Enhancing Mental Health Support through Human-AI Collaboration: Toward Secure and Empathetic AI-Enabled Chatbots

Thesis · September 2024

1 author:

Mara Chu
AutomatedView Technologies

arXiv:2410.02783v1 [cs.CY] 17 Sep 2024

Enhancing Mental Health Support through Human-AI Collaboration: Toward Secure and Empathetic AI-Enabled Chatbots

RAWAN ALMAKINAH, AI in Complex Systems Lab, University at Albany, SUNY, USA
ANDREA NORCINI-PALA, SUNY Downstate Health Sciences University, USA
LINDSEY DISNEY, University at Albany, SUNY, USA
M. ABDULLAH CANBAZ, AI in Complex Systems Lab, University at Albany, SUNY, USA

Access to mental health support remains limited, particularly in marginalized communities where structural and cultural barriers hinder timely care. This paper explores the potential of AI-enabled chatbots as a scalable solution, focusing on advanced large language models (LLMs)—GPT v4, Mistral Large, and LLama V3.1—and assessing their ability to deliver empathetic, meaningful responses in mental health contexts. While these models show promise in generating structured responses, they fall short in replicating the emotional depth and adaptability of human therapists. Additionally, trustworthiness, bias, and privacy challenges persist due to unreliable datasets and limited collaboration with mental health professionals. To address these limitations, we propose a federated learning framework that ensures data privacy, reduces bias, and integrates continuous validation from clinicians to enhance response quality. This approach aims to develop a secure, evidence-based AI chatbot capable of offering trustworthy, empathetic, and bias-reduced mental health support, advancing AI's role in digital mental health care.

CCS Concepts: • **Chatbot, Digital Mental Health, Collaborative Intelligence, Generative Artificial Intelligence, Large Language Models, Federated Learning**;

Acknowledgement

This paper has been submitted for review to the Association of Computing Machinery (ACM) CHI25 conference on Human Factors in Computing Systems.

1 Introduction

In recent years, the global mental health crisis has reached unprecedented levels, with the World Health Organization (WHO) reporting a 25% increase in anxiety and depression during the COVID-19 pandemic alone [30]. Mental health issues have become a significant global concern, with millions of cases going undetected and untreated due to various structural and attitudinal barriers [5, 11, 16]. Structural barriers, such as limited access to mental health services and high costs, prevent individuals from receiving timely and effective care [12]. Attitudinal barriers, including the stigma surrounding mental health and lack of awareness, further exacerbate the problem, leading many to avoid seeking

Authors' Contact Information: Rawan AlMakinah, ralmakinah@albany.edu, AI in Complex Systems Lab, University at Albany, SUNY, Albany, New York, USA; Andrea Norcini-Pala, andrea.norcinipala@downstate.edu, SUNY Downstate Health Sciences University, Brooklyn, New York, USA; Lindsey Disney, ldisney@albany.edu, University at Albany, SUNY, Albany, New York, USA; M. Abdullah Canbaz, mcanbaz@albany.edu, AI in Complex Systems Lab, University at Albany, SUNY, Albany, New York, USA.

Manuscript submitted to ACM

help even when it is available [5, 41]. As a result, untreated mental health conditions often worsen, leading to adverse physical, economic, and emotional outcomes [4, 37].

To address these challenges, digital mental health solutions have emerged as a promising alternative to traditional mental health services. The rapid development of technologies, increased internet accessibility, and widespread use of smartphones have paved the way for innovative approaches to mental health care [12]. Among these solutions, AI-enabled chatbots have gained considerable attention due to their ability to provide cost-effective, scalable, and accessible mental health support. Chatbots such as Woebot [15], Wysa [21], and ViTalk [12] have demonstrated their potential in offering empathic interaction and cognitive-behavioral therapy (CBT), a widely used intervention in treating mental health challanges like anxiety and depression [15, 22]. These AI-enabled systems not only mimic human-like interactions but also offer anonymity and 24/7 availability, making them an appealing option for individuals hesitant to seek traditional mental health services due to stigma or other concerns [6, 19].

The rise of large language models (LLMs), such as GPT-3 and GPT-4, has further revolutionized the field of AI and its application to mental health care [2]. These models, trained on vast amounts of text data, possess the ability to generate coherent, human-like responses in real-time, making them well-suited for use in chatbots aimed at providing psychological support. LLMs can be fine-tuned to understand and respond to specific mental health needs, including crisis intervention and emotional support, allowing them to play a critical role in mental health first aid [27, 36]. By simulating natural conversations, these AI systems have the potential to fill the gaps in mental health services, particularly in underserved communities where access to mental health professionals is limited [15].

However, despite the promising potential of AI-enabled chatbots in mental health, significant challenges remain. Ethical and practical concerns regarding the reliability, security, and trustworthiness of these systems have hindered their full integration into healthcare frameworks [6, 9]. Ensuring that chatbots are not only effective but also empathetic—able to recognize and respond to users' emotional states with reduced bias and enhanced security is critical for gaining public trust and promoting widespread adoption. Moreover, the collaboration between data scientists and mental health practitioners is essential to address these concerns and ensure that AI-enabled chatbots are aligned with evidence-based mental health practices [18].

In this paper, we focus on analyzing large language models (LLMs) for their capacity to comprehend and respond to emergency mental health scenarios, specifically focusing on mental health first aid. Following this analysis, we propose a conceptual design for an AI-enabled chatbot that integrates state-of-the-art advancements in AI, human-computer interaction (HCI), and mental health care practices. This interdisciplinary approach is critical to address the growing mental health needs, where AI and HCI can help overcome accessibility and stigma-related barriers. By adopting a human-AI collaboration approach, we aim to build a trustworthy, secure, and reduced bias AI-enabled chatbot, with mental health professionals in the loop to ensure ethical and practical challenges are met. Our framework also suggests the integration of federated learning and encryption techniques to further enhance the security and reliability of AI chatbots in mental health care systems.

2 Methodology

2.1 Knowledge/Comprehension Measurement

Prompt engineering is the process of carefully crafting the input provided to an AI model in order to generate desired outputs [38]. In the context of using LLMs like ChatGPT, prompt engineering plays a crucial role in guiding the model to respond appropriately and effectively, especially in specialized or sensitive domains like mental health. By tailoring prompts to include specific instructions, context, or even role-based personas, users can control the behavior of the model and improve its performance on complex tasks [34]. This is particularly important when utilizing off-the-shelf AI tools, as these models do not inherently understand the full context of the problem they are addressing without clear guidance.

For example, when generating therapeutic responses, instructing the model to act as a professional counselor through prompt engineering ensures that the tone, language, and structure of the response are aligned with professional standards. Without such prompts, the model might produce generic or inappropriate responses, potentially diminishing the effectiveness of AI-assisted support - or worse, causing harm. Figure 1 demonstrates an example of such a prompt, where specific instructions were given to simulate responses as if from a licensed mental health professional.

