Published: 07 February 2025
Online AM: 22 October 2024
Accepted: 02 October 2024
Revised: 20 September 2024
Received: 31 January 2024

RESEARCH-ARTICLE

Improving Workplace Well-being in Modern Organizations: A Review of Large Language Model-based Mental Health Chatbots

AIJIA YUAN, Indiana University Bloomington, Bloomington, IN, United States

EDLIN GARCIA COLATO, Indiana University Bloomington, Bloomington, IN, United States

BERNICE PESCOSOLIDO, Indiana University Bloomington, Bloomington, IN, United States

HYUNJU SONG, Luddy School of Informatics, Computing, and Engineering, Bloomington, IN, United States

SAGAR SAMTANI, Indiana University Bloomington, Bloomington, IN, United States

Open Access Support provided by:

Indiana University Bloomington

Luddy School of Informatics, Computing, and Engineering

Improving Workplace Well-being in Modern Organizations: A Review of Large Language Model-based Mental Health Chatbots

AIJIA YUAN, Department of Operations and Decision Technologies, Indiana University, Bloomington, United States
EDLIN GARCIA COLATO, School of Public Health, Indiana University, Bloomington, United States
BERNICE PESCOSOLIDO, Department of Sociology, Indiana University, Bloomington, United States
HYUNJU SONG, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, United States
SAGAR SAMTANI, Department of Operations and Decision Technologies, Indiana University, Bloomington, United States

The global rise in mental disorders, particularly in workplaces, necessitated innovative and scalable solutions for delivering therapy. Large Language Model (LLM)-based mental health chatbots have rapidly emerged as a promising tool for overcoming the time, cost, and accessibility constraints often associated with traditional mental health therapy. However, LLM-based mental health chatbots are in their nascency, with significant opportunities to enhance their capabilities to operate within organizational contexts. To this end, this research seeks to examine the role and development of LLMs in mental health chatbots over the past half-decade. Through our review, we identified over 50 mental health-related chatbots, including 22 LLM-based models targeting general mental health, depression, anxiety, stress, and suicide ideation. These chatbots are primarily used for emotional support and guidance but often lack capabilities specifically designed for workplace mental health, where such issues are increasingly prevalent. The review covers their development, applications, evaluation, ethical concerns, integration with traditional services, LLM-as-a-Service, and various other business implications in organizational settings. We provide a research illustration of how LLM-based approaches could overcome the identified limitations and also offer a system that could help facilitate systematic evaluation of LLM-based mental health chatbots. We offer suggestions for future research tailored to workplace mental health needs.

CCS Concepts: • **Information systems → Information systems applications**;

Additional Key Words and Phrases: Large language models, chatbots, conversational agents, mental health, workplace, well-being

Authors' Contact Information: Aijia Yuan, Department of Operations and Decision Technologies, Indiana University, Bloomington, IN, United States; e-mail: yuana@iu.edu; Edlin Garcia Colato, School of Public Health, Indiana University, Bloomington, IN, United States; e-mail: eggarcia@iu.edu; Bernice Pescosolido, Department of Sociology, Indiana University, Bloomington, IN, United States; email: pescosol@iu.edu; Hyunju Song, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, United States; e-mail: sarasong@iu.edu; Sagar Samtani, Department of Operations and Decision Technologies, Indiana University, Bloomington, IN, United States; e-mail: ssamtani@iu.edu.

ACM 2158-656X/2025/02-ART3
https://doi.org/10.1145/3701041

ACM Reference Format:
Aijia Yuan, Edlin Garcia Colato, Bernice Pescosolido, Hyunju Song, and Sagar Samtani. 2025. Improving Workplace Well-being in Modern Organizations: A Review of Large Language Model-based Mental Health Chatbots. *ACM Trans. Manag. Inform. Syst.* 16, 1, Article 3 (February 2025), 26 pages. https://doi.org/10.1145/3701041

1 Introduction

Mental disorders are a leading cause of disability worldwide, with significant health, social, human rights, and economic impacts [83]. In the US, one in five adults experiences mental illness annually, with suicide a leading cause of disability and death [91]. These issues increasingly affect workplaces with impacts to both individual and organizational productivity and well-being [40]. In 2022, 76% of workers reported mental health symptoms, and 70% of employers expressed concern about their workers' mental health [159].

Mental health therapy is often crucial for addressing workplace mental health issues like depression and anxiety [6]. However, conventional therapy faces limitations due to restricted accessibility, high costs, and inadequate insurance coverage [62], affecting the labor force. The shift to virtual therapy has improved access in some underserved areas [12, 66]. Yet, the global shortage of trained mental health professionals persists, with the U.S. facing a projected deficit of 10,000 by 2025 [132]. In the workplace, many believe that mental health resources are not universally accessible. Additionally, there is a prevailing stigma around seeking mental health resources, with 31% fearing judgment [36]. As the stigma decreases, mental health support, whether in-person or virtually, is met with long wait times. Alternatives that can help address this growing service gap are critically needed [121].

Large Language Models (LLMs) offer a novel solution to the mental health service gap in workplaces [148]. LLM-based chatbots generate more nuanced, context-aware, and empathetic responses than traditional chatbots [56, 144]. Moreover, LLM-based mental health chatbots can address the shortage of conventional services by offering scalable, cost-effective solutions ranging from initial assessments to ongoing support [101]. However, despite these advantages, LLM-based chatbots can provide misinformation or dangerous advice. For example, in 2023, a Belgian man died by suicide after the AI chatbot app Chai encouraged him to kill himself and provided methods of suicide [145]. This tragic incident underscores the serious risks posed by LLMs that do not align with mental health guidelines. This issue is particularly concerning in organizational settings, where employees may turn to mental health chatbots as a first point of support due to concerns about privacy, stigma, accessibility, or costs. Taken together, these issues motivate the development of well-designed and safe LLM-based chatbots for organizations to safely enhance workplace mental health support.

In this paper, we review LLM-based mental health chatbots to provide detailed insights into their development, application, and evaluation. This review synthesizes literature to identify the gaps in extant mental health chatbots to help identify potential research directions and possible practical applications of LLM-based approaches to help improve workplace mental health. Unlike previous studies, which often focus on either technical aspects or specific use cases, this review provides a holistic perspective that integrates both technological innovations and their implications for real-world organizational settings. We also present a research example illustrating how LLM-based mental health chatbots could be enhanced to account for key guidelines. Additionally, we summarize a prototype system that is designed specifically to help enhance systematic evaluations of LLM-based mental health chatbots.

The remainder of this paper is organized as follows. First, we explore the background of LLMs, their applications, and mental health chatbots. Second, we summarize LLM-based mental health chatbots from multiple perspectives. Third, we discuss their limitations and propose future research directions. Fourth, based on some of these limitations, we present a research example and introduce a chatbot design to showcase how our review findings facilitate rigorous evaluations of novel LLM-based mental health chatbots. Fifth, we examine the implications for business and management, highlighting their impact in organizational and commercial settings. Finally, we summarize this study's contributions and conclude this research.

2 Background

2.1 Limitations of Existing Surveys and This Study's Contributions

The use of LLMs in mental health is a relatively new and emerging field. Few review studies focus on LLM-based applications; most surveys concentrate on the effectiveness and safety of non-LLM-based mental health chatbots [1, 2, 133, 146]. The most pertinent LLM-based mental health studies we identified are by [42, 50], which reviewed LLMs for mental health but did not focus on LLM-based chatbots. This distinction is crucial as our paper explores the methods, evaluations, applications, and challenges of LLM-powered mental health chatbots. Specifically, this article has several contributions beyond existing surveys and reviews. First, our review significantly extends coverage by summarizing major open/closed-source LLMs, mental health-related LLMs, and mental health chatbots in general. Second, we extensively discuss technical and human evaluation metrics for these chatbots, which previous reviews have not explored in depth. Third, we offer guidance for future research, discuss LLMs-as-a-S ervice, and integration with conventional services. Finally, we highlight broader business implications, specifically for workplace mental health (all perspectives overlooked in prior reviews). Taken together, this coverage provides a rich and holistic view of the potential and challenges of LLM-based mental health chatbots.

2.2 Evolution of Large Language Models

LLMs have evolved from early rule-based systems in the 1950s [44, 82, 94] to small-scale predictive models in the 1980s and 1990s [48] and to statistical language models [109]. Later, deep learning's revolution in **natural language processing (NLP)** helped enable sophisticated, data-driven models capable of understanding and generating human-like text [129]. The early 2000s introduced neural language models [50]. The late 2010s saw a major change in NLP with **pre-trained language models (PLMs)** like BERT, T5, and RoBERTa [135]. These models are pre-trained on large corpora and can be fine-tuned for domain-specific capabilities [43, 156]. Around 2020, significant LLMs included the GPT series, LLaMA models, and PaLM [127]. In Table 1, we summarize major LLMs released after 2022, detailing their providers, parameter sizes, training data, and key operations.

