EC7203 - ADVANCED ARTIFICIAL INTELLIGENCE

GROUP PROJECT

By:

EG/2020/3986 Jayasinghe D.M.S.N.

EG/2020/3981 Hettiarachchi P.P.P.

EG/2020/4021 Kavindya P.P.

EG/2020/4034 Kumarasiri L.I.N.

Video Link : https://drive.google.com/drive/folders/1ZkWDPNXEl9YL_m1jMFmEzJKS31ti7gJp?usp=sharing

Table of Contents

List of Figures

List of Tables

1. Introduction

Depression is a mental health disorder that is among the most prevalent in the world, but it has remained undetected because of stigma and lack of awareness and access to timely clinical screening. The conventional approaches to assessment, including paper based surveys or face to face clinical consultation, are useful but not always available to patients who might be reluctant to turn to professional assistance. The PHQ-9 (Patient Health Questionnaire) is a clinically valid instrument that is extensively applied to screen patients and assess the severity of depression (Kroenke, n.d.).

The AI Powered Depression Level Analyzer represents a chat-based application that recognizes the level of depression by analyzing the clinically validated PHQ-9 questionnaire. The interaction between the patients and the chatbot resembles a natural conversation, and the counselors can later access the answers and results via a special portal. This project aims to develop an AI supported tool that can make the depression screening process more interactive, accessible, and encouraging, as well as help provide counselors with the correct information on the mental health state of the patients.

1.1 . Problem Statement

Depression often goes undetected due to stigma, lack of awareness, and limited access to mental health services. Although the PHQ-9 questionnaire is a clinical instrument of standard, it is not an interactive tool due to its fixed format. An interactive and AI oriented solution that will simplify the process of depression screening and make it more user-friendly is required.

1.2. Scope

- The interface of the system is in the form of chats and a patient is expected to answer questions of PHQ-9 in the natural dialogue.
- The app makes sure that the nine PHQ-9 questions have been answered by the time the session is done.
- The system is only limited to screening purposes and does not offer medical diagnosis and treatment.

2. System Architecture & Workflow

The architecture of the AI Powered Depression Level Analyzer is inspired by an agentic workflow, with various elements collaborating sequentially and modularly to produce not only a conversational interaction, but also the analysis of depression levels.

The agentic workflow of this project coordinates the conversation, implements the semantic search that involves a knowledge base and calls the RAG pipeline to produce responses rich with context and makes sure that PHQ-9 questions are answered so that the results could be sent to the classifier agent to analyze the level of depression (Anna Gutowska, 2025). It makes this system flexible, modular, and reliable to the extent that it ensures that the user engages.

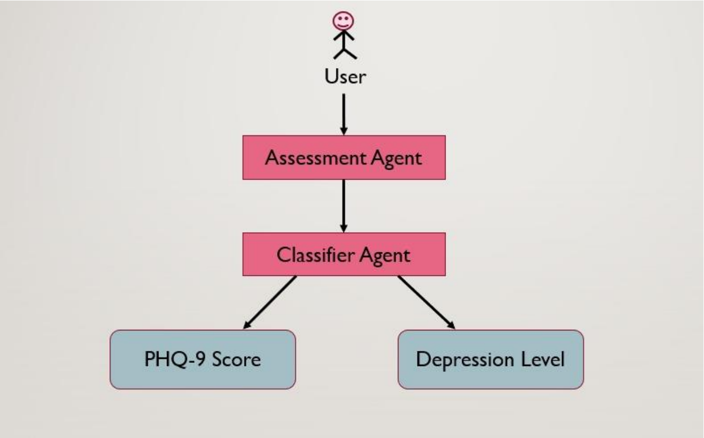


Figure 2.1. - Agent Workflow

The workflow of the system starts with the interaction between the system and the user, during which patients can communicate with the chatbot in a conversational and friendly way. The Agent layer oversees all the inputs and serves as the main control unit, deciding what to do with the input: triggering a semantic search, maintaining normal dialogue, or asking PHQ-9 questions. Where it is required, semantic search is carried out on the knowledge base to retrieve information that is contextually relevant so that responses are not only conversational, but informative.

