Dialogue Generative Model for Mental Therapy

*Senior Project Final Report

Project Advisor - Minho Lee

Askhat Sametov Eldana Sovetova Makhabbat Bakytzhan
Daniiar Suiunbekov
{askhat.sametov, eldana.sovetova, makhabbat.takishova, daniiar.suiunbekov}@nu.edu.kz
Nazarbayev University

***Index Terms*—Mental health, generative models, cosine similarity, open-domain question-answering, gpt2, transformers**

I. Executive Summary

The increasing demand for accessible mental health resources has inspired the development of intelligent systems capable of offering immediate support. In this study, we present a mental health chatbot designed to provide empathetic and contextually relevant assistance to users. We initiated our approach by employing foundational NLP techniques, related to open-domain question answering methods, to establish a baseline understanding of conversational dynamics. Subsequently, we explored the application of generative models to capture the nuanced and sensitive nature of mental health queries. Our methodology started with the full fine-tuning of smaller models, which provided initial insights into the training process of LLMs. Building upon this foundation, we leveraged the Low-Rank Adaptation (LoRA) [1] technique to fine-tune larger, more complex models, thus harnessing their superior generative capabilities without the computational expense of training from scratch. The success of our models can be attributed to the properly formatted and curated dataset, which was crucial in training the chatbot to understand and respond to a diverse range of mental health queries. The final chatbot demonstrates promising capabilities in delivering instant, reliable, and empathetic interaction, marking a significant step forward in digital mental health assistance. Our work not only showcases the potential of hybrid NLP applications in mental health scenarios but also paves the way for further innovations in therapeutic conversational agents.

II. Introduction

Millions of people around the world suffer from mental health problems, yet there is often a shortage of human therapists and barriers that prevent people from seeking help. For instance, some individuals find it difficult to receive mental health treatment due to factors such as cost, fear, shame, humiliation, etc. As a result of this, we intend to create an AI-driven, easily-accessible, and efficient mental health support system that offers real-time support. Our goal is to improve people's mental health and create technological tool that complements traditional therapy services. We aim to contribute to mental health therapy by utilizing current technical advancements in deep learning and natural language understanding. Our solution is mainly focusing on the usage of modern generative models, especially Large Language Models. We will perform additional training to make these models suitable for mental health domain. Since training these models is computationally heavy, we employed techniques that allow efficient training that produce reasonable results. In this report, we present the related works similar to our topic, our approaches in the creation and the technical processes behind the development of TherapifyAI, as well as the results of the executable trained system with analysis for future improvement.

III. Related Work

The burgeoning intersection of artificial intelligence (AI) and mental health has seen a surge in innovative approaches to address psychological well-being. Central to this evolution are generative language models, with the Generative Pretrained Transformer (GPT) series at the forefront. GPT-2, a predecessor to GPT-3, serves as a crucial milestone in the development of large-scale language models.

GPT-2, with its 1.5 billion parameters, demonstrated a remarkable leap in natural language understanding and generation [2]. This model's success paved the way for subsequent advancements, leading to the creation of GPT-3, which boasts a staggering 175 billion parameters. GPT-2, recognized for its capacity to generate coherent and contextually relevant text across diverse domains, laid the groundwork for the exploration of generative models in various applications, including mental health therapy [2].

The work by Desiree Bill and Theodor Eriksson, in their paper "Fine-Tuning a LLM Using Reinforcement Learning from Human Feedback for a Therapy Chatbot Application," delves into the nuanced process of adapting large language models (LLMs) for therapeutic dialogue generation [3]. By leveraging reinforcement learning from human feedback, their approach acknowledges the importance of continuous refinement through interaction with human experts. This methodology not only improved the model's performance but also addressed the challenges associated with generating empathetic and contextually appropriate responses in a therapeutic setting. In a parallel exploration, Patel et al.'s work on "Combating Depression in Students using an Intelligent ChatBot:

A Cognitive Behavioral Therapy" emphasizes the practical application of chatbot technology within the framework of Cognitive Behavioral Therapy (CBT) for addressing depression in students [4]. The integration of CBT principles into an intelligent chatbot offers a promising avenue for scalable mental health support. Further expanding the landscape, Sharma et al.'s paper, "Digital Psychiatry - Curbing Depression using Therapy Chatbot and Depression Analysis," delves into the realm of digital psychiatry by combining therapy chatbots with depression analysis tools [5]. This work underscores the potential synergy between advanced language models and targeted therapeutic interventions, presenting a holistic approach to mental health care. A thorough study by Abd-alrazaq and Alajlani highlights the growing enthusiasm for using chatbots for mental health applications. Their research methodically examines current empirical studies to outline the scope of chatbot usage in mental health treatments. By thoroughly searching bibliographic databases and carefully examining 53 individual studies, the review summarizes findings from 41 different chatbots. The review's key finding is about how chatbots function. Although written language is still the main form of input, output modes now include a mix of written, spoken, and visual elements, increasing accessibility and user interaction. The results of their research reveal that chatbots frequently take control of conversations, leading interactions with users, ultimately promoting participation and directing discussions [6].

The proposed project aims to build upon these foundational works, especially the insights gained from GPT-2 and the fine-tuning methodology proposed by Bill and Eriksson, to develop a Dialogue Generative Model for Mental Therapy powered by GPT-3 technology. The colossal parameter count and diverse training data of GPT-3 offer unparalleled potential for capturing nuanced language patterns and contextual understanding [2]. By integrating the findings from Patel et al. and Sharma et al., the goal is to create a responsive, context-aware, and evidence-based conversational agent tailored specifically for mental health therapy. This collaborative approach draws inspiration from the success of GPT-2, the refinement methodologies proposed by Bill and Eriksson, and the practical applications explored by Patel et al. and Sharma et al., creating a comprehensive foundation for the proposed project. By building upon the gained insights from the related works, this project seeks to advance the state-of-the-art in therapeutic chatbot applications. Through the integration of cutting-edge technology, this endeavor aligns with the broader mission of leveraging artificial intelligence for positive societal impact and well-being.

IV. Project Approach

This project introduces an advanced web-based mental health chatbot, engineered to provide immediate and empathetic support to users. The chatbot leverages the possibilities of powerful LLMs by extending to mental health domain through careful training process.

A. Software Architecture

We build the web application using Python programming language, Flask framework for the backend, and HTML, CSS, and JavaScript for the front end.

The messaging system uses a stack data structure. It stores incoming messages in a stack and generates answers for the messages from the top of the stack to the bottom in a sequence. The model receives the messages, generates answers, and sends them back to the user.

The application makes asynchronous requests to the server to improve efficiency and consider cases when a user sends several request messages in a row.

B. Model Training and Third-pary components

The base of our chatbot is the gpt2-large model, which has been fine-tuned using LoRA (Low-Rank Adaptation), a technique that allows us to effectively enhance the model's capabilities without the extensive computational cost associated with training large models. The training process was executed on a GPU-enabled environment using the PyTorch framework and the PEFT library, which facilitated efficient and scalable optimization.

V. Project Execution

A. Cosine Similarity

We initially trained an LSTM model for intent classification, paired with the 'all-mpnet-base-v2' model from Sentence-Transformers for user query vectorization. Then we used cosine similarity to align user queries with relevant pre-existing questions in our dataset, the answer from the most similar question was given as the response for the given query.

Such approach was successful in handling straightforward queries and provided accurate responses for high-frequency questions. But it struggled with nuanced queries and those involving less common topics. The model's initial accuracy was promising but heavily relied on the dataset's diversity and volume.

B. Open-Domain Question Answering

This approach involved a two-staged system with a Retriever model that fetched data from Wikipedia and a BERT-based [7] Reader model that extracted answers from given Wikipedia passages. This system was constrained by high inference times and was limited to strictly answering questions without conversational capability.

C. Generative Models

Realizing the limitations of the previous approaches, we decided to adopt a more sophisticated model that could generate responses rather than retrieving them. For this we imported pretrained GPT-2 smallest model, which has 124 million parameters, and its corresponding vectorizer, with approximately 50 thousand tokens. With the transition to using GPT-2, the architecture of our chatbot underwent significant enhancements to accommodate the increased computational demands. GPT2-small was fully fine-tuned using the prepared medical dataset (at that time we did not find appropriate dataset for mental health domain). The full fine-tuning process involved training the whole model, which took significant amount of time and resources. Unlike the static responses in the previous setup, GPT-2 allowed for more flexible and varied responses. The fine-tuned GPT-2 model significantly improved interaction quality, making conversations more engaging and contextually relevant.

