A Comprehensive Literature Review on [LLMs For Mental Health Therapeutics]

Roop Kumar   Thi Thuy Dung Tran   Chatdanai Pangwisate

Kasey Kelly   Naga Sowjanya Barla

Abstract

This study presents a comprehensive review of the capabilities, limitations, and ethical implications of using large language models (LLMs) in mental health therapeutics. We explore the evolution of LLMs, from early models like ELIZA to modern architectures such as GPT-4 and FLAN-T5, and evaluate their effectiveness across mental health prediction and support tasks including stress, depression and suicide risk detection. Drawing on fine-tuning strategies, zero- and few-shot prompting and instruction learning, the study analyzes how models such as Alpaca, FLAN-T5 and GPT variants perform using benchmark datasets like Reddit Suicide Watch and EmpatheticDialogues. We highlight the trade-offs between general-purpose and fine-tuned models in accuracy, empathy, and real-time responsiveness. Special focus is given to the privacy, ethical and safety risks inherent in deploying AI chatbots in clinical settings. The study concludes with actionable recommendations for prompt engineering, dataset curation, personalization and crisis response safeguards to improve the reliability of LLMs for mental health applications.

Contents

1 Introduction

Mental health is a critical component of public health, with profound implications for individuals and societies worldwide. According to the U.S. National Institute of Mental Health (NIMH), 22.8 percent of U.S. adults experienced mental illness in 2021, while the World Health Organization (WHO) reports that mental health disorders account for 30 percent of the global non-fatal disease burden, making them a leading cause of disability. These statistics highlight the urgent need for innovative solutions to address the growing mental health crisis, particularly in light of the global shortage of mental health professionals.

Large Language Models (LLMs) represent a groundbreaking advancement in artificial intelligence, characterized by their ability to process and generate human-like text at an unprecedented scale. Built on sophisticated architectures such as the Transformer, these

models have demonstrated remarkable capabilities in understanding context, generating coherent responses, and performing complex language tasks. The evolution of LLMs, from early statistical models to modern systems like GPT-4 and LLaMA, has been driven by innovations in model architecture, training techniques, and computational efficiency.

The Transformer architecture, introduced by Vaswani et al. (2017), serves as the foundation for most contemporary LLMs. Its key innovation, the self-attention mechanism, enables models to dynamically weigh the importance of different words in a sequence, overcoming the limitations of earlier recurrent neural networks (RNNs) and long short-term memory (LSTM) models. Unlike RNNs, which process text sequentially and struggle with long-range dependencies, Transformers parallelize computations and capture global context more effectively. This architectural superiority has led to significant improvements in tasks requiring sophisticated language understanding, such as sentiment analysis, dialogue generation, and mental health support applications.

This study aims to address this gap by providing the first detailed review of LLM applications in mental health care, using public datasets including the Reddit Mental Health Dataset Low et al. (2020) and EmpatheticDialogues Rashkin et al. (2019) to clarify four key areas.

1. Datasets, models, and training techniques used in mental health applications.

2. Applications and conditions targeted by LLMs, along with validation measures.

3. Discrepancies between current tools and their practical implementation in clinical settings.

4. Ethical and privacy considerations surrounding the use of LLMs in sensitive mental health contexts.

2 History of LLMs

article lipsum

2.1 History of Large Language Models

Several researchers in artificial intelligence (AI) have attempted to get computers to interpret visual data since the 1950s. Achieving this objective appeared to be rather simple in the early days of AI. One early development is *ELIZA*, a computer program that enables a machine to engage in natural language conversation, developed by Weizenbaum (1966). ELIZA is regarded as one of the first examples of a natural language model and was a rule-based chatbot created to mimic a conversation with a therapist.

A Japanese engineer named Kunihiko Fukushima was motivated by the findings of Hubel and Wiesel (1959) to create the cognitron, one of the first deep neural networks, and the neocognitron, its successor, in the 1970s. Fukushima (1969) claimed that he had some success teaching the neocognitron to detect handwritten numbers, but the particular learning strategies he employed did not appear to work well for more difficult visual tasks. However, the neocognitron served as a significant source of inspiration for subsequent deep neural network approaches, such as convolutional neural networks, or ConvNets as most industry participants refer to them.

Neural network applications to problems involving ordered sequences, like words, date back to the 1980s, when recurrent neural networks (RNNs) were introduced Elman (1990). These networks were motivated by theories about how the brain processes sequences. This consecutive processing of a sentence and subsequent brain activations that represent it

serve as a loose inspiration for recurrent neural networks. The RNN differs primarily in that its hidden units have extra "recurrent" connections; each hidden unit is connected to both the other hidden units and to itself (dashed arrows).

A Swiss research team came up with a solution in the late 1990s: each unit in a recurrent neural network should have a more complicated structure, with unique weights that decide what information can be "forgotten" and what can be sent on at the next time step. The more complicated units were dubbed *long short-term memory* (LSTM) units. Although the nomenclature may be perplexing, the idea is that these units enable more "short-term" memory to be retained throughout the sentence processing process. Like the ordinary weights in a conventional neural network, the specialized weights are also learned using back-propagation Gers et al. (2000).

When the Internet rose in the 2000s, many researchers began to create large-scale datasets, sometimes referred to as "web corpora". Because of their capacity to efficiently handle huge volumes of data, statistical language models (LMs) had largely overtaken symbolic models in the majority of NLP tasks by 2009. Around 2012, neural networks gained popularity in image processing, and these models were then modified for language modeling. Google switched to Neural Machine Translation (NMT) for its translation service in 2016. Sequence-to-sequence (seq2seq) deep LSTM networks were used to accomplish this transformation before the development of transformers.

Google researchers introduced the transformer design in their ground-breaking article, "Attention Is All You Need," at the 2017 NeurIPS conference Vaswani et al. (2017). By mainly expanding upon the attention mechanism created in 2014, this effort sought to improve the seq2seq technology that was first launched in 2014. BERT was first implemented in 2018 Devlin et al. (2019) and quickly gained traction. While the original transformer model contains encoder and decoder blocks, BERT utilizes simply the encoder component.

Although the decoder-only model GPT-1 was first presented in 2018 Radford et al. (2019), GPT-2, which was released in 2019 Radford et al. (2019), attracted a lot of interest. GPT-2 was initially suppressed by OpenAI because of worries about its potential for malicious usage. The model was further developed with the release of GPT-3 in 2020 Brown et al. (2020), but as of 2024, it can only be accessed through an API; there is no way to download and execute it locally. But the introduction of the ChatGPT browser-based interface in 2022 captivated the audience and attracted a lot of media attention. Launched in 2023, GPT-4 was praised for its multimodal capabilities and increased accuracy, which some refer to as the "holy grail" OpenAI (2023). The architecture and number of parameters for GPT-4 have not been revealed by OpenAI.

