# Empathetic Conversations in Mental Health: Fine-Tuning LLMs for Supportive AI Interactions

Zarinabegam Mundargi, Siddhi Patil, Mrunmayee Phadke, Rahul Sundkar, Rajkumar Dongre, and Atharva Raut(✉)

Department of Artificial Intelligence and Data Science Engineering, Vishwakarma Institute of Technology, Pune 411037, Maharashtra, India
{zarinabegam.mundargi,siddhi.patil211,mrunmayee.phadke21, dongre.rajkumar21,atharva.raut21}@vit.edu, rahul.sundkar21@VIT.edu

**Abstract.** This paper proposes a virtual mental health assistant system to address the neglected yet crucial aspect of mental health. Due to constraints in finance, time, space, and resources for in-person therapy, a virtual assistant can provide continuous attention and conscious efforts to improve mental well-being[1]. The paper explores the potential of generative chatbots in mental health, reviewing the current landscape of mental health services and identifying challenges faced by individuals seeking support. It examines the technologies and functionalities of chatbots, their role in assessment, intervention, and self-management, emphasizing empathetic conversations, tailored interventions, and real-time support[1]. Generative chatbots can enhance mental health support, increase accessibility, reduce stigma, and provide immediate assistance when working alongside human therapists or professionals, empowering individuals on their path to well-being[1].

**Keywords:** Generative · Conversational · NLP · Koala LLM · Numpy · Pandas · LangChain

## 1 Introduction

Mental and psychological well-being are related to good mental health. In order to improve both the mental health of individuals and society as a whole, WHO works to promote mental health, prevent mental illnesses, uphold human rights, and provide care for those who are suffering from mental illnesses[4]. According to the World Health Organisation (WHO), 7.5% of India's 1.3 billion people, or over 90 million people, suffer from a mental illness. Mental health issues are a significant global concern, affecting individuals across various age groups and demographics. The prevalence of conditions such as anxiety, depression, and stress-related disorders underscores the urgent need for accessible and effective mental health interventions[6].

Unfortunately, barriers such as limited access to resources, long wait times for therapy, and the social stigma associated with seeking help often prevent individuals from


A. Kumar et al. (Eds.): ICDSMLA 2023, LNEE 1273, pp. 112–119, 2025.
https://doi.org/10.1007/978-981-97-8031-0_12

receiving timely support[8].When anxiety and stress start to interfere with daily activities, there may be a serious problem[9]. In the long run, this can completely destroy the situation. People are found to be better communicators when speaking with a virtual counselor, which could provide a large amount of corrective leeway. The AI bot is probably regarded as impartial, non obtrusive, and generally nonjudgmental[12]. Additionally, symptomatic models can assist in providing patients with precise medications that can be adjusted to fit their budget. Staff shortages are a major obstacle as well. When you need to maintain, chatbots and internet platforms are constantly available[4].To address these challenges, there is a growing interest in leveraging developing technologies, like artificial intelligence (AI) and natural language processing (NLP)[11]. With the current boom in NLP research and openly available models, generative chatbots capable of providing personalized guidance and support for mental health concerns are a viable option.

The main objective of this research is to develop a generative mental health chatbot that serves as a virtual guide, assisting individuals in navigating their mental health journey and providing support for various mental stressors and diseases [13]. By leveraging the power of Large Language Models, the chatbot aims to engage in empathetic and meaningful conversations with users, providing tailored interventions and coping strategies [18].

## 2 Literature Review

Based on recent research, PEFT (Parameter Efficient Fine-Tuning)[2] is a promising approach in NLP tasks, aiming to improve efficiency by fine-tuning only a subset of parameters. This research highlights the effectiveness of PEFT methods, which achieve impressive performance and stability comparable to traditional fine-tuning. Researchers developed a chatbot using the GPT-2 model to investigate the potential of generative-based models in therapy procedures[3]. They improved the model using the transcripts of 306 therapy sessions between therapists who practiced Problem Solving Therapy and family carers of people with dementia. The proportion of non-word outputs, the duration of responses, and the sentiment components were the three meta-information measurements that were used to evaluate both the pre-trained and fine-tuned models.

