Therapist AI: Fine tuning Llama 2 to create a mental health chatbot

Tom Barzilay, Gilad Vardy-zer, Yonatan Tintpulver
School of Computer Science
Tel Aviv University
tombarzilay,giladv,tintpulver1@mail.tau.ac.il

Abstract

This paper presents an approach to developing an interactive psychology therapist using the LLM Llama 2. The proposed therapist aims to engage users in treatment like conversations by generating responses that are not only contextually relevant but also empathic, questioning, and reflective. It should mimic a real therapist session. Leveraging the capabilities of Hugging Face's version of Meta's Llama 2, we demonstrate the potential of creating an AI-driven psychology therapist that can effectively understand and address users' psychological concerns.

1 Introduction

Artificial intelligence has significantly improved in the field of natural language processing, enabling a wide range of applications, including chatbots that engage in conversations nearly indistinguishable from human interactions. We say nearly, because it is still easy to tell the difference. With chatbots we get an emotion free robotic kind of response, which makes any application of a chatbot with a need of emotion problematic. They are great for Q&A, giving information, planing trips, but when humans need an emotional companion, they are at fault. A great example is the psychology field, and specifically an AI therapist that can provide empathic support to individuals grappling with emotional and psychological challenges.

Mental health is an integral aspect of overall well-being, yet many individuals face barriers to accessing professional psychological support. Geographical constraints, financial limitations, introverts, people who don't feel comfortable with opening up to strangers or with the one sided sharing, and the persistent stigma surrounding mental health care can deter people from seeking assistance when they need it. Using the new NLP technology, today we can create a pseudo therapy session using ai. This breaks down the barriers and allows for an enhanced mental support system for all those who need it.

This paper introduces an innovative approach that harnesses the vast opensource nlp environment recently created, that allows even undergraduates to study the field, to create a responsive and empathic psychology therapist. Our aim is to develop an AI-driven therapist capable of understanding users' concerns and providing responses that are empathic, questioning, reflective, and contextually relevant. Such a therapist can serve as a complement to traditional therapy or offer an initial point of contact for individuals seeking emotional support. It should simulate a real therapist session.

At the heart of our research is the LLM Llama 2 language model, developed by Meta(Touvron et al., 2023) which was released earlier this year. We are using huggingFace's model[1] in order to use Meta's Llama2 model. LLM Llama 2 has demonstrated good capabilities in maintaining context and generating coherent responses in context-rich conversations. It serves as the foundational technology for our AI psychology therapist. We fine tuned the given model so that we get a more real like conversation.

To train our therapist, we employ the Annomi(Wu et al., 2022) dataset, a valuable resource comprising diverse conversations related to mental health and emotional well-being from real therapy sessions. This dataset encompasses a wide spectrum of emotional scenarios, enabling our model to learn and adapt effectively to various psychological contexts. The conversations were graded by experts in the field and the score given was good/bad. We used only the good conversations to train our model.

We also used 2 other therapy sessions that were released to the public to add diversity.

We used hugging face's peft implementation[2],

[1] https://huggingface.co/blog/llama2
[2] https://github.com/huggingface/peft

specifically LoRA(Hu et al., 2021) in order to fine-tune the 7 billion parameters LLM in a much shorter and more efficient way. We saw that it works well in this implementation (Gema et al., 2023) and decided to use it as well. We used bits and bytes(Dettmers et al., 2022) to improve the calculation time.

In our quest to evaluate the quality of the responses generated by our therapist and ensure they embody the empathic, questioning, and reflective nature expected of a psychology therapist, we turn to the EMP-EVAL(Amjad et al., 2023) framework. This framework provides a structured evaluation mechanism to gauge the therapist's performance in terms of empathy, questioning, and opinion sharing. We graded the three values and averaged the score on 1000 different user inputs.

In this paper, we dive into the methodology underpinning the development of our AI-driven psychology therapist, present the results of our experiments and our thought process. By harnessing the capabilities of LLM Llama 2 and drawing insights from the Annomi dataset and EMP-EVAL framework, we aspire to contribute to the burgeoning field of AI-driven mental health solutions. Our ultimate goal is to provide a more accessible avenue for individuals to seek emotional support and enhance their overall well-being.