The quick development of natural language processing is reflected in this evolution, which has produced increasingly complex and powerful language models.

2.2 The advent of RNN with LSTM and Transformer Architecture with Self Attention

Understanding the foundations of deep learning and natural language processing (NLP) is important before exploring self-attention and the transformer. Neural networks are used in the machine learning branch of deep learning to extract relationships and patterns from data. A group of algorithms known as neural networks are made to identify patterns and create connections in data. Besides, NLP is a subfield of AI that studies how people and computers communicate using natural language.

NLP models have historically been created to extract statistical patterns and linguistic structures from large amounts of textual data. By capturing the connections between words, phrases, and sentences, these models hope to produce writing that is both logical and pertinent to its context. NLP has benefited greatly from the emergence of several important models, including decision trees, support vector machines (SVMs), maximum entropy models (MEM) Berger et al. (1996), hidden Markov models (HMM) Rabiner (1989), and n-gram models Jelinek and Mercer (1980). These models serve as the basis for traditional NLP modeling, propelling developments in text categorization, information extraction, speech recognition, machine translation, and sentiment analysis.

In the past, NLP and computer vision operated as distinct fields within artificial intelligence. However, the fundamental models resulting from these breakthroughs have also found use in the field of computer vision due to the recent developments in LLMs and their application in processing large volumes of data. Because the Self-Attention mechanism and transformer design have a theoretical underpinning, this convergence makes it possible to integrate research efforts in both domains Vaswani et al. (2017). The development of LLMs based on transformer architecture and the self-attention mechanism has transformed NLP. By expanding the boundaries of natural language understanding, these sophisticated LLMs have changed the field of natural language processing. It is essential to further investigate the topic and gain a thorough grasp of the self-attention mechanism and transformer architecture in order to lead this fascinating wave of LLMs and realize their full potential.

With its particular advantages over the traditional convolutional neural network (CNN) and recurrent neural network (RNN) architectures, the transformer architecture stands out. The model can process all items in parallel by using self-attention, which allows it to dynamically allocate its attention to various input sequence segments. The model's ability to collect long-range dependencies and contextual information is greatly improved by this capability Vaswani et al. (2017). Furthermore, a review of transformer development shows significant advancements, such as the application of NLP approaches in many fields, the development of hybrid architectures, and the extension of these techniques to computer vision applications Dosovitskiy et al. (2020). The creation of innovative LLMs like GPT Radford et al. (2019) (Generative Pre-Trained Transformer) has been made possible by these developments.

2.3 Limitations of RNNs and LSTMs

Recurrent Neural Networks (RNNs) can be categorized into discrete-time and continuous-time variants Grossberg (2013). A defining feature of RNNs is their cyclic connections, allowing the network to update its current state based on both input and previous states Salehinejad et al. (2017). In certain applications, fully connected RNNs Elman (1990) and selective RNNs Šter (2013), which employ common recurrent cells like sigma cells, have shown promise. However, RNNs struggle to capture long-range dependencies between distant input tokens Šter (2013). Although RNNs are designed to model sequential data, they often struggle with learning long-term dependencies due to issues like vanishing and exploding gradients. Due to these limitations, self-attention mechanisms were introduced as an alternative to traditional RNN-based approaches.

2.4 How Transformers Revolutionized NLP with Multi-Head Self-Attention

By employing multiple attention heads and introducing so-called self-attention to substitute RNNs in the encoder and decoder, the Transformer architecture expanded the mechanism Vaswani et al. (2017). For NMT, and more recently for language modeling Radford et al. (2018) and other downstream tasks Strubell et al. (2018), this design quickly emerged as the de facto state-of-the-art architecture. The Transformer distributes a token's attention across the entire input sequence multiple times. This mechanism allows the model to intuitively capture various semantic and syntactic attributes. Because of this feature, a field of study has emerged that focuses on the attention mechanisms and interpretation of Transformer-based networks Raganato and Tiedemann (2018). According to new research on machine translation (MT) Voita et al. (2019), all attention heads are optional, whereas a small number are specialized for a particular function, such as concentrating on rare words or syntactic dependency relations, and greatly enhance translation performance. However, recent studies have tried to align the mathematical definition of self-attention with the linguistic assumption that attention would be most helpful in a limited local context, such as when translating sentences Hao et al. (2019). For example, in order to enhance the focus on local positional patterns, Shaw et al. Shaw et al. (2018) substitute relative position encoding for the Transformer's sinusoidal position encoding. In order to bias the attention weights towards local locations, a number of studies alter the attention formula Yang et al. (2018). Convolutional modules are used to substitute portions of self-attention, increasing the computational efficiency of the networks as a whole Wu et al. (2019). In order to avoid redundancy among the many attention heads and to promote local attention patterns, Cui et al. (2019) mask out specific tokens when computing attention. All of these contributions have demonstrated the value of locality and the potential for using lightweight convolutional networks to achieve competitive results with fewer parameters Wu et al. (2019). Our study focuses exclusively on the original Transformer architecture, investigating the replacement of fixed attention patterns with learnable self-attention mechanisms in the encoder.

2.5 Therapy Chatbots Built on Pre-Trained LLMs

The development of mental health chatbots has evolved significantly over time, incorporating advancements in artificial intelligence and natural language processing. This progression illustrates how advancements in AI and natural language processing have been harnessed to create more effective and empathetic mental health chatbots, improving accessibility and support for individuals seeking mental health assistance.

The use of content analysis and social interaction patterns has been instrumental in identifying and predicting mental health risks. Meanwhile, large language models (LLMs) in recent years, including GPT-4, PaLM, FLAN-T5, and Alpaca show that big pre-trained models can potentially handle a variety of tasks in zero-shot scenarios, or problems that were not encountered during training. Question answering Omar et al. (2023), logic reasoning Wei et al. (2022), machine translation Brants et al. (2007), and other tasks are examples. Based on hundreds of billions of parameters, several tests have shown that these LLMs have begun to demonstrate the ability to comprehend human common sense beneath natural language and perform appropriate reasoning and inference in accordance with it Bubeck et al. (2023). Among other uses, the ability of LLMs to

comprehend human mental health states using natural language is one specific question that has not yet been addressed. Mental health conditions, including anxiety, major depressive disorder, suicidal thoughts, and others Coppersmith et al. (2015), have been the subject of copious research in the last ten years. Social media's real-time nature and archival features frequently aid in reducing retrospective bias. The wealth of social media data also makes it easier to identify, track, and maybe anticipate risk variables over time. In addition to monitoring and identifying threats, social media platforms could be useful avenues for providing communities at risk with timely support Kruzan et al. (2022).

