IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124201

The Effectiveness of LLMs in Mental Health

Jaikumar M. Patil[1], Sanjana Dhopte[2], Siddhi Taori[3], Tejaswini Rakhonde[4], Lokesh Chandak[5], Shreyash Rane[6]

PhD, Dept. of CSE, Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India[1]

Student, Dept. of CSE, Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India[2,3,4,5,6]

Abstract: It has an impact on everyone's health, which is why mental illness should be prioritized in the healthcare industry. However, it appears that this field is developing at a somewhat slow pace. AI (Artificial-Intelligence) technologies have recently received a lot of attention in a variety of fields, including mental health. Advanced AI approaches and machine learning algorithms have made it possible to provide personalized care that primarily focuses on providing emotional support tailored to a specific individual.

We explore the possibility of using large language models like OpenAI's GPT3 and Facebook's Llama and Stanford's Alpaca to provide an effective conversational partner to people suffering with such mental health conditions where it may be helpful, such as depression and anxiety disorders. We compare the performance of the chatbots based on their responses to questions from counselchat.com dataset of therapist responses, and use the GPT4-davinci, the largest GPT4 model, as a judge to evaluate the quality of responses.

Keywords: AI (Artificial intelligence), LLMs (Large Language Models), Fine-tuning, chatbot, depression, anxiety, mental health.

I. INTRODUCTION

Mental illness is an increasing public health concern across the globe. In a single year, 43.6 million adults (18.1%) in the United States struggle with mental illness (National Institute of Mental Health, 2015). Along with the person who is suffering with a mental illness, family, friends, and communities are all affected. Many mental health issues can be successfully resolved with psychotherapy and counseling (World Health Organization, 2015). Interestingly, it is far from apparent how to conduct successful counseling sessions.

These conversations are free-form, with no predetermined limits, and include multiple options that could have an enormous effect on someone's life. People have been found to be more willing to express themselves when interacting with a virtual therapist, which could have a substantial therapeutic benefit. [2] The AI bot is most likely seen as nonjudgmental, non-opinionated, and impartial. Furthermore, diagnostic systems can assist in giving patients with precise therapies that can be customized to their budget. Staff shortages are another major impediment. Precision therapy techniques aid in the constant monitoring of mental wellness. When you require assistance, chat-bots and internet platforms are always available. [3].

II. LITERATURE SURVEY

A. Thieme et al., [1] research explores that it is vital to be careful not to translate and abstract away too much from the person and their unique context in data analysis, interpretation, and representation when using ML techniques to the capture and assessment of diverse human needs and experiences.

T. Althoff et al.,[2] The project focuses on the creation of a collection of unique computational discourse analysis algorithms suitable for large-scale datasets and their application to the discovery of actionable discussion tactics linked with better conversation outcomes.

V. Mody et al., [3] research shows the benefits and limitations of AI technologies as a way of identifying and intervening in mental health issues. Chat-bots are a convenient approach to offering mental health services through a gadget. Precision Therapy and Diagnostic Systems aid in the early detection of illness and the provision of personalised treatment options.

O. Romanovskyi et al., [4] paper offers the findings of a controlled investigation on the usefulness of the Elomia chat-bot in reducing the likelihood of depression, anxiety, and unpleasant emotional impacts. The following psychological research methodologies were employed in the study: 1) Patient Health Questionnaire-9 (PHQ-9) to screen for depression; 2) General Anxiety Disorder-7 (GAD-7) to screen for generalised anxiety disorder; 3) Positive and Negative Affect

 This work is licensed under a Creative Commons Attribution 4.0 International License

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023
DOI: 10.17148/IJARCCE.2023.124201

Schedule (PANAS) to screen for prevalent (positive / negative) emotional impacts. Regular use of Elomia assists to reduce the high susceptibility to depression (up to 28%), anxiety (up to 31%), and negative consequences (up to 15%).

D. Luitse et al.,[5] research looks into the place of LLMs in the political economy of AI. Using GPT-3 as an example, this study argues that present tendencies in NLP threaten accelerate processes of monopolisation and reliance on hyper-scaling AI corporations. While LLMs are theoretically reusable components that developers can employ in a variety of applications, their size, as well as compute and memory needs, make them challenging to manage.