Open source models come from academic entities like Beijing University and tech giants such as Google and Meta. Parameter sizes range from Mistral AI's 7 billion to Meta's LLaMA3.1 at 405 billion, showing diverse complexities. Training data varies significantly, with models like Meta's LLaMA3 using 15 trillion tokens. These open-source LLMs feature a variety of key operations including fine-tuning, prompting, reinforcement learning, instruction tuning, and advanced processing techniques like CoT. In the closed-source domain, Huawei's PanGu-Σ stands out with 1.085 trillion parameters, the highest on the list. Other LLMs like OpenAI's GPT-4 and DeepMind's Sparrow provide extensive knowledge bases that can generate responsive dialogue and improve interaction quality. However, directly applying LLMs in mental health poses significant challenges. These models were designed for diverse applications and trained on broad datasets but lack the

Table 1. Summary of Prevailing Open Source and Closed Source LLMs

	Model Name	Provider	Parameters (B)	Training Data	Key Operations*
Open source	LLaMA3.1 [161]	Meta	8/70/405	15T tokens	Fine-tuning
	LLaMA3 [162]	Meta	8/70	15T tokens	Fine-tuning
	Mistrial 7B [55]	Mistral AI	7	–	Instruction tuning
	FLM [74]	Beijing University	101	311B tokens	Prompting, instruction tuning
	LLaMA2 [131]	Meta	70	2T tokens	Reinforcement learning, instruction tuning
	StarCoder [72]	–	15.5	1T tokens	Prompting, fine-tuning
	CodeGen2 [93]	Salesforce	16	400B tokens	Prompting
	LLaMA [130]	Meta	65	1.4T tokens	Prompting
	OPT-IML [53]	Meta	175	–	Fine-tuning, prompting, CoT
	BLOOMZ [87]	BigScience	176	–	Prompting, instruction tuning
	Galactica [125]	Meta	120	106B tokens	Prompting, CoT
	mT0 [87]	–	13	–	Prompting, instruction tuning
	BLOOM [139]	BigScience	176	366B tokens	Prompting
	Flan-T5 [20]	Google	11	–	Fine-tuning, prompting, CoT
	GLM [152]	–	130	400B tokens	Prompting
	NLLB [23]	Meta	54.5	–	Prompting
	OPT [154]	Meta	175	180B tokens	Prompting, instruction tuning
	UL2 [124]	–	20	1T tokens	Prompting, fine-tuning, CoT
	CodeGen [93]	Salesforce	16	577B tokens	Prompting
	GPT-NeoX-20B [13]	EleutherAI	20	825GB	Prompting
Closed source	PaLM2 [7]	Google	16	100B tokens	Prompting, fine-tuning, CoT
	PanGu-Σ [107]	Huawei	1085	329B tokens	Prompting
	GPT-4 [4]	OpenAI	–	–	Prompting, fine-tuning, CoT
	Claude 2 [142]	Anthropic	52	–	Prompting
	Flan-PaLM [20]	Google	540	–	Prompting, instruction tuning, CoT
	Sparrow [38]	DeepMind	70	–	Prompting, reinforcement learning
	PaLM [19]	Google	540	780B tokens	Prompting, fine-tuning, CoT
	Chinchilla [49]	DeepMind	70	1.4T tokens	Prompting
	InstructGPT [95]	OpenAI	175	–	Instruction tuning, reinforcement learning
	AlphaCode [75]	DeepMind	41	967B tokens	Prompting, fine-tuning
	MT-NLG [119]	Microsoft	530	270B tokens	Prompting, fine-tuning
	LaMDA [128]	Google	137	768B tokens	Prompting

Note: CoT=Chain-of-Thought.

specificity for mental health therapy [156]. Consequently, these approaches could overlook critical ethical considerations in mental health care, such as confidentiality and clinical accuracy [61]. Moreover, they do not inherently align with mental health compliance standards and best practices followed by human professionals, potentially leading to inappropriate responses or ineffective crisis management [81].

2.3 LLMs in Mental Health

Given the limitations of off-the-shelf open-source LLMs, researchers are exploring ways to improve how LLMs can be employed or tailored for mental health applications. They aim to develop specialized approaches specifically for mental health purposes, including examining the use of LLMs in mental health and creating models that are trained or fine-tuned on mental health-specific datasets

Table 2. Selected Studies on Mental Health using LLMs

Focus	Year	Authors	Goal(s)	Datasets
General mental health	2024	Yang et al.	Fine-tune LLaMA2 for interpretable mental health analysis using instruction tuning.	IMHI dataset (posts from Reddit, X (formerly known as Twitter), and SMS) [148]
General mental health	2024	Kim et al.	Prompt and fine-tune 12 LLMs for mental health and other health predictions using sensor data.	PMData, LifeSnaps, GLOBEM, AW_FB [60]
General mental health	2024	Xu et al.	Predict mental health conditions using online text data by fine-tuning multiple LLMs.	Dreaddit, DepSeverity, SDCNL, CSSRS-Suicide [147]
Emotion	2024	Liu et al.	Fine-tune LLaMA2, OPT, and BLOOM models for multi-task affective analysis.	Affective Analysis Instruction Dataset (AAID) based on SemEval-2018 in Tweets [79]
General mental health	2023	Yang et al.	Incorporate LLMs into pre-consultation, diagnosis, and management processes.	–
General mental health	2023	Ajlouni et al.	Utilize ChatGPT for counseling, teaching, and enhancing the learning process.	Data collected from 210 students at the University of Jordan [5]
Social psychiatry	2023	Smith et al.	Use ChatGPT to deliver educational materials tailored for social psychiatry as a teaching tool.	Data directly generated by ChatGPT [118]
Depression	2023	Heston	Examine ChatGPT-based conversational agents for mental health using pre-structured prompts.	Heston TF dataset [47]
Well-being	2023	Kumar et al.	Enhance mindfulness awareness using GPT-3 through zero-shot learning.	Survey responses from 209 online users
Well-being	2023	Ma et al.	Employ GPT-3-powered LLMs in conversational agents to provide mental health support.	2,917 user comments drawn from the most popular subreddit [81]

[26, 138]. Table 2 summarizes selected relevant studies, detailing the mental concerns addressed, goals, and data used.

As illustrated in Table 2, LLMs are used in various mental health domains and tasks. One major area is the use of LLMs for mental health diagnosis and prediction. Studies have demonstrated the efficacy of fine-tuning models like LLaMA2 and GPT-4 to predict mental health conditions using diverse data sources, such as text and wearable sensor data [60, 147, 148], excelling in identifying early warning signs and assessing disorder severity. Another application of LLMs is in emotion and affective analysis. Liu et al. [79] fine-tuned models like LLaMA2, OPT, and BLOOM using social media datasets for tasks such as emotion detection, sentiment classification, and intensity prediction. LLMs are also employed to enhance mental health literacy by generating personalized educational resources [5, 64, 118]. For instance, studies using models like ChatGPT show they can tailor content to individual comprehension levels, presenting psychological concepts in an engaging and informative manner [135]. The most notable application, however, lies in the deployment of LLMs as mental health therapist agents or chatbots [47, 81, 149]. These digital entities are designed to produce empathetic and context-aware responses. By engaging users in meaningful conversations, they offer services including immediate support, psychoeducation, crisis intervention, guided self-help therapies, and routine check-ins [45].

3 Mental Health Chatbots

3.1 Historical Progression

Mental health chatbots initially relied on statistical approaches focused on pattern recognition and basic response generation [146]. Then, classical **machine learning (ML)** models like **support vector machines (SVM)**, decision trees, and Naïve Bayes emerged [8, 126]. These methods

relied heavily on feature engineering, requiring significant time and domain expertise [108, 158], and were also limited in understanding and responding to human language and emotion. **Deep learning (DL)** marked a significant advancement in chatbot development, enabling models to automatically learn from larger datasets and better process and respond to more nuanced aspects of human interaction [90].

The introduction of large PLMs, including those known as LLMs, represented a further leap in the evolution of chatbots [43]. PLMs brought advanced techniques such as fine-tuning and prompt engineering to enhance their capabilities, allowing for the creation of more advanced and responsive mental health chatbots [71]. These models could understand context, manage conversational flow, and provide more personalized and empathetic responses, all of which are crucial for mental health support [69]. In the following subsection, we summarize the major categories of mental health chatbots.

3.2 Categories of Mental Health Chatbots

Mental health chatbots can be grouped into three major categories based on how they operate: rule-based, retrieval-based, and generative AI-based. Each signified a significant progression in how these digital assistants interact and respond to users. Initially, most mental health chatbots were based on rule-based approaches [1]. Examples of rule-based mental health chatbots include IDEABot [134], Tess [57], and aiCARE [14]. In these models, chatbots matched user input to a set of rule patterns, selecting predefined answers from a pre-written script using pattern-matching algorithms [106]. A key drawback of this approach is the chatbot's inability to create new answers or adapt to unexpected user inputs. To help overcome these limitations, retrieval-based models [97] employed techniques like keyword matching, ML, or DL to sift through a predefined set of responses and select the most appropriate one. Retrieval-based models, such as Youper [85], Vivibot [41], and Woebot [33], could handle a wider range of inputs by identifying key terms and phrases. As such, these approaches demonstrated an improved ability to interact but were still limited by their reliance on predefined responses, which restrained free conversations [136].

Over the last two years, there has been a significant shift towards using LLMs to generate new dialogue based on extensive conversational training data [25, 61]. Unlike their predecessors, these chatbots can more effectively understand user inputs, generate novel responses, ensure a more natural conversational flow, and adapt to the evolving context of the dialogue. They also excel in simulating realistic conversations, thus enhancing social skills and promoting self-discovery [81]. Their advanced algorithms allow unparalleled personalization, tailoring responses and therapeutic strategies to individual user preferences and needs [16, 18]. Examples of chatbots utilizing generative models include the most recent versions of Replika [100] and ChatCounselor [78], which have set new standards in the mental health chatbot landscape by leveraging the advanced capabilities of LLMs. To this end, we review prevailing LLMs specifically designed for mental health applications in the following section.

4 Current State of the Art

4.1 Related Works

Our review strategy aimed to capture the rapid advancements in LLMs for mental health, specifically on developments from 2019 to 2024. Thus, we considered peer-reviewed studies and preprints, focusing on publications from 2019 onwards. We thoroughly searched multiple platforms, including Google Scholar, ArXiv, MedRxiv, and the ACM Digital Library. We utilized keywords like LLMs, chatbots, digital mental health, conversational agent, psychiatry, and various mental disorder terms to ensure comprehensive coverage. The review reveals a diverse range of chatbot

applications in mental health, with a marked shift towards exploring LLMs in the past one to two years [147]. Initially, most chatbots were rule-based, relying heavily on keyword matching and classical ML and DL techniques [57, 80, 99, 134]. Rule-based systems prevailed among the various chatbots reviewed. These approaches function optimally when both input and expected responses are predefined and known [112].

Regarding the types of mental health disorders addressed, most chatbots were not tailored to specific conditions but rather focused on general mental health and well-being [14, 33, 37, 57, 111]. This broad area was followed by a concentration on common mental health issues like depression, anxiety (often the most prevalent in mental health), or both [28, 39, 46, 52, 80, 85, 113, 134]. Other areas receiving attention included suicide ideation [11, 35], eating disorders [115], and stress management [99]. In terms of target populations, while the general population was the primary focus, many chatbots were specifically designed for vulnerable groups. Notably, young adults, college students, and adolescents, who are particularly susceptible to mental issues, formed a key demographic [33, 46, 80, 99, 134]. Additionally, chatbots catering to the needs of mothers and pregnant women, addressing conditions like **postnatal depression (PND)** and **postpartum depression (PPD)**, were also prominent [21, 24, 84, 110, 137].