Mental health issues pose a substantial burden to both individuals and societies across the globe. For example, more than 20 percent of American adults, according to a recent survey, may have at least one mental illness in their lifetime America (2022), and 5.6 percent may have experienced a severe psychotic disease that substantially hinders functioning of Mental Health (2023). The productivity losses from depression and anxiety alone cost the world economy some 1 trillion dollar a year on Mental Illness (2023). The multi-task arrangement has also been investigated in other studies Benton et al. (2017), including the simultaneous prediction of anxiety and depression Sarkar et al. (2022). These models, however, have limited versatility because they are bound to predefined job sets. From a different angle, the use of chatbots for mental health services has been the subject of additional research Cameron et al. (2017). The majority of chatbots are merely rule-based, although they can be strengthened by more sophisticated models Abd-alrazaq et al. (2019). Even while research on enabling AI for mental health is expanding, it's crucial to remember that current methods might occasionally introduce bias and even provide users detrimental advice Chen et al. (2019).

Researchers and practitioners have moved toward larger and more potent language models (e.g., GPT-3 and T5) following the tremendous success of Transformer-based language models like BERT Devlin et al. (2019) and GPT Radford et al. (2018). In the meanwhile, scholars have suggested instruction finetuning, a technique that applies different prompts to various activities and datasets. This method directs a model to execute several tasks inside a single, cohesive framework during the training and generation stages Wei et al. (2022). With tens to hundreds of billions of parameters, these instruction-finetuned LLMs, like GPT-4, PaLM, FLAN-T5, LLaMA, and Alpaca show promising performance on a range of tasks, including question answering, logic reasoning, machine translation Brants et al. (2007), and more. The potential of these LLMs in health-related sectors has been investigated by researchers Jiang et al. (2023). For example, Singhal et al. (2023) improved PaLM-2 on medical domains and obtained 86.5 percent on the MedQA dataset. Likewise, Wu et al. (2019) improved LLaMA on medical publications and shown encouraging outcomes on other biomedical QA datasets. The use of LLMs to address public health issues was investigated by Jo et al. (2023). Jiang et al. (2023) refined a medical language model across a variety of clinical and operational prediction tasks after training it on unstructured clinical notes from the electronic health record. According to their assessment, a model like this can be used in a number of therapeutic activities. The field of mental health has comparatively less work. A few studies investigated LLMs' capacity for emotion reasoning and sentiment processing. Nearer to our research, Lamichhane (2023) and Amin et al. (2023) evaluated ChatGPT's (GPT-3.5) performance on a variety of classification tasks, including stress, depression, and suicide detection. The findings demonstrated that while ChatGPT offers some promise for use in mental health applications, there is still much space for development, as seen by the at least 5–10 percent accuracy and F1-score performance discrepancies. The potential

reasoning ability of GPT-3.5 for reasoning tasks (such as potential stresses) was further assessed by Yang et al. (2023a). Despite the promising results of LLMs in zero-shot settings, most prior studies have primarily focused on this approach and have not extensively explored strategies to further optimize their performance in mental health applications. Despite the promising results of LLMs in zero-shot settings, most prior studies have primarily focused on this approach and have not extensively explored alternative strategies to enhance performance in mental health applications. To bridge this gap, Yang et al. (2023b) introduced Mental-LLaMA, a suite of LLaMA-based models fine-tuned on domain-specific mental health datasets. These models are designed to address a variety of mental health-related tasks, including the detection of depression, suicidal ideation, and psychological stress. However, current research in this area remains limited, with evaluations largely confined to LLaMA and GPT-3.5, indicating a significant opportunity for future work to expand across newer LLM architectures and methodologies.

3 Privacy Concerns

The integration of large language models (LLMs) into mental health support systems introduces significant privacy challenges that must be addressed to ensure ethical and secure use. These challenges primarily involve data storage, usage, and access, necessitating robust measures to mitigate potential risks Volkmer et al. (2024); Iwaya et al. (2022).

3.1 Storage of Data

Storing sensitive user data is a critical concern in LLM-driven mental health applications. Conversations often contain deeply personal information, making secure storage essential. Many AI mental health platforms use cloud-based storage solutions. which, if not properly protected, are susceptible to breaches. Regulations such as the General Data Protection Regulation (GDPR) in Europe impose strict guidelines on data storage, including requirements for anonymization and data minimization. However, compliance varies, and lapses can lead to significant privacy violations Iwaya et al. (2022).

3.2 Use of Data

The utilization of user data by LLM-based mental health platforms raises ethical and privacy concerns. These platforms often use user interactions to refine their algorithms through continuous learning, which can enhance personalization, but also introduce risks related to unintended data exploitation. There is concern that companies may repurpose anonymized data for research or commercial purposes without explicit user consent. Ethical frameworks suggest that informed consent should be obtained before user data is used beyond direct therapeutic engagement Guo et al. (2024).

3.3 Access to Data

Determining who has access to user data is a fundamental issue in LLM-based mental health applications. Although users expect confidentiality, service providers may need access to monitor AI performance and ensure the accuracy of mental health responses. However, concerns arise when third-party entities, such as external researchers,

are granted access to anonymized datasets, raising the potential for data misuse. Mechanisms such as differential privacy and role-based access controls are being explored as potential solutions to mitigate these risks Iwaya et al. (2022).

3.4 Ensuring Privacy in LLM-Based Mental Health Support

Addressing privacy concerns in LLM-driven mental health support requires a multifaceted approach. Firstly, end-to-end encryption should be employed to protect user interactions from unauthorized access. Secondly, federated learning techniques, which allow AI models to be trained without transferring data to centralized servers, can help reduce privacy risks. In addition, regulatory compliance frameworks must be continuously updated to reflect evolving risks and ensure that AI-driven mental health applications adhere to the highest privacy standards. Finally, greater transparency with respect to data collection and processing, along with clear and informed user consent policies, are essential to maintain trust in AI-based mental health services Volkmer et al. (2024); Iwaya et al. (2022).