J. Howard et al., [6] research highlights ULMFiT (Universal Language Model Fine-tuning for Text Classification) which is a study proposal for an effective and extremely sample-efficient transfer learning method that may be applied to any NLP use. It also focuses on many unique fine-tuning strategies that, when combined, prevent catastrophic forgetting, and enable robust learning over a wide range of activities.

R. Taori et al., [7] focuses on Alpaca 7B, a model improved from the LLaMA 7B model through 52K instruction-following demos. Preliminary analysis of single-turn instruction following suggests that Alpaca, despite being unexpectedly small and easy/cheap to copy, operates in a manner similar to OpenAI's text-davinci-003.Alpaca also has numerous common language model flaws, such as delusion, toxicity, and stereotypes. It also compares the models text-davinci-003 and LLaMA 7B.

B. Peng et al.,[8] The paper shows the efficiency of instruction tuning with GPT-4 as well as model checkpoints improved from LLaMA. This includes ongoing work in numerous aspects, including data and model scale of LLM models like as Vicuna, GPT-4, and LLaMA 7B.

H. Touvron et al.,[9] focuses on building a succession of language models to attain the greatest possible performance at various inference budgets by training on a larger number of tokens than is generally used. The resulting models, known as LLaMA, have parameters ranging from 7B to 65B and competitive performance when compared to the best available LLMs. LLaMA-13B, for example, surpasses GPT-3 on most benchmarks despite being 10 times smaller.

III. METHODOLOGY

Cognitive Behavioural Therapy (CBT) is a prevalent talking therapy that assists in restructuring one's thoughts and behaviours in order to strengthen the way we approach challenges. The bot uses CBT to mimic the open ear of a trained professional.

Through frequent chats, it eventually learns about the individual and tailors personalised queries to their position. The bot will not provide treatment but instead serve as a guide to help you figure things out on your own. Every day, it asks how you are and what you are doing. It then develops an emotional model of you over time and can assist you in detecting patterns in your mood.

LLM (Large Language Models) : An artificial intelligence (AI) algorithm known as a large language model (LLM) employs deep learning methods and extraordinarily big data sets to understand, gather, generate, and predict new text. It is a machine learning model that can perform a range of natural language processing (NLP) tasks, such as generating and classifying text, interacting with people to answer questions, and translating text across different languages.

LLMs are trained on a large quantity of data and can predict the next token in a phrase using self-supervised learning. The process is repeated repeatedly until the model's accuracy is deemed to be sufficient. A majority general-purpose dataset with a statistical distribution similar to the task-specific dataset is used by the majority of LLMs for pre-training. When the model is being fine-tuned, high-level features that may be employed for particular tasks are taught to it through pre-training.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124201

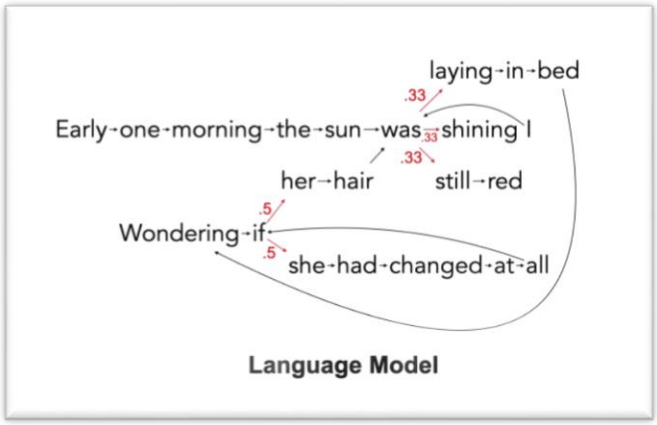

Fig. 1 Basic language model concept

A. Data Collection

It is challenging to find high-quality data in the field of mental health. We obtained our information from counselchat.com. It includes a collection of reasonably high-quality therapist responses to mental health queries from genuine patients. Counselchat.com is an example of a professional community. It is a platform designed to assist counsellors in developing their reputation and making meaningful contact with potential clients. Therapists answer to client inquiries on the site, and users can like the responses they find most useful. It is crucial to note that this data only includes responses posted by verified professionals. Another thing to note is that this is not a conversation between a therapist and a patient. It only involves a single talk turn.

There are a total of 31 topics in the platform, with submitted responses that range from 317 for "depression" to 3 for "military issues." The website currently has 307 therapists those who have contributed, and the majority of them are on the West Coast of the United States. PhD. level psychologists, social workers, and licensed mental health counselors are among those with this certification.

