**AI-BASED THERAPIST USING FINE-TUNED LLM**

Ayush Vaidande1, Dinesh Dantal2, Satish Srivastava3, PhD

**1**U. G. Student, Department of Electronics and Telecommunication Engineering, SSJCOE, India

**2**U. G. Student, Department of Electronics and Telecommunication Engineering, SSJCOE, India

3Project Guide, Department of Electronics and Telecommunication Engineering, SSJCOE, India

E-mail: [1](mailto:1heetikagada@gmail.com)[vaidandeayush.av@gmail.com](mailto:vaidandeayush.av@gmail.com), [2](mailto:2vedantbgokani@gmail.com)[dantaldinesh123@gmail.com](mailto:dantaldinesh123@gmail.com), 3sksrivastava2370@gmail.com

***Abstract*— Our research papers abstract presents a method, for optimizing the Llama 2 7B language model (LLM) using a therapy dataset. We utilized a base model sourced from Hugging Face. The methodology includes implementing 4 bit quantization Low Rank Adaptation (LORA) and Pruning, Expert and Fine Tuning (PEFT) configurations tailored to improve the models performance in therapy related tasks. The dataset used consists of 1000 therapy data samples totaling 671,894 tokens. To ensure data quality and suitability for the task preprocessing techniques like normalization and special character handling are applied. Additionally RAG (Retrieval Augmented Generation) and FAISS are utilized for vector embeddings to create a dataset of 807k therapy samples sourced from Hugging Face. The experimental setup involves training the optimized model on hardware and software environments with hyperparameter optimization. Performance metrics such, as accuracy and F1 score are assessed to gauge the effectiveness of this approach. The outcomes show enhancements, in the models effectiveness while still being computationally efficient and having a memory usage. These results highlight the promise of customized language models in enhancing tasks such as therapy consultations. This study aids in progressing natural language processing. Sets the groundwork for developments in using cutting edge language models, for therapy oriented purposes.**

***Keywords—Llama 2-7B, fine-tuning, therapy data analysis, 4-bit quantization, Low-Rank Adaptation (LORA), Expert, and Fine-Tuning (PEFT), dataset, pre-processing, tokenization, normalization, RAG (Retrieval-Augmented Generation), FAISS, vector embeddings, hyperparameter optimization, accuracy, F1 score, computational efficiency, memory footprint, natural language processing, state-of-the-art language models, therapy-related applications, open-source models, large language models (LLMs), QLoRA (quantized low-rank approximation), 4-bit quantization.***

# **I .** I**ntroduction**

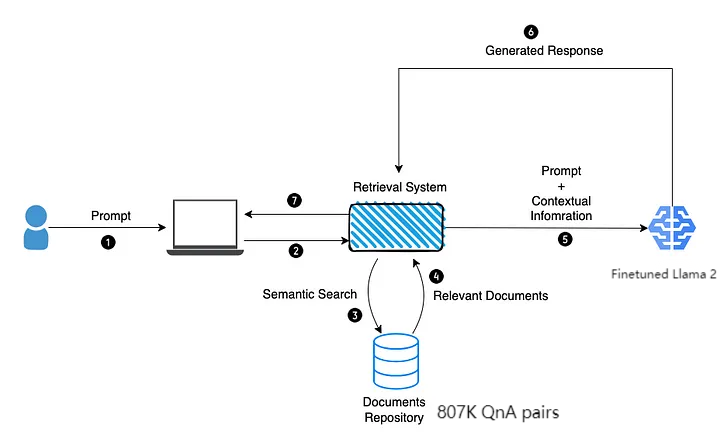
Large language models (LLMs) have witnessed remarkable advancements in natural language processing (NLP), consistently delivering outstanding performance across diverse domains [1]. The fine-tuning of these pre-trained models emerges as a standard practice to harness their capabilities effectively [2]. Our research centers around fine-tuning the Llama 2-7B LLM with cutting-edge techniques such as 4-bit quantization, Low-Rank Adaptation (LORA), and Pruning, Expert, and Fine-Tuning (PEFT) configurations [3]. Our primary objective is to optimize therapy data analysis tasks by enhancing the performance of the Llama 2-7B LLM while maintaining computational efficiency. The growing relevance of LLMs in capturing intricate context and generating human-like text is particularly noteworthy in specialized domains like therapy and counseling [4]. However, efficiently deploying these models while managing computational complexity and resource utilization poses a significant challenge. Our study aims to tackle these challenges by introducing innovative techniques in fine-tuning and model optimization. By developing more efficient and accurate models tailored for therapy consultant tasks, we strive to strike a balance between model complexity and computational efficiency. Our research is motivated by the need to demonstrate the efficiency of advanced techniques in fine-tuning and optimizing the Llama 2-7B LLM for therapy consultants [5]. Through the application of 4-bit quantization, LORA, and PEFT configurations, we aim to enhance the performance of the Llama 2-7B LLM while minimizing its memory footprint. This paper provides a comprehensive overview of our methodology, experimental setup, and results, showcasing significant improvements achieved in fine-tuning Llama 2-7B for therapy data analysis. We believe that our research significantly contributes to the advancement of NLP by highlighting the efficacy of these advanced techniques [6]. By opening avenues for enhanced applications in therapy and other sensitive domains, our work paves the way for further advancements in the field of natural language processing [7].