4 Efficacy of Mental Health Chat Bots

Recent advancements in large language models (LLMs) have opened up more relevance in the field of digital mental health, offering scalable solutions for early detection, diagnosis and support. But the efficacy of the models is a very important consideration.

This section explores the data sources, fine-tuning methods, performance trade-offs, and ethical implications of using AI-driven chatbots in mental health contexts. We highlight the impact of training data biases, discuss the balance between speed and accuracy and examine real-world risks including misinformation, user trust and demographic disparities.

4.1 Model Training and Biases

Chatbots and AI models designed for mental health support are predominantly trained on datasets derived from social media platforms, psychotherapy transcripts, and electronic health records (EHRs). A majority of models in the literature have been fine-tuned specifically for mental health prediction tasks using benchmark datasets such as *Dreaddit*, *DepSeverity*, *SDCNL*, and *CSSRS-Suicide* Arriba-Pérez and García-Méndez (2024). These datasets are often annotated by human experts and sourced primarily from Reddit, enabling models to perform tasks such as binary stress classification, depression detection, and suicide risk estimation. While these data sources provide scalability, they also introduce significant demographic and cultural biases.

Bias remains a persistent challenge. Studies have highlighted how the overreliance on Reddit-based data skews model generalizability, often failing to capture the nuances of underrepresented groups Li et al. (2023); Gbollie et al. (2023). Diagnostic models trained on such datasets may inadvertently reinforce users' distress rather than challenge cognitive distortions, potentially increasing false positives. Furthermore, a lack of diversity in training data has led to measurable racial and gender biases in chatbot responses Arriba-Pérez and García-Méndez (2024). Even expert-annotated corpora are susceptible to embedded stereotypes and confirmation bias, raising concerns about the validity of ground-truth labels Greco et al. (2023).

4.2 Trade-Offs in Performance and Design

Performance trade-offs are evident across model size, inference method, and processing strategy. Larger general-purpose models such as GPT-4 and FLAN-T5 offer faster response times, while smaller fine-tuned models like Mental-Alpaca often achieve higher task-specific accuracy Xu et al. (2024). Improved prompt engineering can enhance inference speed, but often at the cost of response quality. Stream-processing models allow for real-time interaction but demand continuous updating, whereas batch-processing models deliver greater accuracy by processing data in aggregate,albeit with delayed outputs McGorry et al. (2025).

Fine-tuning strategies play a central role in improving performance. Approaches such as instruction tuning and LoRA (Low-Rank Adaptation) have been applied to enhance LLMs for cognitive behavioral therapy (CBT) tasks. These models optimize predictions using cross-entropy loss, reducing divergence from human-labeled data Na (2024).

4.3 Accuracy, Coherence and Logical Structure

Transformer-based models have shown high accuracy in mental health tasks,with some detecting depression at rates exceeding 96%, and chatbot-based models for cognitive impairment reaching over 80% accuracy and 85% recall Greco et al. (2023); McGorry et al. (2025). Fine-tuned models such as Mental-FLAN-T5 consistently outperform zero-shot models like GPT-4 by up to 4.8% in balanced accuracy Xu et al. (2024). However, zero-shot performance remains inconsistent, often fluctuating between 50% and 83% based on prompt structure.

In terms of coherence, rule-based conversational agents like Woebot and Wysa tend to outperform generative models in logical consistency and structure Stade et al. (2024). While GPT-4 responses are generally fluent, they can be overly generic and lack context-specific insight. In contrast, fine-tuned models offer more domain-relevant outputs. Notably, GPT-based models are sensitive to prompt phrasing, leading to variability in output quality and occasional overgeneralization,particularly in nuanced cases such as suicide risk Sejnowski (2023).

4.4 Risk of Misinformation and Ethical Implications

Numerous studies caution against the misuse of LLMs in high-stakes contexts. Despite their sophistication, generative models are prone to producing "falsely reasonable" yet incorrect outputs. This poses a significant risk in mental health settings, where the consequences of misinformation can be severe. The absence of robust, systematic evaluation frameworks for mental health reasoning further compounds the issue Greco et al. (2023). In addition, the use of sensitive training data raises major concerns about privacy, consent, and ethical safeguards Arriba-Pérez and García-Méndez (2024).

The potential for discriminatory outputs is fairly high given the lack of fairness assessments across different subgroups of population. Vulnerable users, including those in crisis, are particularly susceptible to receiving inaccurate or even harmful advice.

4.5 Real-Time Performance vs. Quality

While real-time models like GPT-4 excel in responsiveness, fine-tuned models such as Mental-FLAN-T5 yield more accurate and context-sensitive results with lower computa-

tional demands Xu et al. (2024). Enhancing model accuracy through prompt engineering often increases processing time. However, instruction-fine-tuned models manage to balance inference speed with depth of response more efficiently than zero-shot systems.

4.6   Handling Ambiguity in User Prompts

Fine-tuned models have demonstrated superior contextual understanding of vague or ambiguous queries related to stress or depression. In contrast, zero-shot models frequently default to binary or overly simplified interpretations. Few-shot learning methods have shown to improve GPT-4's performance by around 4.1% on ambiguous mental health queries Xu et al. (2024), indicating the need for more adaptable model architectures in this space.

4.7   User Willingness and Perceptions

User receptivity remains high, with 81.1% of university students reporting willingness to engage with mental health chatbots,though only 4% have done so in practice Gbollie et al. (2023). Younger demographics (18–35) show higher adoption rates, while users in acute distress often prefer human counselors due to trust and safety concerns.

Preferences are influenced by multiple factors. Stigma around seeking therapy makes some users favour AI over traditional face-to-face approaches. Conversational AI models like GPT-4 are generally preferred due to flexible and natural interaction. Rule-based model responses can often be repeated due to a fixed response database. However, concerns around privacy, transparency, and the potential for AI-generated errors remain key barriers to adoption.

4.8   Outlook and Efficacy

Emerging models such as CBT-LLM exemplify domain-specific fine-tuning tailored for psychological support. Evaluations show that such models outperform generic LLMs in both fluency and therapeutic relevance Na (2024). Despite their promise, several researchers question whether LLMs truly "understand" mental health concerns or merely reflect user input in plausible ways,a phenomenon likened to a "reverse Turing test" Sejnowski (2023).

A broader body of research suggests that while LLMs may augment therapists and automate diagnostics, they are not yet reliable standalone tools. Studies have shown marked symptom improvement using chatbot interventions (Hedge's g = 0.64 for depression and 0.7 for distress) Stade et al. (2024), but little impact on long-term psychological well-being. Lastly, transformer-based models like BERT continue to demonstrate high efficacy for mental health classification, though their success remains closely tied to data quality and fine-tuning depth Greco et al. (2023); Li et al. (2023).