To evaluate the capabilities of AI-enabled chatbots in replicating the expertise of mental health professionals, we adopted a comparative approach using two large language models (LLMs): ChatGPT and Mistral. The goal was to measure the models' ability to comprehend and provide supportive responses to complex mental health issues.

We developed specific *prompts* to include the *persona (or profiles)* for the LLMs to simulate the behavior of mental health professionals. These profiles included contextual information designed to reflect the therapeutic language and professional demeanor of licensed counselors. The purpose of creating such profiles was to ensure that the LLMs responded with the level of professionalism and empathy required in mental health interactions. Figure 1 presents an example of a prompt crafted for one of the data points, incorporating the therapist's background, education, and the patient's condition. It is important to note that each prompt used in our analysis featured a unique persona tailored to the therapist's qualifications and experience, as noted in the dataset. This personalized approach aimed to enhance the LLMs' ability to communicate empathetically and utilize appropriate mental health terminology, drawing from the 'CounselChat' dataset [7].

> **You are a female having a Doctorate in Behavioral Health, a Licensed Professional Clinical Counselor, and a Certified Counselor in Mental Health in a Primary Care Setting. You received the following question from a person suffering from a mental health issue. What would be your response? Do not provide an unnecessary response. Reply as if you were talking to the person directly. The question is:** "I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here. I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it. How can I change my feeling of being worthless to everyone?

Fig. 1. LLMs Prompt Example

2.2 Question Set Selection

The CounselChat dataset consisted of user-submitted questions on a wide range of mental health topics, with responses provided by certified counselors. For this study, we focused on questions related to eight crisis-related categories: anxiety, depression, grief and loss, self-harm, stress, anger management, trauma, and military issues. This filtering process resulted in approximately 700 questions.

To refine the analysis further, we identified the ten most-viewed questions on the platform, which predominantly fell within the categories of anxiety, depression, anger management, and military issues. These top 10 questions were then selected for manual analysis.

2.3 Analysis of Response Generation and Comparison for Top 10 Questions

For this part of the analysis, we utilized two of the most popular models: OpenAI's *ChatGPT v4* and its open-source competitor, *Mistral Large*. The selected questions were input into both models, using the respective prompts shown in Figure 1. The responses generated by the LLMs were compared to those of human counselors, with a focus on assessing strengths and weaknesses in empathy, comprehension, and therapeutic value.

2.4 Preliminary Qualitative Analysis of Therapist vs LLM Generated Responses

We began our analysis with a manual comparison of responses generated by ChatGPT and Mistral against those provided by human therapists. This preliminary qualitative assessment allowed us to explore the differences in tone, empathy, and overall communication style. It quickly became evident that while AI-generated responses can be coherent and professional, they often lack the nuanced emotional intelligence found in human responses.

Key distinctions emerged from this analysis: (1) The AI responses tended to have a polished but somewhat generic tone, delivering psychoeducational content in a way that felt mechanical and impersonal. This lack of a "human touch" made them easily identifiable as machine-generated. (2) In contrast, therapist responses were more concise and dialogic, incorporating questions aimed at fostering engagement and building a 'working therapeutic relationship.' This approach aligns with established mental health practices that emphasize checking in with the client to create a more interactive and supportive dialogue, as opposed to the didactic and instructional style often seen in AI responses. (3) Despite these differences, certain similarities were present. All responses—whether from the therapist, ChatGPT, or Mistral—featured empathetic statements and included essential psychoeducational content related to anxiety, as demonstrated in Appendix A.

These qualitative findings prompted the need for a more detailed quantitative analysis to further explore the extent of these differences, which we carried out using cosine similarity.

2.5 Quantitative Analysis of Therapist vs LLM Generated Responses

To quantify the differences between the LLM-generated responses and those of human therapists, we conducted a Cosine Similarity Analysis. This technique compares two sets of text data—therapist vs. LLM responses—by converting them into vector form and measuring the cosine of the angle between them. A cosine similarity score of 1 indicates identical vectors, while a score of 0 represents completely dissimilar vectors [8].

For this analysis, we employed Term Frequency-Inverse Document Frequency (TF-IDF) to calculate cosine similarity. TF-IDF is particularly well-suited for this task because it assigns weight to words based on their importance in individual documents (therapist or LLM responses) and their rarity across the entire dataset. In the context of mental health responses, where certain key terms like "empathy" or "anxiety" carry significant meaning, TF-IDF helps capture both the specific therapeutic language used by human therapists and the critical elements of AI-generated responses. This ensures that the comparison isn't simply based on word counts, but rather on the significance of the terms used in context. The formula for cosine similarity is:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| ||\mathbf{B}||}$$

Using Python, we calculated cosine similarity scores between the therapist responses and those generated by ChatGPT and Mistral for selected questions. The results, shown in Figure 2, reveal a clear gap between the LLMs and human therapists. The majority of the cosine similarity scores fall between 0.21 and 0.48, indicating that while the LLMs can

	Therapist	GPT	Cosine Similarity
0	If everyone thinks you're worthless, then mayb...	I'm really glad you reached out. I can hear ho...	0.3126
1	Hello, and thank you for your question and see...	I'm really sorry to hear that you're feeling t...	0.4835
2	It sounds like you are having some pretty seve...	I'm really sorry you're feeling this way. It's...	0.4149
3	Seasonal Affective Disorder (S.A.D.) is a term...	I 'm sorry to hear you're feeling this way. It...	0.3290
4	Light therapy is very helpful. You are not alo...	I'm really glad you reached out. What you're d...	0.2177
5	Anxiety can cause such a feeling of discomfort...	It sounds like you're going through a challeng...	0.3799
6	There are lots of things you can do, but first...	It sounds like you're feeling a lot of pressur...	0.3809
7	I think you have identified something importan...	It's understandable to feel this way, especial...	0.2326
8	Anger is a normal emotion, no different than j...	I understand that you're feeling frustrated an...	0.2853
9	You have been through so much and it sounds li...	It's completely understandable to feel overwhe...	0.3124

	Therapist	Mistral	Cosine Similarity
0	If everyone thinks you're worthless, then mayb...	I'm really glad you reached out, and I want yo...	0.3498
1	Hello, and thank you for your question and see...	I'm really glad you reached out to share what ...	0.4212
2	It sounds like you are having some pretty seve...	I'm really glad you reached out to share what ...	0.4087
3	Seasonal Affective Disorder (S.A.D.) is a term...	It sounds like you may be experiencing Seasona...	0.3522
4	Light therapy is very helpful. You are not alo...	It sounds like you may be experiencing Seasona...	0.2786
5	Anxiety can cause such a feeling of discomfort...	I'm really glad you've reached out to discuss ...	0.3913
6	There are lots of things you can do, but first...	I'm really glad you reached out, and I want yo...	0.3394
7	I think you have identified something importan...	It's not uncommon to feel this way, but it's i...	0.2744
8	Anger is a normal emotion, no different than j...	I appreciate you reaching out and sharing this...	0.3024
9	You have been through so much and it sounds li...	First, I want to acknowledge your strength and...	0.4256

Fig. 2. Cosine Similarity Analysis of Therapists and LLMs Responses for 10 Questions

generate syntactically correct and coherent responses, they often fail to capture the deeper emotional and empathetic elements present in human therapeutic communication.

