# Addressing the Challenges of Mental Health Conversations with Large Language Models

# Addressing the Challenges of Mental Health Conversations with Large Language Models


Shuvam Shiwakoti
Virginia Tech
Blacksburg, Virginia, USA
shuvam@vt.edu

Siddhant Bikram Shah
Northeastern University
Boston, Massachusetts, USA
shah.siddhantb@northeastern.edu

Imran Razzak
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE
University of New South Wales
Sydney, Australia
imran.razzak@mbzuai.ac.ae

Surendrabikram Thapa
Virginia Tech
Blacksburg, Virginia, USA
surendrabikram@vt.edu

Usman Naseem
Macquarie University
Sydney, New South Wales, Australia
usman.naseem@mq.edu



## Abstract

Virtual Mental Health Assistants offer a promising solution to address the growing demand for accessible and scalable mental healthcare. However, existing dialogue generation models struggle with the complexities inherent in mental health conversations. In this paper, we explore the limitations of current Medical Dialogue Generation models by conducting experiments on the large language model ChatMGL. We propose modifications to ChatMGL, including finetuning the model on a mental health dataset without proximal policy optimization and incorporating dialogue act labels, to enhance its ability to handle the complex nature of mental health dialogues. Our results demonstrate that these modifications outperform baseline models in terms of ROUGE and BERT scores. Our work suggests that specialized fine-tuning and incorporating domain-specific knowledge can improve the efficacy of virtual assistants for mental health support.



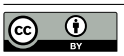
## CCS Concepts

• **Information systems** → *Web searching and information discovery*; *Information retrieval*; **World Wide Web**; • **Computing methodologies** → **Natural language processing**; Machine learning; • **Applied computing** → *Consumer health.*

## Keywords

Large Language Models, Mental Health, Social Media, Computational Social Science

**ACM Reference Format:**
Shuvam Shiwakoti, Siddhant Bikram Shah, Imran Razzak, Surendrabikram Thapa, and Usman Naseem. 2025. Addressing the Challenges of Mental Health Conversations with Large Language Models. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3701716.3718374



*WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia*



ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3718374


## 1 Introduction

Mental health issues represent a significant and growing challenge globally, with millions of people affected by conditions such as depression, anxiety, and other psychological disorders [9]. The World Health Organization (WHO) estimates that one in four people will be affected by mental or neurological disorders at some point in their lives, with many of these individuals lacking access to the necessary care and support [19]. Although treatment strategies such as therapy and medication exist, constraints like high costs, the scarcity of mental health professionals, and social stigmas associated with seeking help have created barriers to timely and accessible care. Consequently, many individuals either remain undiagnosed or fail to receive consistent therapeutic support, thereby exacerbating the mental health crisis.

In recent years, the rapid evolution of artificial intelligence (AI) has opened new avenues for addressing these gaps in care. Virtual Mental Health Assistants (VMHAs) have gained traction as a cost-effective and readily accessible approach to supplement traditional mental health services [15, 18]. This signifies a revolutionary trend in healthcare, where AI is increasingly being leveraged to augment human expertise and address systemic inefficiencies. These AI-driven systems can serve as a preliminary point of contact, offering immediate, round-the-clock assistance to individuals who might otherwise not seek or obtain help due to geographic, financial, or social barriers. Preliminary research indicates that VMHAs, when well-designed, can not only deliver evidence-based therapeutic exercises—such as Cognitive Behavioral Therapy (CBT) [16] and psychoeducation [17]—but also foster empathetic, user-centered dialogues that play a critical role in mental health support [11]. These systems can also help reduce the stigma associated with seeking mental health care by offering anonymity and privacy to users, encouraging more people to seek help [1]. Moreover, VMHAs can assist in early detection and intervention, identifying signs of mental distress before they escalate into more severe conditions, thereby potentially reducing the burden on healthcare systems [2].