5   Prompt Engineering

5.1   Introduction to Prompt Engineering

Prompt engineering refers to the deliberate design and manipulation of the inputs provided to large language models (LLMs) in order to optimise performance across various

tasks. The body of research reviewed below collectively highlights that the structure, clarity, and contextual richness of a prompt significantly affect an LLM's ability to generalise, reason, and generate high-quality responses. A critical analysis of these studies reveals an evolving understanding of how prompt structure influences LLM outputs, not only in simple tasks but also in complex, reasoning-intensive scenarios, offering key implications for their use in sensitive applications such as mental health support.

5.2 Prompt Clarity and Task Generalisation

Garg et al. (2021) examine the in-context learning capabilities of transformer-based models, focusing on their ability to perform simple functions like sorting and pattern recognition. Their findings underscore that prompt structure is particularly crucial for enabling generalisation in tasks that fall just beyond training data, what they call "medial" tasks. Notably, as task complexity increases, the model's performance degrades unless the prompt is structured to support more abstract reasoning. This introduces an important insight: while LLMs can simulate learning through prompt engineering, this ability is fragile and highly dependent on how well the prompt scaffolds the task Garg et al. (2021). Brown et al. (2020) extend this idea through the concept of few-shot learning, showing that providing clear, structured examples within prompts dramatically improves performance. They find that including multiple relevant task examples (as opposed to zero- or one-shot prompts) increases contextual understanding and reduces ambiguity. Together, these studies suggest that prompt engineering is not just about making inputs readable, it is about making the task learnable within the constraints of the LLM's architecture (Brown et al. (2020)). This has direct implications for mental health chatbots, where ambiguous user input must be interpreted with nuance and sensitivity.

5.3 Advanced Prompting strategies: Reasoning and Reflection

While Garg and Brown focus on prompt clarity and structure in general task completion, more recent work by Wang et al. (2022) and Yao et al. (2023) pushes this further by exploring how prompting can shape the reasoning process of LLMs. Wang et al. (2022) Chain-of-Thought (CoT) prompting introduces intermediate reasoning steps that mimic human-like deliberation. Their results show that CoT prompting not only improves task performance but also maintains accuracy as task complexity increases, a contrast to Garg et al. (2021) finding that LLM performance diminishes under complexity. This suggests that introducing a structured reasoning process within prompts can compensate for limitations in generalisation, further supporting the need for more sophisticated prompt strategies (Wang et al. (2022)). Building on this, Yao et al. (2023) propose Tree-of-Thought (ToT) prompting, which enables the model to explore multiple solution paths before selecting the most appropriate one. Unlike CoT, which tends to follow a linear reasoning path, ToT encourages branching thought processes and reflective evaluation. This aligns with decision-making frameworks used in cognitive-behavioural therapy (CBT) and other mental health contexts, where multiple perspectives and reflective thinking are essential. The findings from ToT experiments demonstrate that LLMs can be prompted not just to answer questions, but to think through them in a structured, human-like way, critical for engaging users in therapeutic dialogue (Yao et al. (2023)).

5.4 Reasoning-Action Separation in Interactive Tasks

Further refining this approach, Yao et al. (2022) introduce ReAct prompting, which separates reasoning from action. Here, the model first reasons through the problem, then generates a final output. This division ensures that responses are not just reactive but deliberative, enhancing LLM performance in interactive tasks like question-answering. In the context of mental health, this separation is particularly important: it mirrors the therapeutic process of validating emotions before suggesting actions, a key to empathetic engagement (Yao et al. (2022)).

5.5 Evolving Understanding of Prompt Engineering

Taken together, these studies highlight a central inference: high-quality LLM outputs are not merely a function of the model's size or training data, but of how intelligently they are prompted. While early research like that of Garg et al. (2021); Brown et al. (2020) emphasised prompt clarity and example inclusion, newer approaches like CoT, ToT, and ReAct show that prompting can simulate cognitive processes such as reflection, evaluation, and decision-making. This progression reflects a broader shift in prompt engineering, from focusing solely on inputs to shaping the model's thinking process.

5.6 Implications for Mental Health Chatbots

These findings carry profound implications for the design of LLM-based mental health chatbots. Individuals in psychological distress often struggle to express themselves clearly or to provide structured input. This makes user-initiated prompt quality highly variable. If the chatbot relies solely on direct user input, its responses risk being generic, inappropriate, or even harmful. To address this, the chatbot must act as an intermediary prompt engineer, restructuring, expanding, or interpreting user inputs using internal reasoning frameworks. Techniques such as CoT and ReAct can be embedded into the system to simulate deeper understanding, allowing the model to "think through" user concerns before responding (Wang et al. (2022); Yao et al. (2022)). For example, instead of responding directly to "I feel off," the model could internally generate a series of interpretive questions, reflections, and emotional validations before offering a meaningful reply. Additionally, the insights from ToT prompting can be used to generate multiple potential responses and select the one that best aligns with therapeutic goals, such as validation, reflection, or gentle guidance (Yao et al. (2023)). In this way, the chatbot mimics the thoughtful deliberation of a trained therapist, increasing the likelihood of meaningful engagement. Finally, prompt engineering becomes not just a tool for improving performance but a safeguard. By structuring internal reasoning processes, LLMs can avoid surface-level interpretations and respond in ways that are more aligned with therapeutic best practices, improving both the emotional and clinical efficacy of the chatbot.
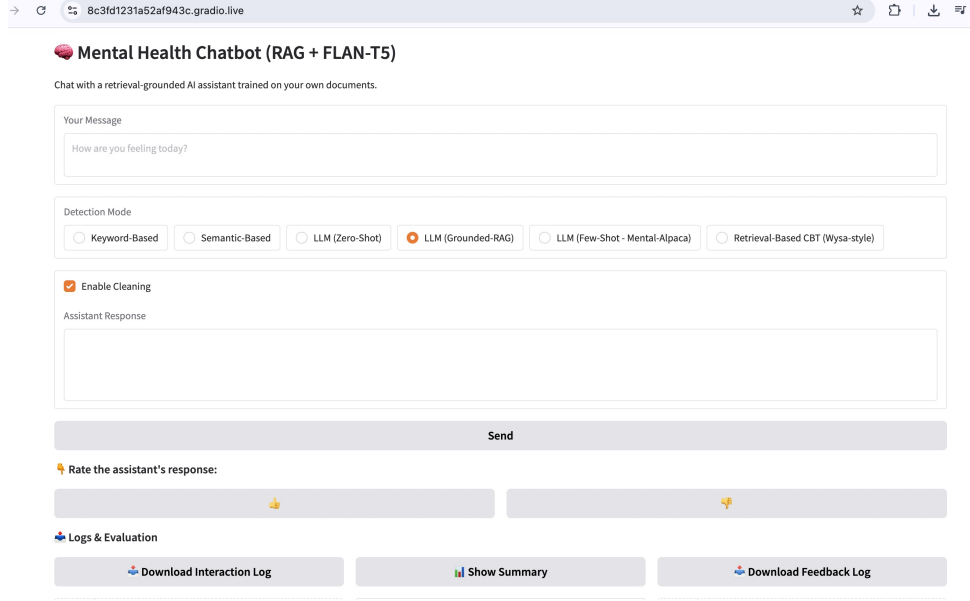
Figure 1: Comparison tool of Different Types of Mental Health Chatbots

