Talk To Me, Your Virtual AI Therapist: Advancing AI-Driven Psychotherapeutic Engagement with Sentiment Analysis

Stanford CS224N Custom Project

George Birikorang
Department of Computer Science
Stanford University
george25@stanford.edu

Nathan Paek
Department of Computer Science
Stanford University
nathanjp@stanford.edu

Zoe Lynch
Department of Computer Science
Stanford University
zlynch@stanford.edu

Abstract

Not everyone has the opportunity to go to therapy due to cost, stigma, and fear of vulnerability. While an AI therapist could potentially provide much needed mental health support, current large language models (LLMs) such as ChatGPT and Claude do not perform therapy particularly well. In order to create an AI therapist, we first evaluated three pre-trained LLMs (Llama-3-8b, GPT-3.5 Turbo, and Mistral-7B-v0.1) on a real-life patient-therapist conversation dataset to serve as our base model. After observing that GPT-3.5 performed best in terms of therapy response score, a quality metric for therapy conversations, as well as sentiment analysis, we fine-tuned a more resource-efficient variant, GPT-3.5-0125, on both a leading therapy conversation dataset (Amod (2022)) and a sentiment analysis dataset covering six emotions prevalent in therapy. Our fine-tuned model achieved a 21% relative improvement in therapy response score and a 17.9% relative improvement in sentiment analysis accuracy. Our results demonstrate the feasibility of equipping an LLM to act as an AI therapist that is capable of providing higher quality therapeutic and empathetic responses, thereby taking a step towards more accessible and effective mental health support.

1 Key Information to include

- Mentor: Kamyar Salahi
- External Collaborators (if you have any): No
- Sharing project: No

2 Introduction

Not everyone has the opportunity to go to therapy due to reasons such as cost, stigma, and fear of vulnerability. With loneliness already declared by the Surgeon General as national health crisis (General (2024)), the National Academies of Sciences and Medicine (2020) have called for viable alternatives to provide quick and effective therapeutic support. In a national survey conducted by Martinengo et al. (2022) during the pandemic, 47% of adults expressed an interest in talking to a chatbot if needed, and 23% of adults had already talked to a mental health chatbot. This highlights the growing demand for and potential benefit of an AI therapist.

Unfortunately, current large language models (LLMs) like ChatGPT and Claude are not designed for mental health therapy and as such do not perform particularly well on that task. In this project, we investigate the performance of three LLMs (Mistral, Llama, and GPT-3.5) using a quality metric, therapy response score, defined by Pérez-Rosas et al. (2019). In evaluating how these models perform on a therapy conversation dataset provided by Pérez-Rosas et al. (2019), we found that GPT-3.5 performs the best in terms of both therapy response score and sentiment analysis accuracy. As such, we used a more resource efficient variant of the model, GPT-3.5-0125, as our base model for an AI therapist. After fine-tuning the model on both a leading therapy conversation dataset (Amod (2022)) and a sentiment analysis dataset covering six emotions prevalent in therapy, we achieved a 21% relative improvement in therapy response score and a 17.9% relative improvement in sentiment analysis accuracy. By equipping our AI therapist with higher quality therapy conversation skills and better sentiment analysis to provide more empathetic responses, we aspire to make therapy more accessible to people who might otherwise quietly suffer with loneliness and other mental health issues.

3 Related Work

For our project, having a concrete metric to evaluate therapeutic responses is crucial. NLP and psychology researchers such Kilbourne et al. (2018) have long recognized the need for better automated evaluation metrics to facilitate faster AI training and scalable mental health services. Pérez-Rosas et al. (2019) provided a linguistic approach to identify high-quality versus low-quality counseling conversations. Their work emphasizes the importance of using linguistic features such as lexical entrainment, topic focus on behavioral change, balanced turn-taking, and positive sentiment in analyzing counseling quality. As such we utilize the metric of Pérez-Rosas et al. (2019) for evaluating therapy quality instead of more general linguistic metrics such as n-grams (Can et al. (2012)), or audio-based prosody patterns (Xiao et al. (2014)).