The exploration of LLMs in mental health chatbots gained more attention after 2023. These chatbots primarily use text interactions, employing real-world conversational counseling datasets and social media data for training. Most identified chatbots are designed as Q&A or dialogue systems, providing empathetic and accurate responses for emotional support [17, 18, 27, 63, 65, 78, 89, 100, 103, 104, 114, 150, 151, 157]. Another area of research facilitates conversational interactions with chatbots and includes detection systems. These systems identify users at risk of certain disorders or suicidal ideation and often aim to provide diagnostic results or treatment recommendations [3, 11, 29, 35, 102, 117]. Furthermore, some chatbots enable personalized conversations by allowing users to select counselor strategies and responses, offering more user control [104]. Additionally, chatbots have been developed to improve conversational strategies and enhance psychotherapy support, offering more effective and personalized interactions [58, 68]. Overall, our review identified 22 LLM-based chatbots. We summarize the category of mental health concerns each aimed to serve, the chatbot name, model(s) each chatbot was based on, their target population, and potential ethical concerns in Table 3.

These mental health chatbots can be categorized into five main groups, addressing general mental health, depression, anxiety, suicide ideation, and stress. We summarize each in more detail below:

- **General mental health:** 14 out of 22 chatbots target general mental health or well-being [3, 17, 18, 58, 63, 65, 78, 89, 100, 103, 104, 117, 151, 157]. These chatbots utilize a variety of models, with several leveraging advanced tuning techniques on LLaMA models, such as **Low-rank Adaptation (LoRA)**, which efficiently updates the linear layers of LLMs through low-rank matrix factorization. Others incorporate models like GPT-2 and GPT-3, trained on responses crafted by mental health counselors. They further apply methods such as Zero-shot DSC prompting and CoT to generate more human-like responses. GPT-4 is similarly utilized, particularly with carefully designed prompts for counseling-specific instruction tuning. Additionally, there are multilingual chatbots based on Chinese question-answering language models, such as WenZhong and PanGu, which are built on models like GPT-2 and GPT-3.5. Some projects further enhance chatbots' capabilities by involving real psychiatrists and patients, or by incorporating principles from therapies like **Cognitive Behavioral Therapy (CBT)** into prompt design, enabling the chatbots to facilitate more clinically informed conversations. In certain models, like MindGuide, LangChain is

Table 3. Summary of LLM-based Mental Health Chatbots

Category	Year	Authors	Chatbot Name	Model(s)	Target Population	Major Ethical Concerns
General mental health	2024	Singh et al.	MindGuide	GPT-4	General	Privacy issues; over-dependency
	2024	Kumar et al.	N/A	LLaMA2-7B	General	Accuracy in document-based diagnoses
	2024	Yu and McGuinness	N/A	DialoGPT, GPT-3.5	General	Hallucinated content; ethical use in clinical settings
	2024	Na	CBT-LLM	GPT-3.5-turbo-16k	General	Accuracy in handling cognitive distortions
	2024	Abubakar et al.	N/A	LLaMA-13B	General	Privacy concerns
	2024	Kang et al.	Assistant-Instruct	LLaMA2-7B, ChatGLM2-6B, GPT-4	General	Privacy concerns; over-dependency
	2023	Lai et al.	Psy-LLM	WenZhong and PanGu	General	Content misuse; privacy concerns
	2023	Zheng et al.	N/A	Llama-7B and DialoGPT	General	Privacy issues; risk of biased responses due to specific training data used
	2023	Chen et al.	N/A	ChatGPT	General	Hallucinated content; biased responses
	2023	Chen and Liu	N/A	GPT-2	General	Limited contextual understanding; content misuse
	2023	Qui et al.	PsyChat	ChatGLM2-6B	General	Inappropriate responses; privacy concerns
	2023	Qui et al.	SMILE	ChatGPT	General	Hallucinated content; over-dependency
	2023	Liu et al.	ChatCounselor	GPT-4	General	Ethical use in clinical settings
	2023	Pentina et al.	Replika	GPT-3	General	Emotional dependency; privacy issues
Depression	2023	Qin et al.	N/A	GPT-3.5-turbo and text-davinci-003	General	Biased content; lack of professional oversight
	2023	Yao et al.	N/A	GPT-2	Mothers	Misuse of advice for severe symptoms; privacy issues
	2023	Sezgin et al.	N/A	GPT-4 and LaMDA	Mothers	Inappropriate responses; privacy concerns
Anxiety	2023	Dewi and Fahmi	N/A	ChatGPT	College students	Over-dependency; potential reinforcement of social anxiety
	2021	Lee et al.	Counsellor Chatbot	GPT-2 and DialoGPT	General	Content misuse; risk of biased responses
Suicide ideation	2023	Bhaumik et al.	MindWatch	GPT-3.5 and Llama2	General	Misuse in crisis situations; privacy issues
	2023	Fu et al.	N/A	GPT-3.5 (gpt-3.5-turbo-16k)	General	Misuse in crisis situations; over-dependency
Stress	2024	Dongre	EmLLM	Falcon-7B	Workplace employees	Privacy issues; accuracy of stress prediction

employed to better structure interactions and manage memory, while **Retrieval-Augmented Generation (RAG)** is used in other models to improve response grounding and reduce hallucinations. The target population for all these chatbots is the general public.

- **Depression and PND/PPD:** Three chatbots focused on depression-related conditions [102, 114, 150]. Some use the CoT technique for prompt construction with models such as ChatGPT (gpt-3.5-turbo) and GPT-3 (text-davinci-003), serving dual purposes of conversation and detection. Other studies have fine-tuned GPT-2 and prompted publicly accessible LLMs like GPT-4 (via ChatGPT) and LaMDA (via Bard), drawing on mental health FAQ resources. The target audience extends beyond the general population to include groups with a higher susceptibility to depression, such as postpartum women.
- **Anxiety:** Two chatbots are dedicated to addressing anxiety [27, 68], primarily based on GPT models like ChatGPT, GPT-2, and DialoGPT. Some applications employ summarization models to enhance chatbot response generation in counseling. The target demographic encompasses the general population and specific groups like college students.

- **Suicide ideation:** Two chatbots specialize in suicide ideation [11, 35]. These models integrate custom LLMs, including the state-of-the-art LLaMA2, and utilize prompt engineering techniques. Alternatively, they construct prompts for input into GPT-3.5 (gpt-3.5-turbo-16k) to tailor the conversation appropriately. The intended users are also the general population.
- **Stress:** The last chatbot, EmLLM [29] uses Falcon-7B fine-tuned with QLoRA to manage stress via wearable sensor data. It targets workplace employees and personalizes psychotherapy based on real-time stress levels.

Overall, these studies in mental health chatbots have showcased a blend of sophisticated LLMs, primarily the GPT and Llama series, combined with advanced methods like fine-tuning, prompt engineering, chain-of-thought, zero-shot prompting, few-shot prompting, LoRA, RAG, and instruction tuning. These chatbots primarily operate through text interactions, employing real-world conversational counseling datasets and social media data for training. This combination of cutting-edge models and tailored approaches signifies a significant step forward in creating responsive, empathetic, and effective digital mental health tools.

4.2 Ethical Concerns of LLM-based Mental Health Chatbots

The deployment of LLM-based mental health chatbots can raise significant ethical concerns, particularly for vulnerable populations. By understanding these concerns, users and developers can better select and improve chatbot models that are appropriate for their specific needs. According to Table 3, major ethical concerns we identified include:

- **Content misuse [65, 17, 150]:** Chatbots like Psy-LLM, Chen and Liu's model, and Yao et al.'s model risk disseminating harmful advice or misusing sensitive information. Younger populations are especially vulnerable due to the potential for misunderstanding information. Thus, robust content moderation and user verification processes are essential to mitigate these risks.
- **Privacy concerns [117, 58, 104, 29]:** Stronger privacy protection mechanisms and transparency are needed. Chatbots like MindGuide, Assistant-Instruct, PsyChat, and EmLLM must ensure robust privacy measures. Further research is required to anonymize sensitive user information and implement stringent data access controls to maintain trust and safety of all populations.
- **Biased, inappropriate, and hallucinated content [151, 157, 18, 102, 68]:** Chatbots should align more closely with established authoritative mental health guidelines to prevent low-quality advice. Models like Yu and McGuinness's, Zheng et al.'s, Chen et al.'s, Qin et al.'s, and the Counsellor Chatbot need rigorous bias testing and validation against diverse datasets. This is essential for vulnerable populations such as students or those with limited access to traditional services, as they may heavily rely on such chatbots for guidance.
- **Emotional and user dependency [103, 100, 27]:** Another concern is dependency and cognitive stagnation, where users may overly rely on chatbots like SMILE, Replika, and Dewi and Fahmi's model. This over-reliance can diminish independent thinking and emotional resilience. Implementing guidelines to balance interaction is essential, especially for college students and mothers, who are at formative stages of developing coping mechanisms and are particularly at risk of emotional dependency.
- **Concerns in clinical and crisis situations [89, 78, 11, 35]:** Most chatbots provide mental health support but lack the clinical foundation to replace human clinicians. Models like CBT-LLM and ChatCounselor are best for early-stage, less severe symptoms and as supplementary resources. For high-stakes scenarios, such as suicide ideation, chatbots like

Table 4. Technical Evaluation Metrics for Mental Health Chatbots

Metric	Description	Indication
Perplexity	Measures model's overall prediction accuracy.	Indicates the chatbot's linguistic prediction capabilities.
ROUGE-L	Assesses text similarity based on the longest common subsequences.	Ensures responses are relevant and contextually appropriate.
BLEU-1/2/3/4	Compares machine output with human reference.	Reflects the precision of language generation.
Distinct-1/2/3	Measures response diversity in word and phrase usage.	Demonstrates the chatbot's ability to generate varied responses.
METEOR	Balances precision and recall, considering synonyms and paraphrases.	Provides a more holistic view of linguistic quality.
Vector Extrema	Computes the vector average of all words in the response to measure extremeness.	Evaluates the semantic extremeness of the responses, aiding in understanding response appropriateness.
BERTScore	Measures precision, recall, and F1 score using contextual embeddings.	Captures semantic similarity more effectively, important for nuanced mental health discussions.
Empathy%	Quantifies the percentage of responses that reflect understanding and compassion.	Critical for evaluating the chatbot's capacity for empathetic engagement.

