CuringBot

Rui Ji, Johnny Yang, Prithvik Gowda

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, leveraging domain knowledge carried in training data and delivering the first step of support.

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 to, 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 to that can assist real therapists with some observations that they have missed.

2 Related Work

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).

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 for fine-tuning 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.

In its raw form, it has been cleaned to contain only text sentences with two columns, Context and Response. The Context column contains the statements or questions 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. Here is an example of the data format:

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, of which we will use 20000 random samples due to computing resource constraints. We combine both into a single file and split them into train and test samples (80/20), where we compare the fine-tuned model outputs with test data.

3.2 Fine-tune GPT 2

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

The data was read from a CSV file and extracted each line into "User: {content}" or "Assistant: {content}" for role separation. The data was then cleaned by spacing out all punctuations so they were recognized as individual tokens, and tokens were added to distinguish the beginning and end of the sentences.

As we trained and started our conversation with the chatbot, 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 elements like '<s>' to specify the end of the sentence and for the model to state a response.

To understand the process of fine-tuning and handling the GPT-2 model, we used limited samples of data and generated a conversation, which can be seen in the example below:

User: I've been feeling so sad and overwhelmed lately.

Assistant: I'm here if you need to talk.

User: Why I feel like there is nobody loves me

Assistant: I'm here for you.

User: Will you listen to me?

Assistant: I'm listening.

User: So, how should I do to relieve my stress

Assistant: If you are feeling depressed, ask yourself the following questions: Why do I feel this way?

User: Ok, feels better, what's next

Assistant: Please go on.

User: Should I go for some exercise?

Assistant: Sure.

User: You are so helpful, Thank you!

Assistant: Hi there.

User: Bye

Assistant: I think I can do that.

The training was run for 5 epochs with a batch_size of 128 and sample_size of 10000.

3.3 Future Scope

As we can see, the conversation is trivial due to the limited amount of data used for training. We made a couple of observations with this initial attempt at training the model with a smaller dataset. It needs to learn its role as an assistant, and it simply generates text after the question.

In the next phase, we aim to train the model with a much larger dataset to make a model learn to respond as an assistant and recognize the special tokens in the training set to be able to read the history and generate responses correspondingly. We will also consider making the model more conversational so that less prompt engineering is required, as we saw in our initial attempt at building this system.

For evaluation, we will develop a Feed Forward Neural Network Language model as a baseline to compare how the modern language model architecture like GPT-3 built using Transformers compare with the previous generation language models.

As this model is trained on a specific kind of data to perform a particular task, we will use the perplexity score (equation below) as it helps us understand how well the model generalizes to unseen data.

$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)...p(w_n|w_1,w_2,..,w_{n-1}) = \prod_{i=1}^{n} p(w_i|w_1,...,w_{i-1})$$

References

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Chung, Neo Christopher, George Dyer, and Lennart Brocki. "Challenges of large language models for mental health counseling." arXiv preprint arXiv:2311.13857 (2023).

L. Brocki, G. C. Dyer, A. Gladka, N. C. Chung, Deep learning mental health dialogue system, in: 2023 IEEE International Confer- ence on Big Data and Smart Computing (BigComp), IEEE, 2023. doi:10.1109/bigcomp57234.2023.00097.

J. M. Liu, D. Li, H. Cao, T. Ren, Z. Liao, J. Wu, Chatcoun- selor: A large language models for mental health support (Sep. 2023). arXiv:2309.15461, doi:10.48550/ARXIV.2309.15461.

G. Nicol, R. Wang, S. Graham, S. Dodd, J. Garbutt, 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) (2022) e40242. doi:10.2196/40242.

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

A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, T. Althoff, Human- ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support (Mar. 2022). arXiv:2203.15144, doi:10.48550/ARXIV.2203.15144.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. ACM Trans. Intell. Syst. Technol. Just Accepted (January 2024). https://doi.org/10.1145/3641289

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

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

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

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

R. Iyer, M. Ostendorf, and M. Meteer. Analyzing and predict- ing language model improvements. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 1997.