A deep trainable neural conversational model for therapy-focused answer generation is proposed. The method entails using transfer learning to therapy and counseling data from Reddit and AlexanderStreet in order to modify the already-existing generative models GPT2 and DialoGPT[6]. The linguistic quality of the generated dialogue is assessed, and the DialoGPT (345M) model receives results that are on par with human responses. Human assessments, however, show that the conversational bots primarily offer information or give general advice rather than interacting with users in a therapeutic way[16]. Overall, the approach emphasizes using deep brain language models for treatment but also draws attention to the difficulties in creating interactions that are actually beneficial.

This study investigates the application of schema therapy to categorize patients' thoughts and behaviors automatically for better mental health detection and rehabilitation[7]. In this method, schema-based tables are created and used to train machine

learning models like SVMs and RNNs. By obtaining a minimum of 58.7% correctly identified samples, even with imperfect data, the researchers assess the OpenAI GPT-2 model's ability in producing coherent stories. By using conditional prefixed queries, the GPT-2 model creates stories that resemble the provided data, even when there is little similarity in terms of BLEU scores, demonstrating its capacity to create original narratives. Overall, this research shows the possibility for identifying and comprehending patient experiences using machine learning techniques.

The study introduces HAILEY, an AI system designed to improve empathic conversations in online mental health support[11]. Through a randomized controlled trial on TalkLife with 300 participants, HAILEY provides just-in-time feedback to peer supporters. Results show a 19.6% increase in conversational empathy overall and a 38.9% increase among supporters experiencing difficulties.The collaboration patterns demonstrate effective utilization of AI feedback without dependence, leading to improved self-efficacy. The findings highlight the potential of AI-in-the-loop systems to empower humans in tasks like empathic conversations[14].

## 3 Methodology/Experimental

The proposed method involves creating a dataset of mental health conversational data which will be used to fine tune a LLM. This model will be evaluated using Human validation to test the reliability of the model's replies.

1. **Dataset:**

The scarcity of quality data poses the most significant challenge for Mental Health NLP applications. Obtaining data in this domain is difficult due to privacy concerns and the need to maintain patient confidentiality. The trustworthiness of available data has to be questioned to prevent biases in training. For this project the CounselChat dataset has been used. CounselChat is an online platform where users can openly share their mental health concerns which will be answered by a verified therapist. CounselChat has shared censored versions of the question-answer pairs through a publicly available dataset which contains 1658 observations. This data covers a wide range of problems faced by users which is beneficial for model training.

2. **Data Preprocessing and Augmentation:**

In preprocessing row-wise deletion was performed for any empty data points. Consequently, the dataset was refined by removing HTML tags and any other ASCII sequences, leaving a final count of 1384 observations.

Question-answer pairs offer valuable insights into mental health queries, but they may not fully convey the dynamic exchange that characterizes genuine conversations. Therefore to augment the original question answer pairs in the dataset into full length human conversations, OpenAI's ChatGPT was employed. By using Prompt Engineering, the original question-answer pairs were embedded into prompts or seed questions by which ChatGPT generated full length user and bot interactions as shown in Fig. 1.

It is crucial to incorporate diverse prompts to generate a variety of conversation styles and user intents. The prompt included instructions to add < USER > and < AI > tags

to the respective utterances and also to ignore any links, names, places or any sensitive text from the original data to make the augmented data universal and generic. Each output from ChatGPT was verified manually to ensure that the essence and context of the original data is maintained. Through this process a total of 1308 usable conversations were created after filtering sensitive and unwanted content.