However, the success of VMHAs depends heavily on their ability to generate contextually appropriate, empathetic, and clinically accurate responses—a task that has proven to be challenging [12]. Current Medical Dialogue Generation (MDG) models, while effective in various medical domains, often struggle with the complexities inherent in mental health conversations. In many existing frameworks, template-based or rule-based frameworks force the assistant into a rigid conversation pattern, limiting its adaptability and empathy. Additionally, the lack of semantics could also result in inappropriate dialogues. The complexity can overwhelm existing MDG models, leading to responses that may lack coherence, emotional depth, or even relevance, thus diminishing the effectiveness of VMHAs and potentially causing harm to the user [7].

To address these issues, we perform experiments with ChatMGL[1] and modify it to create a specialized VMHA system to effectively handle the nuanced demands of mental health dialogues. ChatMGL is a large language model (LLM) built upon GPT-2 that uses supervised fine-tuning and reinforcement learning (RL) with proximal policy optimization (PPO) and is specifically trained to answer data science questions. While it excels in generating responses for in-domain queries, its architecture and training datasets are not optimized for mental health conversations, which may affect its performance in this domain. To this end, we propose a set of modifications: supervised fine-tuning on a carefully designed mental health dataset and incorporating dialogue-act labels that capture the unique structure and emotional undercurrents of therapeutic sessions. By evaluating these adaptations, we show how ChatMGL can be transformed into a more context-aware and empathetic system, capable of generating clinically relevant and supportive responses. Our research seeks to bridge these gaps by refining a generalist model in a manner that enhances its ability to handle the nuanced and sensitive nature of mental health dialogues, thereby improving the overall efficacy of VMHAs.

## 2 Related Work

The development of Virtual Mental Health Assistants (VMHAs) has seen significant advancements in recent years, driven by the growing need for accessible and scalable mental health care solutions. Several systems have been introduced, each leveraging different technological frameworks and therapeutic approaches to provide effective mental health support. Wysa [20] is one of the leading VMHAs, recognized for its ability to deliver empathetic and evidence-based conversations. The system is designed to recognize specific patterns in user input that indicate stress, anxiety, or depression. Wysa employs pre-trained language models fine-tuned on mental health-related dialogues to ensure that its responses are both contextually relevant and emotionally supportive. The chatbot also integrates evidence-based therapeutic techniques such as Cognitive Behavioral Therapy (CBT) and Dialectical Behavior Therapy (DBT), providing exercises and tools to help users manage their mental health. Another prominent system is Woebot [4], a chatbot that primarily uses structured conversations to deliver CBT techniques. Woebot's framework combines decision-tree logic with machine learning models that are specifically trained on psychological dialogue data. It's reinforcement learning component further enhances its ability to tailor interactions based on user feedback, improving the overall therapeutic experience. READER(Response-act Guided Reinforced Dialogue Generation for Mental Health Counseling)[14] highlights the limitations of existing dialogue generation models in the mental health domain due to their lack of contextual understanding and leverages transformer-based reinforcement learning (TRL) with Proximal Policy Optimization (PPO) to jointly predict dialogue acts and generate semantically rich responses. Unlike traditional open-domain or task-oriented VMHA systems, it focuses on creating a hybrid conversational flow to effectively engage clients in counseling sessions.

Replika [10] offers a different approach by focusing on building long-term, emotionally supportive relationships with users. The system uses sequence-to-sequence models and recurrent neural networks (RNNs) to generate contextually appropriate responses, adapting to the user's evolving emotional needs over time. However, the system's generalist nature has limited its effectiveness in specific mental health contexts. Tess [5] uses a modular architecture combining rule-based systems and machine learning algorithms to deliver psychological support via text messaging. The system incorporates modules for different therapeutic techniques, such as CBT, DBT, and psychoeducation, allowing it to provide a wide range of mental health interventions. Tess is often integrated into broader mental health care programs, where it complements traditional therapy by providing continuous, on-demand support. Youper [21] is another AI-powered mental health assistant that integrates various therapeutic techniques into its dialogue management system. Combining rule-based logic with machine learning, Youper engages users in brief daily conversations to monitor their mood and provide tailored interventions. Ginger takes a hybrid approach, combining AI-driven chatbots with human mental health professionals to offer comprehensive mental health care. The AI component of Ginger is responsible for initial assessments and immediate support, using pre-trained language models fine-tuned on mental health data. Although Ginger was originally launched in 2011 as one of the first online mental health platform, today it is merged with the Headspace platform [6].

