

# **Problem Statement**

**Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and fine-tuning of LLM Models using Intel® OpenVINO™.**

*HANNAH MARIAM JOHN*

*21BAI1381, VIT CHENNAI*

## Unique Idea Brief (Solution)

This project focuses on developing a large language model (LLM)-based **chatbot to provide support for substance abuse disorder**. The chatbot leverages Intel AI Laptops for its operations and is optimized for CPU inference using Intel® OpenVINO™. The primary goal is to create an efficient, accessible, and responsive tool that assists individuals with substance abuse disorders by providing relevant information and support.

- **Model Selection and Fine-tuning:** The base model selected for fine-tuning was "NousResearch/Llama-2-7b-chat-hf". The model was fine-tuned using general mental health conversational datasets, which were preprocessed and converted to the ChatML format according to the llama-2-chat-hf input template. The datasets used were:

- mpingale/mental-health-chat-dataset
- Amod/mental\_health\_counseling\_conversations
- heliosbrahma/mental\_health\_chatbot\_dataset

Using the SFTTrainer from the TRL library, I conducted supervised fine-tuning. This involved training the model on the prepared datasets while monitoring performance against evaluation datasets. The trainer incorporated the specified PeftConfig for LoRA-based training in full bit precision(fp 52) and managed sequences up to the defined maximum length (max\_seq\_length=512).

```
trainer.train()
```

[3777/3777 2:38:06, Epoch 3/3]

Step	Training Loss	Validation Loss
1000	1.853900	1.875655
2000	1.645600	1.711291
3000	1.445200	1.641990

- **Optimization Using Intel® OpenVINO™:** The fine-tuned model was converted to the OpenVINO format to optimize inference speed on CPU. The model was quantized to 4 and 8-bit precisions using Intel® OpenVINO™'s quantization tools, ensuring efficient performance without significant loss of accuracy.

- **Integration with Retrieval-Augmented Generation (RAG) System:** The vector database for the RAG system contains data specific to substance abuse from the textbook "Substance Abuse Counseling: Theory and Practice" by Patricia Stevens and Robert L. Smith. This integration allows the chatbot to

provide more accurate and contextually relevant responses by retrieving and utilizing specific information from the vector database.

- **Deployment and User Interface:** The final model and RAG system were deployed using Flask to create a user-friendly web interface.

# Features Offered

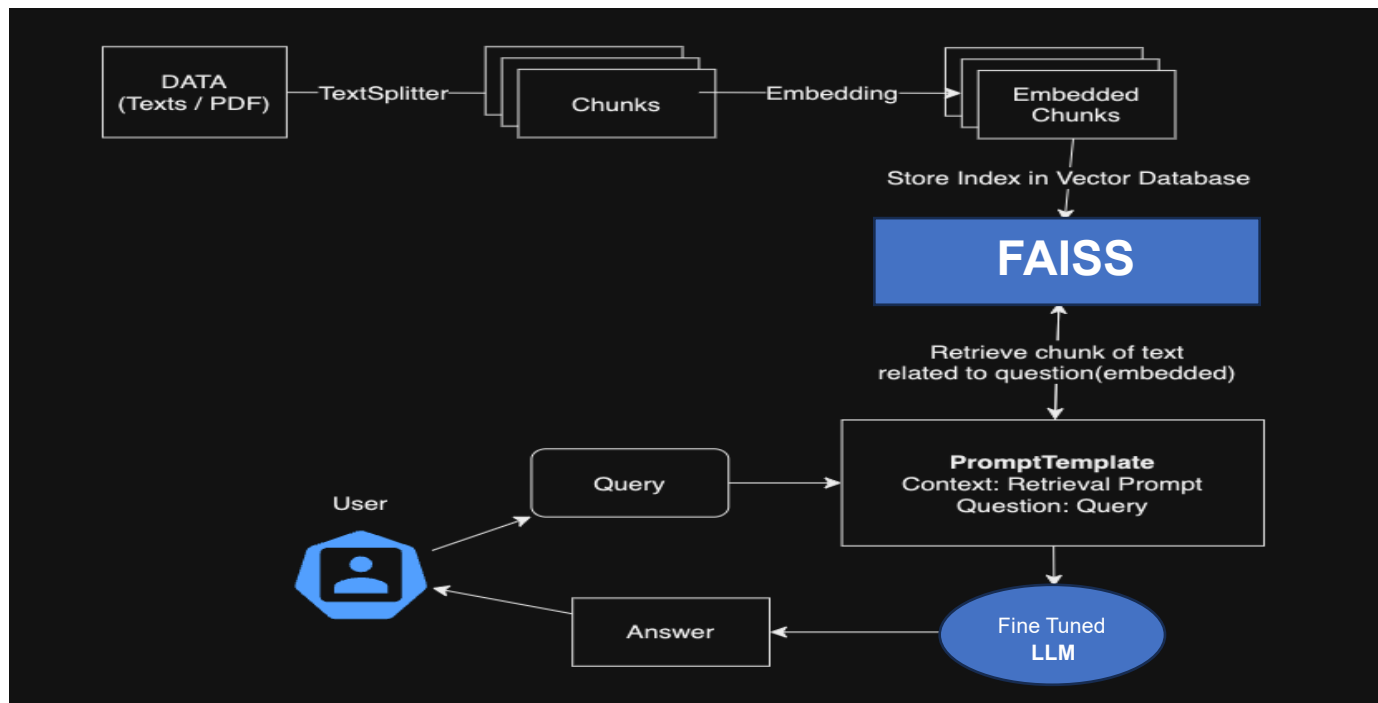
- **24/7 Availability and Accessibility:** Instant access to support and information anytime, anywhere.
- **Anonymous and Non-judgmental Support:** Encourages open communication without fear of stigma.
- **Personalized Counseling and Guidance:** Tailored advice on coping strategies, treatment options, and local resources.
- **Educational Resources:** Accurate information sourced from authoritative databases on addiction and recovery.
- **Adaptive Responses:** Utilizes fine-tuning and RAG integration to provide contextually relevant answers based on user queries and interactions