Here is an example of this model's answers:

Question: "My husband is taking 40 mg of prozac and is really depressed and has thought of suicide what do we do?"

Generated answer of fine-tuned model: "if your husband is going through this he needs treatment and his medication is really not enough to help him."

VI. Second Semester Methodology

A. Data Preprocessing

After thorough investigation of different datasets, we finally decided to utilize 'marmikpandya/mental-health' dataset available at HuggingFace. It comprises a collection of approximately 13,000 entries, each containing a mental health-related query and corresponding professional responses.

To facilitate training with a large language model the data was concatenated into a single column, merging both the input queries and the output responses. This transformation was essential to streamline the input format for LLM training, allowing the model to better understand the context and flow of conversation in mental health consultations.

To ensure compatibility with the memory constraints of our machine, the dataset underwent a trimming process. Entries exceeding 1000 characters were excluded from the dataset to prevent memory overflow and to maintain processing efficiency during model training.

An essential aspect of preparing the dataset was the removal of noise and sensitive information to uphold privacy standards and improve data quality.

This included:

- Removing Hyperlinks and Numerical Data: Regular expressions (regex) were utilized to identify and remove entries containing website links and numbers. This step was crucial to eliminate external references and irrelevant numerical data that could detract from the model's learning focus.
- Name Removal Using Named Entity Recognition (NER): To further protect privacy and prevent the model from generating personal information, names identified through NER techniques were removed from the dataset. The following NER model was used for this task: 'dbmdz/bert-large-cased-finetuned-conll03-english'.

The dataset was reduced to 10,551 rows. This cleaned dataset not only meets the technical constraints of our machine but also aligns with ethical guidelines by excluding personal and sensitive information. This dataset was then tokenized using tokenizer of our model.

Some of the removed queries were used as test set, which resulted in 531 query-answer pairs. Test set would be used for evaluation purposes.

The data preparation phase was important in setting a strong foundation for the subsequent development stages of our mental health chatbot. By carefully curating and refining the dataset, we have ensured that the training material is both high-quality and compliant with privacy standards, which is essential for the responsible deployment of AI in mental health applications. The next phase will involve the actual training of the LLM, where the prepared dataset will be used to teach the model the nuances of professional and empathetic responses to mental health inquiries.

B. Parameter Efficient Fine-Tuning (PEFT)

To enhance the GPT-2 model without substantially increasing the number of trainable parameters, we implement Low-Rank Adaptation (LoRA). This technique introduces low-rank matrices that transform the attention module weights in a parameter-efficient manner [1]. Initially, we applied LoRA to the 'bigscience/bloomz-560m' model [8], which is a multilingual large language model with 560 million parameters.

The LoRA configuration is set with several parameters to optimize learning:

- Rank (r): Determines the size of the low-rank matrices. A higher rank allows for more complex adaptations but increases the numbers of trainable parameters.

- LoRA Alpha: A scaling factor that adjusts the magnitude of changes applied by the low-rank matrices.

- LoRA Dropout: Applied to the low-rank matrices to prevent overfitting by randomly zeroing parts of the data during training.

- Bias Mode: Configured to only train the bias parameters introduced by LoRA, maintaining pre-trained biases of the original model u ntouched.

The LoRA parameters for the 'bigscience/bloomz-560m' model are as folows:

- Rank (r): 4
- LoRA Alpha: 1
- Target Modules: ['query_key_value']
- LoRA Dropout: 0.05

This configuration resulted in 393,216 trainable parameters out of a total of 559,607,808.

The training process is designed to leverage computational efficiency. An automatic batch size finder is employed to optimize the use of available computational resources. The learning rate is set higher than typically used in full fine-tuning. It is set to 5e-5.

Training is done over 20 epochs, allowing the model sufficient iterations to learn from the data while balancing the computational costs associated with longer training periods.

The adapted PEFT model is trained using the Trainer class from the HuggingFace library. This class manages the training process, including data collation, model updates, and integration of training arguments. The data collator is configured specifically for causal language modeling, focusing on generating coherent text based on the given context without masking any part of the input.