In the chatbot research literature, chatbots have been shown to effectively reduce the severity of mental health concerns for people from different demographic backgrounds, including students with anxiety and stress (Fitzpatrick et al. (2017)), veterans and adolescents who feel stigmatized in sharing their concerns (Ly et al. (2017)), and employees of health care systems who require emotional support (Fitzpatrick et al. (2017)). For example, Fitzpatrick et al. (2017) evaluated the effectiveness of the chatbot Woebot in delivering therapy to university students with symptoms of depression and anxiousness. They found that Woebot significantly decreased depressive symptoms compared to an information-only control group. Other efforts have shown that the effectiveness of chatbots may be influenced by anthropomorphic characteristics such as personification and interactivity (Sannon et al. (2018)). Sannon et al. (2018) examined how these factors influence stress-related self-disclosure to conversational agents. They found that a personified chatbot was more likely than a non-personified chatbot or survey to elicit disclosures about finance-related stressors and chronic stressors, but less likely to elicit detailed disclosures or disclosures about home life. However, as noted by Kretzschmar et al. (2019), some chatbots seem unable to understand the complex use of language associated with a mental health crisis, failing to recognize symptoms and respond appropriately. Our project seeks to remedy this defect by incorporating a deeper understanding of sentiment through fine-tuning on sentiment analysis, which is correlated with more effective therapy (Pérez-Rosas et al. (2019)).

4 Approach

In order to develop an AI therapist, we evaluated three LLMs as candidates for our base model. First, we chose the Llama-3-8b model because of its strong performance on natural language tasks ((Touvron et al. (2023)), despite being smaller at only 8B parameters than other state-of-the-art models. We used the "unsloth/llama-3-8b-bnb-4bit" version of the model and initialized it with 4-bit quantization using the FastLanguageModel class to fit our computational constraints. We then fine-tuned the model using LoRA (Low-Rank Adaptation) with parameters r $= 16$, lora_alpha $= 16$, and lora_dropout $= 0$, and patched it with Unsloth for enhanced performance.

As our second candidate, we selected the Mistral-7B-Instruct-v0.1 model, a fine-tuned version of Mistral-7B that is optimized for instruction-following (Jiang et al. (2023)). Despite having 7B parameters, Mistral-7B utilizes grouped-query attention and sliding window attention for efficient inference speed and memory usage. To work within our Google Colab limitations, we used a

quantized version of the model ("ybelkada/Mistral-7B-v0.1-bf16-sharded"). We modified the code to generate responses with sampling and a maximum length of 200 tokens.

Finally, as our third candidate, we included OpenAI's GPT-3.5 Turbo, a large language model with 175B parameters built on the transformer architecture Ye et al. (2023). We accessed GPT-3.5 via the OpenAI API and prompt-engineered it to generate appropriate responses for each input message.

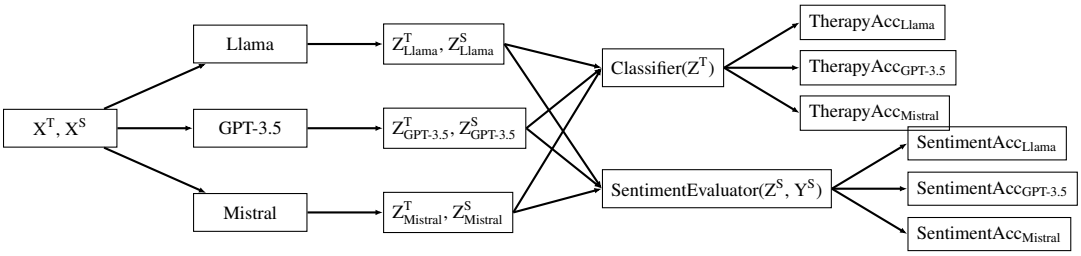
In order to compare how well each of the three LLMs performed, we trained a binary logistic regression classifier on the Pérez-Rosas et al. (2019) dataset to classify whether a therapy response to a given patient context is of high or low quality. Logistic regression is a statistical method for predicting binary outcomes based on input features. In our case, the input features are derived from the therapy conversations, and the binary outcome is whether the therapist's response is classified as high or low quality. The logistic regression model learns a set of weights for each input feature, which are then used to calculate the probability of the response being high quality. The probability was then thresholded at 0.5 to make the final classification decision. Mathematically, the logistic regression model can be expressed as:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} \tag{1}$$

where y is the binary outcome (1 for high quality, 0 for low quality), $x = (x_1, x_2, ..., x_n)$ is the vector of input features, and $\beta = (\beta_0, \beta_1, ..., \beta_n)$ are the weights. If the resulting P is > 0.5, we classify as the response as high quality; else, we classify the response as low quality.