Despite the success of these systems, challenges remain in developing VMHAs that are both effective and widely applicable across different mental health contexts. Given the recent advancements in LLMs, models such as ChatMGL, have promise to develop more effective systems. To address these limitations, our research introduces modifications to the ChatMGL model, including specialized fine-tuning on mental health datasets and the integration of mental health-specific dialogue acts. These enhancements aim to bridge the gap between general-purpose dialogue generation models and the specialized needs of VMHAs, making ChatMGL a more effective tool for virtual mental health care.

## 3 Datasets

We use the HOPE (Mental Health cOunselling of PatiEnts) [8] dataset for our experiments. The dataset is specifically designed for mental health counseling and dialogue-act classification tasks. The HOPE dataset includes approximately 12,800 utterances extracted from 212 mental health counseling sessions, sourced from publicly available YouTube videos. These videos feature a variety of

[1] https://github.com/MChatzakis/ChatMGL

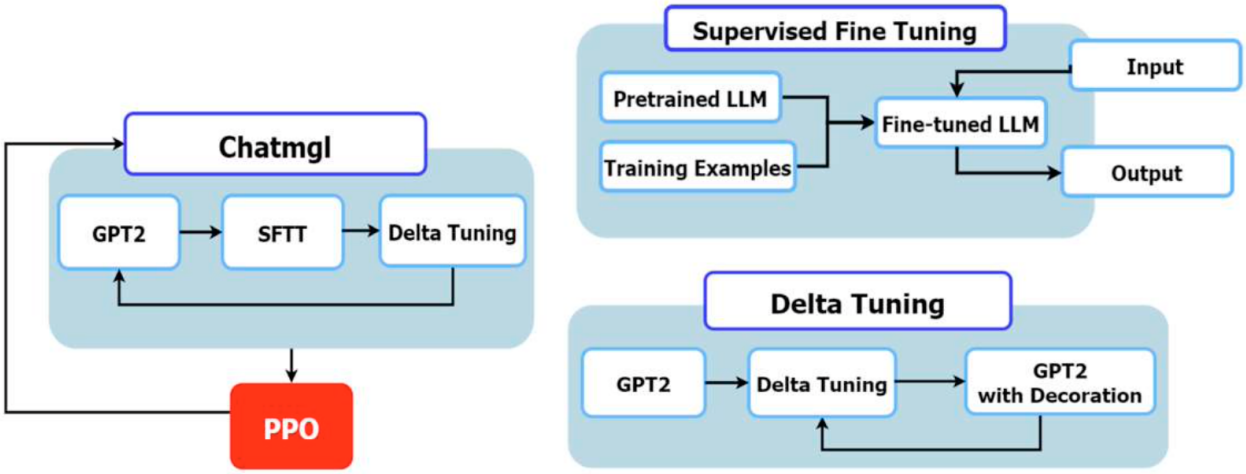



**Figure 1: ChatMGL architecture with our modifications.**

therapists and clients discussing different mental health issues, allowing for a broad representation of counseling styles and methods. Each utterance within the HOPE dataset is annotated with one of twelve unique dialogue-act labels(information-delivery (ID), information request (IRQ), yes/no-question (YNQ), clarification-request (CRQ), opinion-request (ORQ), clarification-delivery (CD), positive-answer(PA), negative-answer (NA), opinion-delivery (OD), greeting (GT), acknowledgment (ACK), general chit-chat (GC)) that were formulated in consultation with mental health professionals, including therapists and counseling experts. These labels aim to capture the nuances and specific needs of therapeutic conversations, distinguishing HOPE from other dialogue datasets. The dialogue-act labels include categories like "empathy," "affirmation," "exploration," and others that are particularly relevant in the context of mental health counseling sessions.