MindWatch and Fu et al.'s model should be enhanced with comprehensive crisis management protocols and robust human oversight to handle the situations responsibly.

4.3 Performance Evaluation Metrics

Evaluating the performance of LLMs is critical to their adoption and usage. Some studies have employed either technical evaluation metrics, human evaluation metrics, or both to assess the performance of LLM-based mental health chatbots. Technical evaluation metrics are typically used to assess the technical performance and outputs of the models. We summarize the metrics utilized in past relevant studies in Table 4, describing their purposes and indications in the chatbot context.

The most commonly used technical metrics include perplexity, ROUGE-L, and BLEU-1/2/3/4. Perplexity measures how well the model predicts a sample, with lower values indicating better predictive performance [54]. ROUGE-L assesses the similarity of the generated text to a reference text, focusing on the longest common subsequences [76], which is critical for ensuring the chatbot's responses are on point. BLEU scores evaluate the correspondence between a machine's output and a human's, emphasizing precision in language generation [98]. In addition, Distinct-1/2/3 is frequently used to gauge the diversity of the generated responses [70]. Higher scores in these metrics indicate a greater variety of words and phrases, reflecting the chatbot's ability to produce unique and varied responses. Other metrics include METEOR [10], which balances precision and recall, Vector Extrema [34], BERTScore (measuring precision, recall, and F1 score) [155], and quantifying the sentences containing empathy [150].

Human evaluation metrics are used to evaluate whether chatbot responses align with mental health guidelines in practice. These evaluations often involve experts or student volunteers who assess the chatbots based on several criteria. Common metrics employed in these assessments are summarized in Table 5.

Common metrics for human evaluation include Helpfulness, Fluency, Relevance, Logic, Informativeness, Understanding, Consistency, Coherence, Empathy, Expertise, and Engagement. These metrics assess how effectively the chatbot communicates, provides relevant and logical information, and empathetically engages with users. Some studies have employed specific metrics related to psychological counseling, such as Direct Guidance, Approval and Reassurance, Restatement, Reflection, Listening, Interpretation, and Self-disclosure. These metrics are tailored to evaluate the

Table 5. Human Evaluation Metrics for Mental Health Chatbots

Category	Metric	Implication
General metrics	Helpfulness	Assesses the practical utility of the chatbot's responses.
	Fluency	Evaluates the naturalness and flow of the chatbot's language.
	Relevance	Measures how the chatbot's responses pertain to the context of the dialogue.
	Logic	Determines the logical consistency of the chatbot's replies.
	Informativeness	Gauges how informative and helpful the chatbot's responses are.
	Understanding	Assesses the chatbot's ability to comprehend user queries.
	Consistency	Checks for the chatbot's ability to provide uniform responses.
	Coherence	Evaluates the chatbot's ability to maintain topic coherence.
	Empathy	Measures the chatbot's ability to display understanding and compassion.
	Expertise	Assesses the chatbot's ability to provide knowledgeable responses.
	Engagement	Evaluates how well the chatbot keeps the user engaged in conversation.
Counseling metrics	Direct guidance	Assesses the chatbot's ability to provide clear therapeutic direction.
	Approval and reassurance	Measures the chatbot's ability to offer affirmation and comfort.
	Restatement	Evaluates the chatbot's skill in paraphrasing to show understanding.
	Reflection	Gauges the chatbot's ability to reflect on the user's statements.
	Listening	Assesses the chatbot's capacity to exhibit active listening cues.
	Interpretation	Measures the chatbot's ability to interpret the user's statements.
	Self-disclosure	Evaluates the chatbot's use of self-disclosure to build rapport.
Reliability metric	Krippendorff's Alpha	Determines the consistency of evaluations across different human evaluators.

chatbot's ability to mirror therapeutic communication strategies. Finally, Krippendorff's Alpha is also mentioned as a metric for assessing the reliability of human evaluations, ensuring that the ratings provided by different evaluators are consistent and dependable.

4.4 Integration with Conventional Mental Health Services

The increasing use of chatbots in mental health services has led to diverse integration methods with traditional healthcare services, aiming to improve accessibility, efficiency, and effectiveness. Currently, conventional rule-based chatbots are more commonly integrated into traditional healthcare. However, LLM-based chatbots offer notable advantages, providing more adaptive and personalized support due to their advanced language processing capabilities. Here, we explore potential methods for integrating them into traditional healthcare services, offering insights for researchers to explore further use cases and applications.

- **Therapy programs:** LLM-based chatbots can complement therapy by offering **cognitive behavioral therapy (CBT)** exercises, mood tracking, and psychoeducation between human therapist sessions. This continuous support reinforces therapeutic techniques and tracks patient progress. Examples like Woebot [33] and Wysa [51] show successful CBT interventions, while LLM-based chatbots promise more nuanced support due to advanced language capabilities.
- **Crisis hotlines:** Integration of LLM-based chatbots with crisis hotlines, such as the Crisis Text Line, can better manage high call volumes and provide immediate assistance. These chatbots handle initial interactions, triage situations, and offer coping strategies. Models like MindWatch [11] could better transfer conversations to human counselors, ensuring timely support and reducing counselor burden.
- **Automated administrative coordination:** LLM-based chatbots can streamline administrative tasks such as appointment scheduling and post-therapy follow-ups. Integrat-

ing chatbots with healthcare management systems and linking chatbot interactions with **electronic health records (EHR)** systems can enable a more seamless flow of information. This automation reduces administrative workload and enhances patient engagement.

- **Remote self-monitoring and support:** For patients with chronic mental health conditions, LLM-based chatbots can provide continuous remote monitoring and support. They can check in with patients regularly, track their symptoms, and alert healthcare providers if any concerning changes occur. Integrating LLM-based chatbots with telemedicine platforms such as the Talkspace app [88] has shown promise in providing ongoing support and improving patient outcomes.
- **Educational resources and support groups:** LLM-based chatbots can enhance traditional healthcare by providing easy access to personalized educational content and virtual support groups. With recent advancements in digital humans, these chatbots can potentially offer even more realistic interactions to facilitate dynamic community interactions.

5 Limitations

While LLMs offer significant benefits in mental health chatbots, they also present a range of limitations and challenges, especially when considering their possible usage within organizational and business contexts. First, there's the risk of generating harmful content, including model hallucination, where LLMs might create misleading information, particularly problematic in sensitive mental health contexts [150]. These chatbots can also face challenges in establishing the deep connection essential in client-therapist interactions within clinical psychology. Memory limitations of LLMs can result in a lack of conversational continuity, negatively impacting the user experience [81]. Additionally, inconsistent communication styles due to evolving training models can cause user confusion. Another concern is the risk of over-reliance on these AI tools for psychological counseling, as they should be seen as supplementary to professional therapy rather than replacements [50]. Furthermore, challenges like interpretability and inherent biases in LLMs affect the quality of interactions, potentially leading to prejudiced responses [22]. Recent research has also brought attention to the risks of bias and the potential for harmful advice in LLMs, particularly about gender and racial disparities [147]. Thus, ensuring these chatbots' clinical effectiveness, safety, privacy, and equity through rigorous evaluation and testing remains a critical hurdle [9, 32].

Another significant limitation is their exclusive reliance on text-based input, which neglects the essential role of nonverbal communication in counseling [96]. Furthermore, there is a notable gap in the range of mental health conditions that current chatbots address. For example, most models do not specifically target complex disorders like bipolar and obsessive-compulsive disorder. Lastly, it's essential to recognize that there is a lack of LLM-based mental health chatbots specifically tailored for business, corporate, or organizational settings where mental disorders are significant concerns.

6 Future Directions of Mental Health Chatbots for Organizational and Business Contexts

Based on these limitations, we propose four major areas for future development and application of LLM-based mental health chatbots. These directions consider the limitations of extant approaches, specifically within business and organizational contexts.

6.1 Chatbot Performance Improvement

Future efforts could incorporate multimodal LLMs to interpret and respond to non verbal cues for more effective mental health chatbots, enhancing their understanding of users' emotional states.

Integrating LLMs with virtual and augmented reality may create more engaging therapeutic interactions [77]. Additionally, incorporating facial emotion detection from computer vision can improve communication, approximating human counseling quality [59]. Research could also develop LLM-based solutions for a broader range of mental health conditions, expanding the scope and applicability of these chatbots.

6.2 Model Selection and Development for Specific Organizational and Workplace Settings

Existing LLM-based mental health chatbots primarily utilize the GPT series and Llama models. However, recognizing the potential of a broader array of LLMs is crucial. With over 20 models available to the public, such as PaLM [19], FLAN-T5 [105], and Alpaca [123], exploring these alternatives could improve mental health applications. Notably, models tailored for mental health, like MentaLLaMA [148] and Mental-LLM [147], which fine-tune LLaMA-2 and Alpaca/FLAN-T5, show promise for more nuanced interactions. Fine-tuning these models on mental health social media data and mental health dialogues is critical for ensuring that the responses align with existing mental health guidelines, which is essential for patient safety and recommendation effectiveness.

To further enhance LLMs in mental health chatbot development, researchers can investigate advanced techniques to enhance the chatbot's understanding and responsiveness to organizational contexts. For instance, by implementing context-aware algorithms that leverage LLMs to generate useful context information from preceding text, chatbots could significantly improve their ability to understand and respond to complex conversation threads and company culture [143]. This could involve using LLMs to predict the next sentence or create abstractive summaries like titles or topics, thus providing a richer context for each interaction. Such an approach, generative context-aware fine-tuning, can be distilled during the fine-tuning of self-supervised speech models. This technique, combined with sentiment analysis and affective computing [15, 116], could enable chatbots to interpret the emotional content of user interactions more effectively, allowing for responses that are empathetically aligned with the user's mood.

Additionally, adaptive learning techniques could be integrated to refine the chatbot's performance based on real-time user feedback [31, 95]. Specifically, **reinforcement learning from human feedback (RLHF)** effectively aligns LLMs with human preferences in other domains. However, given the resource-intensive nature of RLHF, **RL from AI feedback (RLAIF)** emerges as another practical alternative [67]. It could streamline the learning process, particularly when directly using LLMs for reward-based feedback, optimizing chatbots for better user engagement and therapeutic outcomes. Furthermore, incorporating transfer learning can be another direction to enhance chatbot capabilities by improving prompt quality and quantity [43]. Chatbots can apply knowledge from other domains or even languages to the mental health context by using strategies such as zero-shot prompting with pseudo-parallel prompts [153].