Upon closer inspection, ChatGPT's scores, which range from 0.2177 to 0.4835, reflect a tendency to provide generalized reassurance and support, often lacking the personalization and interactive elements typical of a therapist's response. Similarly, Mistral's scores, ranging from 0.2744 to 0.4265, suggest that its responses are more formulaic. While Mistral's output often adheres to the structure and tone of a professional response, it lacks the flexibility and emotional responsiveness that are critical in a therapeutic setting.

For instance, ChatGPT's highest similarity score (0.4835) was achieved in a response that echoed supportive language but fell short in delivering the interactive, dialogue-driven communication seen in human therapists. Mistral's highest score (0.4265) reflected its ability to replicate the format of a therapeutic response, but again, the model struggled with adapting its responses dynamically to the emotional needs of the user.

In conclusion, the lower cosine similarity scores for both LLMs emphasize their limitations in matching the depth of empathy, emotional engagement, and intuition characteristic of human counselors. These findings, in conjunction with the manual qualitative analysis, highlight the ongoing challenge of bridging the gap between AI-generated responses and the nuanced therapeutic expertise that defines human mental health care.

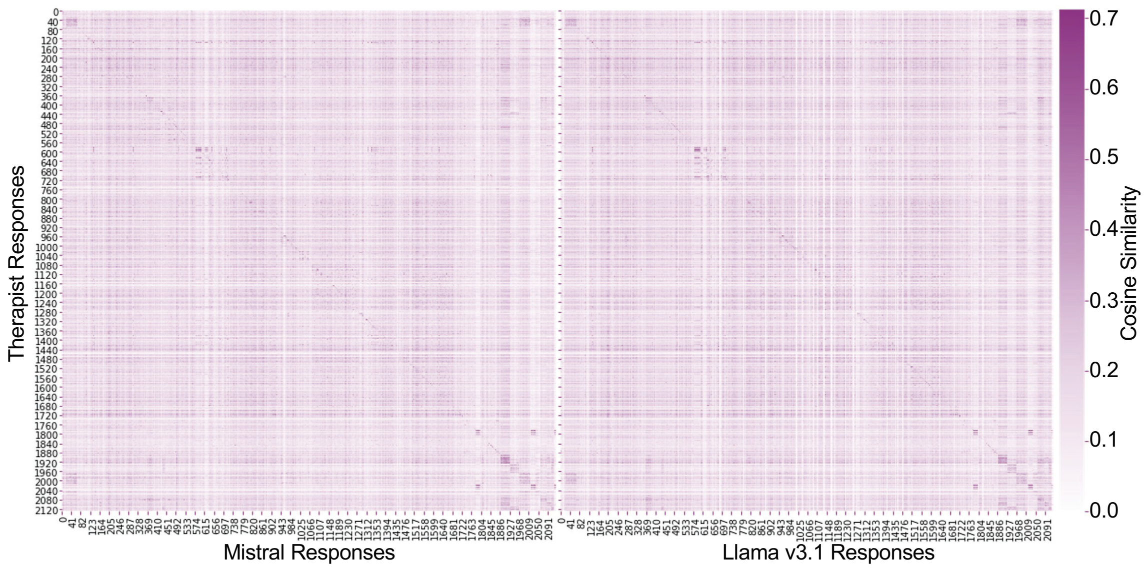

Fig. 3. A Heatmap of Cosine Similarity Analysis of Therapist Compared to LLMs Responses(Mistral on Left, LLamav31 on Right)

Next, we expanded our analysis to encompass a larger dataset of 2,129 questions. These questions, all answered by human therapists, were compared to responses generated by two of the most advanced open-source models: Mistral Large and Llama v3.1 13B. Using cosine similarity, we evaluated how well these models replicate the tone, empathy, and therapeutic insight of licensed professionals. The heatmap in Figure 3 visually depicts the similarity between the models' responses and those of the human therapists, with darker colors indicating higher similarity. This analysis provides a broader view of the models' performance across a large dataset of therapeutic scenarios.

From the heatmap, it is evident that both Mistral and Llama v3.1 display overall low cosine similarity to the human therapist responses across the dataset of 2,129 questions. The predominance of lighter shades across the map indicates that these models struggle to replicate the nuanced therapeutic communication provided by licensed professionals. Despite their technological sophistication, both models fall short of consistently capturing the depth of empathy, emotional attunement, and interactive conversational style that are critical in mental health support. This aligns with earlier findings where LLMs generated responses that were more structured and syntactically correct but lacked the human touch and emotional complexity. There are, however, a few instances where both models show higher cosine similarity, represented by the darker areas in the heatmap. These pockets suggest that in certain cases, Mistral and Llama v3.1 managed to produce responses that were more aligned with the human therapists' answers, possibly when delivering straightforward psychoeducational content. These higher similarity scores may indicate that the models can effectively mimic the surface-level structure of therapeutic language, particularly in responses that involve more factual or direct information. Nevertheless, these instances are rare, indicating the models' limitations in consistently generating responses that match the emotional depth and personalization characteristic of human therapists.

Moreover, the heatmap shows no significant difference between Mistral and Llama v3.1 in terms of their performance. Both models exhibit similar patterns of low similarity, suggesting that while they are among the most advanced open-source models, they share common challenges when applied to emotionally charged and complex therapeutic scenarios. This reinforces the idea that while these models are capable of producing relevant content, they often lack the adaptability and emotional intelligence that is central to effective therapy, highlighting the continued gap between AI-generated and human therapist responses.

The heatmap in Figure 4 presents a cosine similarity cross-comparison between responses generated by Mistral and Llama v3.1 across the 2,129 questions. The diagonal line of darker shading indicates instances where the same question was input into both models, with relatively higher cosine similarity between their responses. However, for the majority of the responses, the lighter shades across the heatmap reflect generally low similarity, suggesting that while both models are highly advanced, they generate distinct responses for the same questions. This indicates that although Mistral and Llama v3.1 might follow similar linguistic structures, their approach to providing answers varies, resulting in different outputs. The overall variation between these two models underscores their differing methods in handling the same mental health-related questions, despite being advanced in their design and capabilities.