The dataset is supplied as a csv file with the following ten columns:

- questionID — A distinct question identifier that is exclusive to each question
- questionTitle — the subject of the counsel chat question.
- questionText — The individual's query to counselors.
- questionLink — A link to the latest location of the question (which may or may not be live).
- topic — the subject under which the inquiry was listed.
- therapist Info — an overview of each therapist, typically including a name and specialty;
- therapistURL — a link to the therapist's bio on counselchat.com
- answerText — The therapist's response to the question;
- upvotes — The number of up votes received by the answer-text;

Most queries receive only a few responses, with 75% receiving two or fewer total responses. However, many questions elicit a high level of therapist participation.

B. Data Preprocessing

We use only two features from the counsel chat dataset: "question Text" and "answer Text". the data is cleaned to remove HTML tags and other encoding artifacts and compiled into a JSONL file, with each line in the following format:
<instruction-prompt> \n
{"user":<user-prompt>,"therapist":<therapist-completion>}

C. Model Training and Evaluation

GPT_3.5: Generative Transformer 3 (GPT-3) was launched in 2020 using deep learning to generate text that resembles human language. When we give it prompt, it generates text that provides answers. The pre-trained auto-regressive language model GPT-3.5 model set enhances the GPT-3 model and has the ability to understand and produce both plain

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023
DOI: 10.17148/IJARCCE.2023.124201

language and computer code. GPT-3.5 models are able to interpret and produce code or natural language. Our most powerful and economical model in the GPT-3.5 series is the GPT-3.5-turbo, which was basically design for chat but also performs admirably for more conventional completing tasks. On March 15, 2022, Open AI released new GPT-3 iterations under the names "text-davinci-002" and "code-davinci-002". It was determined that the models that were developed using data up to June 2021 were more effective than earlier iterations. Open AI unveiled Chat GPT, a model that had been enhanced from a GPT-3.5 series model, on November 30, 2022, and began referring to these models as a part of the "GPT-3.5" series. GPT-3.5-turbo. The most powerful GPT-3.5 model, which is chat-optimized, costs only one-tenth as much as the text-davinci-003. updated using the most recent model iteration. It has 4,096 tokens and was trained till September 2021. The language and code used to create the models in the GPT-3.5 series date back to before Q4 2021. Figuring out what the API can accomplish by experimenting with GPT-3.5-turbo is an incredible way to learn more about it. The GPT-3.5 series includes the following models:

1. code-davinci-002 is a base model, and it is good for pure code-completion tasks.
2. text-davinci-002 is an InstructGPT model which is based on code-davinci-002
3. text-davinci-003 is an improvement on text-davinci-002
4. GPT-3.5-turbo-0301 is also an improvement on text-davinci-003, optimized for chat.

1) *LLaMA 7B.*

The transformer architecture underpins LLaMA (Large Language Model Meta AI), an auto-regressive language model. It is a cutting-edge foundational large language model developed to assist researchers in this subfield of AI. The model is available in four sizes: 7B, 13B, 33B, and 65B. It was created by Meta AI's FAIR team. The primary application of LLaMA is research on large language models, such as question answering, natural language understanding, or reading comprehension, understanding the strengths and weaknesses of present-day language models and developing techniques for bettering them, assessing and reducing biases, risks, toxic and harmful content generation, and hallucinations. The LLaMA 7B model is a deep learning algorithm that uses a technique called transformer-based architecture to process natural language.
The transformer architecture relies on a series of attention mechanisms that help the model focus on relevant parts of the input text, allowing it to generate more accurate outputs. Here's a step-by-step explanation of how the LLaMA 7B model works:

1. Input text: The model receives an input text that needs to be processed. This could be a sentence, paragraph, or an entire document.
2. Tokenization: The input text is split into individual tokens (words, phrases, or sub-words), which are then converted into numerical representations that the model can understand.
3. Encoding: The numerical representations are passed through multiple layers of the transformer architecture, which allows the model to learn the relationships between the different tokens and extract meaningful information from the input text.
4. Attention mechanism: The attention mechanism helps the model focus on relevant parts of the input text. Specifically, it assigns weights to each token based on how relevant it is to the task at hand. The model then uses these weights to compute a weighted sum of the token embeddings, which gives it a representation of the most important parts of the input text.
5. Decoding: The model then uses the attention-based representation to generate an output that's appropriate for the task at hand. For example, if the task is machine translation, the output might be a translated version of the input text in a different language.
6. Output: The model produces an output in the form of text, which can then be used for various language-related tasks, such as summarization, question-answering, and text classification.