In order to test the same setup with more complex model, we shifted our focus towards gpt2-large model [9]. The LoRA configuration setup was the same as for bloomz model, except for the target modules which were set as 'c_attn' of the GPT-2 architecture, which is critical for the model's ability to focus on relevant parts of the input data when making predictions. This configuration resulted in 737,280 trainable parameters out of total 774 million parameters.

VII. Results and Evaluation

A. Evaluation Metrics

We concentrated on the quality of generated responses to mental health queries by employing metrics that measure semantic similarity and fluency. The chosen metrics were embeddings cosine similarity and BERT-score [10] for semantic analysis and GPT2 perplexity for fluency. The selection of BERT-score over traditional metrics such as BLEU or ROUGE was driven by the necessity for a deeper semantic understanding in model responses. BLEU [11] and ROUGE [12] primarily focus on n-gram overlap, which might not fully capture our models' performance since there can be various answers or response paraphrases to the same query. BERT-Score, on the other hand, utilizes contextual embeddings from BERT models, allowing for a more nuanced comparison of semantic similarities between the reference responses from professionals and the predictions made by our models.

BERT-Score computes the cosine similarity of BERT embeddings between words in model-generated responses and reference texts. This metric provides three key statistics: Precision, Recall and F1 score.

Since BERT-Score computes word-level cosine similarity, we also need the evaluation metrics for evaluating how similar ground truth answers are to model predictions in sentence-level. For this we computed cosine similarity based on sentence embeddings and averaged the results across the test set. These sentence embeddings were computed with the model from this papers: [13], [14].

Perplexity measures how well a probability model predicts a sample, it is used here to assess the fluency of the responses. A lower perplexity score indicates that the model is more confident and fluent in its responses.

B. Human Evaluation

For the evaluation of our dialogue generative models for mental therapy, we employed a human evaluation approach. Human evaluation is a widely recognized method for assessing the quality and effectiveness of natural language processing models, particularly in tasks such as dialogue generation where human judgment plays a crucial role in assessing the appropriateness and relevance of responses.

To conduct the evaluation, we designed a structured survey that consisted of 15 identical questions from both dialogue generative models. These questions were carefully selected to cover a range of scenarios and topics commonly encountered in mental health therapy conversations. Each question was accompanied by the responses generated by each model.

We recruited 20 participants from diverse backgrounds, including individuals with varying levels of familiarity with mental health concepts and technologies. Participants were instructed to rate the responses provided by each model on a scale from 1 to 10 based on several criteria, including empathy, relevance, effectiveness in addressing the user's needs, and overall quality of response.

To ensure the reliability and validity of the evaluation results, we implemented several measures:

1. Randomization: The order of questions and the sequence of models (i.e., which model's response was presented first) were randomized to minimize potential biases.

2. Anonymity: Participants were assured of the anonymity of their responses to encourage honest and unbiased feedback.

3. Training and Instructions: Participants were provided with clear instructions on how to complete the survey and were given a brief overview of the purpose of the evaluation to ensure understanding.

4. Data Analysis: Upon completion of the survey, the responses were aggregated and analyzed to identify trends, patterns, and areas of consensus or divergence among participants.

Results of Survey:

Upon completion of the human evaluation survey, we collected and analyzed the ratings provided by the 20 participants for the responses generated by each generative model. The survey aimed to assess the effectiveness, relevance, empathy, and overall quality of the chatbot responses in addressing the mental health-related questions posed.

The analysis of the survey results revealed several key findings:

1. High Ratings for Empathy and Supportiveness: Participants consistently rated both dialogue generative models highly for their empathetic and supportive responses. The models demonstrated an understanding of the users' emotional states and provided comforting and encouraging messages.

2. Relevance to User Needs: The responses generated by both models were perceived as highly relevant to the users' needs and concerns. Participants noted that the chatbots addressed the specific questions posed by users effectively and provided practical advice and coping strategies.

3. Effectiveness in Providing Information: Both dialogue generative models were praised for their effectiveness in delivering accurate and helpful information regarding mental health disorders, coping mechanisms, and self-help techniques. Participants found the responses informative and valuable in increasing their understanding of mental health issues.

4. Recommendations for Professional Help: Participants appreciated the models' ability to recommend seeking professional help when necessary. The chatbots appropriately identified situations where professional intervention was warranted and encouraged users to consider consulting a therapist or counselor for further assistance.