2 Our Approach

2.1 Preliminaries and Related Work

There are ethical concerns regarding using AI in the psychology domain. Some ethical research has been done(Fiske et al., 2019) and more is needed, but in this article we focused on a possible implementation.

AI can be used to help during or in close to real time during a therapist session to help the therapist(de Mello and de Souza, 2019),(Morande, 2022).

After the initial search, we didn't find a specific implementation in order to address this idea and so this paper came to be.

2.2 Our Data

We will separate the data we used into 4 parts and explain each part separately.

2.2.1 Training Data

In our research we found a couple of sources for training data. At first we searched for records of transcripts of therapy sessions, and then we found Annomi. We decided to mostly use the Annomi dataset as it was already partially preprocessed and the inputs were given a grade, meaning we didn't need to guess which therapy sessions we found were considered good practice and which were not. We wanted to add 2 therapy sessions[3][4] to make our training data more diverse, and so we chose the best 2 therapy sessions we found, where we saw that the therapist was also interpreting the information given and was asking good follow up questions.

The preprocessing of the 2 therapy sessions were important. The therapy session was written as heard, meaning sometimes the therapist answered one worded answers such as "ok" or "hmmmm" while the client was speaking. This was a problem for 2 reasons: a chatbot conversation is not a real-time answer, meaning that we get the full client statement before we answer, and so "hmmm" is not a good answer, as the client expects a real answer to his question. Also, we didn't want to teach our bot that a one worded answer such as "ok" is a good answer, since it's not empathic, it doesn't interpret the coversation and it stops the flow of the conversation.

Sometimes we had text appear in a parentheses such as sounds that were heard, or a small explanation, which was not part of the therapists answer.

In order to prepare the text, we had to remove the short answers and the parentheses. we also had to distinguish between the words of a client and the words of the therapist for the training so we used a special token to denote the start of the therapist words and the end of the therapist words. All we needed to do was connect the sentence before to the sentence after while removing the unwanted parts.

We created a conversation in the length of 4000 tokens (the amount Llama2 can take as an input), conversations longer than that were split into parts, each one as a different conversation. Each conversation was built from multiple inputs from a client and answers from our "therapist" for each input. If the 4000 tokens was at the center of a sentence, we removed the client's answer and the therapists response to the next conversation, that way we kept

[3] https://julac-cuh.primo.exlibrisgroup.com/discovery/fulldisplay?vid=852JULAC_CUH:CUH&docid=alma991026221669703408&context=L&lang=en

[4] https://eportfolio.ucalgary.ca/vroom/wp-content/uploads/sites/140/2017/07/Counselling-Session-and-Analysis-.pdf

the flow and didn't disrupt a sentence.

We evaluated three different checkpoints of the Lamma architecture in our research, since We wanted to test how the amount of training data can affect our results.

Pretrained (as found in huggingface)

Trained on 80% of the Annomi conversations that were graded good.

Trained on 100% of Annomi conversations that were graded good and the two therapy conversations stated above.

2.2.2 Evaluation

We used 1000 inputs from RSDD[5] a Reddit dataset containing sentences written by real people regarding their degraded mental health. These inputs were fed to our trained models and the answers were saved for an evaluation.

2.2.3 Evaluator training data

We used the dataset that came with EMP-EVAL to train it's model. We then used the trained model to grade the answers our model gave to the user inputs

3 Llama 2

Llama 2 is a large language model developed by Meta AI. It is available for free for research use. It's trained on a massive dataset of text and code, including publicly available online data sources. Llama 2 has 7 billion, 13 billion, or 70 billion parameters, depending on the model size. It outperforms other open source language models on many external benchmarks, including reasoning, coding, proficiency, and knowledge tests (according to meta officials).

It's trained on a diverse dataset of text and code, which gives it a broad knowledge base (2 trillion tokens of data from publicly available sources and over one million new human-annotated examples). It's open source, which means that it can be used and modified by anyone. It's available in multiple programming languages, which makes it easy to use for different projects.

Llama 2 is an auto-regressive language model that uses an optimized transformer architecture. The transformer architecture is a neural network architecture that is well-suited for natural language processing tasks. It is composed of a stack of self-attention layers, which allows the model to learn long-range dependencies between words.