We processed the HOPE dataset to incorporate it into the ChatMGL model. The initial dataset contained irrelevant columns, which were removed, leaving only three key columns: Type (indicating whether the speaker is a patient or therapist), Dialogue_ID (the conversation identifier), and Utterance. An extra column 'target' was also created. Each sentence was used as the input for the models, and the sentence that follows it was used as the desired response. To create the 'target' column, we copied the content from the 'utterance' column, shifting it by one position. To process the data, two different methods were employed. The first data loading method (V1) implemented a "Constructing question-answer pairs" approach. It paired successive patient (P) and therapist (T) statements, treating each P statement followed by a T statement as a question-answer pair. Dialogues were identified by splitting the ID field, ensuring new conversations were recognized and grouped separately. However, grouping too many utterances into one conversation led to a limited number of training samples. To address this, a second data loading method (V2) was introduced, enhancing data by treating each patient's utterance as a separate training sample. This increased the number of training samples by creating multiple samples from a single dialogue. For each sample, the patient's cumulative utterances were paired with the therapist's cumulative responses, effectively increasing the data and improving the model's performance on context information.

## 4 Methodology

### 4.1 ChatMGL

ChatMGL is an LLM designed specifically to answer questions in the domains of Data Science and STEM (Science, Technology, Engineering, and Mathematics). The model is built on the GPT-2 architecture, which has been fine-tuned for specialized tasks. ChatMGL employs a dual training strategy that combines supervised fine-tuning with Reinforcement Learning (RL), particularly utilizing PPO. The model is primarily trained on large datasets consisting of data science and STEM-related questions.

The training process involves supervised fine-tuning with specific question-answer pairs to enhance the model's adaptability to these domains. In addition to supervised fine-tuning, ChatMGL employs a reinforcement learning (RL) strategy using Proximal Policy Optimization (PPO) [13], which iteratively refines the model based on feedback. While PPO is a valuable technique for training RL agents, the reward signal is relatively straightforward in our specific environment, reducing the need for the full exploration and exploitation balance PPO typically provides. PPO's role here primarily serves to calibrate ChatMGL's responses rather than undertake extensive policy optimization. A flow diagram illustrating the core components of our system is shown in Figure 1, highlighting each stage of the pipeline from data input to final answer generation.

### 4.2 SFTT

Huggingface's Supervised Fine-Tuning Trainer (SFTT)[2] is used for adapting pre-trained models to supervised tasks on smaller datasets.

[2]https://huggingface.co/docs/trl/en/sft_trainer

**Table 1: Evaluation table**

| Model | Beam Size 1 Rouge 1 | Beam Size 2 Rouge 1 | Beam Size 4 Rouge 1 | BERT F1 | Meteor |
|---|---|---|---|---|---|
| GPT2 | 10.61 | 6.235 | 5.135 | 0.641 | 0.0441 |
| Baseline - ChatMGL | 4.655 | 2.419 | 2.154 | 0.620 | 0.0311 |
| Delta Tuning | 12.249 | 7.006 | 5.624 | 0.635 | 0.0499 |
| SFTT_V1 | 10.765 | 5.384 | 3.511 | 0.689 | 0.0947 |
| SFTT_V2 | 11.509 | 5.148 | 3.161 | 0.687 | 0.0772 |

**Table 2: Paramater table**

| Model | Learning Rate | Beam Size 1 Rouge 1 | Beam Size 2 Rouge 1 | Beam Size 4 Rouge 1 | BERT F1 | BERT Recall | BERT Precision |
|---|---|---|---|---|---|---|---|
| GPT2-Medium (V1) | $1 \times e^{-4}$ | 35.504 | 29.544 | 29.620 | 0.782 | 0.771 | 0.793 |
| GPT2-Large (V1) | $1 \times e^{-4}$ | 36.027 | 30.667 | 28.650 | 0.783 | 0.773 | 0.793 |
| GPT2-Large (V1) | $1 \times e^{-5}$ | 36.332 | 30.601 | 27.820 | 0.780 | 0.767 | 0.796 |
| GPT2-Large (V2) | $1 \times e^{-4}$ | 21.718 | 11.062 | 7.443 | 0.740 | 0.72337 | 0.759 |

SFTT implements memory management techniques like parameter-efficient fine-tuning (PEFT) to increase memory overhead during training. It simplifies the customization of models for various natural language processing tasks such as text classification, token classification, and sequence-to-sequence tasks.