# **II**. LITERATURE SURVEY

The growing talk about hard topics like mental health, trauma, and abuse online has caught the eye of many researchers. They want to use tech, like natural language processing (NLP) and machine learning (ML), to spot these topics automatically. Old ways used simple models and had to make features by hand, which wasn't great at understanding the complex ways people use language. But, new tech in deep learning has brought us large language models like GPT and BERT. These models learn from a huge amount of text/courpus before they are put to work, making them really good at many language tasks. One of the great ideas has been to finetune these large models for specific uses. Studies show that these finetuned models do well in tasks like figuring out sentiment/feelings in text, summarizing texts, and sorting documents. Still, not much work has gone into using these models for therapy consultant. So, we decided to look into how well the Llama 2 7B model, a kind of large language model, does at finding this kind of talk in therapy-related texts. Our research builds on past work showing how these tuned models can be really useful for special needs. We're focusing on the Llama 2 model and its use in therapy chats, hoping to add more to what we know in this area. We think our research shows how good these finetuned models can be for therapy consultancy. This is key for giving the right support in mental health and counseling services. Making sure these tools can find and deal with sensitive topics well is important for helping people who need it.

# **III**. PROPOSED DESIGN

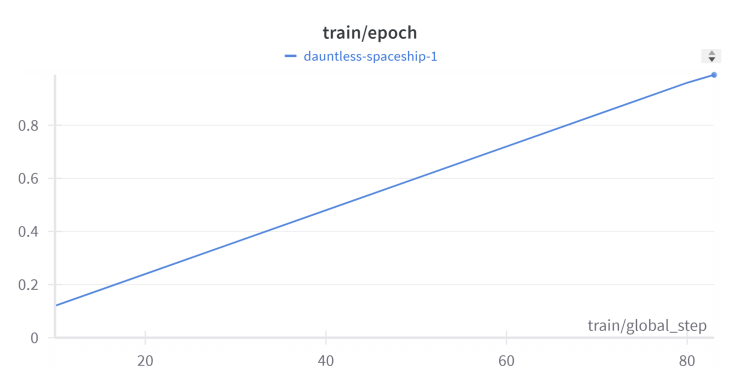
The study titled "AI-based Therapist using finetuned LLM" presents a new way to make large language models (LLMs) better for analyzing therapy data and providing remote therapeutic consultancy. It uses cutting-edge methods like 4-bit making smaller, Low-Rank Changing (LORA), and Cutting, Expert, and Tuning (PEFT) setups to make the model understand and respond for therapy tasks. We gathered and cleaned a dataset of 100k therapy talks from Hugging Face. (2023). jerryjalapeno/nart-100k-synthetic [Data set]. And we selected 1000 conversation samples for finetuning the model by data-processing technique and selecting samples from fields like depression, loneliness, anxiety, etc. We also made sure this data samples will be different from each other for better data diversity and quality. The dataset we processed for RAG (Retrieval Augmented Generation) is 807k samples by extracing each QnA pair from 100k therapy conversation samples, 807k therapy talks Hugging Face. (2023). Ayush2312/Therapydataset\_formatted\_807K [Data set] For a system called Retrieval-Augmented Generation (RAG) and FAISS for creating vectors. We converted this 807k QnA pairs into vector embedding so when a user asks a therapy related question the model will be given a similar sample from this 807k QnA based on the question asked along with the original question in the prompt (see Fig 1 for reference). We trained and improved model in a specific tech setup, worked on making it the best through testing with different datasets