2.6 Insights and Implications for Human-AI Interaction in Mental Health

The comparative analysis of LLM-generated responses, both through cosine similarity and qualitative methods, offers valuable insights into the current limitations and potential of AI in interacting with human emotions in mental health

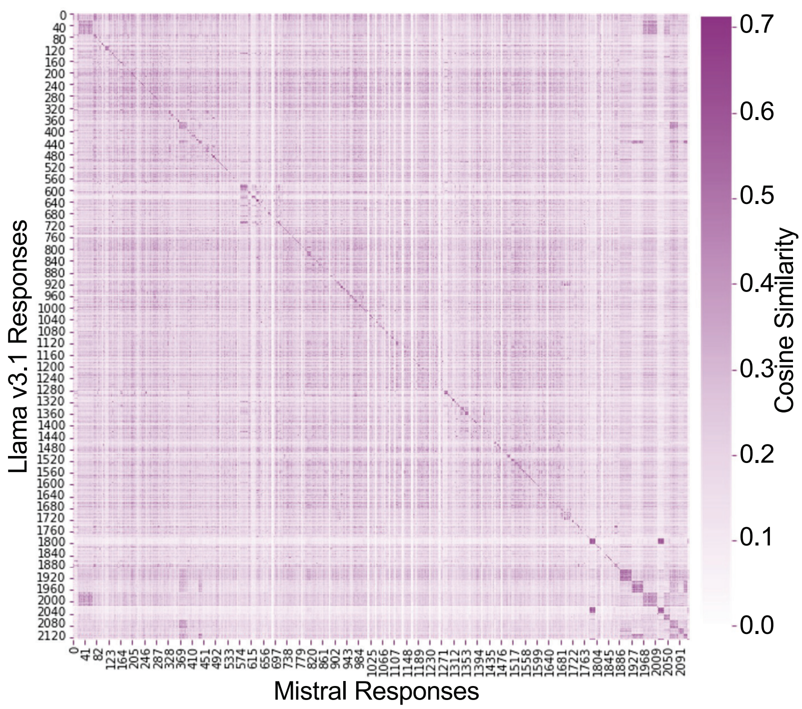

Fig. 4. Cosine Similarity Cross Comparison of LLama v3.1 vs Mistral Responses

contexts. The consistently low cosine similarity scores across models highlight a key shortcoming: while AI models such as Mistral and Llama v3.1 can generate structured, coherent responses, they often lack the deeper emotional attunement and psychological nuance characteristic of human therapists. This gap is especially critical in therapeutic settings, where understanding and addressing complex emotions requires more than delivering factual information—it requires empathy, personalization, and adaptive interaction.

The use of detailed personas and profiles was intended to guide LLMs toward more human-like interactions, helping them adopt the tone and therapeutic language expected of mental health professionals. While these personas aided in shaping the responses toward a more professional demeanor, the cosine similarity analysis reveals that the models still struggled to capture the depth and emotional responsiveness that are critical in therapeutic dialogues. The qualitative analysis further supports this, showing that AI responses, although well-structured, often lacked the emotional sensitivity and adaptive feedback seen in human therapists' responses. This suggests that current LLMs are more successful at mimicking surface-level professional behavior than at engaging meaningfully with the underlying emotional complexities of clients in distress.

This limitation, however, does not entirely undermine the role of AI in mental health support. While LLMs may not yet be capable of replicating the full range of human emotional intelligence, their ability to provide accurate, structured responses could serve as a foundation for more specialized therapeutic tools. Importantly, the findings underscore the need for continuous human oversight in AI-enabled mental health applications. Without clinical guidance, there is a significant risk that LLMs may deliver responses that, while 'correct' in form, fail to meet the emotional needs of users or could even cause harm. Future research must explore whether AI systems can be refined to better detect and respond to emotional cues, or if their role should be limited to more supportive, adjunct functions in a therapeutic setting.

Looking ahead, the implications for Human-AI interaction in mental health are clear. To develop more effective AI systems, future research should focus on improving models' ability to process not only the linguistic content of user input but also the emotional undertones that are essential in therapeutic exchanges. This may involve integrating more sophisticated emotional recognition capabilities and utilizing multi-modal data, such as voice tone or facial expressions, to provide a fuller understanding of the user's emotional state. Such advancements would move AI closer to providing responses that align not just with the content but also with the emotional needs of individuals, making AI a more effective tool in mental health care.

2.7 Feasibility Analysis of Needs for AI-Enabled First Aid Mental Support

Building on the insights regarding the limitations and potential of current AI models in mental health interactions, we now turn to the feasibility of developing an AI-enabled chatbot that can effectively provide first aid mental health support. To ensure such a system can meet the unique demands of mental health care, we must consider key factors such as trustworthiness, security, bias mitigation, empathy, and privacy.

The development of an AI-enabled chatbot for mental health support must adhere to a rigorous set of criteria to ensure its effectiveness, reliability, and acceptance by both users and healthcare professionals. As mental health continues to be a growing global concern, exacerbated by the COVID-19 pandemic, the role of digital tools in providing scalable, accessible, and private support has become increasingly important. To design a **feasible AI-enabled chatbot**, it is essential to address key concerns surrounding trustworthiness, security, bias mitigation, empathy, and privacy, shown in Figure 5.

In the context of mental health, **trustworthiness** is paramount. Users must be able to trust the chatbot to provide accurate, timely, and ethically sound responses to their needs. This trust can be bolstered by incorporating a **human-in-the-loop** system, wherein human oversight is used to monitor, validate, and guide the AI's decisions and interactions. Such oversight ensures that the chatbot behaves in ways that align with human values and ethical standards, mitigating potential harm and fostering user confidence.

Additionally, the chatbot must be **secure**, especially given the sensitive nature of mental health data. Mental health interactions involve the disclosure of personal, often vulnerable, information. Therefore, robust **cryptographic techniques** are required to safeguard this data from unauthorized access or misuse. By employing state-of-the-art encryption, the chatbot can ensure that all communications between users and the system remain confidential, protecting users' privacy and enhancing their willingness to engage.

A crucial component of the system's design is its ability to be **reduced bias**. Bias in AI systems can perpetuate harmful stereotypes or lead to inequitable outcomes, particularly in mental health care, where individualized and empathetic responses are critical. To reduce bias, the system must leverage a diverse range of data sources. By increasing the **number of nodes participating** in a federated learning framework, the chatbot can learn from a broader dataset, ensuring that its responses are not disproportionately influenced by any one population or context. This approach helps mitigate the risk of bias and enhances the system's fairness and inclusivity.

Equally important, the chatbot must demonstrate **empathy** in its interactions. The ability to offer empathetic responses is central to mental health support, where users need to feel understood and validated. By using an **evidence-proved chatbot**, one that has been rigorously tested and validated against existing mental health interventions, the system can ensure that its responses are both scientifically grounded and empathetic. This not only enhances user satisfaction but also increases the chatbot's effectiveness in delivering meaningful support.

Finally, **privacy** is a critical concern in any digital mental health intervention. Users must feel confident that their interactions with the chatbot remain private and secure. **Federated learning** offers a solution to this challenge by enabling the chatbot to learn from user interactions without storing sensitive data in a central location. This decentralized approach allows the system to improve over time while minimizing the risk of data breaches and ensuring that user privacy is respected at all stages.

In summary, the feasibility of an AI-enabled chatbot for mental health support depends on its ability to meet these five key criteria: trustworthiness, security, bias mitigation, empathy, and privacy. The integration of **human-in-the-loop oversight**, **cryptographic security**, **diverse data sources**, **evidence-based interactions**, and **federated learning** creates a robust framework that addresses the unique challenges of providing digital mental health support. The proposed system, as outlined in Figure 5, represents a comprehensive approach to designing a trustworthy, secure, reduced bias, empathetic, and privacy-preserving chatbot capable of addressing the mental health needs of diverse populations.