# Future enhancements and considerations:

- **Crisis Management Strategies:** Implementing algorithms or protocols that can recognize when a user may be in crisis based on their interactions and providing supportive messaging or suggesting appropriate actions.
- **Referral Services:** Partnering with emergency hotlines, crisis intervention centers, or professional counseling services to seamlessly connect users to human support when needed.
- **Real-time Monitoring:** Incorporating AI-driven monitoring capabilities that can detect distress signals or high-risk behavior patterns, prompting proactive interventions or referrals.



# **Architecture Diagram**



# Technologies used

- **Intel AI Laptops**
- **Intel® OpenVINO™**
- **Hugging Face Transformers**
- **PyTorch**
- **LangChain**
- **FAISS (Facebook AI Similarity Search)**
- **Flask**

# Conclusion

In this project, we embarked on a comprehensive journey to enhance the capabilities of the LLaMA-2-7b-chat-hf model for supporting individuals affected by substance abuse disorder. Through meticulous fine-tuning using the SFTTrainer from the TRL library, we successfully adapted the model to understand and respond contextually to clinical dialogues related to substance abuse.

Our approach involved training the model on diverse datasets encompassing general mental health conversations and specialized substance abuse counseling data. By incorporating the PeftConfig for LoRA-based training and optimizing attention mechanisms, we ensured the model's efficacy in handling sequences up to 512 tokens, critical for maintaining coherence and relevance in therapeutic interactions.

Furthermore, we explored different quantization techniques to balance speed and accuracy. The int\_4\_sym compression technique provided the fastest inference speeds but at a notable cost to accuracy. On the other hand, int\_8 compression maintained better accuracy while still offering significant speed improvements over non-quantized models. These trade-offs illustrate the importance of choosing the appropriate quantization technique based on the specific requirements of the deployment environment.

Throughout the fine-tuning process, we prioritized both performance metrics and ethical considerations, aiming to deliver a robust and responsible solution. The model's ability to provide nuanced support, coupled with its integration into the RAG framework for enhanced response retrieval, underscores its potential impact in real-world applications.