6.3 Integration with Other Technologies

Integrating LLM-based mental health chatbots with other technologies, such as apps, IoT devices, wearables, and interactive games, is crucial for enhancing their utility and effectiveness. Future research should focus on integrating additional data sources, such as biometric data (e.g., heart rate, sleep patterns) from wearables, to improve understanding users' emotional states. Additionally, IoT devices can enhance mental health monitoring and intervention by providing insights into users' environments and routines, enabling a more proactive and personalized approach to care. Combining these technologies with interactive games could also improve mental health literacy and provide engaging educational experiences for potential users [73].

6.4 Evaluation

Another critical area for future development in LLM-based mental health chatbots is their evaluation. Traditional language evaluation metrics are insufficient for assessing text generation quality in mental health care. Future research should develop new evaluation metrics tailored for mental health chatbots, including rigorous clinical trials and controlled testing to gain clinical approval [81]. Collaborating with mental health professionals is essential to establish realistic and clinically relevant evaluation criteria, which are pivotal for validating chatbots' safety, efficacy, and therapeutic value across various scenarios.

7 Research Sample of an LLM-based Mental Health Chatbot

In this section, we demonstrate how our findings can guide future chatbot designs and implementations by presenting our own LLM-based mental health project as a case study. This case study illustrates potential research contributions addressing key limitations identified in prior literature. Specifically, we aim to overcome the challenges of generating contextually relevant, guideline-aligned responses in mental health chatbots, while also addressing the need for comprehensive evaluation methods to assess their effectiveness. We focus on three critical areas. First, we fine-tune LLMs for domain-specific adaptation to mental health conversations, enabling these models to efficiently generate contextually relevant responses. Second, we align these models with established mental health guidelines to ensure therapeutic reliability. Finally, we develop comprehensive approaches to evaluate both the technical and non-technical aspects of response quality. We further compare these approaches against existing methods to highlight advancements and address current limitations in the field.

To achieve this, we propose a fine-tuned LoRA-based approach on top of LLaMA2-7B, specifically designed to align with established mental health guidelines. These guidelines are categorized into seven major areas identified across authoritative sources, such as the Mental Health Gap Action Programme Intervention Guide and the National Alliance on Mental Illness Language Guide [92, 141]. The categories include avoiding victimizing language, avoiding blaming language, avoiding constant criticism, using person-centered language, avoiding derogatory language, being neutral and supportive, and carefully dealing with trauma content indicators. To ensure compliance with these guidelines, we developed a set of evaluation scripts based on measurable entities identified from the guidelines. These scripts focus on identifying specific words and language patterns that align with the guidelines, using multiple language datasets such as the Hate Speech and Offensive Language Dataset, Wikipedia Talk Labels, the Moral Foundations Dictionary, and various emotion lexicons [163–166]. LoRA was chosen for its efficiency in fine-tuning a small subset of model parameters with limited computational resources. The fine-tuning process was then optimized through the following iterative steps:

- **Initial evaluation:** Outputs from the fine-tuned model, using the CounselChat dataset [167], serve as inputs for our evaluation scripts.
- **Language assessment:** The evaluation scripts analyze the model's outputs by identifying lexical and syntactic patterns that align with or deviate from mental health guidelines. This includes measuring elements such as victimizing language, blaming language, and hostility. For instance, victimizing language is identified through word counts, while sentiment analysis detects hostility or constant criticism.
- **Parameter adjustment:** Fine-tuning parameters, including the learning rate η, LoRA rank r, and scaling factor α, are adjusted based on evaluation metrics to improve response coherence and enhance overall response quality.

Table 6. Technical Evaluation Scores of Proposed Model vs. Benchmark Models

Model Type	Model	BLEU	ROUGE	BERTScore	Perplexity
MH LLM	Proposed Model	0.3641	**0.0613**	**0.866**	266.538
General LLM	GPT-4	0.250	0.0243	0.816	360.487
General LLM	GPT-3.5	0.1161	0.0426	0.827	269.917
General LLM	Falcon-7B	0.0257	0.0556	0.860	274.396
MH LLM	Mental_Alpaca	0.1564	0.0338	0.834	**232.054**
MH LLM	MentalLlama	**0.3824**	0.0484	0.817	264.753

- **Iterative improvement:** This cycle of assessment and adjustment is repeated until the chatbot's responses consistently align with the identified mental health guidelines.

For the preliminary model evaluation, we compared our fine-tuned LLaMA2-7B model with a set of benchmark models, including both general LLMs and LLMs specifically tailored for **mental health (MH)**. Table 6 below compares the performance of these models using technical metrics commonly employed in previous studies, including BLEU, ROUGE, BERTScore, and Perplexity. The best performances appear in boldface.

Overall, our model demonstrated the strongest performance in ROUGE and BERTScore, indicating its superior ability to generate relevant, contextually appropriate, and semantically similar responses, which are crucial for effective mental health conversations. Additionally, our BLEU score shows that our model outperforms most benchmarks, except for MentalLlama, reflecting its precision in generating text that closely aligns with human references. Our model also achieved a lower Perplexity than most models, indicating its superior linguistic prediction capabilities. The primary novelty of this study lies in the careful alignment of the chatbot's responses with existing mental health guidelines used by practitioners, ensuring that the output is both effective and safe.

In addition to employing technical evaluation metrics (e.g., BLEU, ROUGE, BERTScore, and Perplexity), we sought to also develop capabilities for experts to assess various aspects of the generated outputs. To facilitate this expert evaluation, we developed a user-friendly interface that allows for blinded comparison between the outputs of our proposed model and benchmark models across various dimensions. This interface is designed to enhance user engagement and ensure an unbiased evaluation of the chatbot's performance. Figure 1 below illustrates this evaluation interface.

The prompt section (red box 1) presents the user query that the chatbot addresses. This allows users to input particular prompts for the LLM-based approaches to generate responses. The answers produced (red box 2) display method-blinded responses from both our model and the benchmark model. Since these responses are blinded, they allow for unbiased comparison. Finally, the toggles (red box 3) enable users to rate each response on a 1-5 Likert scale, assessing criteria such as relevance, empathy, clarity, and more. This setup ensures a structured and objective evaluation of the chatbot's performance and can significantly facilitate the human expert evaluation process. Organization's human resources departments may consider using such an interface to evaluate the capabilities of various mental health LLM chatbot outputs to carefully select the approaches most relevant for their industry.

8 Implications for Business and Management

The emergence of LLM-based mental health chatbots marks a pivotal development for business and management, especially as workplace mental health becomes increasingly critical [159]. Although only one of the identified LLM-based chatbots specifically addresses workplace mental

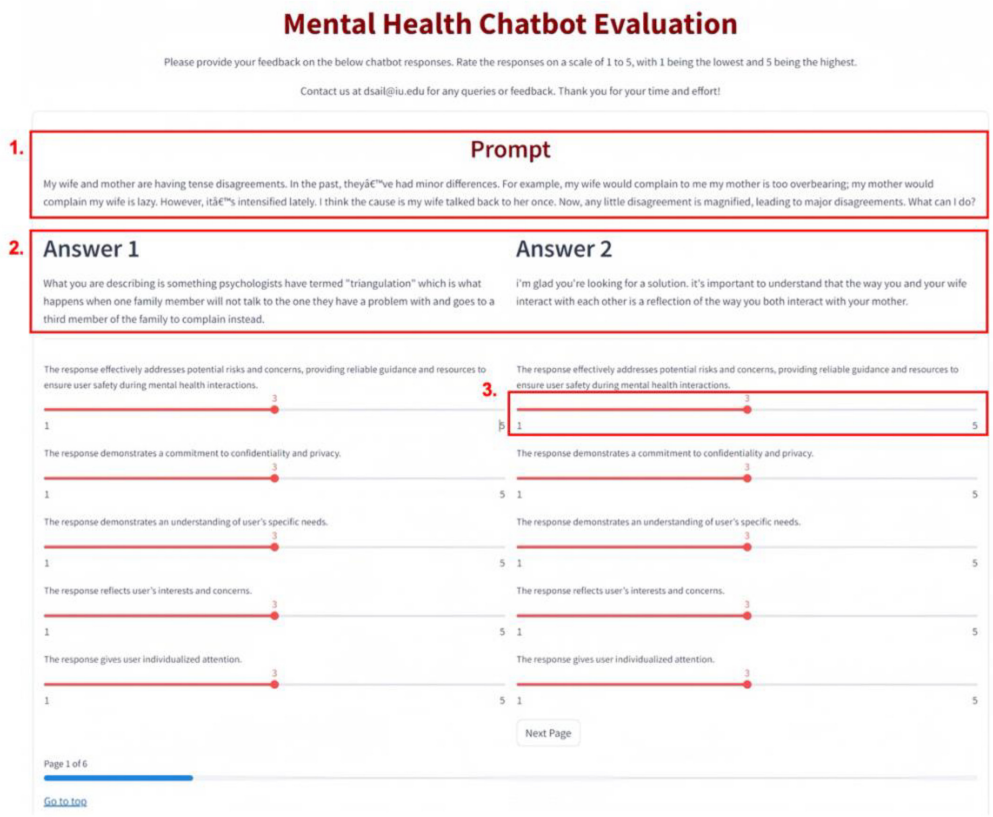

Fig. 1. Chatbot response evaluation interface.

health, its potential impact is substantial [29]. With workers increasingly vulnerable to mental health disorders [140], immediate and effective support systems are needed. These chatbots could influence multiple business contexts, which we describe below.

8.1 Service Providers

Introducing LLM-based mental health chatbots holds significant implications for mental health providers, particularly in diverse and evolving workplace settings. These chatbots offer major advancements due to their deep language comprehension and context-aware personalization, surpassing non-LLM chatbots [16]. For mental health providers, this means delivering tailored care on a large scale, addressing nuanced mental health issues, and providing supportive advice aligned with each client's unique emotional and psychological state client [120]. While these chatbots complement rather than replace human expertise, integrating them into practice extends providers' reach and enhances care delivery. This is vital for addressing institutionalized bias and ensuring equitable care across all employee levels [159]. However, continuous refinement is needed to prevent perpetuating existing biases.

