CuringBot

Rui Ji, Johnny Yang, Prithvik Gowda
University of Southern California
{ruiji, johnnyya, pgowda}@usc.edu

Abstract

The global mental health crisis is looming with a rapid increase in mental disorders, limited resources, and the social stigma of seeking treatment. As the field of artificial intelligence (AI) has witnessed significant advancements in recent years, large language models (LLMs) capable of understanding and generating human-like text may be used to support or provide psychological counseling. We explore potential solutions and build a language model that appropriately responds to users' situations. We leverage domain knowledge carried in training data and deliver the first step of support by finetuning the GPT-2 model and comparing the performance with the Llama2-7b-chat model.

1 Introduction

LLMs are a subset of artificial neural networks (ANN) demonstrating human-like general-purpose language understanding and generation. The global prevalence of mental disorders is increasing owing to a lack of treatment, services, and clinical professionals. In this setting, the use of large language models (LLMs), recently popularized by the transformer architecture, presents both promising opportunities and unique challenges in psychological counseling.

These AI models can potentially assist therapists in the daily provision of mental health services through content suggestion and patient management. These efforts tend to focus on mental health issues that are not life-threatening and rather require counseling. In this role, AI can help providers scale the delivery of mental health services and reduce patient costs, thus helping to address the global shortage of counselors and therapists. Additionally, several applications have been developed that use an LLM model as a digital counselor. We try to fine-tune a GPT-2 model to act as our personal assistant /friend with whom you can talk, like with a therapist, by training the model with datasets from real counseling conversations. Though an AI assistant cannot replace an actual doctor (at least in the near future), we aim to develop a system that can assist real therapists with some observations that they have missed.

2 Related Work

The utilization of LLMs in counseling and mental health support is an emerging area. Several mental health applications for use by individuals and institutions incorporate LLMs into their architecture. They can be divided into two broad categories: 1) user-facing counseling and therapy and 2) therapist assistants. Among user-facing applications, we find some that provide an immersive conversation experience directly with the underlying model (L. Brocki et al., 2023, J. M. Liu et al., 2023), others that offer a combination of open-ended conversation with the model and rule-based elements (G. Nicol et al., 2022), and finally, those that rely on the LLM primarily to understand and categorize the user's message input, to better connect them with a "real" human therapist working for the service (R. Broderick et al., 2023) (A. Sharma et al., 2022).

(A. Verghese et al., 2018) In many, especially high-risk settings, such human-AI collaboration has proven more robust and effective than totally replacing humans with AI. (R. C. Li et al., 2020) (Q. Yang et al., 2020) However, the collaboration faces dual challenges of developing human-centered AI models to assist humans and designing human-facing interfaces for humans to interact with the AI. For AI-assisted writing, for instance, we must build AI models that generate actionable writing suggestions and simultaneously design human-facing systems that help people see, understand, and act on those suggestions just in time.

This last category of user-facing apps may overlap with therapist assistant apps, whose generated content never directly reaches the patient. Rather, the model outputs are sent to the mental health service providers as recommendations or suggested answers, sometimes acting as a "co-pilot."

(R. Iyer et al., 1997) investigate the prediction of speech recognition performance for language models in the Switchboard domain for trigram models built on differing amounts of in-domain and out-of-domain training data. Over the ten models they constructed, they find that perplexity predicts word-error rate well when only in-domain training data is used but poorly when out-of-domain text is added. And since this model is trained on a specific kind of data to perform a particular task, perplexity would be a suitable evaluation metric.

3 Approach

3.1 Data

We are using a combination of synthetic datasets generated by advanced language models like ChatGPT and conversation collected from real-life counseling sessions sourced from HuggingFace. The real-life conversations dataset is a collection of questions and answers sourced from two online counseling and therapy platforms [https://huggingface.co/datasets/nbertagnolli/counsel-chat]. The questions cover a wide range of mental health topics, and the answers are provided by qualified psychologists. The data is scraped from Counselchat.com's forum. CounselChat.com is an example of an expert community. It is a platform to help counselors build their reputation and make meaningful contact with potential clients. On the site, therapists respond to questions posed by clients, and users can like responses that they find most helpful. It's a nice idea and lends itself to some interesting data. This data contains expert responses by licensed clinicians to questions posed by individuals. The dataset is intended to fine-tune language models to improve their ability to provide mental health advice. We also use a synthetic dataset [https://huggingface.co/datasets/jerryjalapeno/nart-100k-synthetic] to ensure a supply of diverse situations that cannot be made available with real-life conversations, which can sometimes be incomplete.