To train our logistic regression classifier for therapy response, we first extracted a variety of linguistic features from the therapy conversations in the Pérez-Rosas et al. (2019) dataset. These features were then combined into a TF-IDF feature vector representation for each conversation, which served as the input to the logistic regression model. We used the TfidfVectorizer to transform the text data into TF-IDF features. The model was trained using the logistic regression algorithm, which optimizes the model parameters to minimize the binary cross-entropy (BCE) loss. After training, our logistic regression classifier achieved a final accuracy of 0.9038 on the test set.

Using the therapy response classifier, we compared how well each of the three LLMs generated therapeutic responses when prompted with real therapy session contexts and asked to play the role of an expert, professional therapist. As shown in the diagram below, we fed the therapy context X^T (patient messages) from the Pérez-Rosas et al. (2019) dataset into each model and generated corresponding therapist responses (Z^T_{Llama}, $Z^T_{\text{GPT-3.5}}$, Z^T_{Mistral}).



We then evaluated these responses using our logistic regression classifier, which labels each response as either high or low quality. The overall therapy accuracy score for each model was calculated as:

$$\text{score} = \frac{\text{number of highs}}{\text{number of highs} + \text{number of lows}} \times 100.$$

In addition to evaluating the three LLMs' generated therapy responses, we evaluated their sentiment analysis performance. As shown in the diagram, we input a sentiment analysis dataset compiled by Dair-AI (2022) (X^S) spanning six emotions (sadness, fear, anger, surprise, love, and joy) into each model and prompted them to return the most relevant emotion for each text sample. The models' predicted emotions (Z^S) were then compared to the ground truth labels (Y^S) using our Sentiment Evaluator function, which calculates the sentiment accuracy as:

$$\text{score} = \frac{\text{number of correct matches}}{\text{total number of predictions}} \times 100.$$

The predicted emotions were normalized by removing punctuation, numbers, and non-letter characters, and converting the text to lowercase to match the label format. The predicted emotions were normalized by removing punctuation, numbers, and non-letter characters, and converting the text to lowercase to match the label format.

Our decision to incorporate sentiment analysis into our approach was motivated by the concept of transfer learning. Transfer learning is a technique where knowledge gained from solving one problem is applied to a different but related problem. By leveraging the learned features and representations from a source task, the model can improve its performance on a target task with limited training data or computational resources. In their paper, Pérez-Rosas et al. (2019) highlight sentiment analysis as one of the crucial linguistic cues found in high-quality counseling. By fine-tuning the model on this related source task (sentiment analysis), we aimed to provide the AI therapist with a better understanding of emotions and sentiment, which could then be transferred to the task of generating therapeutic responses. This approach allows the model to leverage its acquired knowledge of sentiment to generate more empathetic and contextually appropriate responses, thereby improving its performance on the therapy task.

5 Experiments

5.1 Data

We used three datasets in our experiments. The first two are therapy conversation datasets, and the third is a sentiment analysis dataset.

The first therapy dataset, provided by Pérez-Rosas et al. (2019), contains patient-therapist conversations in the form (X^T, Y^T), where X^T represents the patient context statement and Y^T represents the corresponding therapist response. The superscript T denotes that it is a therapy dataset. We used this dataset to train our logistic regression classifier for evaluating the quality of generated therapist responses.

The second therapy dataset is the Mental Health Counseling Conversations dataset from Hugging Face (Amod (2022)). This dataset is also in the form (X^T, Y^T). To create our test set for evaluating our models on the therapy task, we set aside 10 conversations out of 100 from this dataset. We then fine-tuned GPT-3.5 on the remaining conversations from the Amod dataset, followed by fine-tuning the resulting model on the entire Pérez-Rosas et al. (2019) dataset.

The sentiment analysis dataset Dair-AI (2022) is in the form (X^S, Y^S), where X^S is a sentence and Y^S is the corresponding sentiment label. The superscript S denotes that it is the sentiment dataset. The labels are integers ranging from 0 to 5, each corresponding to a specific emotion in the list ["sadness", "joy", "love", "anger", "fear", "surprise"]. To form our sentiment dataset for fine-tuning and evaluation, we set aside 10% of the data for testing and used the remaining 90% for fine-tuning our model. All datasets were preprocessed using the TfidfVectorizer for text vectorization.