3 Federated Learning as a Potential Solution

Google introduced Federated Learning in 2016, a recent trend in which a central server supervises a group of edge nodes (clients) that collaboratively learn a shared machine-learning (ML) model without sharing their sensitive data [25, 33, 35]. All decentralized participating nodes learn the shared ML model when a trusted curator aggregates their locally computed model's parameters to a centralized server and then sends back the resulting model, as depicted in Figure 6, preserving the privacy of their data [17, 26, 35]. Federated Learning has been a trend in different fields, especially healthcare, due to patient data-sensitive nature [35]. Although studies have proposed frameworks asserting

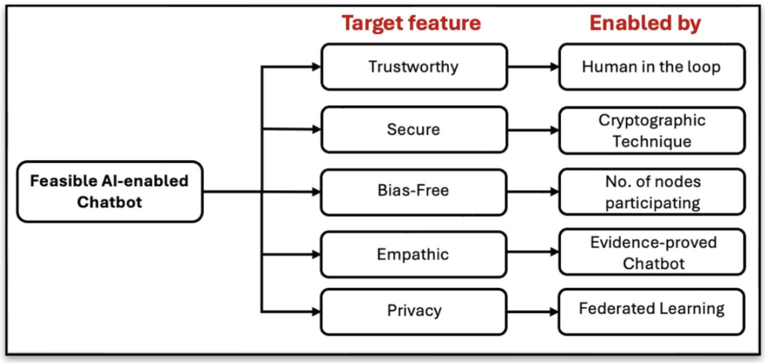

Fig. 5. A summary map of the proposed framework features

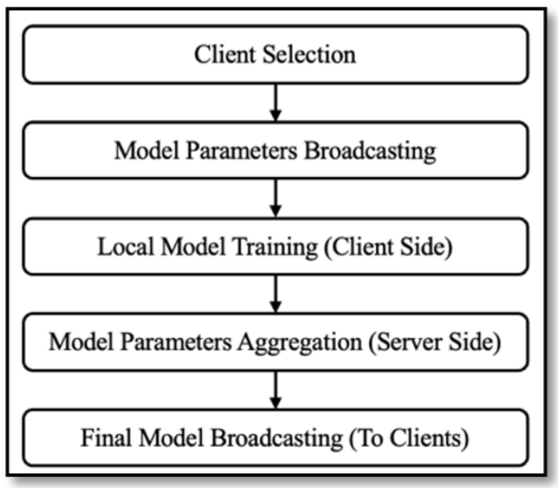

Fig. 6. Federated Learning Steps

that federated learning is a promising approach to utilize and protect the privacy of patient data silos that reside on health institutional servers, it suffers from multiple challenges [17, 25, 35].

Two broad classifications of federated learning challenges are training-related concerns and privacy and security concerns [25]. Communication exhaustion or cost due to training iterations and the heterogeneity of data (non-identically distributed data) and devices used in the training process are instances of training-related concerns. Whereas membership inference attacks, data poisoning attacks, model poisoning attacks, byzantine attacks, and backdoor attacks are instances of privacy and security concerns [23, 25, 26, 40]. Therefore, practical and effective techniques to alleviate the above problems are urgently needed when considering federated learning in building AI-enabled chatbots as mental health support.

For example, implementing a cryptographic technique in federated learning, such as homomorphic encryption, is required to preserve data privacy [17, 29]. Homomorphic encryption in federated learning can preserve data privacy by giving participating clients the ability to perform operations on ciphertext, producing outcomes like those gained when performing the same operations on the original plaintext [29, 40]. Also, Minimizing the communication rounds among edge nodes and the aggregator server as well as the size of messages sent in these rounds could solve the communication cost in federated learning [23]. Another technique is applying [13] suggestion of incorporating human intelligence with AI intelligence to validate model construction, spot unobserved data flaws, evaluate data bias, and proactively identify, control, and mitigate risk. The Human AI collaboration must include multi-stakeholder participation to assess all aspects of AI [31].

In this paper, we propose a framework for building an AI-enabled chatbot that utilizes federated learning in training the AI model and magnifies the role of the human element in the loop. Stakeholders must assess the state of AI use for mental health. As [18] advised, it is the responsibility of stakeholders, including all of us, to make an active move to transform mental health support by incorporating the mental health clinicians' expertise with the expertise of data scientists and other scientists. Our ultimate goal is to create a reliable, evidence-proved, reduced bias, and empathic AI-enabled chatbot that can serve as first aid for mental health, especially in underrepresented and underserved communities.

4 First Aid Mental Health

Although some researchers have raised concerns that AI-enabled chatbots may lack empathy and personalization, with these limitations highlighted as obstacles in previous studies [6, 28], recent findings challenge this view. A study by