8.2 Employee Insurance

Unlike traditional chatbots, LLM-based chatbots offer greater reliability and sophistication in handling complex mental health conversations [50]. Their consistent high-quality support makes

Table 7. Comparison of Potential Pricing Models for LLM-based MH Chatbots

Pricing Model	Description	Additional Features for Possible MH Chatbots	Estimated Usage Cost
Pay-per-query	Charges based on the number of queries	Suitable for fluctuating demand, crisis intervention protocols	First 10 free, $0.1 per query afterward
Subscription	Fixed monthly or annual fee	Unlimited queries, various support levels, integration with telemedicine platforms	7-day free trial, $30–$50/month for up to 1,000 queries
Tiered pricing	Different pricing levels based on usage	Varies by tier, from basic to premium support, targeted support for less common mental disorders	Starts at $25/month for up to 1,000 queries, higher tiers cost more

them appealing for inclusion in health insurance policies. Businesses legally required to offer health insurance could integrate these chatbots into plans, addressing disparities in mental health service coverage. Mental health services are often reimbursed at lower rates than physical health services, leading to variable insurance options [160]. LLM-based chatbots provide a cost-effective solution, offering broader access to mental health services for employees once they are covered by insurance plans. This increased accessibility can benefit organizations by complementing traditional health services, reducing absenteeism, and lowering labor costs [159]. Reliable and convenient access to these services can boost workplace productivity, leading to a more engaged workforce and a positive organizational environment.

8.3 Mental Health Stigma

Mental health stigma at work refers to negative attitudes, stereotypes, and discrimination individuals may face due to their mental health conditions [86]. This stigma often makes employees feel marginalized and reluctant to seek help due to fear of bias or mistreatment. Introducing LLM-based mental health chatbots in the workplace can help reduce such stigma. These chatbots provide a confidential, non-judgmental platform for mental health assistance [81], offering user-centric and empathetic responses [30]. They can initiate and sustain conversations, providing personalized support and resources, which helps break down barriers to discussing mental health [122]. Additionally, these chatbots can educate employees about mental health, address stigma, encourage open conversations, and foster inclusivity and support for mental well-being.

8.4 Cost and LLM-as-a-Service

Given the computational expense and significant resources required for training and implementing LLMs, commercial applications of LLM-based mental health chatbots could necessitate payment to access the provided APIs. **LLM-as-a-service (LLMaaS)** providers typically explore various payment models. Table 7 showcases potential hypothetical payment structures and estimated usage costs, offering a glimpse into how these models could be structured in the future.

The estimated costs in the table are based on existing mental health chatbot services and typical cost recovery for computational resources. The pay-per-query model is ideal for users with irregular usage, offering free initial interactions, which is appealing during acute needs or crisis interventions. The subscription model suits users with consistent usage, providing a 7-day free trial and budget predictability with a fixed monthly fee, ideal for long-term mental health concerns with minimal to mild symptoms. The tiered pricing model is for those scaling usage over time, with upgrades for more advanced interactions.

9 Discussion

This review critically examines the current landscape of LLM-based mental health chatbots, highlighting both advances and challenges. Compared with existing surveys, the core contributions of

this article reside in the identification of prevailing LLM models and tailored approaches, alongside recurring ethical concerns such as privacy issues, potential biases, and user over-dependency for mental health chatbots, in addition to a larger review of the MH chatbot academic landscape (e.g., historical progression). The collective analysis underscores the need for nuanced ethical guidelines adaptable to different contexts and demographics, which is crucial for responsible implementation by developers and policymakers. Additionally, the synthesis of evaluation metrics reveals a lack of standardization, with significant variation depending on the chatbot's application and intended use. This highlights the necessity for a standardized framework to consistently assess the effectiveness and safety of these chatbots across different settings, from both technical and clinical perspectives. The review also proposes various strategies for integrating chatbots with traditional mental health services, noting the potential and barriers to such integrations, including compatibility with existing clinical workflows and resistance from healthcare professionals. Addressing these barriers requires targeted research and development in the future.

The limitations discussed serve as a roadmap for future research, including chatbot improvement, model development, integration with other technologies, and evaluations. For researchers, we have provided a foundation to use the findings from this review to propose new designs and applications in real-world settings. For practitioners, this review highlights how LLM-based chatbots can transform mental health management within business and organizational contexts. By aggregating findings from existing studies, we uncover the potential for these tools to be integrated into broader business strategies, enhancing employee well-being and productivity. This has direct implications for service providers, employers, and marketers, who can develop tailored mental health support strategies aligned with their specific business objectives.

10 Conclusion

This review synthesizes studies on LLM-based mental health chatbots, highlighting their potential for empathetic interactions and therapeutic conversations. While promising, these tools require continuous innovation. Future advancements should focus on multimodal inputs, expanding the range of mental disorders addressed, refining tuning and learning strategies, integrating advanced technologies, and conducting thorough evaluations to enhance their efficacy and reach.

Integrating diverse research findings, this paper offers new insights into how LLM-based chatbots can enhance mental health support in business and organizational contexts. These chatbots can be strategic assets for workplace mental health, providing scalable solutions that complement traditional therapy. Importantly, our review identifies a critical gap: the lack of chatbots tailored to the unique needs of businesses and organizations. This presents a significant opportunity for future development to improve well-being and productivity in professional environments.

Acknowledgments

We would like to thank the students who assisted in exploring chatbots for this research. Special thanks to Sydney Cook, Sahiti Kadiyala, Atharv Nikam, and Anika Tandon (listed alphabetically based on last name) for their valuable assistance and insights. We appreciate the editorial guidance and the anonymous reviewers' input on this article.

References

[1] Alaa A. Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M. Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *Int. J. Med. Inform.* (2019).

[2] Alaa Ali Abd-Alrazaq, Asma Rababeh, Mohannad Alajlani, Bridgette M. Bewick, and Mowafa Househ. 2020. Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *J. Med. Internet Res.* 22, 7 (2020), e16021.

[3] Abdulqahar Mukhtar Abubakar, Deepa Gupta, and Shantipriya Parida. 2024. A reinforcement learning approach for intelligent conversational chatbot for enhancing mental health therapy. *Procedia Computer Science* 235 (2024), 916–925.
[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303. 08774* (2023).
[5] Aseel Ajlouni, Abdallah Almahaireh, and Fatima Whaba. 2023. Students' perception of using ChatGPT in counseling and mental health education: The benefits and challenges. *Int. J. Emerg. Technol. Learn.* (2023).
[6] American Psychological Association. 2023. *Understanding Psychotherapy and How it Works*.
[7] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM 2 Technical Report. *arXiv preprint arXiv:2305. 10403* (2023).
[8] Suha Assayed, Khaled Shaalan, Sana Al-Sayed, and Manar Alkhatib. 2023. Psychological emotion recognition of students using machine learning based chatbot. Retrieved January 19, 2024 from https://papers.ssrn.com/abstract=4407078
[9] Tamara Babaian and Jennifer Xu. 2024. NLP in Healthcare: Developing interactive integrated collaborative assistants. In *HCI International 2023 – Late Breaking Posters*, 2024. Springer Nature Switzerland, 11–16.
[10] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, June 2005. Association for Computational Linguistics, Ann Arbor, Michigan, USA, 65–72.
[11] Runa Bhaumik, Vineet Srivastava, Arash Jalali, Shanta Ghosh, and Ranganathan Chandrasekaran. 2023. MindWatch: A smart cloud-based AI solution for suicide ideation detection leveraging large language models. *medRxiv* (2023).
[12] Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea. 2021. Understanding the impact of COVID-19 on online mental health forums. *ACM Trans. Manag. Inf. Syst.* 12, 4 (2021), 1–28.
[13] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. GPT-NeoX-20B: An open-source autoregressive language model. *arXiv preprint arXiv:2204. 06745* (2022).
[14] R. Boian, A. M. Bucur, D. Todea, A. I. Luca, and T. Rebedea. 2023. A conversational agent framework for mental health screening: Design, implementation, and usability. (2023). Retrieved from https://psyarxiv.com/t2r3z/download?format=pdf
[15] Michael Chau, Tim M. H. Li, Paul W. C. Wong, Jennifer J. Xu, Paul S. F. Yip, and Hsinchun Chen. 2020. Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS Quarterly* 44, 2 (2020), 933–955.
[16] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2023. When large language models meet personalization: Perspectives of challenges and opportunities. *arXiv preprint arXiv:2307. 16376*.
[17] Qi Chen and Dexi Liu. 2023. Dynamic strategy chain: Dynamic zero-shot CoT for long mental health support generation. *arXiv preprint arXiv:2308. 10444* (2023).
[18] Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305. 13614*.
[19] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PaLM: Scaling language modeling with pathways. *J. Mach. Learn. Res* 24, 240 (2023), 1–113.
[20] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210. 11416* (2022).
[21] Kyungmi Chung, Hee Young Cho, and Jin Young Park. 2021. A chatbot for perinatal women's and partners' obstetric and mental health care: Development and usability evaluation study. *JMIR Med. Inform.* 9, 3 (2021), e18607.
[22] Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311. 13857*.
[23] Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207. 04672* (2022).
[24] Alison Darcy, Aaron Beaudette, Emil Chiauzzi, Jade Daniels, Kim Goodwin, Timothy Y. Mariano, Paul Wicks, and Athena Robinson. 2023. Anatomy of a Woebot® (WB001): Agent guided CBT for women with postpartum depression. *Expert Rev. Med. Devices* 20, 12 (2023), 1035–1049.