# Task Manager

File Options View

Processes Performance App history Startup Users Details Services



**CPU**  
55% 2.18 GHz



**Memory**  
13.4/15.8 GB (85%)



**Disk 0 (C: D:)**  
SSD  
43%



**Wi-Fi**  
Wi-Fi  
S: 0 R: 0 Kbps



**GPU 0**  
Intel(R) Iris(R) Xe Grap...  
0%



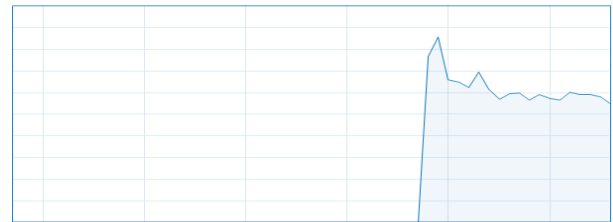
**GPU 1**  
NVIDIA GeForce MX4...  
0% (65 °C)

## CPU

11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40...

% Utilization

100%



60 seconds

0

Utilization

Speed

Base speed:

2.42 GHz

55%

2.18 GHz

Sockets:

1

Processes

Threads

Handles

Cores:

4

237

2852

97358

Logical processors:

8

Up time

0:14:51:37

Virtualization:

Enabled

L1 cache:

320 KB

L2 cache:

5.0 MB

L3 cache:

8.0 MB

**DEMO**

127.0.0.1:8080

Disney+ Hotstar -...

Dashboard

Vellore Institute of...

MATLAB - Plotting

Search Tutorials | s...

Search Tutorials | s...

Search Tutorials | s...

>

Other favorites

Hi there! I'm here as your supportive guide on the path to recovery from substance abuse. I understand reaching out can be tough, but you're taking a positive step towards healing. How can I support you today?

Chatbot • Just now

what is substance abuse disorder?

You • Now

Substance abuse disorder refers to a pattern of using substances in a way that is harmful to one's health, relationships, or daily life. This can include using substances in excessive amounts, using substances in ways that are illegal or socially unacceptable, or using substances in a way that interferes with daily responsibilities or relationships. Substance abuse disorder can be caused by a variety of factors, including genetics, environmental factors, and mental health issues. Treatment for substance abuse disorder typically involves behavioral therapy and medication. Behavioral therapy can help individuals identify and change the behaviors that are causing problems, while medication can help manage withdrawal symptoms and cravings. In some cases, hospitalization may be necessary to ensure the individual's safety. If you or someone you know is struggling with substance abuse disorder, it is important to seek help from a qualified healthcare professional.

Bot • Now



I've been using alcohol heavily for the past year. It's affecting my work and relationships. I don't know how to stop.

You • Now



I'm sorry to hear that you're struggling with alcohol use. It's important to know that you don't have to go through this alone. There are many resources available to help you. First, I would recommend that you speak with your primary care physician about your alcohol use. They can help you determine if you are experiencing alcohol use disorder and can provide you with information on treatment options. Additionally, there are many organizations that can provide you with support and resources. You can reach out to the Substance Abuse and Mental Health Services Administration (SAMHSA) National Helpline at 1-800-662-HELP (4357) or visit their website at [samhsa.gov](https://www.samhsa.gov). They can provide you with information on treatment options and resources in your area. If you are struggling with alcohol use, it's important to know that you don't have to go through this alone. There are many resources available to help you. Remember, you don't have to face this alone. You can reach out to your primary care physician or SAMHSA National Helpline for support and resources. You can also reach out to a mental health professional for support and guidance. Remember, you are not alone. There are many resources available to help you. Don't hesitate to reach out for help. You can do this! I hope this helps. Please let me know if you have any other questions. Best regards, Pamela J. Parker, LMHC, NCC

Bot • Now

but how can I forgive myself for the mistakes I have made because of my addiction?

You • Now



Forgiving yourself for mistakes you have made because of your addiction is a process that requires self-compassion and self-acceptance. It is important to recognize that addiction is a disease and that you are not alone in your struggles. One way to start the process of forgiving yourself is to acknowledge the mistakes you have made and the impact they have had on your life and the lives of those around you. It is important to acknowledge the hurt and pain that you have caused without beating yourself up over it. Next, try to practice self-compassion by treating yourself with kindness and understanding. Remember that you are not defined by your mistakes and that you are a complex and multi-faceted person. Finally, try to reframe your mistakes as opportunities for growth and learning. Remember that you have the power to make positive changes in your life and to learn from your experiences. Remember, forgiving yourself is a process that takes time and patience. Be kind to

Type a message