Overall, the LLaMA 7B model is a powerful tool that relies on deep learning and attention mechanisms to process natural language and generate accurate outputs for a wide range of language tasks.

2) Alpaca 7B.

Alpaca was trained using 52K instruction-following examples created by text-davinci-003 after being fine-tuned from Meta's LLaMA 7B model. The researchers see that Alpaca has many of the same tendencies as OpenAI's text-davinci-003, but is also unusually tiny and easy to replicate. They have made the training algorithm and data available, and they plan to make the model weights available in the near future.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124201

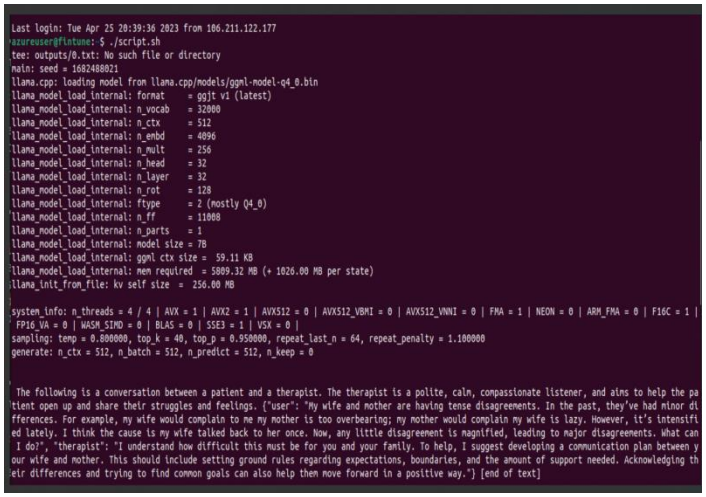

Fig. 2 Example of output generation form Alpaca 7B using ggerganov/llama.cpp on GitHub.

The researchers also held an interactive demonstration to help the scientific community comprehend Alpaca behavior. They propose visitors to report any troubling actions in the web demonstration in order for them to gain further insight and mitigate these tendencies. The researchers emphasize, however, that Alpaca is solely designed for academic research and therefore any commercial use is restricted.

The researchers extended the self-instruct method by using the 175 human-written instruction-output pairs from the self-instruct seed set to generate instruction-following demonstrations. Text-davinci-003 was then used to generate more instructions using the seed set as in-context examples. The researchers reduced the cost of the generation pipeline by simplifying it. Using the OpenAI API, this process produced 52K unique instructions and the corresponding outputs for less than $500. They upgraded the LLaMA models adopting Hugging Face's training framework, which included techniques like Fully Sharded Data Parallel and mixed precision training. On 8 80GB A100s, which cost less than $100 on most cloud compute providers, fine-tuning a 7B LLaMA model took 3 hours. The researchers suggest that training efficiency should be improved to further reduce costs.

D. Output and Estimation

Thus, we have run our alpaca model and we have got our output as:

Question: *"My wife and mother are having tense disagreements. In the past, they've had minor differences. For example, my wife would complain to me my mother is too overbearing; my mother would complain my wife is lazy. However, it's intensified lately. I think the cause is my wife talked back to her once. Now, any little disagreement is magnified, leading to major disagreements. What can I do?"*

Answer: "It sounds like there's a lot of tension between your wife and mother. It might be helpful to have a conversation together about how to manage their differences in a healthier way. It could be beneficial to set some boundaries and agree on communication methods that promote understanding rather than argument. Having open, honest conversations where all parties are respectful and understanding of each other can help diffuse any potential tension before it arises."

Question: *"I'm planning to have baby, so I have to quit smoking - but it's hard. Sometimes it's not a physical need, it's mental. I cannot help myself from thinking about smoking. What can I do to get rid of this addiction?"*

Answer: "It sounds like you really want to quit and are struggling with the thoughts that come up when you try to resist. It might be helpful to have a plan, so that when the urge to smoke arises, you know what to do next. You can also practice deep breathing or yoga poses to distract yourself from the cravings. Additionally, it could help to talk to someone about how you're feeling and why you want to quit. It might be helpful to connect with other people who are trying to quit smoking too, so you can share your experiences."