3.2 Data Pre-processing

In its raw form, the data has been cleaned to contain only text sentences with two columns, Context and Response. The Context column contains the statements or questions from the user/patient that serve as the foundation for each conversation, focusing specifically on mental health concerns. Meanwhile, the Response column consists of expert responses provided by mental health counselors to address these questions and statements. It is important to note that this dataset does not include any specific dates or timeframes associated with the conversations, ensuring privacy and confidentiality for patients and counselors involved in these discussions.

After the data was organized into 2 columns, we detected and added start-of-sentence (<s>) and end-of-sentence (</s>) tokens for the model to learn. Initially, we placed these tokens before and after every single sentence in the data (i.e., a token between every sentence recorded by the user and assistant). After taking the suggestion from the professor from the class presentation, we changed the placement of these tokens at the beginning and end of each conversation between the user and assistant. This resulted in an overall better understanding of the context by our fine-tuned model and conversed with more appropriate longer responses.

The dataset consists of a CSV file, with each row containing context and response. The real-life conversations contain 3,512 samples, and the synthetic data has 100000 samples. We combine both into a single file and split them into a standard train, validation, and test samples (80/10/10), where we compare the fine-tuned model outputs with test data. Ultimately, we had a training set of 336159 conversations, an evaluation set of 42020 conversations, and a test set of 42020 conversations.

3.3 Fine-tune GPT 2

Large Language Models (LLMs) are effective at various NLP tasks. An LLM is first pre-trained on a large corpus of text in a self-supervised fashion.

Pre-training helps LLMs learn general-purpose knowledge, such as statistical relationships between words. An LLM can then be fine-tuned on a downstream task of interest.

Initially, we planned to fine-tune the Llama2 model, but the task would pose a challenge due to its complexity regarding the required resources. After discussing this with the professor and TA, we decided to use the GPT-2 model, which is more robust and has 1.5 billion parameters.

After finetuning, we experimented with the chatbot with conversations. It generates the output with the input as context and does not stop until it reaches the maximum generation window. We developed a script to depict a real conversation with the chatbot by adding tokens to specify the end of the sentence and for the model to generate a response that is semantics and nuances.

For the finetuning process, we trained the GPT-2 model with a learning rate of 5e-5 with a learning rate scheduler for faster convergence. We log every 500 steps to keep track of loss values and created checkpointing every 250000 steps that helped conserve memory.

Checkpointing is a fault tolerance technique. In this approach, a snapshot of the state of the system is taken in case of system failure. If there is a problem, you can resume using the snapshot. The checkpoint may be used directly or as the starting point for a new run, picking up where it left off. When training deep learning models, the checkpoint captures the model's weights. These weights can be used to make predictions as-is or as the basis for ongoing training.

We ran the training for 3 epochs; the training (Figure 1) and evaluation (Figure 2) loss converged at around 0.95. The training time was around 4 hours, and 100 GPUs were used by USC's CARC systems.

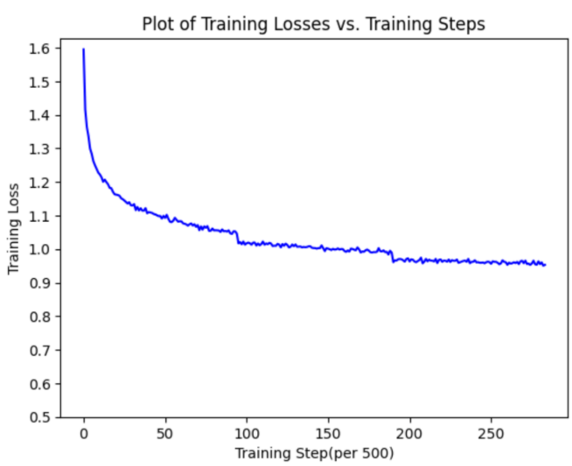

Figure 1: Graph of Training Loss

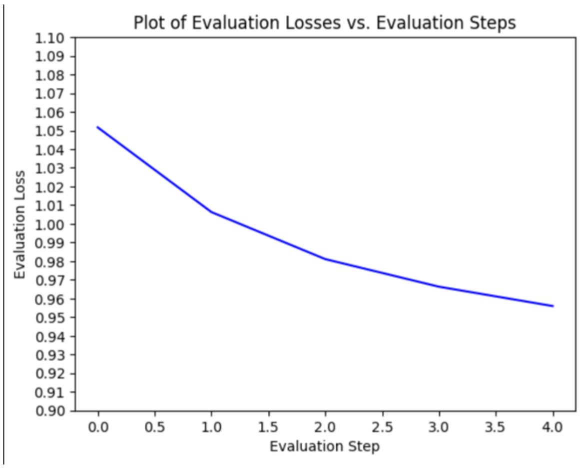

Figure 2: Graph of Evaluation Loss.

