (Springer, 2021).
98. Gaur, M. *et al.* Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*, 514–525 (2019).
99. Pirina, I. & Cöltekin, Ç. Identifying depression on Reddit: The effect of training data. In Gonzalez-Hernandez, G., Weissenbacher, D., Sarker, A. & Paul, M. (eds.) *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, 9–12, DOI: [10.18653/v1/W18-5903](#) (Association for Computational Linguistics, Brussels, Belgium, 2018).
100. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K. & Mitchell, M. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 31–39, DOI: [10.3115/v1/W15-1204](#) (Association for Computational Linguistics, Denver, Colorado, 2015).
101. Ji, S., Li, X., Huang, Z. & Cambria, E. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Comput. Appl.* 34, 10309–10319 (2022).

102. Garg, M. *et al.* CAMS: An annotated corpus for causal analysis of mental health issues in social media posts. In Calzolari, N. *et al.* (eds.) *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6387–6396 (European Language Resources Association, Marseille, France, 2022).
103. Rashkin, H., Smith, E. M., Li, M. & Boureau, Y.-L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Korhonen, A., Traum, D. & Márquez, L. (eds.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381, DOI: [10.18653/v1/P19-1534](https://doi.org/10.18653/v1/P19-1534) (Association for Computational Linguistics, Florence, Italy, 2019).
104. SATHVIK, M. & Garg, M. Multiwd: Multiple wellness dimensions in social media posts. *TechRxiv* (2023).
105. Garg, M., Shahbandegan, A., Chadha, A. & Mago, V. An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts. In Rogers, A., Boyd-Graber, J. & Okazaki, N. (eds.) *Findings of the Association for Computational Linguistics: ACL 2023*, 11960–11969, DOI: [10.18653/v1/2023.findings-acl.757](https://doi.org/10.18653/v1/2023.findings-acl.757) (Association for Computational Linguistics, Toronto, Canada, 2023).
106. Demszky, D. *et al.* Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547* (2020).
107. Lahnala, A. *et al.* Exploring self-identified counseling expertise in online support forums. In Zong, C., Xia, F., Li, W. & Navigli, R. (eds.) *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4467–4480, DOI: [10.18653/v1/2021.findings-acl.392](https://doi.org/10.18653/v1/2021.findings-acl.392) (Association for Computational Linguistics, Online, 2021).
108. Sharma, A., Miner, A., Atkins, D. & Althoff, T. A computational approach to understanding empathy expressed in text-based mental health support. In Webber, B., Cohn, T., He, Y. & Liu, Y. (eds.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276, DOI: [10.18653/v1/2020.emnlp-main.425](https://doi.org/10.18653/v1/2020.emnlp-main.425) (Association for Computational Linguistics, Online, 2020).
109. Talklife: Peer support network for mental health - support group. <https://www.talklife.com/>. Accessed: 2024-01-01.
110. Shreevastava, S. & Foltz, P. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 151–158 (2021).
111. Pérez, A., Parapar, J., Barreiro, Á. & Lopez-Larrosa, S. Bdi-sen: A sentence dataset for clinical symptoms of depression. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2996–3006 (2023).
112. Zhang, Z., Chen, S., Wu, M. & Zhu, K. Symptom identification for interpretable detection of multiple mental disorders on social media. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, 9970–9985 (2022).
113. National violent death reporting system (nvdrs). <https://www.cdc.gov/violenceprevention/datasources/nvdrs/index.html>. Accessed: 2023-12-31.
114. Shen, Y., Yang, H. & Lin, L. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstmbased model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6247–6251 (IEEE, 2022).
115. Bradley, M. M. & Lang, P. J. Affective norms for english words (anew): Instruction manual and affective ratings. Tech. Rep., Technical report C-1, the center for research in psychophysiology ... (1999).
116. Warriner, A. B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behav. research methods* 45, 1191–1207 (2013).
117. Mohammad, S. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Gurevych, I. & Miyao, Y. (eds.) *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), 174–184, DOI: [10.18653/v1/P18-1017](https://doi.org/10.18653/v1/P18-1017) (Association for Computational Linguistics, Melbourne, Australia, 2018).