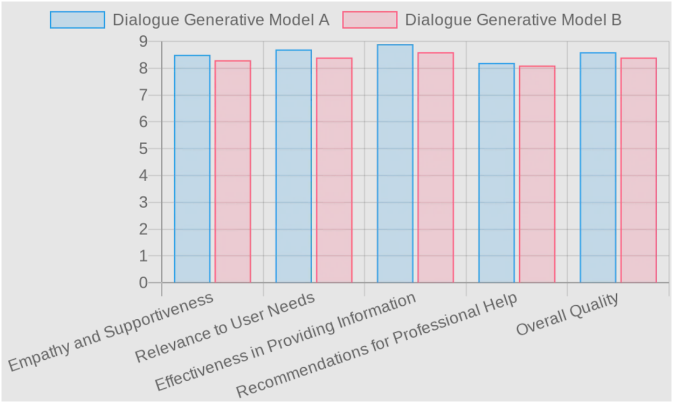

Fig. 1. Comparison of Models

5. Consistency and Reliability: Across the survey responses, there was a high degree of consistency in the ratings provided for each dialogue generative model. This consistency indicates that both models consistently delivered responses that met or exceeded participants' expectations in terms of quality and effectiveness.

Overall, the survey results demonstrated the effectiveness of our dialogue generative models for mental therapy in providing empathetic, relevant, and supportive responses to users' mental health-related queries. The positive feedback from participants validates the utility and potential impact of AI-driven mental health support systems in complementing traditional therapy services and addressing the barriers to accessing mental health care. These findings will inform further refinements and optimizations to enhance the performance and usability of our models, ultimately contributing to improved mental health outcomes for individuals worldwide.

This bar chart represents the average ratings of two dialogue generative models across five evaluation criteria. The ratings are based on the responses of 20 participants who rated the models' performance in terms of empathy and supportiveness, relevance to user needs, effectiveness in providing information, recommendations for professional help, and overall quality. The chart provides a visual comparison of the average ratings for each criterion between the two models, allowing for easy assessment of their relative strengths and weaknesses.

C. Results Analysis

Overall, bloomz and gpt2-large models that were fine tuned using LoRA were compared using above mentioned evaluation metrics. To show the baseline results, we included comparison with the zero-shot results of the gpt2-large model prior to fine-tuning. It should be noted that zero-shot results for the bloomz model were not included in the comparison. This exclusion was due to the observation that the zero-shot predictions from the base bloomz model typically yield empty sentences or results that are substantially irrelevant.

Table I shows the performance of three models on evaluation metrics. We can see that fine-tuned gpt2-large model provided the best results on each metrics.

Cosine Similarity: This metric assesses the degree of similarity between sentence embeddings of the ground truth and the model's predictions, with a higher score indicating greater similarity. Here, the fine-tuned bloomz model achieved a cosine similarity score of 61.80, outperforming the gpt2-large baseline model, which scored 49.44. The gpt2-large model, fine-tuned with LoRA, showed a significant improvement with a score of 66.67, suggesting that fine-tuning has effectively aligned the model's outputs more closely with the semantic space of the ground truth answers.

BERT-Score F1: The BERT-Score F1 is a measure of the weighted overlap between the predicted and reference token embeddings, factoring in precision and recall. The gpt2-large model exhibits a strong performance with a BERT-score F1 of 74.69, surpassing the fine-tuned bloomz model's score of 74.35. The gpt2-large baseline's lower score of 71.14, showed that fine-tuning has bridged this gap significantly.

Perplexity: Perplexity measures how well a probability model predicts a sample, with a lower score indicating better predictive performance. The perplexity for the bloomz model stands at 46.63, which indicates a competent level of prediction accuracy. The gpt2-large baseline, however, registered an extremely high perplexity score (8.6e+41), suggesting a very poor model performance in generating fluent sentences. In contrast, the fine-tuned gpt2-large model achieved a perplexity of 15.27, demonstrating a substantial improvement and suggesting that the predictions from this model are the most fluent.

The task-specific fine-tuning using LoRA has been demonstrably beneficial for both the bloomz and gpt2-large models. Notably, the gpt2-large model shows the most significant gain in performance post fine-tuning, which underscores the efficacy of LoRA as a fine-tuning method for this application. For the practical use of developing a mental health chatbot, the reduced perplexity and increased similarity and alignment with ground truth answers are particularly crucial. These improvements suggest that users are likely to receive responses that are more coherent, contextually appropriate, and semantically rich.