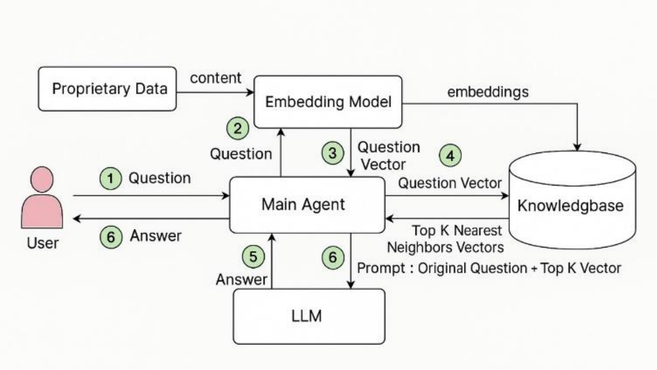


Figure 2.2. - RAG workflow

The figure 2.2. shows that how RAG system works: when a user asks a question, the system will first search a knowledge base to find the most relevant data, extend the original query with the retrieved information, and feed the query plus the retrieved information to the LLM, which will generate an answer that contains context. The GPT-based model produces user-friendly and empathetic responses, so the conversation should be natural, introducing PHQ-9 questions step-by-step. PHQ-9 In the quest to ensure integrity of assessment, the workflow requires completion of all the nine questions of PHQ-9 before the end of the session otherwise reliable and valid data was not collected. After collecting responses, mistral v0.3 (7B) is used to classify levels of depression into Minimal, Mild, Moderate, Moderately Severe or Severe by processing them through the fine-tuned mistral v0.3 (7B) model, which has been trained on PHQ-9 labeled data. That process is done by classifier agent. Then, the results of PHQ-9 and the summary of the session are safely stored in MongoDB and can be reviewed and monitored by the counselors.

2.1. Used Techniques

1.Document Preprocessing and Ingestion.

- PyPDFLoader / DirectoryLoader / TextLoader. - Add unstructured files (text files and PDFs) to the system.
- RecursiveCharacterTextSplitter
    - Breaks long documents into readable small chunks.
    - Relevant to semantic search and effective embedding.

2. Semantic Representation Embeddings.

- OpenAI Embeddings (OpenAIEmbeddings)
    - Text chunks are converted into dense vectors (Edirisinghe, 2024).
    - Embeddings are semantic representations of a text, allowing the use of semantics to do search.
    - These embeddings are then saved in the MongoDB Atlas Vector Search, and can be retrieved.

3. Vector Search (RAG Retrieval Layer)

- MongoDB Atlas Vector Search
    - The source, chunk index, and category are salient embeddings and metadata stored as the acts.
    - Semantic similarity search is supported.
    - Provides the k best chunk of document to use in the RAG pipeline.

4. Retrieval-Augmented Generation (RAG).

- Recovers the chunks of knowledge based on the query. This allows a more accurate, context-specific response.

5. Model Fine Tuning

- Data Preparation- Used a dataset that had PHQ-9 responses and related outputs (Miraz, 2025). The sample was preprocessed as Hugging Face Dataset, with every sample formatted into an appropriate instruction-response format, in a clean and uniform format to be fine-tuned.
- Data Augmentation - Was paraphrased with the Hugging Face model Vamsi/T5_Paraphrase_Paws to augment the dataset size and variety.
- Model Selection & Fine tuning - Mistral-7B v0.3 model with 7B parameters was selected to be fine-tuned. Parameter-Efficient Fine-Tuning (PEFT) which is used with LoRA (Low Rank Adaptation) to update only a small number of additional parameters was configured with a maximum sequence length of 512 tokens to handle the training samples. This methodology made the amount of trainable parameters much smaller, so that resource-efficient training did not compromise model performance.

- Training Setup - The model was trained with small batch sizes because of memory limitations with GPUs and a batch size of 2 and gradient accumulation to stabilize training. Faster training with mixed precision (fp16) was also enabled and gradient checkpointing saved memory. This model was trained by AdamW in 15 epochs with periodic logging and a checkpoint save to check progress.