6 Development and Assessment of Mental Health Bots

6.1 Bot Comparison Tool

We built this comparison to help understand the different types of mental health chatbots described in the literature. By analyzing models like keyword-based bots, semantic-based bots, zero-shot models, grounded models and CBT-based bots, we evaluated their strengths and weaknesses in handling mental health queries. This comparison aligns with the findings from the literature, showcasing how different approaches, from simple keyword matching to complex retrieval-augmented generation and fine-tuned models, offer varying degrees of empathy, accuracy, and context-awareness but the responses offered were questionable in a lot of contexts.

6.2 Dataset Impact Comparison

This dataset presents a comparative analysis of three chatbot models built upon the GPT-2 base architecture (nsbarla/gpt2-base), each fine-tuned with different training data to explore their responsiveness to emotionally sensitive prompts. One model was trained on a Reddit dataset (nsbarla/gpt2-reddit), another on empathy-rich conversations from Hugging Face (nsbarla/gpt2-empathy) and the third remained as the unaltered baseline. All three models were exposed to the same set of ten distress-related prompts, and their outputs were recorded for comparison. The responses are unique across the models, offering insight into how different training influences affect tone, empathy, and coherence. The empathy-tuned model generally strives to offer supportive or compassionate replies, although some are shallow or lack contextual depth. The Reddit-style model frequently mirrors informal, narrative-driven responses, occasionally incorporating anecdotal elements reflective of real-world user conversations. Conversely, the vanilla model often produces short, vague, or incomplete outputs, indicating limited emotional

understanding and generative capacity. Collectively, the dataset reveals distinct stylistic and qualitative differences, shedding light on how training data shapes a model's ability to engage with sensitive or personal topics.

Prompt	Base Response	Reddit Response	Empathy Response
I'm feeling really down lately. I don't want to wake up.	I was really	So I'm just going to be taking a shower and eating food.	I am sorry. I hope you can find a new job soon.
No one listens to me when I say I'm not okay.	No, I don't care.	No one listens to me when I say I'm not okay.	I'm sorry, I would be so mad at you.
I feel like I'm a burden to everyone around me.	I don't want to be on the wrong side of the spectrum. I want to be a person who wants to be respected and respected and respected by all of the other people who don't really understand how I'm supposed to be. I want to be a person who believes in a person who believes in	I feel like I'm an annoying burden to everybody around me.	Oh no
I don't know if I can keep going.	I can	I am a social-exchange student and I have a really low self-esteem.	I have a small dog that's just a little bit older. I can't do much about it.
I want to feel better, but I can't see a way out.	I want to	I'm doing this because I don't want to be seen as a piece of shit. I just want to be happy. I want to be happy. I want to be happy.	I was in a housework meeting and saw a big group of people trying to make a sandwich.

6.3   Fine-tuned Chatbot with LoRa

The Mistral-7B-Instruct-v0.1 model, fine-tuned on a mental health dataset and optimized using LoRA and 4-bit quantization, is now available on Hugging Face under the model ID nsbarla/mental-health-mistral-lora. It was trained on Hugging Face's platform using GPUs, a process that took over 2-3 hours due to the model's size and the applied optimizations. This fine-tuned model is designed to offer more efficient and contextually relevant responses for conversations related to mental health, making it an ideal resource for applications aiming to support users in need of empathetic interactions and advice. The Mistral Mental Health Bot was trained on a combined dataset consisting of Empathetic Dialogues and Counsel Chat datasets. The Empathetic Dialogues dataset includes conversations designed to encourage empathetic and supportive responses, while the Counsel Chat dataset contains counseling-style dialogues aimed at addressing mental health concerns.

Figure 2: Chatbot with LoRa

7 Critical Analysis

In the development of a mental health chatbot designed to assist individuals experiencing depression, there were several limitations and challenges due to the use of a free platform, the short timeframe of the project (12 weeks), and the complexity of designing an AI-driven tool for such sensitive use cases. This section critically evaluates the chatbot's performance, highlighting its strengths, limitations, and areas for potential improvement.

7.1 Limitations of the Free Platform

The decision to use a free platform to build the chatbot imposed significant constraints on its functionality and development. Free platforms typically offer limited resources compared to paid services, which can impact both the scale and quality of the final product. These platforms often restrict access to advanced model features, such as fine-tuning and customisation, and have fewer capabilities for managing large datasets. For instance, a paid service might provide the ability to train the model on larger and more specific datasets, particularly those related to mental health, which would enhance the chatbot's ability to provide contextually appropriate and empathetic responses. The limited functionalities of the free platform used in this project prevented the implementation of advanced features such as dynamic backpropagation, which is essential for improving the chatbot's understanding of context and refining its output over time. Further, the created was built off of a chatbot with generic, and limited training data, meaning that our ability to fine tune this to the required specifications added another layer of complexity.

7.2 Data Collection and Model Training Challenges

Mental health chatbots require specialized training data that includes therapeutic dialogue, expert-curated mental health advice, and crisis intervention scripts. However, without access to a robust dataset, the chatbot may struggle to provide meaningful and helpful responses. In particular, the chatbot is likely to give generic answers that lack the nuance needed to address the unique experiences of individuals with depression. Furthermore, without fine-tuning the model on mental health-specific data, the chatbot may fail to recognise subtle cues or signs of a more severe mental health crisis, such as suicidal ideation, which could potentially jeopardise the user's safety.