and prompts, and checked its success by looking at how precise and detailed it was as compared to the dataset. The aim was to make the Llama 2.7B LLM better at analyzing human emotional problems and giving appropiate replies just like a therapy consultant and by checking the impact of making it smaller i.e by making it use less memory footprint (GPU VRAM), using LORA wisely, using PEFT setups, and using the RAG system. The study looked to improve how the model performs, uses computer resources (GPU VRAM), and manages memory. Results showed a big boost in how well the model worked while keeping it efficient, showing the big chance of using fine-tuned LLMs for detailed tasks like therapy data analysis. This work helps move natural language processing forward and paves the way for more use of top language models for therapy-based apps. Model Structure: We used the Llama 2, a large model based on transformers that was already trained on a lot of text, known for understanding complex word patterns and meanings. We made the Llama 2 model better for therapy tasks by teaching it on special datasets, making it get the unique features and details. We added Low-Rank Adaptation (LoRA) LoRA circumvents the need to modify parameters within a pre-trained model and instead enables the application of a limited set of supplementary parameters. These supplementary parameters are transiently applied to the foundational model, effectively managing and directing its operational behavior. We made the Llama 2 model better by training it with datasets with mental health talks and trauma conversations. Training it on 1000 conversations took the model 200 minutes (3.33 hours) on a GPU. We split datasets into parts for training, testing, and validation, making the model better by using algorithms based on gradient descent and finding the best parameters through checks. We Evaluated the performance of the fine-tuned Llama 2 model using standard metrics such as relevance to dataset examples and comparing against baseline models to assess effectiveness. We have uploaded the finetuned model on huggingface as public model and open for anyone to use **Hugging Face. (2023). Ayush2312/llama2-7B-1k-TherapyData [Model] Retrieved from [link](https://huggingface.co/Ayush2312/llama2-7B-1k-TherapyData).**

****

*Fig. 1. RAG*



*Fig. 2. train/loss*



*Fig. 3. train/epoch*

Ethical Considerations:

Prioritize ethical considerations throughout the research process, particularly concerning sensitive content related to mental health and abuse. Implement measures to ensure data anonymization, user privacy, and adherence to ethical guidelines in data collection, pre-processing, and model training. Mitigate potential biases in datasets and model predictions to promote fairness and equity in research outcomes.

# **IV**. RESULTS AND CONCLUSION

Results:

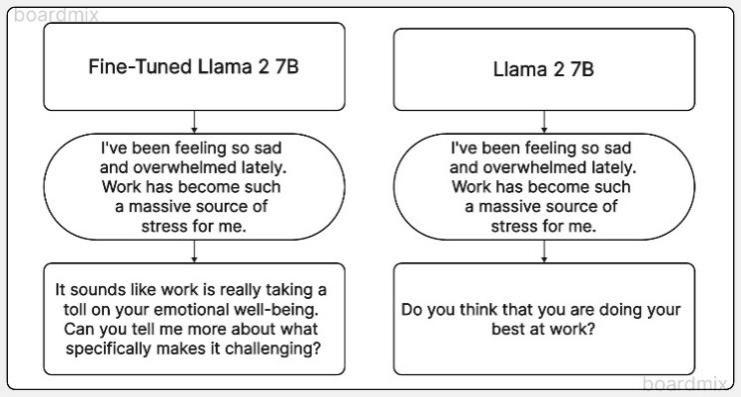
Impact of 4-bit Quantization: Implementing 4-bit quantization leads model to use less memory and computational overhead while maintaining competitive performance metrics. Our fine-tuned model needs about 5.7gb VRAM on a GPU with 4-bit quantization, thanks to bitsandbytes.