- Deployment - For deployment, the fine-tuned Mistral model was exposed through a lightweight inference server using Ollama and securely connected to external applications via an ngrok tunnel. This approach allowed the model to be accessed from anywhere without requiring complex infrastructure, making it easier to integrate with front-end clients during early testing. The deployment flow involved running the model locally through Ollama, assigning it a custom endpoint, and then forwarding that endpoint using ngrok to create a public, authenticated URL. While primarily used for experimentation and validation, this deployment method provided valuable insights into latency, stability, and scalability considerations for later production-ready hosting on cloud services.

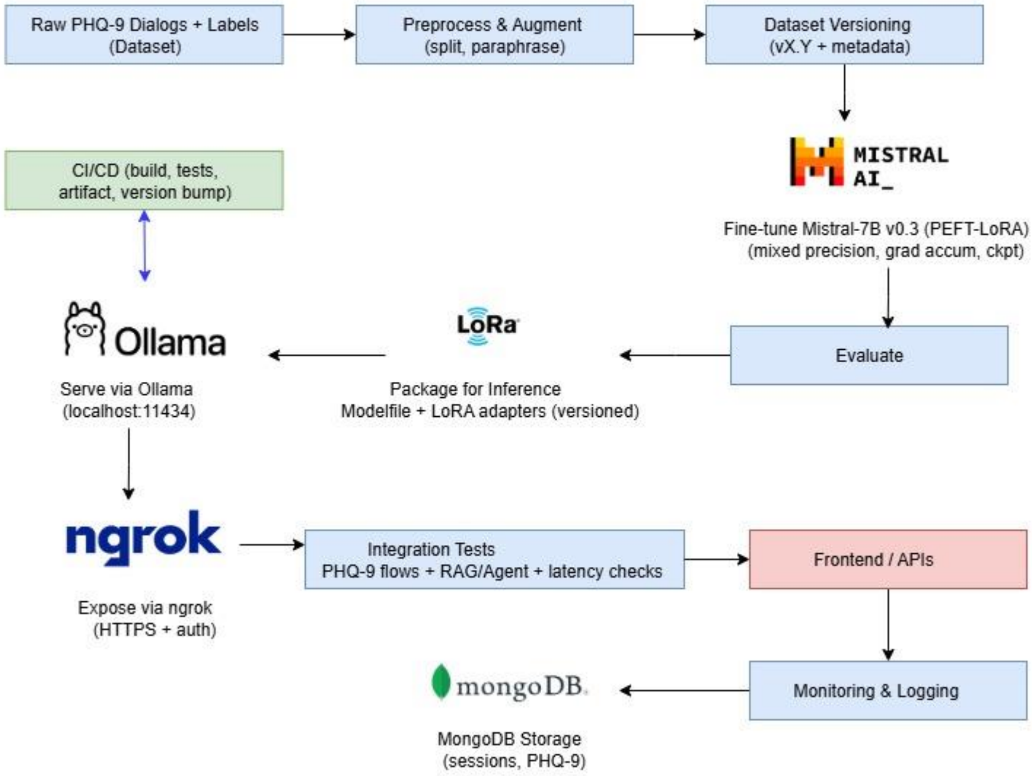

Raw PHQ-9 Dialogs + Labels (Dataset)
Preprocess & Augment (split, paraphrase)
Dataset Versioning (vX.Y + metadata)
MISTRAL AI_
Fine-tune Mistral-7B v0.3 (PEFT-LoRA) (mixed precision, grad accum, ckpt)
CI/CD (build, tests, artifact, version bump)
Ollama
Serve via Ollama (localhost:11434)
LoRa
Package for Inference Modelfile + LoRA adapters (versioned)
Evaluate
ngrok
Expose via ngrok (HTTPS + auth)
Integration Tests PHQ-9 flows + RAG/Agent + latency checks
Frontend / APIs
mongoDB
MongoDB Storage (sessions, PHQ-9)
Monitoring & Logging

Figure 2.3 - MLOps Pipeline

The 2.1. table summarizes the technology stack used in developing the AI-Powered Depression Level Analyzer.