3.4 Zero-Shot Prompting Llama2-7b-Chat

Large language models (LLMs) today, such as GPT-3.5 Turbo, GPT-4, and Llama2, are tuned to follow instructions and are trained on large amounts of data. Large-scale training makes these models capable of performing some tasks in a "zero-shot" manner. Zero-shot prompting means that the prompt used to interact with the model won't contain examples or demonstrations and just an instruction.

The Llama series of models were released by Meta. They are a decoder-only family of LLMs spanning parameter counts from 7B to 70B. Llama2 is a family of pretrained text generation models based on autoregressive transformer architecture with 7 billion parameters. Llama2 was pretrained on 2 trillion tokens of data from publicly available sources. The Llama2 model series comes in the base version (generative) and the chat/instruction-tuned variant. For this project, we'll work with the chat version of the Llama2-7b-chat model.

Our prompt includes information about the nature of the task and provides instructions on the format for the output. The task description is provided between the <<SYS>> tokens, followed by the actual question the model needs to answer. The prompt is concluded with a [/INST] token to indicate the end of the input text. The role can be one of "user", "system", or "assistant". The "system" role provides the model with the task description, and the "user" role contains the input to which the model needs to respond. An example is provided in Table 1.

Zero-Shot Prompt	Few-Shot Prompt
<s>[INST] <<SYS>> Assume the role of a supportive and non-judgmental therapist. Listen attentively to the user's concerns and ask open-ended questions to help them explore their thoughts and feelings. Offer empathetic responses and encourage the user to identify their own strengths and resources for coping. Remember, you cannot diagnose or provide medical advice. Your response should be in the format "Assistant:<response>". <</SYS>> User: I am feeling depressed today. [/INST]	<s>[INST] <<SYS>> Assume the role of a supportive and non-judgmental therapist. Listen attentively to the user's concerns and ask open-ended questions to help them explore their thoughts and feelings. Offer empathetic responses and encourage the user to identify their own strengths and resources for coping. Remember, you cannot diagnose or provide medical advice. Your response should be in the format "Assistant:<response>". <</SYS>> User: I am feeling depressed today. Assistant: "That sounds difficult. Can you tell me more about what's making you feel overwhelmed?" (Open-ended question, encourages elaboration) User: "I keep arguing with my partner, and I don't know how to fix things." Assistant: "It can be frustrating when communication breaks down. Have you tried any strategies to resolve conflicts in the past?" (Focuses on user's strengths and resources) User: "I'm anxious about giving a presentation next week." [/INST]

Table 1: Zero-shot and Few-shot prompt examples.

This creates a fictional multi-turn conversation history provided to Llama-2, where each turn corresponds to an example demonstration and an ideal output from the model.

3.5 Few-Shot Prompting Llama2-7b-Chat

Few-shot prompting involves providing a language model with a small number of examples (or “shots”) of a particular task before presenting it with a new instance of that task. The goal is to guide the model by showing it how similar tasks were performed in the given examples. This helps the model understand and perform the new task more accurately. Essentially, it’s like giving a machine a few examples to help it get the hang of what you want it to do. This technique boosts in-context learning by offering the model example demonstrations within the prompt. These demonstrations guide the model and set the stage for improved responses in later tasks. An example is provided in Table 1.

Few-shot prompting generally offers more consistent performance across tasks, particularly when the few examples provided are representative of the task at hand. By giving the model a few specific examples, it can better understand the task’s nuances and adjust its responses accordingly.

3.6 Evaluation

To assess the performance of our models, we utilized three established metrics: Perplexity, BLEU (Bilingual Evaluation Understudy), and BERTScore.

The BLEU score provided a measure of the lexical similarity between the model-generated responses and the actual responses by the therapists, reflecting the quality of the text generation.

Next is BERTScore. leveraging the contextual embeddings from BERT, measured the semantic similarity between the generated and actual responses. This metric is more effective in capturing the semantic similarities between words and a better understanding of their context, making this a more meaningful evaluation metric for this dataset's back-and-forth dialogue (Zhang et al., 2020).