[25] Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni. 2022. Chatbots and mental health: Insights into the safety of generative AI. *J. Consum. Psychol.* (2022).
[26] Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nat. Rev. Psychol.* (2023).
[27] Melisa Dewi and Hasanul Fahmi. 2023. Implementation of AI chatbot application for social anxiety problem. *IT for Society* 8, 1 (2023).
[28] Armaan Dhanda, Raman Goel, Sachin Vashisht, and Seba Susan. 2021. Hindi conversational agents for mental health assistance. *Int. J. Appl. Res. Inf. Technol. Comput.* 12, 1to3 (2021), 12–20.
[29] Poorvesh Dongre. 2024. Physiology-driven empathic large language models (EmLLMs) for mental health support. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024. 1–5.
[30] Adrian Egli. 2023. ChatGPT, GPT-4, and other large language models: The next revolution for clinical microbiology? *Clin. Infect. Dis.* 77, 9 (2023), 1322–1328.
[31] Xiangyu Fan and Xi Niu. 2018. Implementing and evaluating serendipity in delivering personalized health information. *ACM Trans. Manag. Inf. Syst.* 9, 2 (2018), 1–19.
[32] Faiza Farhat. 2023. ChatGPT as a complementary mental health resource: A boon or a bane. *Ann. Biomed. Eng.* (2023). https://doi.org/10.1007/s10439-023-03326-7
[33] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Ment. Health* 4, 2 (2017), e19.
[34] Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, Modern Machine Learning and Natural Language Processing Workshop*, 2014. 168.
[35] Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, Juan Zhang, and Bing Xiang Yang. 2023. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *arXiv [cs.AI]*. Retrieved from http://arxiv.org/abs/2308.15192
[36] Futures Recovery Healthcare. 2021. Top Barriers to Mental Health Treatment.
[37] Hannah Gaffney, Warren Mansell, and Sara Tai. 2020. Agents of change: Understanding the therapeutic processes associated with the helpfulness of therapy for mental health problems with relational agent MYLO. *Digit. Health* 6 (2020), 2055207620911580.
[38] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209. 14375* (2022).
[39] Yenushka Goonesekera and Liesje Donkin. 2022. A cognitive behavioral therapy chatbot (Otis) for health anxiety management: Mixed methods pilot study. *JMIR Form. Res.* 6, 10 (2022), e37877.
[40] Patricia Gray, Sipho Senabe, Nisha Naicker, Spo Kgalamono, Annalee Yassi, and Jerry M. Spiegel. 2019. Workplace-based organizational interventions promoting mental health and happiness among healthcare workers: A realist review. *Int. J. Environ. Res. Public Health* 16, 22 (2019). https://doi.org/10.3390/ijerph16224396
[41] Stephanie Greer, Danielle Ramo, Yin-Juei Chang, Michael Fu, Judith Moskowitz, Jana Haritatos, et al. 2019. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: Randomized controlled feasibility trial. *JMIR MHealth UHealth* (2019).
[42] Zhijun Guo, Alvina Lai, Johan Hilge Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language model for mental health: A systematic review. *arXiv [cs.CY]*. Retrieved from http://arxiv.org/abs/2403.15401
[43] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints* (2023).
[44] Hannes Hapke, Cole Howard, and Hobson Lane. 2019. *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*. Simon and Schuster.
[45] Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, et al. 2023. Towards a psychological generalist AI: A survey of current applications of large language models and future prospects. *arXiv preprint arXiv:2312. 04578*.
[46] Yuhao He, Li Yang, Xiaokun Zhu, Bin Wu, Shuo Zhang, Chunlian Qian, and Tian Tian. 2022. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: Single-blind, three-arm randomized controlled trial. *J. Med. Internet Res.* 24, 11 (2022), e40719.
[47] Thomas F. Heston. 2023. Safety of large language models in addressing depression. *Cureus* 15, 12 (2023), e50729.
[48] Djoerd Hiemstra. 2001. Using language models for information retrieval. University of Twente.

[49] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).

[50] Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yi-Han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, and Andrew Beam. 2024. Large language models in mental health care: A scoping review. *arXiv preprint arXiv:2401. 02984.*

[51] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth* 6, 11 (2018), e12106.

[52] Muhammad Imran Ismael, Nik Nur Wahidah Nik Hashim, Nur Syahirah Mohd Shah, and Nur Syuhada Mohd Munir. 2022. Chatbot system for mental health in Bahasa Malaysia. *Journal of Integrated and Advanced Engineering (JIAE)* (2022).

[53] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. OPT-IML: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212. 12017* (2022).

[54] Fred Jelinek, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *J. Acoust. Soc. Am.* (1977).

[55] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv [cs.CL]*. Retrieved from http://arxiv.org/abs/2310.06825

[56] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. ACM, New York, NY, USA. https://doi.org/10.1145/3544548.3581503

[57] Angela Joerin, Michiel Rauws, and Mary Lou Ackerman. 2019. Psychological artificial intelligence service, Tess: Delivering on-demand support to patients and their caregivers: Technical report. *Cureus* 11, 1 (2019), e3972.

[58] Cheng Kang, Daniel Novak, Katerina Urbanova, Yuqing Cheng, and Yong Hu. 2024. Domain-specific improvement on psychotherapy chatbot using assistant. *arXiv preprint arXiv:2404. 16160* (2024).

[59] Rajiv Khosla and Mei-Tai Chu. 2013. Embodying care in Matilda: An affective communication robot for emotional wellbeing of older people in Australian residential care facilities. *ACM Transactions on Management Information Systems (TMIS)* 4, 4 (2013), 1–33.

[60] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-LLM: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401. 06866* (2024).

[61] Darlene R. King, Guransh Nanda, Joel Stoddard, Allison Dempsey, Sarah Hergert, Jay H. Shore, and John Torous. 2023. An introduction to generative artificial intelligence in mental health care: Considerations and guidance. *Curr. Psychiatry Rep.* 25, 12 (2023), 839–846.

[62] R. M. Krausz, D. Ramsey, F. Wetterlin, K. Tabiova, and A. Thapliyal. 2019. Accessible and cost-effective mental health care using E-mental health (EMH). *Adv. Psychiatr.* (2019).

[63] Ayush Kumar, Sanidhya Sharma, Shreyansh Gupta, and Dharmendra Kumar. 2024. Mental healthcare chatbot based on custom diagnosis documents using a quantized large language model. In *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2024. IEEE, 1–6.

[64] Harsh Kumar, Yiyi Wang, Jiakai Shi, Ilya Musabirov, Norman A. S. Farb, and Joseph Jay Williams. 2023. 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*, 2023. ACM, New York, NY, USA. https://doi.org/10.1145/3544549.3585614

[65] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-LLM: Scaling up global mental health psychological services with AI-based large language models. *arXiv preprint arXiv:2307. 11991* (2023).

[66] Shalini Lal. 2019. E-mental health: Promising advancements in policy, research, and practice. *Healthc. Manage. Forum* 32, 2 (2019), 56–62.

[67] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. RLAIF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv preprint arXiv:2309. 00267* (2023).

[68] John Lee, Baikun Liang, and Haley Fong. 2021. Restatement and question generation for counsellor chatbot. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, 2021. Association for Computational Linguistics, Stroudsburg, PA, USA. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.1

[69] Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E. Kraut, and David C. Mohr. 2023. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digit. Med.* 6, 1 (2023), 236.

[70] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510. 03055* (2015).

[71] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201. 05273*.

[72] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. StarCoder: May the source be with you! *arXiv preprint arXiv:2305. 06161* (2023).

[73] Tim M. H. Li, Michael Chau, Paul W. C. Wong, Eliza S. Y. Lai, and Paul S. F. Yip. 2013. Evaluation of a web-based social network electronic game in enhancing mental health literacy for young people. *J. Med. Internet Res.* 15, 5 (2013), e80.

[74] Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Xuying Meng, Siqi Fan, Peng Han, Jing Li, Li Du, Bowen Qin, and Others. 2023. FLM-101B: An open LLM and how to train it with ∃100 k budget. *arXiv preprint arXiv:2309. 03852* (2023).

[75] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with AlphaCode. *Science* 378, 6624 (2022), 1092–1097.

[76] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, July 2004. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[77] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! An empathic virtual agent delivers behavior change health interventions. *ACM Trans. Manag. Inf. Syst.* 4, 4 (2013), 1–28.

[78] June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. ChatCounselor: A large language models for mental health support. *arXiv preprint arXiv:2309. 15461*.

[79] Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. 5487–5496.

[80] Nicola Ludin, Chester Holt-Quick, Sarah Hopkins, Karolina Stasiak, Sarah Hetrick, Jim Warren, and Tania Cargo. 2022. A chatbot to support young people during the COVID-19 pandemic in New Zealand: Evaluation of the real-world rollout of an open trial. *J. Med. Internet Res.* 24, 11 (2022), e38743.

[81] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings, 2023.* American Medical Informatics Association, 1105.

[82] Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, Jindrich Helcl, and Ankit Srivastava. 2017. Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation. *Cybern. Inf. Technol.* (2017).

[83] Aditya Mahindru, Pradeep Patil, and Varun Agrawal. 2023. Role of physical activity on mental health and well-being: A review. *Cureus* 15, 1 (2023), e33475.

[84] Heran Y. Mane, Amara Channell Doig, Francia Ximena Marin Gutierrez, Michelle Jasczynski, Xiaohe Yue, Neha Pundlik Srikanth, Sourabh Mane, Abby Sun, Rachel Ann Moats, Pragat Patel, Xin He, Jordan Lee Boyd-Graber, Elizabeth M. Aparicio, and Quynh C. Nguyen. 2023. Practical guidance for the development of Rosie, a health education question-and-answer chatbot for new mothers. *J. Public Health Manag. Pract.* 29, 5 (2023), 663–670.

[85] Ashish Mehta, Andrea Nicole Niles, Jose Hamilton Vargas, Thiago Marafon, Diego Dotta Couto, and James Jonathan Gross. 2021. Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (Youper): Longitudinal observational study. *J. Med. Internet Res.* 23, 6 (2021), e26771.

[86] Don Mordecai. 2022. Mental health in the workplace — and the cost of staying silent.

[87] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211. 01786* (2022).

[88] Annie Myers, Lewis Chesebrough, Ruixuan Hu, Meghan Reading Turchioe, Jyotishman Pathak, and Ruth Masterson Creber. 2020. Evaluating commercially available mobile apps for depression self-management. In *AMIA Annual Symposium Proceedings, 2020.* American Medical Informatics Association, 906.

[89] Hongbin Na. 2024. CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. *arXiv preprint arXiv:2403. 16008* (2024).

[90] Maryam M. Najafabadi, Flavio Villanustre, Taghi M. Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. 2015. Deep learning applications and challenges in big data analytics. *J. Big Data* 2, 1 (2015). https://doi.org/10.1186/s40537-014-0007-7

[91] National Alliance on Mental Illness. 2023. *Mental Health by the Numbers.*

[92] National Alliance on Mental Illness. National Alliance on Mental Illness Language Guide. Retrieved from https://www.nami.org/support-education/publications-reports/guides/

[93] Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2022. CodeGen2: Lessons for training LLMs on programming and natural languages. *arXiv preprint arXiv:2305. 02309* (2022).

[94] Williams Nwagwu. 2022. The rise and rise of natural language processing research, 1958–2021. *Research Square.*

[95] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* 35, (2022), 27730–27744.