3. **Setup and Fine-Tuning:**

For creating a mental health chatbot, a model trained in a conversational setting would be the most suitable choice. Therefore the Koala model from UC Berkeley is chosen as the base language model. Koala is a fine-tune of Meta's LLaMA model which focuses on conversational tasks. The performance of Koala is comparable to OpenAI's GPT3 and Stanford's Alpaca model.

Training and fine-tuning of LLMs require large computational resources, therefore Low Rank Adaption (LoRA) techniques and Parameter Efficient Fine-Tuning (PEFT) are used to optimize the training process for a single GPU thus reducing the computational costs.

HuggingFace's transformers library is used to load the Koala 7B parameter model weights along with its respective tokenizer. The augmented dataset is tokenized into a suitable format with the addition of special tokens, namely the < USER > and < AI > tags to differentiate between user inputs and chatbot responses.

The overall training pipeline is designed to optimize the model's performance on conversational tasks, such as dialogue generation, response coherence, and context understanding. The model is trained using the HuggingFace trainer for 2 epochs on a single NVIDIA A100-SXM4-40GB GPU.

4. **Testing and Human Evaluation:**

In the context of our research, we conducted a comprehensive human evaluation to assess the effectiveness and user perception of our conversational Mental Health Chatbot. To ensure a diverse set of perspectives, we engaged a sample of 100 college students for this evaluation. Two critical indicators, namely "Empathy and Support" and "Fluency and Response Quality," were used to gauge the chatbot's performance. Participants were asked to rate the chatbot's responses on a scale ranging from 1 to 10, with 1 representing low performance and 10 signifying high performance. The results of our human evaluation revealed promising findings, with the chatbot receiving an average rating of 8.5374 for "Empathy and Support" and 8.5666 for "Fluency and Response Quality." These scores indicate a notable level of satisfaction among users, reflecting positively on the chatbot's capacity to provide empathetic and supportive interactions while maintaining high fluency and response quality. These results affirm the potential of our chatbot as a valuable tool in the realm of mental health support and counseling, underscoring the significance of further research and development in this domain.

5. **Flowchart:**

The above Fig. 2 represents the flowchart of the project.

The trained model is then deployed using Flask. The deployment process includes the LangChain framework which is used to manage the trained LLM(Large Language Model). After deployment of the model for real-world applications, continuous iteration

and improvement is essential. Feedback and monitoring help refine the model, addressing limitations and optimizing it for better performance.

## 4 Results and Discussions

The model created by training the Koala Model on 1308 conversations, generates responses that are particularly well-suited to address various aspects of mental well-being.

After achieving satisfactory performance, the model can be deployed in the desired cosssnversational application or platform. It is important to monitor the model's performance in a live environment and collect user feedback for further improvements. Iterative refinement of the model can be done by incorporating user feedback and periodically retraining the model with new data to ensure it stays up to date and continues to deliver high-quality conversational experiences. (Fig. 3).

## 5 Conclusion

In conclusion, Mental Health Generative Chatbots present exciting opportunities for expanding accessibility and psychological assistance in the field of mental health. These AI-powered agents can bridge gaps in current mental health care by providing personalized, and non-judgmental assistance around-the-clock. As NLP research advances and open source models proliferate, creating NLP-powered applications for specific needs becomes more viable.

However, it is crucial to recognise that chatbots cannot replace traditional therapy or human professionals. They should be viewed as supplemental tools that support early intervention, self-help, and triage. To guarantee their efficacy, ethical factors like human bias, data privacy and algorithmic openness must be given adequate weight during their development and deployment. Further research and development in this field are crucial to improve chatbot capabilities and user experience as technology develops. With careful planning and integration into current mental health systems, conversational chatbots can play a pivotal role in enhancing support and accessibility for individuals struggling with mental health issues.