Llama 2 has a context length of 4096 tokens, which is twice as long as the context length of previous transformer models. This allows the model to process more information and generate more comprehensive responses. (This also allowed us to use longer conversations when teaching it.)

Llama 2 can be fine-tuned on a specific task. This fine-tuning helps the model to learn the specific features that are important for that task. Llama 2 is trained to be safe and helpful. This training helps the model to avoid generating harmful or offensive text, and to instead generate text that is helpful and informative. This is especially important when creating a mental health chat Overall, Llama 2 is a powerful and versatile language model that is well-suited for a wide range of tasks. It is open source and available for research and commercial use.

4 Our Fine Tuned Llama 2 model

We chose Llama 2 because of it's text generation, and question answering. We wanted a chatbot that can generate a good response but also answer questions during the conversations. We looked for an opensource LLM with these capabilities and found Llama 2 to be the best answer.

We tested 3 different models in all, 2 that were fine tuned, over 2 system prompts each. This means we have 6 test cases in all that will be shown in the results section.

At first, we wanted to train Llama 2 in a similar way as a masked language model. We started training and found out that it doesn't work. After a short research, we found out that we need a different approach since Llama is a causal language model. We had to use different mechanism to calculate the attention because on the one hand we wanted to prevent the model from generating the patient's, or the user who seeks mental help, words but on the other hand to take these words into consideration. We had to override the Trainer class of hugging face, using the cross entropy loss function with the ignore index parameter, which makes sure the specified token id won't be taken into consideration in the gradient updates. In the preprocessing we made sure that the labels of the "patients" words will have an ignored index (-100) and the ids of the patients words stays as they are -

[5] https://georgetown-ir-lab.github.io/emnlp17-depression/

Model	prompt	Empathy score	Interpretation score	Exploration score	Final score
Base Model	Standard	1090	20	16	232.4
80% Model	Standard	1350	**68**	18	304.4
Full + 2 con	Standard	1415	44	16	307
Base Model	New	1164	28	**286**	358.4
Full + 2 con	New	1475	46	248	412.6
80% Model	New	**1488**	46	266	**421.8**

Table 1: Our result: The grade each model got according to EMP-EVAL. Final grade is calculated as 0.2 * Empathy score + 0.4 * Interpretation score + 0.4 * Exploration score.

which means that statistics that the model learned took into account only the therapist words, but in the calculated attention the patients and the therapists words were taken into account just as we wanted.

The other hyperparameters used were:
max_steps = 100
gradient_accumulation_steps=4
learning_rate = $2e^{-5}$
LoRA parameters:
lora_alpha = 32
lora_dropout = 0.05

We chose to use the 7 billion parameters model because of limited compute resources. We also used peft, specifically LoRA to decrease the storage that was used even more, and to make the fine-tuning of this huge model possible. we also used hugging face's bits and bytes to decrease calculation time. This allowed us a short training time of a few hours each time we trained the model.

We used cross entropy as our loss function.

We started the training with an epoch value of around 100, and noticed that the loss was going down to a value of lower than one. When testing we saw that the model returned empty answers, and we realized the model was overfitted on the training data. After reducing the epoch value we got better results.

The message structure that was used is:

<s>[INST] «SYS»

You are a helpful, respectful and honest assistant...

«/SYS»

<prompt> [/INST] <answer> </s>

We used Llama2 architecture that was pretrained for chat abilities, and that was the input format that this model expects to receive as input, in this format, besides the user input, the model receives a system prompt that affects how the model responses. it gives the model context and instructions about it's knowledge, background and how it should "behave".

We tried 10 different prompts to give the model during training. We picked the top one with the highest score and we also used the standard in order to see the difference.

The standard sys prompt:

"You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature.

If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information."

The top valued prompt was based on the standard and had an extension, we'll refer to it as the "New" prompt:

"You must answer in an empathic way and help the user explore what he is feeling by asking meaningful questions, be aware of what the user is feeling and provide complete and well-thought-out responses without truncation."

5 Results

For easier writing, we will refer to the three models as base, 80% and full.

We see that the base model had the lowest scores, for both prompts.