118. Ma, Z., Mei, Y. & Su, Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *arXiv preprint arXiv:2307.15810* (2023).
119. Van Bulck, L. & Moons, P. What if your patient switches from dr. google to dr. chatgpt? a vignette-based survey of the trustworthiness, value, and danger of chatgpt-generated responses to health questions. *Eur. J. Cardiovasc. Nurs.* zvad038 (2023).
120. Jin, H., Chen, S., Wu, M. & Zhu, K. Q. Psyeval: A comprehensive large language model evaluation benchmark for mental health (2023). [2311.09189](https://doi.org/10.48550/2311.09189).
121. Liu, S. *et al.* Towards emotional support dialog systems. In Zong, C., Xia, F., Li, W. & Navigli, R. (eds.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483, DOI: [10.18653/v1/2021.acl-long.269](https://doi.org/10.18653/v1/2021.acl-long.269) (Association for Computational Linguistics, Online, 2021).
122. O’Leary, K., Schueller, S. M., Wobbrock, J. O. & Pratt, W. “suddenly, we got to become therapists for each other” designing peer support chats for mental health. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–14 (2018).
123. Li, M., Weston, J. & Roller, S. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons (2019). [1909.03087](https://doi.org/10.48550/1909.03087).
124. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert (2020). [1904.09675](https://doi.org/10.48550/1904.09675).
125. Yuan, W., Neubig, G. & Liu, P. Bartscore: Evaluating generated text as text generation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, 27263–27277 (Curran Associates, Inc., 2021).
126. Fu, J., Ng, S.-K., Jiang, Z. & Liu, P. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166* (2023).
127. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E. & Lin, D. (eds.) *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318, DOI: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135) (Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002).
128. Banerjee, S. & Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Goldstein, J., Lavie, A., Lin, C.-Y. & Voss, C. (eds.) *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72 (Association for Computational Linguistics, Ann Arbor, Michigan, 2005).
129. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81 (Association for Computational Linguistics, Barcelona, Spain, 2004).
130. Du, Z. *et al.* Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335 (2022).
131. Zeng, A. *et al.* Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).
132. Harrigan, K., Aguirre, C. & Dredze, M. Do models of mental health based on social media data generalize? In Cohn, T., He, Y. & Liu, Y. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3774–3788, DOI: [10.18653/v1/2020.findings-emnlp.337](https://doi.org/10.18653/v1/2020.findings-emnlp.337) (Association for Computational Linguistics, Online, 2020).

133. Hua, Y. *et al.* Using twitter data to understand public perceptions of approved versus off-label use for covid-19related medications. *J. Am. Med. Informatics Assoc.* 29, 1668–1678 (2022).
134. 104th United States Congress. Health insurance portability and accountability act of 1996 (1996).
135. European Parliament & Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council (2016).

6 Author contributions statement

Study design, data collection, and article screening: Y.H.

Result analysis: Y.H., F.L., K.Y., and Z.L.

Manuscript drafting: Y.H., F.L., K.Y., Z.L, H.N, P.Z, and Y.S.

Critical feedback: Y.S., L.M., and A.B.

Manuscript review: All authors.

Y.H. takes the integral responsibility of the study.

7 Supplementary

7.1 S.1. Screening prompt for ChatGPT-4

The template shown in Figure 3 was employed for screening titles and abstracts in the review's screening phase. After testing various prompts, the one presented here demonstrated the highest accuracy, achieving a perfect score (10 out of 10) in a randomly selected sample. It was crucial to specify that articles should be explicitly focused on LLMs and mental health care. This precision was necessary to avoid the inclusion of articles that were only vaguely related to the topic, a common issue encountered when the criteria were less specific.

You are assisting in a systematic review focused on the intersection of large language models and mental healthcare. Your task is to evaluate titles and abstracts of various papers. Include only original research and case reports. Exclude reviews, editorials, perspectives, letters, and comments. Ensure that each article directly relates to both large language models and mental healthcare.

For each article, assign a code and provide a brief justification:

- 1: Include (meets all criteria for relevance and type)
- 0: Exclude (does not meet criteria or is an excluded type)
- 2: Uncertain (insufficient information or unclear relevance)

Please provide a concise explanation for your decision on each article.

Figure 3. Screening prompt for ChatGPT-4.

7.2 S.2. Data Restriction and Licensing in Related Datasets

Dataset restrictions and licenses establish the guidelines for how data may be utilized, shared, altered, and disseminated. Users must comply with the terms set forth by the restrictions and licenses under which the datasets are released. Below is an explanation of the various restrictions and licenses associated with the datasets we reviewed:

1. Publicly Available: Indicates that the dataset is made freely available to the public and can be used, distributed, or modified without any restrictions, sometimes even for commercial purposes, depending on the exact nature of the public domain status.
2. Research Only: Indicates that the dataset is available only for non-commercial, academic research purposes. It typically restricts commercial use or widespread redistribution without explicit permission.

3. MIT License: A permissive free software license originating at the Massachusetts Institute of Technology (MIT). It allows for almost unrestricted freedom with the work, including commercial use, with the only requirement being to preserve the copyright and license notices.
4. Apache 2.0: A permissive free software license written by the Apache Software Foundation. It allows users to use the software for any purpose, to distribute it, to modify it, and to distribute modified versions of the software under the terms of the license, without concern for royalties.
5. CC-BY-4.0 (Creative Commons Attribution 4.0 International): Allows others to share, copy, distribute, execute, and transmit the work, as well as to remix, adapt, and build upon the work, even commercially, as long as they provide a credit to the original creator.
6. CC-BY-NC-4.0 (Creative Commons Attribution NonCommercial 4.0 International): Similar to CC-BY, this license allows others to remix, adapt, and build upon the work non-commercially, and although their new works must also acknowledge the creator and be non-commercial, they don't have to license their derivative works on the same terms.

Datasets that do not have a specific license mentioned or that the availability of the dataset is not explicitly stated are marked with N/P. It is advisable to contact the dataset providers or creators for more information on usage rights and restrictions.