Table 2.1 – Table of technology stack

Layer	**Technology / Tool**	**Purpose**
Frameworks	LangChain, HuggingFace PEFT, PyTorch	Building agentic workflow, fine-tuning models, and model training.
NLP techniques	Word embedding - OpenAI Embeddings	Converting text into semantic vector representations.
LLMs	GPT Models, Mistral v0.3 (7B, fine-tuned)	GPT for conversation; Mistral for PHQ-9 depression classification.
Databases	MongoDB, MongoDB Atlas Vector Search	Storing session logs, PHQ-9 results, and enabling semantic similarity search.
Document Processing	PyPDFLoader, DirectoryLoader, TextLoader, RecursiveCharacterTextSplitter	Loading documents and splitting into manageable chunks for embedding.
Fine-Tuning Methods	Parameter-Efficient Fine-Tuning (LoRA)	Efficiently adapting the Mistral model to PHQ-9 classification.
Deployment / MLOps	MongoDB Atlas, API Keys	Secure data storage, scalable deployment, and monitoring.

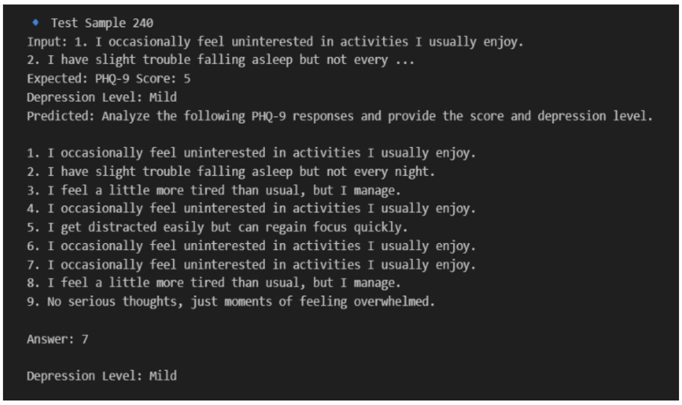
3. Evaluation & Result

Model Evaluation

The model achieved these results, indicating good learning performance and reasonable generalization.

- Training Loss: 0.154900
- Eval Loss: 1.8059
- Perplexity: 6.0854

Figure 3.1 and 3.2 show the testing output and selected real session output respectively. Then Figure 3.3 shows the final output of the application.



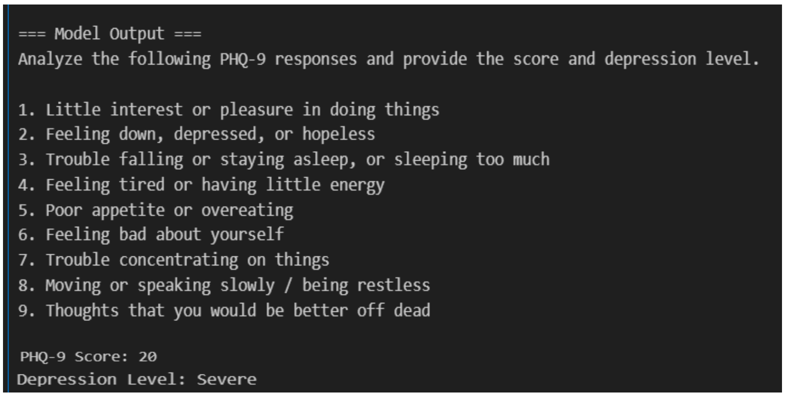
```
🔹 Test Sample 240
Input: 1. I occasionally feel uninterested in activities I usually enjoy.
2. I have slight trouble falling asleep but not every ...
Expected: PHQ-9 Score: 5
Depression Level: Mild
Predicted: Analyze the following PHQ-9 responses and provide the score and depression level.

1. I occasionally feel uninterested in activities I usually enjoy.
2. I have slight trouble falling asleep but not every night.
3. I feel a little more tired than usual, but I manage.
4. I occasionally feel uninterested in activities I usually enjoy.
5. I get distracted easily but can regain focus quickly.
6. I occasionally feel uninterested in activities I usually enjoy.
7. I occasionally feel uninterested in activities I usually enjoy.
8. I feel a little more tired than usual, but I manage.
9. No serious thoughts, just moments of feeling overwhelmed.

Answer: 7

Depression Level: Mild
```