5.2 Evaluation method

To classify the output of our baseline models as either high quality or low quality, we used the evaluation metric defined by Pérez-Rosas et al. (2019) through our logistic regression classifier. Their evaluation metric combines multiple linguistic cues, such as:

- Meta-features: The number of turns by counselor and client, the average words per turn for counselor and client, and the ratio of counselor to client words. Good counselors have a more balanced word exchange with their clients, suggesting that they listen more and let clients to express themselves.
- Linguistic Alignment: Good counselors mirror their patients' language. The similarity in function word usage between counselor and client was calculated as $LSM = 1 - \frac{1}{C}\sum_{c=1}^{C} |p_c(c) - p_p(c)|$, where $p_c(c)$ represents the percentage of words from category c

used by the counselor, $p_p(c)$ represents the percentage of words from category c used by the patient, and C is the total number of function word categories. Additionally, the LSC score was calculated (the counselor's language adaptation to client's over time) using a sliding window approach $LSC(t) = 1 - \frac{1}{C}\sum_{c=1}^{C}|p_c(c,t) - p_p(c,t-1)|$, where $p_c(c,t)$ represents the counselor's percentage of words from category c in window t, $p_p(c,t-1)$ represents the patient's percentage of words from category c in the previous window $t-1$, and the overall LSC is the average over all windows.

- MITI (Motivational Interviewing Treatment Integrity) Behaviors: The counts of MITI behaviors (a standardized list of good counseling behaviors) were used directly as features. Also, ratios like reflections-to-questions and percentage of MITI behaviors ($\frac{\text{MITI behaviors}}{\text{total turns}}$) were calculated.
- Sentiment Analysis: Good counselors effectively infer their patients' sentiment. A sentiment score was assigned to each turn, and average sentiment per turn, sentiment change, and sentiment volatility were calculated

These features were combined into a feature vector representation for each conversation. Unfortunately, the paper does not go into detail about how they were combined. However, we attempted to reconstruct their evaluation metric by training on classifier on their dataset.

As for our evaluation metric for sentiment analysis, please refer to our Approach section.

5.3 Experimental details

After selecting GPT-3.5 as the highest-performing base model for AI therapist based on the therapy evaluation scores, we proceeded to fine-tune it further. The fine-tuning process involved two main stages: fine-tuning the model on the sentiment analysis dataset, and then fine-tuning the fine-tuned model on the Amod dataset for the therapy task.

To set up the fine-tuning process, we used the Hugging Face Transformers library and the OpenAI API. We first tokenized the input data using the GPT-3.5 tokenizer and prepared the datasets for fine-tuning. The Amod dataset was split into training and validation sets, ensuring that the 10 conversations set aside for testing were not included in the fine-tuning process.

In the first stage of fine-tuning, we fine-tuned GPT-3.5-turbo-0125 on the sentiment analysis dataset. The sentiment analysis dataset was split into 90% for training and 10% for testing. Due to time and space constraints on fine-tuning, we used ‘model="gpt-3.5-turbo-0125"‘ instead of ‘model="gpt-3.5-turbo"‘. This more lightweight version of GPT-3.5 provided faster response times due to optimizations in its architecture and inference processes. Additionally, the smaller size of the model made it more suitable for our dataset, as the limited amount of data could more effectively train a smaller model, leading to more a noticeable improvement in performance after fine-tuning.

We used the following hyper-parameters: ‘numepochs=3‘, ‘batchsize=4‘, and ‘learningrate=2e-5‘. These parameters were chosen to balance the trade-off between training time and model performance. We monitored the validation loss during the fine-tuning process and saved the best model checkpoint based on the lowest validation loss.

After fine-tuning on the sentiment analysis dataset, we further fine-tuned the resulting GPT-3.5-turbo-0125 model on the Amod dataset using the same hyper-parameters. To enhance the model's understanding of high-quality therapy conversations, we also fine-tuned it on the entire Pérez-Rosas dataset, which includes specific characteristics such as linguistic alignment, MITI behaviors, and WordNet-Affect semantic information.

This two-stage fine-tuning approach allowed the model to first learn general sentiment analysis before adapting to the specific nuances of high-quality therapy conversations. As a result, the fine-tuned GPT-3.5-turbo-0125 model outperformed Mistral and Llama by a wide margin, demonstrating significant improvements in our evaluation metrics (more details in the results section).