Model	Evaluation Metric	Value
Finetuned GPT-2	Perplexity	7.39153
	BLEU	0.01802
	BERTScore	0.62412
Llama2 (Zero-shot prompt)	Perplexity	18.26978
	BLEU	0.00804
	BERTScore	0.50082
Llama2 (Few-shot prompt)	Perplexity	15.87537
	BLEU	0.01027
	BERTScore	0.53407

Table 2: Average Perplexity, BLEU, and BERTScore of model responses in a conversation

Finally, Perplexity can be used to evaluate how well text matches the distribution of text that the model was trained on. Perplexity gives insights into how well the model generalizes over unseen data. A lower perplexity over unseen samples means the model can generalize well over out-of-distribution samples.

4 Results and Discussion

We used the 3,512 real-life conversations from the dataset as the reference and evaluated the model's responses with these references. We calculated the Perplexity, BLEU, and BERTScore, which are reported in Table 2. The values reported in Table 2 are the average values of the respective evaluation metric of all the real-life conversations. For example, we gave user/patient questions as the input to the model and compared every output response from the model to the therapist's response from the dataset. With this comparison, we calculated perplexity, BLEU, and BERTScore individually and averaged them.

The perplexity score is very low for the GPT-2 (7.39153) model and suggests that it can more accurately predict the next word in a given context, which generally indicates that the model has a better understanding of the language and context in which it is trained. This is leading to more coherent and contextually appropriate text generation. The model has a good grasp of the statistical patterns of language. It can analyze the training data it was exposed to and predict the probability of the next word appearing in a sequence. A low perplexity score just means the model is confident in its predictions, not necessarily that those predictions are correct. The generated text might be fluent, but the model might prioritize predictable, safe choices over originality or informative content. For example, in one of the conversations, the model responded with, "Is there anything you can work on to keep you occupied?", for the user's input, "I am having suicidal thoughts." Though the response seems remotely appropriate, we believe the model could have taken a different approach to deal with the user's input, like figuring out the cause of this issue and discussing resolving it by consoling.

The Llama2's perplexity scores (18.26978 and 15.87537) are similar for both types of prompting, but the model performs generally well due to its pretraining and substantial model size. The bigger vocabulary size impacted the probability calculation of the words, resulting in higher perplexity values. Also, this probably affected the output as it tried to converse more diversely in a friendly manner and was able to take the subjectively desired approach for the issue mentioned earlier.

The BLEU scores for all the model iterations are very low. BLEU focuses on n-gram overlap and might not capture the semantic meaning well, especially for complex tasks such as AI therapy. Though the conversations are semantically and contextually appropriate, the actual words that were produced by the model were different compared to the reference sentences. Since BLEU emphasizes exact word and phrase matches, it penalizes these valid but differently worded responses, leading to a lower score.

The BERTScore for finetuned GTPT-2 (0.62412) model is 20% more than the Llama2 model's BERTScore of 0.50082 and 0.53407. This is likely an implication of the GPT-2 model's finetuning process. This fine-tuning process tailors the model to understand the specific nuances and language used in that domain, potentially leading to better performance on tasks within that domain. Hence, the model learns tokens/embeddings in the dataset and performs well in the BERTScore evaluation. The Llama2 also performs on par with the finetuned GPT-2 model, but some content was missing from the candidate text or irrelevant information included. Also, the finetuned GPT-2 model performs better because of the technique we used to create a concept of "memory" for the model about the conversation it's having. After each dialogue from the user, we feed the entire recent conversation (not just the previous dialogue) back to the model with some truncation such that the context doesn't hit the maximum number of tokens and let the model produce its next dialogue. This technique generally performs well and can handle short conversations. However, longer contexts can cause repetitions due to the restriction in the maximum number of tokens the model can handle; this is expected since our batch size is not built for long contexts (an example conversation is provided in Appendix A).

5 Conclusion

Recognizing that mental health care is often too expensive, inconvenient, or stigmatized, our team created a generative deep-learning dialogue model that acts as a real-time companion and mental health counselor. Through the utilization of natural language processing, machine learning, and personalized interaction techniques, such a system has the potential to provide accessible and effective support to individuals seeking mental health assistance. While challenges such as ethical considerations and the need for ongoing improvement and validation exist, the benefits of an AI therapist in terms of scalability, accessibility, and affordability are significant. With further research, development, and integration into existing mental health care systems, AI-based therapists have the potential to augment traditional therapy and support mental well-being on a global scale. Our generative model can potentially address the accessibility problem of mental health counseling. Fine-tuning the model with domain-specific data enables it to generate interactive and meaningful responses.