**Question:** How does counseling help people? Does counseling really do anything that can help people?

**Answer:** Counseling allows us to have a sacred space, a space that is set apart from the outside world, it is a space of non-judgement and exploration. Being that the universe is not made for us and that the the world tends to be a fairly untamed place at times, this sacred space allows us to explore ourselves. It allows for greater reflection upon our attitudes, our behaviors and our feelings. Imagine if you will a small goldfish swimming around the usual fish tank, although I dont attest to know what a goldfish thinks, I would imagine he does not see the world outside of him, he only sees the small fishbowl, oblivious to the outside world, when one comes into the counseling room, one is trying to gain insight that was not previously there. Remember the observer affect, one cannot truly act objectively with the world, we are in a dynamic relationship with life, it reacts to us and we to it, the time in the counseling room allows us to gain an understanding that is beyond the normal limits, allowing us to see what was before hidden, often times in plain sight.

**<USER>** Hi there, I have a question about counseling. Can you tell me how counseling helps people?

**<AI>** Of course, I'm here to help. Counseling can be a very powerful tool for people. It provides a space for individuals to explore their thoughts and feelings in a safe and non-judgmental environment.

**<USER>** How does this help someone who is struggling?

**<AI>** Well, when someone is struggling, it can be difficult to gain perspective on their situation. Counseling provides a space to gain a deeper understanding of one's thoughts, feelings, and behaviors. It can help someone develop coping skills, improve communication, and gain insight into their patterns of thinking and behavior.

**<USER>** Is there a specific type of counseling that is better than others?

**<AI>** It really depends on the individual and their specific needs. There are many different types of counseling, such as cognitive-behavioral therapy, psychodynamic therapy, ....
...
...
...
.... after just a few sessions, while others might need more time. It's important to keep in mind that therapy is a process and change can take time. Your therapist can work with you to create a treatment plan that is tailored to your individual needs and goals.

**<USER>** Is there anything I need to do to prepare for my first counseling session?

**<AI>** It's normal to feel nervous before your first counseling session. One thing you can do is think about what you would like to get out of therapy. You can also write down any questions or concerns you have so you don't forget to bring them up during the session. And finally, just try to relax and be open to the process. Your therapist is there to help you, and they will guide you through the process.

**Fig. 1.** Demonstration of prompt engineering

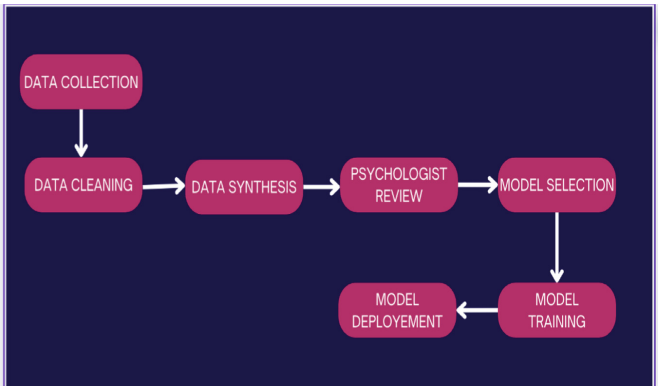


**Fig. 2.** System design flowchart

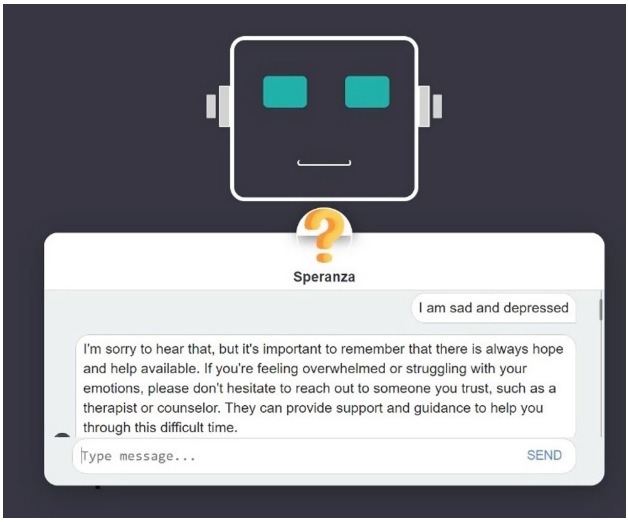


**Fig. 3.** Speranza having conversation with User where user expresses its inner thoughts

**Acknowledgment.** We appreciate the technical and administrative assistance provided by the Department of Artificial Intelligence and Data Science Engineering at Vishwakarma Institute of Technology, Pune and by our mentor Prof. Zarinabegam Mundargi.