### 4.3 Delta Finetuning

Delta Finetuning [3] is a fine-tuning technique that focuses on updating only a subset of a model's parameters rather than the entire model. This method is particularly useful when aiming to adapt a pre-trained LLM to a specific task or domain while minimizing computational costs and avoiding overfitting. Delta tuning operates by identifying and modifying the most relevant parameters that contribute to performance improvements, thereby allowing for efficient resource use. Unlike full fine-tuning, where all model weights are adjusted, delta tuning selectively applies changes, preserving the core knowledge encoded in the model while refining its ability to handle specialized tasks. Delta tuning can achieve significant performance gains with a fraction of the training data and computational effort required for full fine-tuning while retaining the knowledge gained by the LLM during the pre-training stage.

## 5 Experimental Results

We evaluate various hyperparameter settings including batch size, learning rate, and model size (GPT-2 large vs. GPT-2 medium), to evaluate the GPT2 model. Table 2 gives the results for hyperparameter evaluation. It can be observed that, when the learning rate is set to 1e-5 with a batch size of 1, and data is loaded using the Vl dataset class, the model achieves the highest ROUGE and BERT scores. Additionally, in terms of model size, GPT-2 large outperforms GPT-2 medium on text generation downstream tasks. For further model evaluations, we used the optimal configuration (GPT-2 large, learning rate 1e-5, batch size 1).

In the final evaluation, the base GPT2 and ChatMGL are used as the baseline models and the performances for SFTT and Deata Tuning are reported. Table 1 presents the performance of various models. As seen in the table, Delta Tuning and SFTT both outperform the baseline models. Although Delta Tuining lags behind SFTT, it has a considerably low computational cost, given Delta tuning trains models separately with a smaller trainable ratio. The comparison focuses on evaluating the ROUGE score, BERT score, and computational resource cost between supervised fine-tuning and delta tuning. In this test, all hyperparameters and data loading methods are kept consistent across both approaches. The evaluation metrics used in the experiments are detailed in Appendix 5.1.

### 5.1 Evaluation Metrics

In consistency with the existing literature [14], we have used the following metrics:

*5.1.1 BERTScore.* BERTScore is an evaluation metric used to assess the quality of text generation by comparing the similarity between the generated text and reference text at a word or token level, leveraging the power of BERT (Bidirectional Encoder Representations from Transformers) embeddings. Unlike traditional metrics like ROUGE or BLEU, which rely on exact word matches, BERTScore calculates the similarity between tokens in a high-dimensional embedding space, capturing semantic similarities. The metric computes a precision, recall, and F1-score based on these embeddings. For a reference sentence $R = [r_1, r_2, \ldots, r_m]$ and a generated sentence $G = [g_1, g_2, \ldots, g_n]$, where $r_i$ and $g_j$ are the embeddings of the tokens in the sentences, the precision is defined as:

$$\text{Precision} = \frac{1}{|G|} \sum_{g_j \in G} \max_{r_i \in R} \cos(g_j, r_i)$$

where $\cos(g_j, r_i)$ is the cosine similarity between the embeddings of token $g_j$ and $r_i$. Similarly, recall is computed as:

$$\text{Recall} = \frac{1}{|R|} \sum_{r_i \in R} \max_{g_j \in G} \cos(r_i, g_j)$$

The BERTScore F1 is then the harmonic mean of precision and recall. This metric is particularly useful for evaluating generative models where semantic similarity is more important than exact word matches, as it considers the contextual meaning captured by BERT's embeddings.