From the human judgments Fig. 1 we also can see that fine-tuned gpt2 model outperforms the bloomz model (Dialogue Generative Model A), and both models provide pretty good results for mental health queries.

The difference in results of both models can be attributed to the differences in their model architectures, specifically the quantity of parameters within each model. The gpt2-large model is designed with a substantially greater number of parameters, both in terms of the trainable parameters and the total number of parameters throughout the network.

Trainable parameters are critical as they are adjusted during the learning process and are essential for the model to capture the nuances and complexities of the language. A higher count of these parameters allows the gpt2-large model to learn the patterns of dataset representations, leading to a richer understanding and the ability to generate more contextually appropriate responses.

Models	cosine similarity	bert-score f1	perplexity
bloomz	61.80	74.35	46.63
gpt2-large baseline	49.44	71.14	8.6e+41
gpt2-large fine-tuned	66.67	74.69	15.27

TABLE I
PERFORMANCE RESULTS.

The fine-tuned models' strong performance across all metrics establishes this models as a viable option for deployment. However, considering the complexity of mental health dialogues, further qualitative analysis, along with user studies, may be required to ensure the chatbot's responses are empathetic, accurate, and truly beneficial for users seeking mental health support.

D. Examples of Generation

1)
- *Prompt:* I'm facing severe depression and anxiety and I just feel like I'm going through a lot. This really distracts me and I cant get my mind off the things that are bothering me.
- *fine-tuned gpt2-large prediction:* It is very important to seek professional help for your mental health issues, as it can be difficult to find someone who understands you or has experience in treating these disorders. You may also want to consider seeking therapy with an experienced therapist if needed. If all else fails, try talking about what's troubling you openly without judgment from others. Try not to let yourself become overwhelmed by thoughts of self-pitying or negative thinking. And remember: there will always be people out there willing to listen when they need support!
- *fine-tuned bloomz prediction:* Anxiety is often associated with trauma or abuse, but it can also be related to stress from past experiences. This may include recent events such as your death of course (or even something similar), which has been linked directly in some studies to increased rates of suicide attempts. An anxious person tends not always have symptoms at first; they tend instead when they're feeling down for someone else than their own benefit will go up so much before being able...then themselves would start acting out....this could lead them into suffering more serious problems later on.

2)
- *Prompt:* I struggle with depression as well as pretty intense mood swings throughout the month. I experience highs where I feel amazing and energetic and then lows where I lack focus, energy, and generally have a more dark outlook on my life. How can I live a more balanced life?
- *fine-tuned gpt2-large prediction:* Depression is an illness that requires professional help to overcome but it's important for you to seek out support from someone who understands your struggles. It may also be helpful if you try some relaxation techniques such as deep breathing or meditation before bedtime so they don't interfere during those times of low self-esteem. Additionally, practicing mindfulness practices like taking time each day to reflect back over what happened in the past few days could provide additional benefits too. Lastly, seeking therapy through a therapist would allow them to work together towards finding solutions to any underlying issues which might contribute to these symptoms. It sounds like there are many different ways to manage this condition including medication options (which we'll discuss later), lifestyle changes, and even cognitive behavioral therapies.
- *fine-tuned bloomz prediction:* It sounds like you are feeling overwhelmed by your feelings of doubt about yourself! You may be struggling to connect them up in reality because it is so overwhelming for some people that they do not see themselves being truly happy or successful at all - even when their thoughts seem unhealthy (think negative self-talk). This could mean working through these issues together such as: What does this affect me most How long am i going over each day before worrying too much ? What would make sense if we did something different every week from now until we're able...then better than ever....to work out what really matters: whether it's having fun doing things right away rather just relaxing next time you're ready?...or how often should one person spend quality moments

3)
- *Prompt:* I self-harm, and I stop for awhile. Then when I see something sad or depressing, I automatically want to self-harm.
- *fine-tuned gpt2-large prediction:* It's a very difficult thing to do but it is necessary in order to feel better about yourself. It may be helpful if you talk with your therapist about what triggers this behavior and how they can help reduce the frequency of these behaviors. If that doesn't work then try talking more openly about any negative thoughts related to self harm such as "Why me?" or "What am I doing wrong?". You might also consider seeking out support groups where people who have experienced similar experiences are able to share their stories.