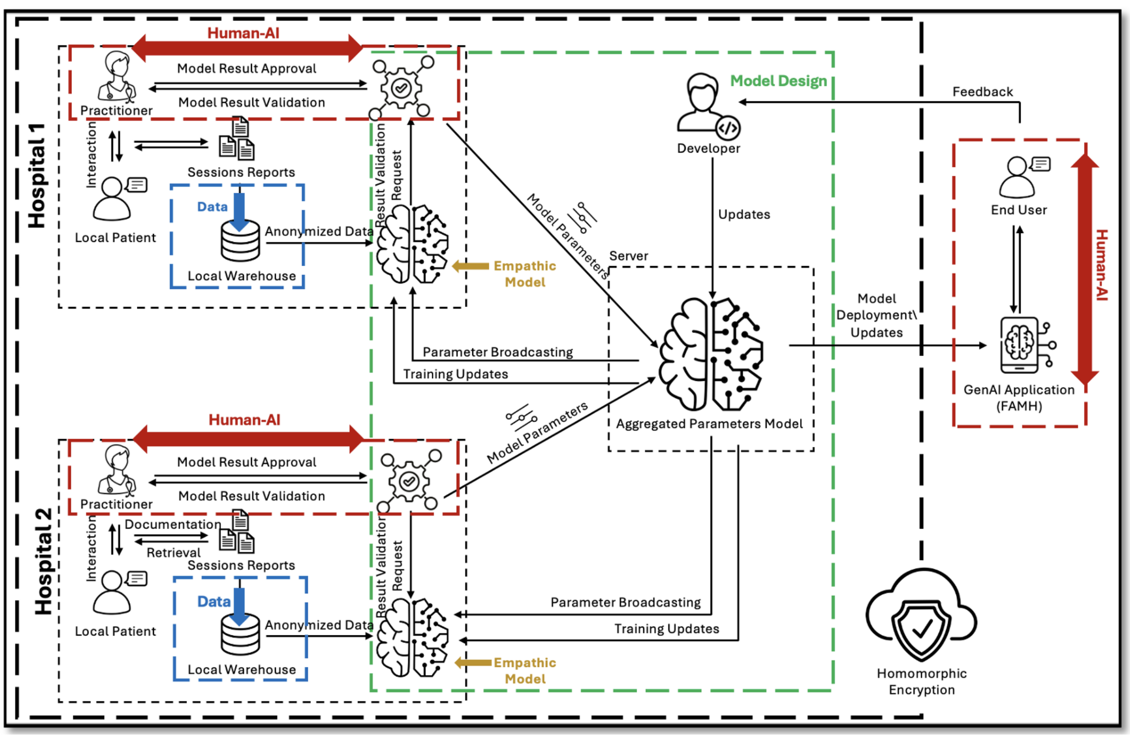

Fig. 7. First Aid Mental Health Framework

[39] asserted that the empathic responses generated by large language models (LLMs) can, in fact, surpass those of humans. An empathic AI-enabled chatbot is grounded in empathy theory from psychology, where cognitive empathy, a learnable and teachable skill, plays a central role in counseling sessions between clinicians and patients [20, 24, 32]. Building on this foundation, we aim to incorporate a model with proven superior empathy into our proposed framework. Specifically, [39] found that ChatGPT-4 outperformed its peers in a between-subject study evaluating the empathic quality of LLM-generated responses.

Figure 7 lays out the main stakeholders involved, the general techniques to be used, and the data and operation flow. We reassert [13] point of view on the importance of human intelligence in the process, where constant, expert feedback is crucial for the success of the entire framework.

The proposed framework's success depends on the number of edge nodes (hospitals and clinics) participating in the federated learning process. Each participating node will enlarge the final dataset used and empower the machine learning model training process. Since federated learning depends on accumulating models' parameters that were trained on unique datasets from multiple and distinct edge nodes, the resulting aggregated model on the centralized server will be trained on diversified data, ensuring more inclusive and unbiased digital mental health support. Having multiple parties (hospitals in our case) who collaboratively train local models with their local datasets and push the model's parameters to an aggregator in federated learning will minimize the bias problem and make the model more generalizable in mental health [25, 33].

Each participating hospital or clinic (for example, hospital 1 and hospital 2 in figure 7) will create reports from doctor/patient sessions, anonymize them, and add them to a local database. It is important to emphasize that all datasets used should follow the same format to ensure the data heterogeneity problem is solved in federated learning. The local data engineer is responsible for confirming compliance with the agreed-upon data format. Moreover, it is important to highlight the role of mental health clinicians and data engineers in diversifying the input data to eliminate bias propagation problem [10] in federated learning. The local empathic machine-learning model will then be trained on these reports and wait for an aggregator request to push the model parameters only while preserving data privacy by not sharing them.

The model parameters will not be pushed to the aggregator until local data scientists and mental health clinicians from the participating edge node validate the local model's result. This step, which incorporates human intelligence, is crucial to making the resulting model more reliable. To eliminate false positive (FP) and false negative (FN) cases, clinician expertise must be seriously aligned with machine learning models. A local data scientist should test the chatbot's performance using evaluation metrics. For example, deepeval [1] could be used to check the chatbot's answer relevancy, hallucination, and bias. This process must take place with the help of a local mental health clinician in order to formulate the different test cases. A local mental health clinician should also validate the model by interacting with it and providing different prompts and scenarios to examine the model responses, rationality, fluency, relevance, and coherence. A semantic textual similarity analysis, for example, using the Multilingual Text Enhancer (MLTE) approach [14], could also be performed to test the semantic similarity of the LLM response to the expected response.

The pushed local model parameters will be used by an aggregating model on a centralized server. The model will enhance its performance and push back its parameters to all participating edge nodes to enhance their performance in turn. Data scientists (developers) will be involved in validating the performance of the central server's resulting model and providing constant updates. The whole process of federated learning will use a reliable and evidence-proven encryption technique. An example of encryption technique is the homomorphic encryption that allows performing operations without decrypting data in advance [3].

The resulting model from the federated learning approach will be used as a first-aid mental health support chatbot. The chatbot will always be available for those who need access to mental health services but suffer from mental healthcare barriers. No data will be collected from the users to enhance the chatbot's performance. To preserve users' privacy and the chatbot's reliability, the only data that will be used to enhance the chatbot's performance is from the participating hospitals. Developers could only collect survey feedback from users to improve the chatbot's features and tackle the pain points associated with using it.

5 Conclusion

The development of AI-enabled mental health support tools presents both opportunities and challenges. This paper has explored these aspects by analyzing the capabilities and limitations of large language models (LLMs) in replicating the therapeutic responses of human mental health professionals. Our analysis, which utilized cosine similarity metrics and qualitative assessment, revealed that although models like ChatGPT-4 and Mistral produce grammatically correct and coherent responses, they often lack the nuanced empathy and emotional depth characteristic of human therapists. This finding underscores a key limitation: while AI models can imitate surface-level responses, they struggle to fully engage with the emotional and psychological complexities of distressed individuals. Moreover, the cross-comparison between LLMs demonstrated that even state-of-the-art models exhibit variations in their output, further complicating their use in sensitive contexts such as mental health.

A key contribution of this paper is the novel framework we proposed for developing a more reliable, empathetic, and secure AI-enabled chatbot for first-aid mental health support. By integrating human oversight—via a "human-in-the-loop" approach—alongside a federated learning system, we aim to address crucial challenges such as bias mitigation, privacy, and trustworthiness. Federated learning ensures that the chatbot can learn from diverse datasets distributed across multiple healthcare nodes, without compromising user privacy. This decentralized approach not only helps reduce bias but also enhances the model's generalizability across different population groups, making it a more inclusive solution for underrepresented and underserved communities.

One of the key findings of this study is that while LLMs have the potential to offer support, their effectiveness depends significantly on expert intervention. Our analysis highlights that incorporating human intelligence at various stages—model validation, data scrutiny, and feedback gathering—can help mitigate issues such as hallucinations, bias propagation, and shallow emotional engagement. The role of clinicians, in particular, is indispensable in ensuring that AI responses align with therapeutic best practices and are free from harmful stereotypes or inaccuracies.

Furthermore, this paper emphasizes the importance of using cutting-edge cryptographic techniques, like homomorphic encryption, to safeguard sensitive mental health data during federated learning. The integration of such privacy-preserving technologies, coupled with constant human oversight, creates a multi-layered framework that prioritizes both security and ethical considerations.

In conclusion, while LLMs offer a promising avenue for mental health support, it is evident that the current models are not yet ready to fully replace human therapists, especially in emotionally sensitive scenarios. However, with the right balance of advanced AI techniques, expert human involvement, and secure, privacy-preserving frameworks, AI-enabled chatbots could become a valuable tool in bridging the mental healthcare gap. The novel framework proposed in this paper presents a pathway toward developing an AI system that is reliable, reduced in bias, empathetic, and privacy-focused, capable of providing first-aid mental health support to diverse populations. As we move forward, a collaborative effort between mental health professionals, data scientists, and users is crucial to making this solution both effective and widely accepted.