*5.1.2 Beam Rouge.* Beam ROUGE is an extension of the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, which evaluates the quality of text summaries by comparing them to reference summaries using overlapping n-grams, word sequences, and word pairs. In the context of Beam ROUGE, the metric is applied to the output of a beam search decoding process in text generation. Beam search is a method used to generate text by keeping track of the top $k$ (beam width) candidate sequences at each step, rather than just the single best option. The ROUGE score is then calculated for each of these beam candidates.

*5.1.3 Meteor.* METEOR (Metric for Evaluation of Translation with Explicit ORdering) is an evaluation metric designed to assess the quality of machine-generated text, particularly in the context of machine translation and summarization. Unlike other metrics that rely solely on exact word matching, METEOR incorporates various levels of matching, including exact, stemmed, synonym, and paraphrase matches, to capture a more nuanced evaluation of the similarity between the generated text and the reference text. The METEOR score is computed by first aligning the generated sentence $G$ and the reference sentence $R$ based on these different types of matches. The metric then calculates precision and recall, similar to ROUGE, but incorporates a penalty function to account for word order differences. The precision $P$ and recall $R$ are given by:

$$P = \frac{\#\text{ of matched unigrams}}{\#\text{ of unigrams in } G}$$

$$R = \frac{\#\text{ of matched unigrams}}{\#\text{ of unigrams in } R}$$

The harmonic mean of precision and recall, called the F-score, is computed with an added weight $\alpha$ to prioritize recall:

$$F_{\text{mean}} = \frac{P \times R}{\alpha \times P + (1-\alpha) \times R}$$

To penalize incorrect word order, METEOR introduces a fragmentation penalty Penalty based on the number of chunks, which are contiguous and ordered sequences of matches:

$$\text{Penalty} = \gamma \times \left(\frac{\#\text{ of chunks}}{\#\text{ of matches}}\right)^{\beta}$$

The final METEOR score is calculated by applying the penalty to the F-score:

$$\text{METEOR} = F_{\text{mean}} \times (1 - \text{Penalty})$$

Where $\alpha$, $\beta$, and $\gamma$ are hyperparameters tuned to optimize the metric's correlation with human judgment. METEOR's ability to consider synonyms, paraphrases, and flexible word order makes it a robust metric for evaluating text generation tasks, offering a more comprehensive assessment than metrics that rely purely on exact matches.

## 6 Conclusion

In this work, we presented a methodical approach to adapt LLMs for the challenging task of mental health dialogue generation. By specifically focusing on modifications to ChatMGL, including specialized fine-tuning and the integration of domain-specific dialogue act labels, we demonstrated measurable improvements in text generation. Our experiments underscore the importance of specifically aligning LLMs with the complexities of therapeutic exchanges, particularly when empathy and emotional nuance are crucial. These findings suggest that specialized fine-tuning and domain-specific knowledge are crucial for developing effective VMHAs that can provide contextually appropriate and supportive mental health care. In the future, researchers could explore the integration of additional mental health-specific modules, such as empathy detection, emotional response generation, and human-in-the-loop feedback mechanisms into VMHAs. As these systems evolve, the incorporation of ethical oversight and domain-specific guidelines will be essential to ensure that virtual mental health assistants operate responsibly and provide a supportive, clinically relevant resource for users seeking help.

## References


[1] Alaa A Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. 2021. Perceptions and opinions of patients about mental health chatbots: scoping review. *Journal of medical Internet research* 23, 1 (2021), e17828.

[2] Shuroug A Alowais, Sahar S Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I Alshaya, Sumaya N Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A Badreldin, et al. 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education* 23, 1 (2023), 689.

[3] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904* (2022).

[4] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e7785.

[5] Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, Michiel Rauws, et al. 2018. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR mental health* 5, 4 (2018), e9782.