This will allow them to learn from each other while still feeling safe enough not to seek treatment themselves. Remember though that there isn't one right way to handle anything so don't give up hope just because someone else has tried another approach!

- *fine-tuned bloomz prediction:* Self-self-destruct is not uncommon in the modern age. It can be very dangerous if you are using it as your only escape from reality. You should always seek professional help so that this does n't happen again! If possible try self-help books like The Inner Journey of Death by Dr Loren Smith which have been helpful before and after suicide attempts.

4)
- *Prompt:* It's not entirely true to say I enjoy being sad, but I always find a way to feel that way. I listen to sad music, read tragic stories, and, in a twisted way, like how bad it makes me feel. I focus on negative aspects of my life even if they aren't legitimate or I just make it seem negative.
- *fine-tuned gpt2-large prediction* It is very difficult for someone who has been through so much pain to be able to have such an easy time with sadness. I don't know what the answer is because there are many things you can do when your heart feels this good about something. You may want to try some relaxation techniques (like deep breathing) which will help calm down your mind while also helping release any stress from your body. You could also consider talking to friends/family members about their feelings as well as seeking out support groups where people share similar experiences. If all else fails, then maybe one day we'll see more positive changes than those described above!
- *fine-tuned bloomz prediction* It sounds as though you are feeling very anxious about your situation. If this is the case then perhaps seeking professional help from someone who can provide supportive listening (evaluations) into what you're hearing may be helpful for you. I would suggest seeing an experienced therapist with whom you'd have some great conversations where he could discuss these feelings.

VIII. Conclusion and Future Work

In conclusion, our project has successfully developed a sophisticated mental health chatbot that integrates advanced natural language processing techniques with the robust capabilities of generative models. Hyperparameter tuning, longer training time, bigger Large Language Models with more parameters, different LoRA configuration settings, larger and better structured dataset, more advanced PEFT techniques similar to LoRA like QLoRA, Prompt-tuning and other improvements can be considered as possible future improvements for our project. To preserve the history of the user's problems, the summary of user prompts with model answers can be used as a context to each new prompt, this would produce results that answer given query by taking into the account current mental state of the user. Additionally, further AI-technologies can be added to our project as well. For example, audio-to-text and text-to-audio pipelines can be used to provide more human-like interactions. Nevertheless, the current progress already presents a valuable starting point for further research in using LLMs in mental health therapy. It is important to mention that such technologies would never fully replace real psychologists and only serve as complementary tools for people who find it difficult to access professional assistance at specific time. While the results are promising, continuous improvement and ethical considerations must guide future developments to ensure that such tools offer reliable and empathetic support, reflecting the complex needs of mental health care.

References

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank adaptation of large language models," 2021.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[3] D. Bill and T. Eriksson, "Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application," p. 12, 2023. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-331920

[4] F. Patel, R. Thakore, I. Nandwani, and S. K. Bharti, "Combating depression in students using an intelligent chatbot: A cognitive behavioral therapy," *2019 IEEE 16th India Council International Conference (INDICON)*, 2019. [Online]. Available: https://doi.org/10.1109/indicon47234.2019.9030346

[5] B. Sharma, H. Puri, and D. Rawat, "Digital psychiatry – curbing depression using therapy chatbot and depression analysis," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018. [Online]. Available: https://doi.org/10.1109/icicct.2018.8472986

[6] A. A. Abd-alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, and M. Househ, "An overview of the features of chatbots in mental health: A scoping review," *International Journal of Medical Informatics*, vol. 132, p. 103978, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505619307166

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018.

[8] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf *et al.*, "Crosslingual generalization through multitask finetuning," *arXiv preprint arXiv:2211.01786*, 2022.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[10] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[11] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, "Bleu: a method for automatic evaluation of machine translation," 2002, pp. 311–318.

[12] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[13] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, and A. Panchenko, "ParaDetox: Detoxification with parallel data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).*

Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6804–6818. [Online]. Available: https://aclanthology.org/2022.acl-long.469

[14] J. Wieting, T. Berg-Kirkpatrick, K. Gimpel, and G. Neubig, "Beyond BLEU:Training neural machine translation with semantic similarity," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019.