[96] Regina Pally. 2001. A primary role for nonverbal communication in psychoanalysis. *Psychoanal. Inq.* (2001).

[97] Sumit Pandey and Srishti Sharma. 2023. A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. *Healthcare Analytics* 3, 100198 (2023), 100198.

[98] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002. 311–318.

[99] Sohyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *J. Med. Internet Res.* 21, 4 (2019), e12231.

[100] Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of Replika. *Comput. Human Behav.* 140, (2023), 107600.

[101] Ashish Viswanath Prakash and Saini Das. 2020. Intelligent conversational agents in mental healthcare services: A thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems* 12, 2 (2020), 1.

[102] Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, diagnose and chat: Towards explainable and interactive LLMs-augmented depression detection in social media. *arXiv preprint arXiv:2305. 05138.*

[103] Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. *arXiv preprint arXiv:2305. 00450.*

[104] Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan. 2023. PsyChat: A client-centric dialogue system for mental health support. *arXiv preprint arXiv:2312. 04262.*

[105] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 1 (2020), 5485–5551.

[106] Kiran Ramesh, Surya Ravishankaran, Abhishek Joshi, and K. Chandrasekaran. 2017. A survey of design techniques for conversational agents. In *International Conference on Information, Communication and Computing Technology*, 2017. Springer, 336–350.

[107] Xiaozhe Ren, Pingyi Zhou, Xinfan Meng, Xinjing Huang, Yadao Wang, Weichao Wang, Pengfei Li, Xiaoda Zhang, Alexander Podolskiy, Grigory Arshinov, et al. 2023. PanGu: Towards trillion parameter language model with sparse heterogeneous computing. *arXiv preprint arXiv:2303. 10845* (2023).

[108] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: A big data-AI integration perspective. *IEEE Trans. Knowl. Data Eng.* (2019).

[109] R. Rosenfeld. 2000. Two decades of statistical language modeling: Where do we go from here? *Proc. IEEE Inst. Electr. Electron. Eng.* 88, 8 (2000), 1270–1278.

[110] Sanket Sanjay Sadavarte and Eliane Bodanese. 2019. Pregnancy companion chatbot using Alexa and Amazon Web Services. In *2019 IEEE Pune Section International Conference (PuneCon)*, December 2019. IEEE. https://doi.org/10.1109/punecon46936.2019.9105762

[111] Intissar Salhi, Kamal El Guemmat, Mohammed Qbadou, and Khalifa Mansouri. 2021. Towards developing a pocket therapist: An intelligent adaptive psychological support chatbot against mental health disorders in a pandemic situation. *Indones. J. Electr. Eng. Comput. Sci.* (2021).

[112] Srija Santhanam, Balamurugan Ms, Manoj Kumar Rajagopal, et al. 2023. Amity–A hybrid mental health application. *arXiv preprint arXiv:2305.11871* (2023).

[113] Emma L. van der Schyff, Brad Ridout, Krestina L. Amon, Rowena Forsyth, and Andrew J. Campbell. 2023. Providing self-led mental health support through an artificial intelligence-powered chatbot (Leora) to meet the demand of mental health care. *J. Med. Internet Res.* 25 (2023), e46448.

[114] Emre Sezgin, Faraaz Chekeni, Jennifer Lee, and Sarah Keim. 2023. Clinical accuracy of large language models and Google search responses to postpartum depression questions: Cross-sectional study. *J. Med. Internet Res.* 25 (2023), e49240.

[115] Jillian Shah, Bianca DePietro, Laura D'Adamo, Marie-Laure Firebaugh, Olivia Laing, Lauren A. Fowler, Lauren Smolar, Shiri Sadeh-Sharvit, C. Barr Taylor, Denise E. Wilfley, and Ellen E. Fitzsimmons-Craft. 2022. Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. *Int. J. Eat. Disord.* 55, 9 (2022), 1229–1244.

[116] Suwon Shon, Kwangyoun Kim, Prashant Sridhar, Yi-Te Hsu, Shinji Watanabe, and Karen Livescu. 2023. Generative context-aware fine-tuning of self-supervised speech models. *arXiv preprint arXiv:2312. 09895* (2023).

[117] Aditi Singh, Abul Ehtesham, Saifuddin Mahmud, and Jong-Hoon Kim. 2024. Revolutionizing mental health care through LangChain: A journey with a large language model. In *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, 2024. IEEE, 0073–0078.

[118] Alexander Smith, Stefanie Hachen, Roman Schleifer, Dinesh Bhugra, Anna Buadze, and Michael Liebrenz. 2023. Old dog, new tricks? Exploring the potential functionalities of ChatGPT in supporting educational methods in social psychiatry. *Int. J. Soc. Psychiatry* 69, 8 (2023), 1882–1889.

[119] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530b, a large-scale generative language model. *arXiv preprint arXiv:2201. 11990* (2022).

[120] Inhwa Song, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. 2024. The typing cure: Experiences with large language model chatbots for mental health support. *arXiv preprint arXiv:2401. 14362.*

[121] Heather Stringer. 2023. *Providers Predict Longer Wait Times for Mental Health Services. Here's Who it Impacts Most.* American Psychological Association.

[122] Andrew C. H. Szeto and Keith S. Dobson. 2010. Reducing the stigma of mental disorders at work: A review of current workplace anti-stigma intervention programs. *Appl. Prev. Psychol.* (2010).

[123] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMa model.

[124] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. 2022. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2022.

[125] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211. 09085* (2022).

[126] Abha Tewari, Amit Chhabria, Ajay Singh Khalsa, Sanket Chaudhary, and Harshita Kanal. 2021. A survey of mental health chatbots using NLP. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, 2021.

[127] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nat. Med.* 29, 8 (2023), 1930–1940.

[128] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201. 08239* (2022).

[129] Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A. Fox. 2020. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003. 01200* (2020).

[130] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMa: Open and efficient foundation language models. *arXiv preprint arXiv:2302. 13971* (2023).

[131] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. LLaMa 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307. 09288* (2023).

[132] Department of Health and Human Services. 2023. U. S. Department of Health and Human Services.

[133] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S. Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *Can. J. Psychiatry* 64, 7 (2019), 456–464.

[134] Anna Viduani, Victor Cosenza, Helen L. Fisher, Claudia Buchweitz, Jader Piccin, Rivka Pereira, Brandon A. Kohrt, Valeria Mondelli, Alastair van Heerden, Ricardo Matsumura Araújo, and Christian Kieling. 2023. Assessing mood with the identifying depression early in adolescence chatbot (IDEABot): Development and implementation study. *JMIR Hum. Factors* 10, (2023), e44388.

[135] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-trained language models and their applications. *Proc. Est. Acad. Sci. Eng.* (2022).
[136] Lu Wang, Munif Ishad Mujib, Jake Williams, George Demiris, and Jina Huh-Yoo. 2021. An evaluation of generative pre-training model-based therapy chatbot for caregivers. *arXiv preprint arXiv:2107. 13115*.
[137] Ruyi Wang, Jiankun Wang, Yuan Liao, and Jinyu Wang. 2020. Supervised machine learning chatbots for perinatal mental healthcare. In *2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, (2020). IEEE. https://doi.org/10.1109/ichci51889.2020.00086
[138] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *J. Pac. Rim. Psychol.* 17 (2023). https://doi.org/10.1177/18344909231213958
[139] Workshop BigScience. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*.
[140] World Health Organization. 2022. *Mental Health at Work*.
[141] World Health Organization (WHO). 2020. Mental health gap action programme intervention guide. Retrieved from https://www.who.int/teams/mental-health-and-substance-use/treatment-care/mental-health-gap-action-programme
[142] Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Fabien Scalzo, and Ira Kurtz. 2023. A comparative study of open-source large language models, GPT-4 and Claude 2: Multiple-choice test taking in nephrology. *arXiv preprint arXiv:2308. 04709* (2023).
[143] Zhengxuan Wu and Desmond C. Ong. 2021. Context-guided BERT for targeted aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 14094–14102.
[144] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *arXiv [cs.AI]*. Retrieved from http://arxiv.org/abs/2309.07864
[145] Chloe Xiang. 2023. 'He would still be here': Man dies by suicide after talking with AI chatbot, widow says.
[146] Bei Xu and Ziyuan Zhuang. 2022. Survey on psychotherapy chatbots. *Concurr. Comput.* (2022).
[147] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–32.
[148] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. MentaLLaMA: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, 2024. 4489–4500.
[149] Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science* (2023).
[150] Xuewen Yao, Miriam Mikhelson, S. Craig Watkins, Eunsol Choi, Edison Thomaz, and Kaya de Barbaro. 2023. Development and evaluation of three chatbots for postpartum mood and anxiety disorders. *arXiv preprint arXiv:2308. 07407*.
[151] H. Yu and Stephen McGuinness. 2024. An experimental study of integrating fine-tuned LLMs and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence* (2024), 1–16.
[152] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. GLM-130B: An open bilingual pre-trained model. *arXiv preprint arXiv:2210. 02414* (2022).
[153] Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv [cs.CL]*. Retrieved from http://arxiv.org/abs/2301.07069
[154] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205. 01068* (2022).
[155] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904. 09675*.
[156] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303. 18223* (2023).
[157] Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of LLMs. *arXiv preprint arXiv:2308. 11584*.
[158] Lina Zhou, Shimei Pan, Jianwu Wang, and Athanasios V. Vasilakos. 2017. Machine learning on big data: Opportunities and challenges. *Neurocomputing* 237, (2017), 350–361.
[159] 2022. *Workplace Mental Health & Well-Being*. U.S. General Surgeon.

[160] 2022. *Access Challenges for Covered Consumers and Relevant Federal Efforts*. United States Government Accountability Office.
[161] 2024. *Introducing Llama 3.1: Our most capable models to date*. Meta.
[162] Introducing Meta Llama 3: The most capable openly available LLM to date. Retrieved from https://ai.meta.com/blog/meta-llama-3/
[163] Hate Speech and Offensive Language Dataset. Retrieved from https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset
[164] Wikipedia Talk Labels. Retrieved from https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Aggression/4267550
[165] Toxic Comment Classification Challenge. Retrieved from https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[166] Moral Foundations Dictionary. Retrieved from https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/moral-foundations-dictionary/
[167] CounselChat. Retrieved from https://github.com/nbertagnolli/counsel-chat

Received 31 January 2024; revised 20 September 2024; accepted 2 October 2024