5.4 Results

Table 1 shows the results from evaluating the three LLM base models:

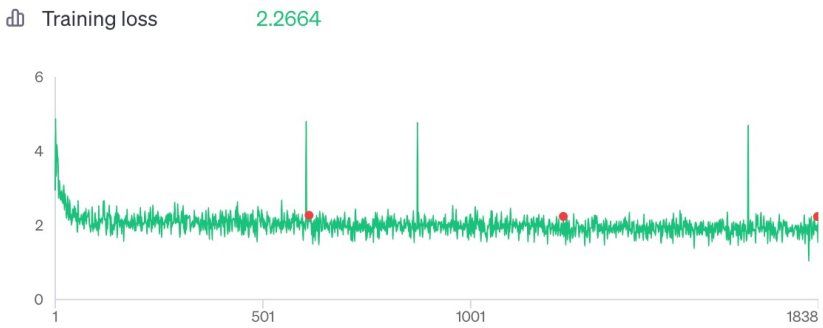


Figure 1: Training loss of 3.5-0125 of finetuned sentiment and Amod -> Trained tokens 2,270,991 Epochs 3 Batch Size 4 LR multiplier 2 seed 16268577.

Table 1: Baseline Model Results for Therapy and Sentiment Analysis

Model	**Therapy Accuracy**	**Therapy Per-response Scores**	**Sentiment Accuracy**
Mistral	3%	3 high, 97 low	39.7%
GPT-3.5	63%	63 high, 37 low	54.2%
GPT-3.5-0125	53%	53 high, 47 low	54.0%
Llama	11%	11 high, 89 low	9.7%

As shown above, both GPT-3.5 and GPT-3.5-0125 outperformed Mistral and Llama in terms of therapy accuracy, per-response scores, and sentiment accuracy. GPT-3.5 achieved a therapy accuracy of 63%, with 63 high-quality responses and 37 low-quality responses, while GPT-3.5-0125 achieved a therapy accuracy of 53%, with 53 high-quality responses and 47 low-quality responses. In comparison, Mistral and Llama had therapy accuracies of only 3% and 11%, respectively. In terms of sentiment accuracy, GPT-3.5 achieved 54.2%, GPT-3.5-0125 achieved 56.0%, while Mistral and Llama achieved 39.7% and 9.7%, respectively. These results are not entirely surprising, given that both GPT-3.5 and GPT-3.5-0125 are more advanced and larger models compared to Mistral and Llama. The slightly lower performance of GPT-3.5-0125 compared to GPT-3.5 can be attributed to the optimizations made in its architecture and inference processes to provide faster response times, which may have led to a minor trade-off in performance. However, the fact that even the best-performing model, GPT-3.5, only achieved a therapy accuracy of 63% and a sentiment accuracy of 54.2% suggests that there is still a lot of room for improvement in generating high-quality therapy responses and accurately analyzing sentiment.

Table 2 displays the results of fine-tuning GPT-3.5-0125:

Table 2: GPT-3.5-0125 Fine-tuning Results for Therapy and Sentiment Analysis

Model	**Therapy Accuracy**	**Therapy Per-response Scores**	**Sentiment Accuracy**
GPT-3.5-0125 fine-tuned on sentiment analysis and then Amod	74%	74 high, 26 low	71.9%

After fine-tuning GPT-3.5-0125 on only the sentiment analysis data, we achieved 91.4%. Fine-tuning GPT-3.5-0125 on both the sentiment analysis and Amod datasets significantly improved performance across all metrics. The therapy accuracy increased to 74%, with 74 high-quality responses and 26 low-quality responses, and the sentiment accuracy improved to 71.9%. This improvement is indicative of the model's ability to better understand and generate contextually appropriate and high-quality therapeutic responses after being fine-tuned on data specifically tailored for sentiment and therapeutic contexts. The accuracy increase compared to the baseline results suggests that the fine-tuning process effectively enhanced the model's capability.

Our Colab, Generated Model Responses, and Sentiment Results - Google Drive https://drive.google.com/drive/folders/1kN8QBQurlhcHyr0B8Ikw8AikLGYWp_oZ?usp=drive_link Drive (2024)

6 Analysis

To get a better understanding of our performance results, we conducted some qualitative analysis by examining the characteristics of the model's outputs in light of the linguistic aspects identified by Pérez-Rosas et al. (2019) as indicators of high-quality counseling. We compared two generated responses to the same context, one classified as high-quality therapy and the other as low-quality, to identify how well each response aligns with these linguistic cues.