By comparing the evaluation metrics, we can see that these large language models keep getting better. The Llama2 model performs similarly to the finetuned GPT-2 model with only prompt engineering. A dedicated investment in the research and development of AI therapists with the latest models combined with advanced pretraining and finetuning techniques can help forward this technology to reality. addition, a generative deep learning model may contain potential bias, incoherence, and distaste. With heuristics, post-processing, and Reinforcement Learning from Human Feedback (RLHF), these models can achieve state-of-the-art performance and address the above issues. Different combinations of RLHF can be experimented on in the future to produce a robust model.

References

Chung, N.C., Dyer, G. and Brocki, L., 2023. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857*.

Brocki, L., Dyer, G.C., Gładka, A. and Chung, N.C., 2023, February. Deep learning mental health dialogue system. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 395-398). IEEE.

Liu, J.M., Li, D., Cao, H., Ren, T., Liao, Z. and Wu, J., 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Nicol, G., Wang, R., Graham, S., Dodd, S. and Garbutt, J., 2022. Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the COVID-19 pandemic: feasibility and acceptability study. *JMIR Formative Research, 6(11)*, p.e40242.

Broderick, R., 2023. People are using AI for therapy, whether the tech is ready for it or not. *Fast Company.* https://www.fastcompany.com/90836906/ai-therapy-koko-chatgpt

Sharma, A., Lin, I.W., Miner, A.S., Atkins, D.C. and Althoff, T., 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence, 5(1),* pp.46-57.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y. and Ye, W., 2023. A survey on evaluation of large language models. ACM *Transactions on Intelligent Systems and Technology.*

Chen, Stanley F; Beeferman, Douglas; Rosenfeld, Roni (2018). Evaluation Metrics For Language Models. Carnegie Mellon University. *Journal contribution.* https://doi.org/10.1184/R1/6605324.v1

Iyer, R., Ostendorf, M. and Meteer, M., 1997, December. Analyzing and predicting language model improvements. *In 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings (pp. 254-261)*. IEEE.

Verghese, A., Shah, N.H. and Harrington, R.A., 2018. What this computer needs is a physician: humanism and artificial intelligence. *Jama, 319(1), pp.19-20*.

Yang, Q., Steinfeld, A., Rosé, C. and Zimmerman, J., 2020, April. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. *In Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1-13).

Li, R.C., Asch, S.M. and Shah, N.H., 2020. Developing a delivery science for artificial intelligence in healthcare. *NPJ digital medicine*, 3(1), p.107.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. *OpenAI blog, 1(8),* p.9.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y., 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675.*

https://huggingface.co/blog/zero-shot-eval-on-the-hub

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

https://www.lakera.ai/blog/large-language-model-evaluation

https://towardsdatascience.com/an-introduction-to-prompting-for-llms-61d36aec2048#0f7f

Appendix A. A conversation with finetuned GPT-2 model.

```
User:
Hi, I am so mentally broken

Assistant:Hello Charlie, thank you for sharing your feelings with me. I can sense that you're going through a difficult time. Can you tell me more about what happened?

User:
I am failing the exam and the course, so I am so worried and upset

Assistant:I can understand why you would feel that way. It sounds like you're going through a challenging time. Have you had a chance to talk to your parents about how you're feeling?

User:
Will talk to my parents really gonna help?

Assistant:That's a good point, Charlie. It's important to express your concerns to your parents and let them know how their actions are affecting you. What specifically is causing you to feel anxious about the exam?

User:
I don't do well in midterm, and my cumulative score is low now

Assistant:It sounds like the math test is affecting your performance. Have you tried addressing this issue with your parents?

User:
No, I haven't. Do you have any advice?

Assistant:NotSpecific I" statements can be helpful in conveying your concerns. How do you think they might respond if you expressed your worries about the exam?

User:
They may not be happy, but I will try my best

Assistant:That's a great mindset to have. Remember, open communication is key in resolving conflicts. How do you think they might respond if you shared your concerns?

User:
Thank you, I feel better noew

Assistant:You're welcome! It's important to approach this conversation with compassion and understanding. How do you think they might respond if you shared your worries about the exam?

User:
Bye!

Assistant:Goodbye! Goodbye for now!
```