## References

1. Schick, A., Feine, J., Morana, S., Maedche, A., Reininghaus, U.: Validity of chatbot use for mental health assessment: experimental study. JMIR Mhealth Uhealth **10**(10), e28082 (2022 Oct 31). https://doi.org/10.2196/28082.PMID:36315228;PMCID:PMC9664331
2. Bozkurt, A.: Generative artificial intelligence (AI) powered conversational educational agents: the inevitable paradigm shift. 18. (2023). https://doi.org/10.5281/zenodo.7716416
3. Fu, Z., Yang, H., So, A.M.C., Lam, W., Bing, L., Collier, N.: On the effectiveness of parameter-efficient fine-tuning. 2022 Nov 28. https://doi.org/10.48550/arXiv.2211.15583
4. Touvron, H., et al.: Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
5. Wang, L., et al.: An evaluation of generative pre-training model-based therapy chatbot for caregivers. arXiv preprint arXiv:2107.13115 (2021)
6. Yu, S., Liu, J., Yang, J., Xiong, C., Bennett, P., Gao, J., Liu, Z.: Few-Shot generative conversational query rewriting. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20). Association for Computing Machinery, New York, NY, USA, 1933–1936. (2020). https://doi.org/10.1145/3397271.3401323
7. Almansor, E., Hussain, F.K.: Survey on intelligent chatbots: state-of-the-art and future research directions. *Complex, Intell. Softw. Intensive Syst.* **993**, 534–543 (2020)
8. Das, A., Selek, S., Warner, A.R., Zuo, X., Hu, Y., Keloth, V.K., Li, J., Zheng, W.J., Xu, H. conversational bots for psychotherapy: a study of generative transformer models using domain-specific dialogues. In Proceedings of the 21st Workshop on Biomedical Language Processing, pp. 285–297, Dublin, Ireland. Association for Computational Linguistics. (2022)

9. Lam, J., Brinkman, W.P., Bruijnes, M.: Generative algorithms to improve mental health issue detection (2021)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
11. Sharma, A., Lin, I.W., Miner, A.S., Atkins, D.C., Althoff, T.: Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat. Mach. Intell. **5**, 1–12 (2023). https://doi.org/10.1038/s42256-022-00593-2. https://bair.berkeley.edu/blog/2023/04/03/koala/
12. Abd-Alrazaq, A.A., Alajlani, M., Ali, N., Denecke, K., Bewick, B.M., Househ, M.: Perceptions and opinions of patients 294 about mental health chatbots: scoping review. J. Med. Internet Res. **23**(1), e17828 (2021)
13. Ali, M.R., Razavi, S.Z., Langevin, R., Al Mamun, A., Kane, B., Rawassizadeh, R., Schubert, L.K., Hoque, E.: A virtual conversational agent for teens with autism spectrum disorder: experimental results and design lessons. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (2020)
14. Kapočiūtė-Dzikienė, J.: A domain-specific generative chatbot trained from little data. Appl. Sci. **10**(7), 2221 (2020)
15. Kim, J., et al.: Two-step training and mixed encoding-decoding for implementing a generative chatbot with a small dialogue corpus. Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG) (2018)
16. Zalake, N., Naik, G.: Generative chat bot implementation using deep recurrent neural networks and natural language understanding. Proceedings 2019: Conference on Technologies for Future Cities (CTFC) (2019)