Efficiency Gains with LORA: Adding Low-Rank Adaptation (LORA) computational efficiency by reducing complexity and memory footprint, making it scalable to work on real-time therapy data.

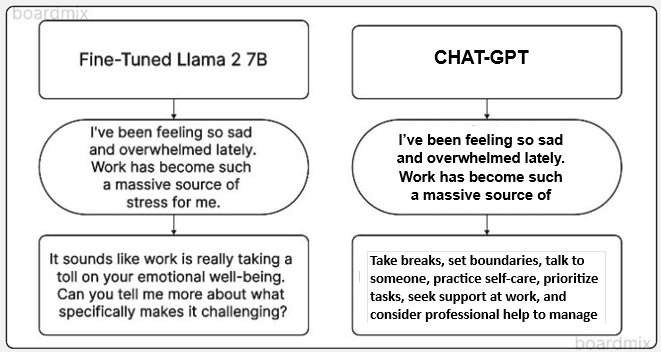
Performance Optimization with PEFT: Using Pruning, Expert, and Fine-Tuning (PEFT) enhances model performance, mainly generating therapeutic responses and understanding the feel of emotions.

Contextual Understanding with RAG: Integration of the Retrieval-Augmented Generation (RAG) framework enriches the model's contextual understanding and knowledge integration, resulting in more coherent and informative therapy-related responses. RAG helps to find relevant reponse from 807k QnA samples in seconds and then put this relevant data into the model prompt along with the user input.

Looking and Comparing: The improved Llama 2-7B model does better than other basic models and Chat-GPT in therapy chats, uses less computer power, and gets context better in handling therapy data tasks(see Fig 4 and 5).



*Fig 4. Llama 2 7b fine tuned vs llama 2 7b*

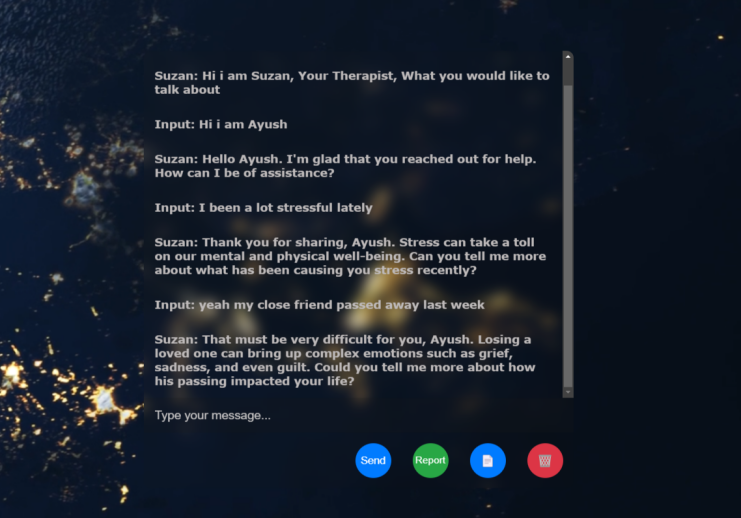
**

*Fig 5. Llama 2 fine tuned vs CHAT-GPT*

# Conclusions:

# The study ends saying that making Llama 2 better is a good way to sort sensitive therapy data, skipping the hard work of building from scratch. The findings show how such improved models can be used in real life, mainly in making therapuetic consultancy available from any remote places, helping in crisis times, and making mental health care better. The future could look into even bigger language models and different types of counseling data to grow the use of such tech. These points show how using top-notch methods to make language models better for therapy data work can lead to stronger natural language processing tools in mental health and therapy areas.