We can see that the Empathy score was relying on both the model and the prompt and we have some mixed grades. We see the highest score given to the 80% model with the new prompt, while close behind was the full model with the new prompt.

We can see close, low scores for the interpretation for all 6 tested options. The highest score given to the 80% model with the standard prompt.

we can see a much higher grade for the new prompt rather than the standard prompt on the exploration score, with the base model having the highest score.

In all, we see that the prompt had a great influence on the output. The New prompt had a higher score in all three models than the standard prompt.

Included is a client input and two outputs, we chose the model with the lowest score and the highest score: "I dont know why I just cried. I was making cookies and I just started feeling overwhelmed with stress."

Base prompt, Base model: " it's completely normal to feel overwhelmed and emotional at times, especially when we are engaging in activities that we enjoy. It's important to acknowledge and validate your feelings, and to take care of yourself during these moments. Here are some suggestions that may help: 1. Take a break: Sometimes, taking a step back and giving yourself permission to take a break can help you feel better. You can take a few deep breaths, go for a walk, or do something else that helps you relax. 2. Practice self-compassion: Be kind to yourself and try to reframe your emotions. Remind yourself that it's okay to feel overwhelmed and that you are doing your best. 3. Talk to someone: Reach out to a trusted friend, family member, or mental health professional and talk to them about how you're feeling. Talking through your emotions can help you"

New Prompt, 80% model: " it's completely normal to feel overwhelmed and emotional when cooking, especially when it's a task that requires attention to detail and precision. It's possible that the act of cooking triggered something within you, perhaps a memory or a feeling that you weren't expecting. It's important to acknowledge and validate your emotions, and to give yourself permission to feel whatever you're feeling. It's okay to take a break and step away from the task if you need to. Sometimes, taking a few deep breaths, going for a walk, or engaging in a relaxing activity can help you regain your composure and perspective. Can you tell me more about what happened while you were cooking that made you feel overwhelmed? What were you thinking about or focusing on during that time?"

Here you can see a big difference with the exploration of the bot and the improvement in the interpretation. Both are empathic but the first output doesn't have a chat flow and stops and directs the client to talk with a real therapist after only one message.

6 conclusions

We see a 36% increase in our final grade between the base prompt and the best grade. Since the base model has the lowest score for each prompt, we see that the fine tuning worked and we achieved a better score.

Since we see all three models have a lower score with the standard prompt, we see that the prompt had a better influence on the final score than the fine tuning in our specific research. We can improve the fine tuning and we can improve the prompt, but until further tests, we can't say for sure which change will have a greater affect.

Since both the full and the 80% with the new prompt have a close empathy score to the standard prompt, we see the bigger increase because of the fine tuning of the models and not because of the prompt. We see that only stating the the answer has to be empathic, doesn't really boost this score and so maybe changing the prompt to explain how to answer in an empathic way can also increase the score.

The Interpretation score was close all round but we see a better result with the trained model. We see that the prompt had a great influence on the output. The New prompt had a higher score in all three models than the standard prompt, with the highest score given to the 80% model. Since all the scores were really close between both prompts, we can assume that the prompt didn't have any real affect on the interpretation score, but we think that the fine tuning of the models did have an influence, to some extent. This means that if the prompt would have had a part stating that it should also interpret the answer of the client, maybe we would have seen this score go up. This could have affected the exploration score so we should be very specific to make sure one isn't improved while the other is downgraded.

Interpretation is also the hardest part of the answer, since the model is already built in a way that makes it empathic, and exploring is not hard - asking questions to gather more data and asking for explanations of the previous answers, can greatly improve this score while asking harder not so spe-

cific questions in order to move the conversation in another different subject, or trying to connect the dots can be harder. This also makes interpretation harder - the model needs to connect different sentences that were said, add logic and understand human behavior to give a good interpretation.

The "New" prompt stating that the model had to ask meaningful questions to help the client, boosted the Exploration score, meaning that the chat has an ability to create a real conversation and not only oneliners. This is a important aspect of a chatbot and an important aspect of a therapy session. However, the Interpretation score was not boosted at all, and so finding a better prompt that also tells the model to reference this aspect when answering can also greatly increase the final score.