> **Context:** I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here. I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it. How can I change my feeling of being worthless to everyone?
> **Predicted High Quality Response:** Feeling worthless is a difficult emotion that many people struggle with. That being said it is something that can be worked on and many people find value in themselves and others over time with treatment. It sounds like many of these feelings you have are private to you. It is imperative that you find someone that you feel comfortable with to verbalize some of your feelings.
> **Predicted Low Quality Response:** Feeling worthless can feel like it has a grip on you that you can't overcome. You may know inside that you have things to celebrate and offer but just can't be convinced. The important thing is that it can be done, but not all at once. It is more of a path than a destination. While some things may be able to help such as focusing on positive feedback from other areas of your life than the one that is sparking these feelings, in the end it is some internal changes you will need to make which may require some experimentation to determine what will make the difference. Sometimes this can be quite challenging. You may want to meet with a counselor at least in part to have someone that you are being accountable to try some of these changes.

In the high quality response, the model response evinces several positive characteristics that align with the linguistic cues identified by Pérez-Rosas et al. (2019). It demonstrates balanced turn-taking interaction by directly addressing the specific concern raised by the patient ("Feeling worthless is a difficult emotion...") and offering a concise and focused response. This corresponds to the finding that during high-quality counseling, counselors achieve a more balanced word exchange with clients. The model response also shows a positive sentiment trend, focusing on the potential for improvement ("it is something that can be worked on and many people find value in themselves and others over time with treatment"). Also, we can see that many times the response mirrors the patient's language with similar phrases such as "feelings" and "feeling worthless" (as in linguistic alignment). Lastly, the model's response touches on the topic of behavior change and commitment ("It is imperative that you find someone that you feel comfortable with to verbalize some of your feelings"), as opposed to resistance and persuasion.

In contrast, in the low quality response, the model evinces some shortcomings in terms of linguistic cues. Although it acknowledges the difficulty of overcoming feelings of worthlessness, the response is much longer and less focused compared to the high-quality example; this demonstrates imbalanced turn-taking. Also, the positive trend is overshadowed by the emphasis on the challenge of making internal changes (e.g., "in the end it is some internal changes you will need to make which may require some experimentation to determine what will make the difference. Sometimes this can be quite challenging."). Finally, the response also shows less linguistic alignment with the patient's language.

7 Conclusion

In our goal to develop an emotionally-intelligent AI therapist, we first evaluated the performance of three pre-trained LLMs (Llama-3-8b, GPT-3.5 Turbo, and Mistral-7B-v0.1) as our base model on a corpus of patient-therapist therapy conversations. We used a logistic regression classifier to evaluate the quality of generated therapist responses based on linguistic and semantic features from

high-quality therapy sessions, and performed sentiment analysis using a dataset of data covering six common emotions in therapy. Our results showed that GPT-3.5 Turbo performed the best in terms of both therapy response generation and sentiment analysis. As such, we fine-tuned GPT-3.5 Turbo-0125 (a more lightweight version of 3.5, which still outperformed Mistral and Llama) using a two-stage approach. First, we fine-tuned the model on the sentiment analysis dataset, aiming to improve its understanding of emotions and sentiment. We conjectured that this would enable the model to generate more empathetic and contextually appropriate responses in a therapeutic setting. We then fine-tuned the sentiment-enhanced model on two therapy datasets: the Amod dataset and the Pérez-Rosas dataset.

The fine-tuning process yielded promising results. The model achieved a therapy accuracy of 74%, a 21% relative increase, and a sentiment accuracy of 71.9%, a 17.9% relative increase. These results highlights the effectiveness of our approach in improving the model's ability to generate high-quality therapy responses and accurately analyze sentiment, which in turn demonstrates the feasibility of equipping an LLM to act as an AI therapist. We hope our work allows the research community to take a step forward in providing more accessible and effective mental health support for all.

7.1 Limitations

Our work has several limitations. The two datasets we used for therapy and sentiment analysis to fine tune are limited and may not fully capture the diversity/breadth of real-world therapeutic interactions which affects the model's generalizability. Also, the logistic regression classifier for high and low quality therapy conversations is not 100 percent accurate so a few of the classifications may be incorrect. Additionally, using GPT-3.5-0125, chosen for its faster response times and reduced computational needs, may have slightly lower performance compared to the regular GPT-3.5 model which impacts response accuracy. Lastly, the AI therapist's responses still lag behind human therapists in empathy and sentiment and contextual understanding due to these data and model limitations which raises ethical concerns for real-world deployment.