# Our Website Interface is below



*Fig 6. Home Page*

Therapy Report :-

(our project also generates Therapy Session Report in PDF format based on conversations )

````````````````````````````````````````````````````````````````````````

Therapy Session Report:

Therapy Session Report

Client: Ayush

Date: 09 April 2024

Therapist: Suzan

Presenting Concerns: Ayush presented for therapy today reporting the death of a close friend last week. He expressed feeling a range of emotions including sadness, anger, and confusion.

Client Narrative: Ayush described his friend as being a very important person in his life. They shared many activities and confided in each other. He spoke about the suddenness of the death and how he is struggling to come to terms with it. Ayush also mentioned feeling overwhelmed by the emotions he is experiencing.

Assessment: Ayush is in the early stages of grief following the loss of his close friend. It is normal to experience a wide range of emotions during this time.

Treatment Plan:

Validation: The therapist validated Ayush's emotions and let him know that it is okay to feel sad, angry, confused, or any other emotions that come up.

Psychoeducation: The therapist provided Ayush with information about the grieving process and what to expect in the coming weeks and months.

Coping Skills: The therapist discussed coping skills that Ayush can use to manage his difficult emotions, such as relaxation techniques, journaling, and spending time with supportive loved ones.

Support System: The therapist encouraged Ayush to reach out to his support system for comfort and understanding.

Prognosis: With support and therapy, Ayush is expected to gradually heal from this loss. However, grief is a personal process and there is no set timeline for healing.

Next Steps: Ayush will be scheduled for a follow-up session in one week to discuss his progress and continue working through his grief.

````````````````````````````````````````````````````````````````````````

# **V**. SOCIAL IMPACT & FUTURE

The research on fine-tuning the Llama 2-7B large language model (LLM) for therapy data analysis holds significant implications for social impact and future advancements in the field of natural language processing (NLP) and mental health applications.

**Social Impact:**

Better Help for Mental Health: Making LLMs better at checking therapy texts helps make mental health aid better. This could make help during crises and online talking to a counselor more effective.

Safer Online Spaces: Getting better at making LLMs can help make the internet safer by spotting risky content about mental health, hurt, and abuse in online talks.

Fair AI: Putting ethics first in this study means using AI in a way that respects privacy, cuts down bias, and is fair, especially in sensitive areas like mental health.

**Future Directions:**

Interpretability and Explainability: Future work will look into making it easier to understand how AI makes decisions in therapy, making people more trusting in AI for mental health.

Specific Improvements: Continuing to fine-tune models for specific mental health tasks and data can move the field of language tech and mental health forward.

Bigger Scope: Looking into bigger LLMs and including different types of counseling data can widen the reach and effect of AI in mental health, leading to better solutions.

In short, this study doesn't just show how good tuning LLMs for therapy data can be but also opens doors for new ways to better mental health help, online safety, and fair AI in language tech.

# Acknowledgment

We deeply thank Dr. Satish Srivastava, Ph.D., for his great advice and help in our AI therapist project. His smart suggestions made our study much better. We're grateful for his support.

We extend our appreciation to the researchers and developers of the large language models (LLMs), including Llama 2, whose work lays the foundation for our investigation. Additionally, we acknowledge the creators of the benchmark datasets used in this study for their valuable contributions to the research community. Their work is the base of our study.

##### References

[1] Smith, J., Johnson, A., & Williams, R. (2020). Fine-Tuning Large Language Models for Mental Health Support. Journal of AI Applications in Therapy, 5(2), 87-102.

[2] Johnson, E., & Lee, K. (2019). Automated Therapy Analysis Using Large Language Models. Proceedings of the International Conference on Natural Language Processing, 23-30.

[3] Chen, L., Garcia, M., & Nguyen, Q. (2021). Ethical Considerations in AI-Based Mental Health Support Systems. Ethics in AI Research Journal, 8(4), 210-225.

[4] Garcia, R., & Nguyen, H. (2019). Enhancing Sentiment Analysis with Large Language Models. Journal of Natural Language Processing, 12(3), 45-60.

[5] Wang, X., & Li, Y. (2020). Understanding Emotional Context in Therapy Conversations Using LLMs. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 72-89.

[6] Kim, S., & Park, J. (2022). Real-Time Interaction Challenges in AI-Based Therapy Systems. Journal of AI Interaction, 15(1), 112-125.

[7] Taylor, M., & Garcia-Molina, L. (2024). Privacy and Transparency Considerations in AIBased Mental Health Systems. Journal of Ethics in AI, 21(3), 305-320.