[6] Headspace. 2024. Headspace App. https://www.headspace.com/

[7] Ellen E Lee, John Torous, Munmun De Choudhury, Colin A Depp, Sarah A Graham, Ho-Cheol Kim, Martin P Paulus, John H Krystal, and Dilip V Jeste. 2021. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 6, 9 (2021), 856–864.

[8] Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 735–745.

[9] Usman Naseem, Surendrabikram Thapa, Qi Zhang, Junaid Rashid, Liang Hu, and Mehwish Nasim. 2023. Temporal Tides of Emotional Resonance: A Novel Approach to Identify Mental Health on Social Media. In *Proceedings of the 11th International Workshop on Natural Language Processing for Social Media*. 1–8.

[10] Replika. 2024. Replika Chatbot. https://replika.com/

[11] Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 2650–2656.

[12] Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 2436–2449.

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[14] Aseem Srivastava, Ishan Pandey, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Response-act guided reinforced dialogue generation for mental health counseling. In *Proceedings of the ACM Web Conference 2023*. 1118–1129.

[15] Surendrabikram Thapa and Surabhi Adhikari. 2023. ChatGPT, bard, and large language models for biomedical research: opportunities and pitfalls. *Annals of biomedical engineering* 51, 12 (2023), 2647–2651.

[16] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–50.
[17] Myrthe L Tielman, Mark A Neerincx, Marieke Van Meggelen, Ingmar Franken, and Willem-Paul Brinkman. 2017. How should a virtual agent present psychoeducation? Influence of verbal and textual presentation on adherence. *Technology and Health Care* 25, 6 (2017), 1081–1096.
[18] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry* 64, 7 (2019), 456–464.
[19] WHO. 2024. *Mental Disorders.* https://www.who.int/news-room/fact-sheets/detail/mental-disorders
[20] Wysa. 2024. Wysa Chatbot. https://www.wysa.com/
[21] Youper. 2024. Youper Chatbot. https://www.youper.ai/

# A Appendix

## Limitations

The limitations of this study include the inherent biases present in the pre-trained language model, which may lead to suboptimal performance in providing mental health support across different populations. The dataset used for fine-tuning, while valuable, may not represent the full diversity of mental health issues encountered in real-world settings, leading to potential gaps in the model's ability to address less common or culturally specific concerns. Additionally, the fine-tuning techniques employed, though effective in improving task-specific performance, may result in overfitting, particularly when the dataset lacks variability or is limited in size. This could hinder the model's capacity to generalize well to new data or unforeseen contexts. Furthermore, the model's reliance on text-based inputs might limit its understanding of more nuanced mental health symptoms that could be better captured through multimodal inputs such as tone, facial expressions, or behavior patterns, reducing its effectiveness in real-world applications. Finally, as this is an initial attempt, extensive validation across diverse clinical settings remains necessary to assess the model's robustness and reliability.

## Ethical Concerns

The development of a virtual mental health assistant using a pre-trained LLM primarily may invoke issues of privacy, confidentiality, and data security. Since mental health data is highly sensitive, there is a significant risk that user data may be exposed or misused if adequate safeguards are not in place. Ensuring that personally identifiable information is either anonymized or securely protected during both the fine-tuning process and the use of the assistant is essential to mitigate this risk. Additionally, biases inherent in both the model and the dataset used for fine-tuning could result in the model providing skewed or potentially harmful advice, especially for underrepresented groups. This raises concerns about fairness and inclusivity, as the model may fail to offer equitable mental health support to all users.

Another critical concern is the potential for over-reliance on the virtual assistant by users. While such tools can provide valuable support, they are not substitutes for professional mental health care. Users might perceive the system as an authoritative source, leading them to follow advice that may not be suitable for their specific condition. This creates risks of inappropriate self-diagnosis or inadequate treatment, especially in the absence of professional supervision. To address these concerns, it is crucial to maintain transparency about the limitations of the model and ensure that users are directed to qualified professionals for more serious or complex mental health issues. Ensuring that the system's outputs are clearly communicated as supplementary and not a replacement for human care is vital for ethical deployment.