7.2 Future Work

Future work could focus on several key areas to enhance AI-driven therapeutic models. Diversifying and expanding datasets to include a broader range of therapeutic conversations and emotional expressions from various linguistic backgrounds would improve model generalizability. Developing more sophisticated models, such as those incorporating multimodal data like voice, facial expressions, etc., could better capture the complexity of human emotions and therapeutic interactions. Techniques like transfer learning and domain adaptation could be explored to leverage knowledge from related tasks.

8 Ethics Statement

Our project faces several ethical challenges common to AI-driven mental health support systems. Kretzschmar et al. (2019) emphasize the need for evidence-based support and highlight the limitations of chatbots like Woebot, Wysa, and Joy in providing personalized and contextually aware support. We recognize that inappropriate or unqualified AI responses could exacerbate already unfavorable emotional states or situations. Although our project focuses on training a model using a public dataset, we will rigorously evaluate our AI therapist on effectiveness and safety before deploying it as a product. Coghlan et al. (2023) further expand on these ethical considerations raised by Kretzschmar et al. (2019), identifying four main concerns: the level of human involvement, the need for an adequate evidence base, data collection and use, and the handling of unexpected disclosure of crimes. They argue that chatbots are far from being able to recreate the therapeutic alliance that builds up over time that exists between patients and human therapists. We recognize that our AI therapist should not be misused as a replacement for professional help, especially for serious mental health problems that require a licensed professional. To mitigate this risk, we will clearly communicate the limitations of our system to users, in line with the ethical principles of non-maleficence and beneficence. Also while our current project does not involve direct user interaction, we acknowledge the need to carefully consider the issue of unexpected disclosure of crimes in any future deployment of our AI therapist, as raised by Coghlan et al. We also recognize the potential for chatbots to dehumanize care and

replace human therapist jobs, and are committed to making sure our AI therapist augments rather than supplants human care.

References

Amod. 2022. Mental health counseling conversations.

Doğan Can, Panayiotis G. Georgiou, David C. Atkins, and Shrikanth S. Narayanan. 2012. A case study: detecting counselor reflections in psychotherapy for addictions using linguistic features. In *Proc. Interspeech 2012*, pages 2254–2257.

Simon Coghlan, Kimberley Leins, Sarah Sheldrick, Marc Cheong, Piers Gooding, and Simon D'Alfonso. 2023. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digit Health*, 9:20552076231183542.

Dair-AI. 2022. Emotion dataset.

Google Drive. 2024. `https://drive.google.com/drive/folders/1kN8QBQurlhcHyr0B8Ikw8AikLGYWp_oZ?usp=drive_link`. Accessed: 2024-05-22.

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.

Office of the Surgeon General. 2024. Social connection.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Amy M. Kilbourne, Kristin Beck, Benjamin Spaeth-Rublee, Prathibha Ramanuj, Richard W. O'Brien, Nao Tomoyasu, and Harold A. Pincus. 2018. Measuring and improving the quality of mental health care: a global perspective. *World Psychiatry*, 17(1):30–38.

Kira Kretzschmar, Holly Tyroll, Gabriela Pavarini, Arianna Manzini, Ilina Singh, and NeurOx Young People's Advisory Group. 2019. Can your phone be your therapist? young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical Informatics Insights*, 11:1–9.

Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. 2017. A fully automated conversational agent for promoting mental well-being: A pilot rct using mixed methods. *Internet Interventions*, 10:39–46.

Laura Martinengo, Elaine Lum, and Josip Car. 2022. Evaluation of chatbot-delivered interventions for self-management of depression: Content analysis.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935.

Shruti Sannon, Brett Stoll, Dominic DiFranzo, Malte Jung, and Natalya N. Bazarova. 2018. How personification and interactivity influence stress-related disclosures to conversational agents. In *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, New York, NY, USA. ACM.

Engineering National Academies of Sciences and Medicine. 2020. *Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System*. The National Academies Press, Washington, DC.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.08571*.

Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2014. Modeling therapist empathy through prosody in drug addiction counseling. In *Proc. Interspeech 2014*, pages 213–217.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models.