7.3 Ethical and Safety Risks

The ethical considerations in developing a mental health chatbot are paramount, especially when the chatbot is intended to engage with vulnerable individuals. In its current form, the chatbot is prone to several ethical risks, particularly around its handling of sensitive topics like suicide and self-harm. A major concern is the chatbot's tendency to introduce these topics prematurely, even when they have not been mentioned by the user. Such responses can exacerbate distress and disengage users, potentially escalating the situation. The lack of safety mechanisms in the current model makes it difficult to provide a supportive and responsive environment for users in crisis. In more advanced systems, mechanisms like sentiment analysis, emergency alerts, and crisis detection would allow the chatbot to identify when a user may be in immediate danger and refer them to appropriate resources. Without these safety features, the chatbot's current use in mental

health contexts is problematic, as it may inadvertently cause harm by misinterpreting the user's emotional state or failing to escalate critical situations when necessary.

7.4   Personalisation and Continuity of Care

One of the core challenges of the chatbot is its lack of personalisation. Depression is a deeply personal and complex mental health condition, and effective therapeutic tools must be able to adapt to the individual needs of each user. The free platform's limitations hinder the development of such personalised features, as the chatbot currently struggles to maintain continuity in conversations or respond dynamically to the evolving emotional state of the user. In a more advanced system, the chatbot would be able to learn from prior interactions with the user, adjusting its tone and responses to better match the user's emotional context. For example, if the chatbot detects that a user has been expressing consistent feelings of hopelessness or isolation, it could provide more targeted advice or escalate the conversation to a higher level of support. The ability to tailor responses based on the individual's prior inputs is critical in maintaining user engagement and ensuring the chatbot provides ongoing, meaningful support.

7.5   Potential Improvements and Future Directions

To enhance the chatbot's effectiveness, several improvements should be considered in future iterations. Upgrading to a paid platform would allow the chatbot to take advantage of more advanced features, such as model fine-tuning and access to larger, more specialised datasets. With these resources, the chatbot would be able to produce more accurate, contextually relevant, and empathetic responses, particularly in sensitive areas like mental health. Moreover, to improve the model's performance, it is essential to collect more diverse and specialised training data. Collaborating with mental health professionals to curate high-quality datasets, such as therapeutic dialogues, expert-curated resources, and crisis intervention scripts, would ensure the chatbot's responses align with established mental health frameworks. This data would provide the foundation for fine-tuning the model, enabling it to recognise a broader range of emotional cues and handle more complex mental health issues with greater sensitivity. Additionally, incorporating robust safety features is critical. Future versions of the chatbot should be designed with crisis detection mechanisms that can flag critical emotional states, such as suicidal ideation or severe distress, and trigger appropriate responses. These might include providing emergency contact information or automatically escalating the issue to a mental health professional. Integrating sentiment analysis would further enhance the chatbot's ability to gauge the emotional tone of conversations, allowing it to respond more sensitively to the user's state. Personalisation is another key area for improvement. By allowing the chatbot to track and build upon previous conversations, it could provide more individualised support. The chatbot could adapt its responses based on the user's unique emotional context and history, offering a more tailored approach to mental health care. This personalised engagement would be crucial in fostering a sense of connection and support, especially for users who may feel isolated or disconnected. Finally, collaboration with mental health experts throughout the development and fine-tuning stages would ensure that the chatbot adheres to therapeutic guidelines and provides safe, effective support. This input would be invaluable in helping to align the chatbot's responses with best practices in mental health care, ensuring that it offers both practical advice

and compassionate, evidence-based guidance.

8   Conclusion

In summary, while the mental health chatbot developed in this project provides a foundational step toward creating AI-driven support for individuals with depression, its current limitations, stemming from the use of a free platform, limited training data, and a lack of personalisation, restrict its effectiveness in real-world applications. Addressing these limitations by upgrading the platform, refining the model with more specialised data, and incorporating robust safety and personalisation features would significantly enhance the chatbot's ability to provide meaningful, context-sensitive, and safe support. With further development, this chatbot has the potential to become a valuable tool in mental health care, both for individual users seeking support and for therapists managing caseloads.

References

Abd-alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., and Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978.

America, M. H. (2022). The state of mental health in america. https://mhanational.org/issues/state-mental-health-america.

Amin, M. M., Cambria, E., and Schuller, B. W. (2023). Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *arXiv preprint arXiv:2303.03186*.

Arriba-Pérez, F. and García-Méndez, S. (2024). Leveraging llms for real-time mental health predictions. *Arabian Journal for Science and Engineering*.

Benton, A., Mitchell, M., and Hovy, D. (2017). Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

Berger, A., Della Pietra, S. A., and Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Brants, T., Popat, A. C., Xu, P., Och, F. J., and Dean, J. (2007). Large language models in machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., and Kaplan, J. (2020). Language models are few-shot learners. In *Proceedings of NeurIPS 2020*, pages 1877–1901. Accessed: 12 March 2025.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. http://arxiv.org/abs/2303.12712. arXiv preprint arXiv:2303.12712.

Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., Armour, C., and McTear, M. (2017). Towards a chatbot for digital counselling. In *Proceedings of the HCI 2017 Conference.*

Chen, I. Y., Szolovits, P., and Ghassemi, M. (2019). Can ai help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, 21(2):E167–E179.

Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. (2015). From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Cui, Y., Liu, P., Wei, F., Zhou, M., and Wang, H. (2019). A survey on transformer models in natural language processing. In *Proceedings of the 2019 Conference on Natural Language Processing (NLP 2019)*, pages 134–145.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*, arXiv:2010.11929. https://arxiv.org/abs/2010.11929.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Fukushima, K. (1969). Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333.

Garg, S., Tsipras, D., Liang, P., and Valiant, G. (2021). What can transformers learn in-context? a case study of simple function classes. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pages 3505–3514. Accessed: 12 March 2025.

Gbollie, E. F., Bantjes, J., and Jarvis, L. (2023). Intention to use digital mental health solutions. *Digital Health*, 9:1–19.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471.

Greco, C. M., Simeri, A., and Tagarelli, A. (2023). Transformer-based language models for mental health issues: A survey. *Pattern Recognition Letters*, 167:204–211.

Grossberg, S. (2013). Recurrent neural networks. *Scholarpedia*, 8(2):1888.

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., and Li, K. (2024). Large language models for mental health applications: Systematic review. *JMIR Ment Health*, 11:e57400.

Hao, Y., Lu, Z., Liu, J., and Xu, W. (2019). A survey of transformer-based models for natural language processing. In *Proceedings of the 2019 Conference on Natural Language Processing (NLP 2019)*, pages 158–169.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148:574–591.