6 Future Work

While large language models keep advancing and unlocking massive potentials, a possible research area would be using the proposed AI-enabled chatbot by mental health clinicians to assess their patients' conditions. If the proposed chatbot is used as a patient's mental health companion, it could analyze the patient's language patterns to detect early signs of treatment responses as well as make early psychological and pharmacological intervention recommendations. Besides, collecting feedback from the AI-enabled chatbot users through a controlled experiment would increase the feasibility of the model. This would help understand the most important features to incorporate for mental health support during a crisis and the sense of urgency.

References

[1] [n. d.]. *Introduction | DeepEval - The Open-Source LLM Evaluation Framework.* https://docs.confident-ai.com/docs/metrics-introduction

[2] 2022. Language Models are Few-Shot Learners. (2022).

[3] Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. 2019-07-31. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. 51, 4 (2019-07-31), 1–35. https://doi.org/10.1145/3214303

[4] Hatoon S. Alsagri and Mourad Ykhlef. 2020-08-01. Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features. E103.D, 8 (2020-08-01), 1825–1832. https://doi.org/10.1587/transinf.2020EDP7023

[5] L. H. Andrade, J. Alonso, Z. Mneimneh, J. E. Wells, A. Al-Hamzawi, G. Borges, E. Bromet, R. Bruffaerts, G. De Girolamo, R. De Graaf, S. Florescu, O. Gureje, H. R. Hinkov, C. Hu, Y. Huang, I. Hwang, R. Jin, E. G. Karam, V. Kovess-Masfety, D. Levinson, H. Matschinger, S. O'Neill, J. Posada-Villa, R. Sagar, N. A. Sampson, C. Sasu, D. J. Stein, T. Takeshima, M. C. Viana, M. Xavier, and R. C. Kessler. 2014-04. Barriers to mental health treatment: results from the WHO World Mental Health surveys. 44, 6 (2014-04), 1303–1317. https://doi.org/10.1017/S0033291713001943

[6] Luke Balcombe. 2023-10-27. AI Chatbots in Digital Mental Health. 10, 4 (2023-10-27), 82. https://doi.org/10.3390/informatics10040082

[7] Nicolas Bertagnolli. [n. d.]. *Counsel Chat: Bootstrapping High-Quality Therapy Data.* https://towardsdatascience.com/counsel-chat-bootstrapping-high-quality-therapy-data-971b419f33da

[8] Saprativa Bhattacharjee, Anirban Das, Ujjwal Bhattacharya, Swapan K. Parui, and Sudipta Roy. [n. d.]. Sentiment analysis using cosine similarity measure. In *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)* (Kolkata, India, 2015-07). IEEE, 27–32. https://doi.org/10.1109/ReTIS.2015.7232847

[9] Mirko Casu, Sergio Triscari, Sebastiano Battiato, Luca Guarnera, and Pasquale Caponnetto. 2024-07-05. AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications. 14, 13 (2024-07-05), 5889. https://doi.org/10.3390/app14135889

[10] Hongyan Chang and Reza Shokri. [n. d.]. Bias Propagation in Federated Learning. arXiv:2309.02160 [cs, stat] http://arxiv.org/abs/2309.02160

[11] S. Clement, O. Schauman, T. Graham, F. Maggioni, S. Evans-Lacko, N. Bezborodovs, C. Morgan, N. Rüsch, J. S. L. Brown, and G. Thornicroft. 2015-01. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. 45, 1 (2015-01), 11–27. https://doi.org/10.1017/S0033291714000129

[12] Kate Daley, Ines Hungerbuehler, Kate Cavanagh, Heloísa Garcia Claro, Paul Alan Swinton, and Michael Kapps. 2020-11-30. Preliminary Evaluation of the Engagement and Effectiveness of a Mental Health Chatbot. 2 (2020-11-30), 576361. https://doi.org/10.3389/fdgth.2020.576361

[13] Munmun De Choudhury and Emre Kiciman. 2018-09. Integrating Artificial and Human Intelligence in Complex, Sensitive Problem Domains: Experiences from Mental Health. 39, 3 (2018-09), 69–80. https://doi.org/10.1609/aimag.v39i3.2815

[14] Yeli Feng. [n. d.]. Semantic Textual Similarity Analysis of Clinical Text in the Era of LLM. In *2024 IEEE Conference on Artificial Intelligence (CAI)* (Singapore, Singapore, 2024-06-25). IEEE, 1284–1289. https://doi.org/10.1109/CAI59869.2024.00227

[15] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017-06-06. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. 4, 2 (2017-06-06), e19. https://doi.org/10.2196/mental.7785

[16] Cláudia Cristina Fukuda, Maria Aparecida Penso, Deise Matos Do Amparo, Bruno Coimbras De Almeida, and Camila De Aquino Morais. 2016-06. Mental health of young Brazilians: Barriers to professional help-seeking. 33, 2 (2016-06), 355–365. https://doi.org/10.1590/1982-02752016000200017

[17] Robin C. Geyer, Tassilo Klein, and Moin Nabi. 2018-03-01. Differentially Private Federated Learning: A Client Level Perspective. arXiv:1712.07557 [cs, stat] http://arxiv.org/abs/1712.07557

[18] Sarah Graham, Colin Depp, Ellen E. Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V. Jeste. 2019-11. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. 21, 11 (2019-11), 116. https://doi.org/10.1007/s11920-019-1094-0

[19] M D Romael Haque and Sabirat Rubya. 2023-05-22. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. 11 (2023-05-22), e44838. https://doi.org/10.2196/44838

[20] James Hardee. 2003. An Overview of Empathy. (2003). https://doi.org/10.7812/TPP/03-072

[21] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018-11-23. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. 6, 11 (2018-11-23), e12106. https://doi.org/10.2196/12106

[22] Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. 2024-03-26. Large Language Models Produce Responses Perceived to be Empathic. arXiv:2403.18148 [cs] http://arxiv.org/abs/2403.18148

[23] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020-05. Federated Learning: Challenges, Methods, and Future Directions. 37, 3 (2020-05), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[24] Shuya Lin, Lingfeng Lin, Cuiqin Hou, Baijun Chen, Jianfeng Li, and Shiguang Ni. 2023-12-04. Empathy-Based communication Framework for Chatbots: A Mental Health Chatbot Application and Evaluation. In *International Conference on Human-Agent Interaction* (Gothenburg Sweden). ACM, 264–272. https://doi.org/10.1145/3623809.3623865

[25] Priyanka Mary Mammen. 2021-01-13. Federated Learning: Opportunities and Challenges. arXiv:2101.05428 [cs] http://arxiv.org/abs/2101.05428

[26] H Brendan McMahan, Eider Moore, Daniel Ramage, and Seth Hampson. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. (2017).