We were surprised that our 80% model had better results than the full. This could be caused by multiple reasons and needs to be further checked. As an example, if we look at the Annomi dataset and categorize the therapy conversations categorization:

Topic	**#Dialogues**
Reducing alcohol consumption	28 (21.1%)
Smoking cessation	21 (15.8%)
Weight loss	9 (6.8%)
Taking medicine / Following medical procedure	9 (6.8%)
More exercise / Increasing activity	9 (6.8%)
Reducing drug use	8 (6.0%)
Reducing recidivism	7 (5.3%)
Other	48 (36.1%)

Table 2: Taken from Annomi(Wu et al., 2022)

This, considering we used only good conversations (around 90% of all conversations), shows us the different types of conversations we taught the model. This means that if some of the sentences we used from Reddit contained other subjects, such as depression, our model wasn't taught exactly how to cope with those subjects. Using all 100% of the good conversations and adding 2 more conversations, might have caused overfitting, causing our model to not answer well on other subjects.

The training should match the testing, and since we used random 1000 sentences from the reddit mental health dataset, we might have used data that does not really reflect the model's abilities. Having said that, an AI model used for mental health issues, should have a wide variety of subjects to learn from, including depression, dealing with failure etc.

We wanted a model that could simulate a full therapist-client conversation, but only tested our results on a one-liner conversations. More tests are needed to check and create a good full therapy model. Since the model is already trained to be able to hold a conversation this is not an aspect we wanted to test yet.

However the following are a good place to continue the research:

Looking for a better prompt that increases any one of the specific empathy, interpretation and exploration score.

Looking for a better model that increases any of the scores.

Testing the model's ability to hold a full conversation.

Training Llama 2 using Reinforcement learning technique.

6.1 Full disclosure

Mental health is an intense subject that could have bad consequences if it's not approached correctly. It's also complex and needs the ability to understand human behavior to fully understand the connections a person can have between two emotions, two performed actions, the person's history and his personality and so on. For AI to take a place in this world, it would have to be with a close eyes of professional therapist. They should have a part in the training and the evaluation.

Training a mental health AI is hard since there are multiple scenarios that could be discussed at a therapy, not all have a released session, and so teaching the model how to react could be hard. Also there are many different ways to address a certain issue, and so there might not be a right answer and the AI's behavior could be biased, based on the therapist that helped built it.

Our evaluation used AI to give a score. While we did go over some of the answers we received from all three models, we used 1,000 inputs and so we didn't read the majority of the outputs. This could mean that we could have had an output that could have devastating outcome. While giving a score on how well the model preformed, another score should be given - on how bad the output was and if it could lead to a misfortune if a person suffering from poor mental health were to receive this output.

User feedback could also be used to receive some input on how well the outputs were. It could ask the client if he felt that the interpretation was correct, if the conversation helped him or if it was a bad experience and what should change. There should be human eyes going over the data that the model receives throughout the training, even the user feedback, since we don't want a person to worsen the therapy like conversation another client has with the model.

As this chatbot can change the way a person thinks about himself, his life, and can cause the person a worse mental sate than he had before he began the chat, this should have an age restriction, or a legal guardian that would watch closely and make sure everything is alright. A person can be persuaded to hurt himself or others by an AI who he thinks is there to help him.

There is no question whether AI should take a part in therapy sessions, but it should be done with care.

References

Bushra Amjad, Muhammad Zeeshan, and Mirza Omer Beg. 2023. Emp-eval: A framework for measuring empathy in open domain dialogues.

Flávio Luis de Mello and Sebastião Alves de Souza. 2019. Psychotherapy and artificial intelligence: A proposal for alignment.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale.

Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. URL: https://doi.org/10.2196/13216, URL: http://www.ncbi.nlm.nih.gov/pubmed/31094356.

Aryo Pradipta Gema, Luke Daines, Pasquale Minervini, and Beatrice Alex. 2023. Parameter-efficient fine-tuning of llama for the clinical domain.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Swapnil Morande. 2022. Enhancing psychosomatic health using artificial intelligence-based treatment protocol: A data science-driven approach. https://www.sciencedirect.com/science/article/pii/S2667096822000672. International Journal of Information Management Data Insights, 2(2), 100124.

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.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.