Iwaya, L. H., Babar, M. A., Rashid, A., and Wijayarathna, C. (2022). On the privacy of mental health apps. *Empirical Software Engineering*, 28(1):2.

Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, volume 2, pages 381–397.

Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H. A., Laufer, I., Punjabi, P., Miceli, M., Kim, N. C., Orillac, C., Schnurman, Z., Livia, C., Weiss, H., Kurland, D., Neifert, S., Dastagirzada, Y., Kondziolka, D., Cheung, A. T. M., Yang, G., Cao, M., Flores, M., Costa, A. B., Aphinyanaphongs, Y., Cho, K., and Oermann, E. K. (2023). Health system-scale language models are all-purpose prediction engines. *Nature*.

Jo, E., Epstein, D. A., Jung, H., and Kim, Y.-H. (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*, pages 1–16.

Kruzan, K. P., Williams, K. D. A., Meyerhoff, J., Yoo, D. W., O'Dwyer, L. C., De Choudhury, M., and Mohr, D. C. (2022). Social media-based interventions for adolescent and young adult mental health: A scoping review. *Internet Interventions*, 30:100578.

Lamichhane, B. (2023). Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Li, H., Zhang, R., and Kraut, R. E. (2023). Systematic review and meta-analysis of ai-based conversational agents for mental health. *npj Digital Medicine*, 6:236.

Low, D. M., Rumker, L., Torous, J., Cecchi, G., Ghosh, S. S., and Talkar, T. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of Medical Internet Research*, 22(10):e22635.

McGorry, P., Gunasiri, H., Mei, C., Rice, S., and Gao, C. X. (2025). The youth mental health crisis: analysis and solutions. *Frontiers in Psychiatry*, 15:1517533.

Na, H. (2024). Cbt-llm: A chinese large language model for cognitive behavioral therapy. arXiv.

of Mental Health, N. I. (2023). Mental illness. https://www.nimh.nih.gov/health/statistics/mental-illness.

Omar, R., Mangukiya, O., Kalnis, P., and Mansour, E. (2023). Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466*.

on Mental Illness, N. A. (2023). Mental health by the numbers. https://nami.org/mhstats.

OpenAI (2023). Gpt-4 technical report. Technical report, OpenAI. Accessed: 2025-04-30.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *Open AI Blog*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Raganato, A. and Tiedemann, J. (2018). Analysis of encoder representations in transformer-based machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 108–118.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Salehinejad, H., Sankar, S., Barfett, J., Colak, E., and Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*.

Sarkar, S., Alhamadani, A., Alkulaib, L., and Lu, C.-T. (2022). Predicting depression and anxiety on reddit: A multi-task learning approach. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 427–435. IEEE.

Sejnowski, T. J. (2023). Large language models and the reverse turing test. *Neural Computation*, 35:309–342.

Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-attention with relative position representations. In *Proceedings of the 2018 Conference on Neural Information Processing Systems (NeurIPS 2018)*, pages 4642–4650.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Aguera y Arcas, B., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S. S., Barral, J., Webster, D., Corrado, G. S., Matias, Y., Azizi, S., Karthikesalingam, A., and Natarajan, V. (2023). Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Stade, E. C., Stirman, S. W., and Ungar, L. H. (2024). Large language models could change the future of behavioral healthcare. *npj Mental Health Research*, 3:12.

Strubell, E., Ganesh, A., and McCallum, A. (2018). Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 3645–3650.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Voita, E., Wieting, J., Firat, O., Lin, X., and Johnson, M. (2019). On the (un)translatability of encoder representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pages 3814–3825.

Volkmer, S., Meyer-Lindenberg, A., and Schwarz, E. (2024). Large language models in psychiatry: Opportunities and challenges. *Psychiatry Res*, 339:116026.

Šter, B. (2013). Selective recurrent neural network. *Neural Processing Letters*, 38:1–15.

Wang, X., Wei, J., Schuurmans, D., Le, Q., and Chi, E. (2022). Self-consistency improves chain of thought reasoning in language models. In *Proceedings of NeurIPS 2022*, pages 3442–3453. Accessed: 12 March 2025.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2022). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Wu, Z., Shen, Y., Zhang, L., Liu, T., and Wang, X. (2019). Convolutional networks for nlp with dynamic receptive fields. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pages 2155–2162.

Xu, X., Yao, B., and Dong, Y. (2024). Mental-llm: Leveraging large language models for mental health prediction via online text data. In *Proceedings of ACM Interact*.

Yang, K., Ji, S., Zhang, T., Xie, Q., and Ananiadou, S. (2023a). On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Yang, K., Zhang, T., Kuang, Z., Xie, Q., Ananiadou, S., and Huang, J. (2023b). Mental-lama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567 [cs]*.

Yang, Z., Dai, A. M., Yang, Y., Salakhutdinov, R., and Cohen, W. W. (2018). Attention is not explanation. In *Proceedings of the 2018 Conference on Neural Information Processing Systems (NeurIPS 2018)*.

Yao, S., Yu, D., Zhao, J., Shafran, I., and Griffiths, T. (2023). Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2023)*, pages 3671–3682. Accessed: 12 March 2025.

Yao, S., Zhao, J., Yu, D., Shafran, I., and Griffiths, T. (2022). React: Synergizing reasoning and acting in language models. In *Proceedings of NeurIPS 2022*, pages 4562–4573. Accessed: 12 March 2025.
csvsimple attachfile2

A Appendix

A Datasets Used

We used the following datasets to fine-tune and evaluate the chatbot models.

- **Reddit Dataset 1: Depression Posts (TF-IDF Features)**
  A structured dataset containing Reddit posts from depression-related subreddits, with text features pre-processed using TF-IDF vectors. We specifically used the file `depression_post_features_tfidf_256.csv` available via Zenodo:
  https://zenodo.org/records/3941387

- **Reddit Dataset 2: Suicide Risk Posts**
  This dataset includes labeled Reddit posts from 500 users related to suicide risk with annotations, making it useful for identifying emotionally sensitive content. Hugging Face:
  https://huggingface.co/datasets/m4faisal/RedditSuicide/commit/5e57cc7e2ebc706b98f

- **Empathetic Dialogues Dataset**
  A conversational dataset containing 25,000+ short textual dialogues grounded in emotional situations, designed to train models to respond empathetically. Hugging Face:
  https://huggingface.co/datasets/facebook/empathetic_dialogues

These datasets were used to fine-tune different variants of the GPT-2 base model, enabling controlled comparisons of style, emotional engagement, and empathy.