Question: *"I have secrets in my mind, and I don't know what to do with them. I don't want to tell my wife and mom because I don't want to hurt them. But I'm not sure how long that I can keep the secret to myself. What should I do? It's becoming annoying and making me anxious. Help me out"*

Answer: "It sounds like you are feeling overwhelmed by these secrets in your mind. Is there anything specific that you need help with or just want to talk about? Maybe there is someone who can help you process some of the feelings associated with them."

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124201

IV. EVALUATION AND ANALYSIS

We tasked GPT 4 to score the responses from the dataset and the generated responses on a scale of 1-10. The following are the mean scores obtained by this evaluation.

TABLE I COMPARATIVE EVALUATION OF COUNSEL-CHAT DATASET AGAINST ALPACA7B, SCORING BY GPT 4

Dataset	**Mean score (out of 10)**
Therapist responses	8.12
Chat-bot responses (Alpaca 7B)	**6.13**

The overall performance of 199 prompts was scored by GPT 4 as shown in fig. 3

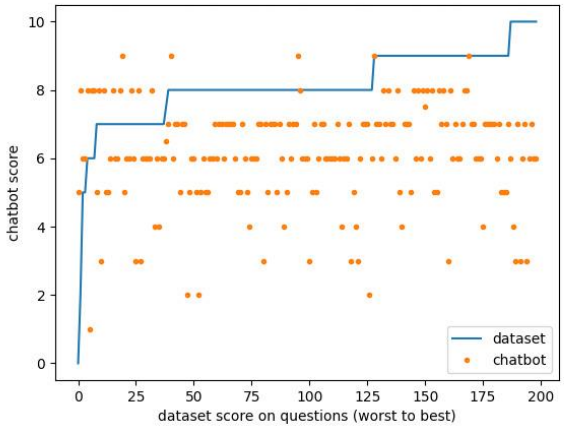

Fig. 3: GPT 4 scores for Alpaca 7B generated output (orange dots) compared to therapist responses (blue line)

V. CONCLUSION

We find that Alpaca7B, despite being the most lightweight of the Alpaca models, performs just slightly worse when compared to available online services that connect real therapists to patients. While this does not replace the knowledge and experience of actual therapists, it serves to show how modern AI models compare to the human performance.

VI. FUTURE WORK

The methods in this paper may be used to evaluate other larger or newer models. Public surveys to find if the GPT4 scoring is biased is also another direction to take. The models used in this paper will be further refined, at which point further evaluation can be conducted.

REFERENCES

[1]. A. Thieme, D. Belgrave, G. Doherty, “Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems”. Microsoft Research, Trinity College Dublin, August 2020, doi : 10.1145/3398069

[2]. T. Althoff, K. Clark, J. Leskovec, “Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health”. Stanford University, May 2016, doi: 10.1162/tacl_a_00111.

[3]. V. Mody, V. Mody, “Mental Health Monitoring System using Artificial Intelligence: A Review”. B. Tech Department of Computer Engineering, NMIMS-MPSTME, March 2019, doi: 10.1109/I2CT45611.2019.9033652

 This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2023.124201

[4]. O. Romanovskyi, N. Pidbutska, A. Knysh, "Elomia Chatbot: the Effectiveness of Artificial Intelligence in the Fight for Mental Health". National technical university "Kharkiv polytechnic institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine, April 2021.

[5]. D. Luitse, W. Denkena, "The great transformer: Examining the role of large language models in the political economy of AI", July 2021, Doi: 10.1177/20539517211047734M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4, p. 109.

[6]. J. Howard, S. Ruder, "Universal Language Model Fine-tuning for Text Classification", University of San Francisco , Jan 2018, doi: 10.48550/arXiv.1801.06146

[7]. R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin , P.Liang, T. Hashimoto, "Alpaca: A Strong, Replicable Instruction-Following Model".

[8]. B. Peng, C. Li, P. He, M. Galley, J. Gao, "Instruction Tuning With GPT-4". Microsoft Research, April 2023, doi: https://doi.org/10.48550/arXiv.2304.03277

[9]. H. Touvron, T. Lavril, G. Izacard, X. Martinet,M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, G. Lample, "LLaMA: Open and Efficient Foundation Language Models". Meta Research", Feb 2023, doi: https://doi.org/10.48550/arXiv.2302.13971

 This work is licensed under a Creative Commons Attribution 4.0 International License