Figure 3.1. – Output for one sample test case

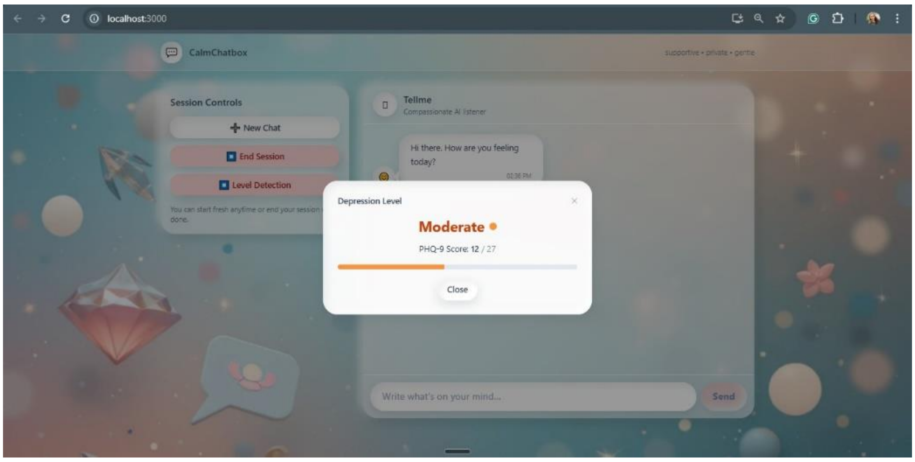


```
=== Model Output ===
Analyze the following PHQ-9 responses and provide the score and depression level.

1. Little interest or pleasure in doing things
2. Feeling down, depressed, or hopeless
3. Trouble falling or staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself
7. Trouble concentrating on things
8. Moving or speaking slowly / being restless
9. Thoughts that you would be better off dead

PHQ-9 Score: 20
Depression Level: Severe
```

Figure 3.2. - Output for real session



localhost:3000
CalmChatbox
supportive • private • gentle
Session Controls
New Chat
End Session
Level Detection
You can start fresh anytime or end your session when done.
Tellme
Compassionate AI listener
Write what's on your mind...
Send
Close

Hi there. How are you feeling today?

02:36 PM

Depression Level

Moderate

PHQ-9 Score: 12 / 27

Figure 3.3. - Final Output on frontend

4. Challenges & Limitation

4.1. Challenges

- Empathic Response Generation: It is necessary to make sure that the GPT model is always supportive and non-judgmental throughout sensitive mental health dialogues.
- PHQ-9 Enforcement: It is necessary to design a workflow that does not allow the user to leave before answering all nine questions of the PHQ-9 and at the same time the interaction must remain natural.
- Complexity of Fine-Tuning: Fine-tuning of the Mistral v0.3 (7B) model with feature-based and parameter-efficient training demanded significant optimization to deal with performance-resource tradeoffs.

4.2. Limitations

- Limited to English Language
- Screening Only: The system can assess depression severity but does not replace clinical diagnosis or treatment.
- The system has not yet undergone large-scale testing or validation in clinical settings.

5. Future Work

- Develop a mobile app version for easier access and user convenience.
- Continuously enrich the knowledge base with updated mental health resources and coping strategies.
- Adapt responses and recommendations based on user history and counselor feedback for more tailored support.

6. Conclusion

The AI-Powered Depression Level Analyzer demonstrates how conversational AI, combined with clinically validated. By integrating agentic workflows, semantic search, and fine-tuned models, the system ensures both accuracy and empathy in user interactions. Future development can focus on publishing for councilors, mobile deployment, personalization, and clinical validation to enhance its effectiveness and impact.

7. References

1.Anna Gutowska, C. S. (2025). *What are agentic workflows?* Retrieved from IBM: https://www.ibm.com/think/topics/agentic-workflows

2.Edirisinghe, S. (2024). *Transformer Model Architecture- Presentation.*

3.Kroenke, D. K. (n.d.). *PHQ-9 (Patient Health Questionnaire-9)*. Retrieved from MD+CALC: https://www.mdcalc.com/calc/1725/phq9-patient-health-questionnaire9

4.Miraz, M. A. (2025, February). *PHQ-9 Student Depression Dataset.* Retrieved from Mendelay Data: https://data.mendeley.com/datasets/kkzjk253cy/1