[27] Paulo Menezes, Julieta Quayle, Heloísa Garcia Claro, Simone Da Silva, Lena R Brandt, Francisco Diez-Canseco, J Jaime Miranda, LeShawndra N Price, David C Mohr, and Ricardo Araya. 2019-04-26. Use of a Mobile Phone App to Treat Depression Comorbid With Hypertension or Diabetes: A Pilot Study in Brazil and Peru. 6, 4 (2019-04-26), e11698. https://doi.org/10.2196/11698

[28] Carlos Montemayor, Jodi Halpern, and Abrol Fairweather. 2022-12. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. 37, 4 (2022-12), 1353–1359. https://doi.org/10.1007/s00146-021-01230-z

[29] Kundan Munjal and Rekha Bhatia. 2023-08. A systematic review of homomorphic encryption and its contributions in healthcare industry. 9, 4 (2023-08), 3759–3786. https://doi.org/10.1007/s40747-022-00756-z

[30] World Health Organization. 2022. *COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide.* https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

[31] Seithikurippu R. Pandi-Perumal, Meera Narasimhan, Mary V. Seeman, and Haitham Jahrami. 2024-06. Artificial intelligence is set to transform mental health services. 29, 3 (2024-06), 155–157. https://doi.org/10.1017/S1092852923002456

[32] Frederic W. Platt and Vaughn F. Keller. 1994-04. Empathic communication: A teachable and learnable skill. 9, 4 (1994-04), 222–226. https://doi.org/10.1007/BF02600129

[33] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. 2020-09-14. The future of digital health with federated learning. 3, 1 (2020-09-14), 119. https://doi.org/10.1038/s41746-020-00323-1

[34] William Schreiber, Jules White, and Douglas C Schmidt. [n. d.]. A Pattern Language for Persona-based Interactions with LLMs. ([n. d.]).

[35] Banuchitra Suruliraj and Rita Orji. 2022-08-10. Federated Learning Framework for Mobile Sensing Apps in Mental Health. In *2022 IEEE 10th International Conference on Serious Games and Applications for Health(SeGAH)* (Sydney, Australia). IEEE, 1–7. https://doi.org/10.1109/SEGAH54908.2022.9978600

[36] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S. Kashavan, and John Blake Torous. 2019-07. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. 64, 7 (2019-07), 456–464. https://doi.org/10.1177/0706743719828977

[37] Marianne Agergaard Vammen, Sigurd Mikkelsen, Åse Marie Hansen, Jens Peter Bonde, Matias B. Grynderup, Henrik Kolstad, Linda Kærlev, Ole Mors, Reiner Rugulies, and Jane Frlund Thomsen. 2016-10. Emotional Demands at Work and the Risk of Clinical Depression: A Longitudinal Study in the Danish Public Sector. 58, 10 (2016-10), 994–1001. https://doi.org/10.1097/JOM.0000000000000849

[38] Juan David Velásquez-Henao, Carlos Jaime Franco-Cardona, and Lorena Cadavid-Higuita. [n. d.]. Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering. 90, 230 ([n. d.]), 9–17. https://doi.org/10.15446/dyna.v90n230.111700

[39] Anuradha Welivita and Pearl Pu. 2024-06-07. Are Large Language Models More Empathetic than Humans? arXiv:2406.05063 [cs] http://arxiv.org/abs/2406.05063

[40] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. 2023-02. A survey on federated learning: challenges and applications. 14, 2 (2023-02), 513–535. https://doi.org/10.1007/s13042-022-01647-y

[41] Amir Hossein Yazdavar, Hussein S Al-Olimat, Tanvi Banerjee, Krishnaprasad Thirunarayan, and Amit P Sheth. 2016. Analyzing Clinical Depressive Symptoms in Twitter. (2016). https://corescholar.libraries.wright.edu/knoesis/1121

Appendices

A A Sample Prompt for A Patient Question, the Human Therapist's Response along with LLM's Outputs

Prompt: You are a Registered Psychotherapist female. You received the following question from a person suffering from a mental health issue. What would be your response? Do not provide an unnecessary response. Reply as if you were talking to the person directly.
The question is: "I just took a job that requires me to travel far away from home. My family and

I really need this job. People keep telling me I have "anxiety" and I'm terrified of having an anxiety attack on the road. This is all new to me. What can I do?"

Therapist Response: Anxiety can cause such a feeling of discomfort in our bodies that it can seem unbearable. The concern that I can hear in what you have written is that the thought of having some anxiety has now left you uncomfortable with any anxiety at all - and that can become a problem, indeed. As opposed to thinking about your anxiety as something much bigger than you, it could be helpful to start breaking it down. What are the things specifically that cause you anxiety about this new job? Is it being away from your family? If so, why?Once you start breaking down what causes you to feel anxiety into smaller "mouthfuls", then we are far better able to understand what has led to discomfort and start to problem solve the issue.

ChatGPT Response: It sounds like you're going through a challenging time, and it's understandable to feel worried about this new situation. Anxiety, especially when it's unfamiliar, can feel overwhelming, but there are ways to manage it.
First, recognize that anxiety is a normal response to new and uncertain situations. It can help to prepare some simple strategies for when you feel it coming on. Breathing exercises, such as taking slow, deep breaths, can help calm your nervous system. You might also try grounding techniques, like focusing on your senses - noticing five things you can see, four you can touch, three you can hear, two you can smell, and one you can taste.
Since you're worried about having an anxiety attack while traveling, plan some breaks or calming activities. Listening to a favorite podcast, practicing mindfulness, or even calling someone you trust could help you feel more grounded.
It could also be helpful to speak with a mental health professional about these feelings. They can work with you to develop a personalized plan to manage anxiety effectively.
Remember, it's okay to feel this way, and there are many ways to get support. You're not alone.

Mistral Response: I'm really glad you've reached out to discuss this. It's completely normal to feel anxious when facing significant changes, like a new job with extensive travel. Let's focus on some strategies to help you manage this anxiety.
Firstly, it's important to understand what an anxiety attack or panic attack feels like. Symptoms can include a racing heart, shortness of breath, dizziness, and overwhelming fear. It's crucial to remember that while these symptoms are unpleasant, they are not dangerous.
Here are some techniques you can use on the road:
1. Deep Breathing: If you start to feel anxious, try taking slow, deep breaths. Inhale for a count of four, hold for four, and exhale for four. This can help calm your body's physical response to anxiety.
2. Grounding Techniques: One example is the 5-4-3-2-1 method. Name 5 things you can see, 4 things you can touch, 3 things you can hear, 2 things you can smell, and 1 thing you can taste. This helps bring your focus back to the present moment.

```
3. Preparation: Plan your trips in advance. Knowing where you're going, how you're getting there,
and where you're staying can help reduce uncertainty and anxiety.
4. Self-Care: Ensure you're getting enough sleep, eating healthily, and staying hydrated. Physical
health can greatly influence your mental state.
5. Stay Connected: Regularly call or video chat with your family. Hearing their voices and seeing
their faces can be comforting.
6. Professional Help: If your anxiety feels unmanageable, consider speaking to a mental health
professional. They can provide strategies and tools tailored to your needs. Some therapists even
specialize in remote or online therapy, which could be helpful given your travel schedule.
Lastly, remember that it's okay to have anxiety, and it's not a sign of weakness. It's a normal
human response to stress. With the right tools and support, you can manage it and succeed in
your new job. You're not alone in this. Let's take it one step at a time together.
```