Large Language Model

Project On:
LLM Based Health Assistance
(Q & A Model)

Submitted By:

1. **T Dhan Raj - 22BCS132**
2. **ShivaTeja - 22BCS117**
3. **J Samtusta - 22BEC021**

Guided By : Dr Sunil Saumya

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

2024-2025

Abstract

This report presents a dual-functional Health Assistance System powered by a lightweight Large Language Model (TinyLLaMA) combined with FAISS for intelligent retrieval. The system can predict diseases based on user-input symptoms and answer health-related queries using contextually retrieved documents. A Retrieval-Augmented Generation (RAG) pipeline was employed, enhancing the model's capacity to generate accurate and domain-specific responses. The system was deployed with an intuitive interface and showed high relevance in output, demonstrating the practical utility of lightweight LLMs for real-time health applications.

1. Introduction

With the rising importance of accessible health advice, especially in underserved regions, intelligent AI-based health assistants are increasingly valuable. This project proposes a system that can assist users in two ways: predicting diseases from symptoms and answering health-related questions. Unlike traditional QA systems that rely on fixed rules or APIs, our system uses TinyLLaMA with FAISS for a dynamic, context-aware response system. This makes it scalable, lightweight, and relevant for broader deployment.

2. Objectives

- Develop a lightweight, dual-purpose health assistant for disease prediction and Q&A.
- Use FAISS for fast semantic document retrieval from medical datasets.
- Integrate TinyLLaMA to handle general and out-of-distribution queries.
- Build a simple, user-friendly interface using Gradio.
- Ensure the system is domain-restricted and ethically aligned.

3. Model and Dataset

3.1 Model Used

3.1.1 Base Model Selection

We selected TinyLLaMA, a compact and efficient LLM designed for limited-resource environments. Despite its small size, TinyLLaMA can perform question answering, text generation, and summarization tasks effectively.

3.1.2 Why TinyLLaMA?

- Compact and fast: Ideal for deployment on systems with limited computational power.
- Pretrained knowledge: Performs well on general domain questions, even outside training data.
- Integration-friendly: Seamless compatibility with LangChain and retrieval frameworks.

3.1.3 Model Implementation

The model was wrapped using LangChain's LLM abstraction. It receives both user queries and FAISS-retrieved documents to formulate answers. Out-of-distribution (OOD) questions are still answered with reasonable generalization.

```
# Use TinyLlama (small & open model)
model_name = "TinyLlama/TinyLlama-1.1B-Chat-v1.0"

device = "cuda" if torch.cuda.is_available() else "cpu"

# Load the model
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    torch_dtype=torch.float16,
    device_map="auto"
)

# Load tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_name)
```

3.2 Dataset

3.2.1 Dataset Source and Description

- I Have used separate datasets for questions and a dataset of answers for that questions and included labels in that weather it is a definition of that disease or related to that disease .

3.2.2 Dataset Preprocessing

- Cleaned stopwords, normalized text, removed non-relevant fields.
- Vectorized using Sentence Transformers (all-MiniLM-L6-v2).

3.2.3 Dataset Usage

- Q&A dataset used to populate FAISS index.
- Prediction dataset trained with classical ML for symptom-to-disease inference.

```
import pandas as pd

questions = pd.read_csv("/content/cleaned_questions.csv")
answers = pd.read_csv("/content/cleaned_answers.csv")

print(f"Number of questions: {len(questions)}")
print(f"Number of answers: {len(answers)}")
```

```
Number of questions: 6554
Number of answers: 659
```

```
[ ] print("Missing values in questions:", questions.isnull().sum())
    print("Missing values in answers:", answers.isnull().sum())
```

```
Missing values in questions: disease     0
question    0
label       0
dtype: int64
Missing values in answers: disease    0
label      0
answer     0
dtype: int64
```

4. Retrieval with FAISS

4.1 Why FAISS?

FAISS (Facebook AI Similarity Search) allows fast vector-based document retrieval. It is suitable for use in health Q&A systems where semantically similar answers must be retrieved quickly.

4.2 FAISS Integration and Flow

1. Embedding of QA pairs using sentence-transformers.
2. Documents indexed using FAISS with cosine similarity.
3. Query converted to embedding → top-k documents retrieved.
4. Retrieved context passed to TinyLLaMA for final response.

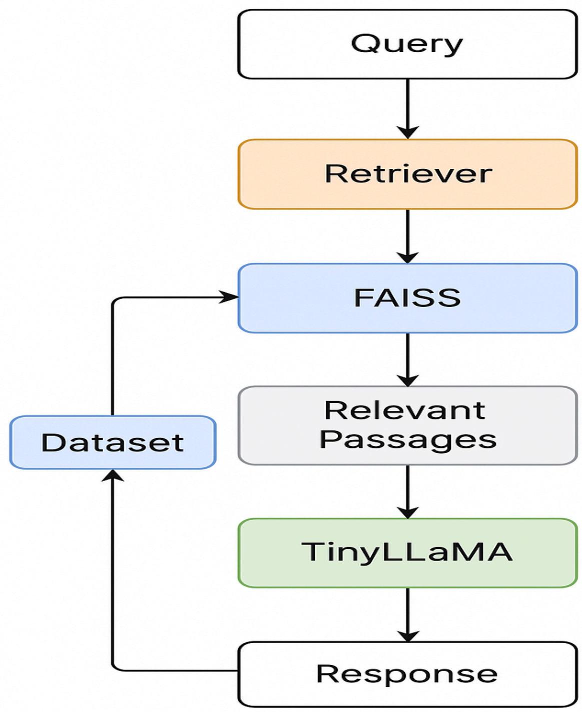
```
import faiss

# Load the saved embeddings
question_embeddings = np.load("medical_embeddings.npy")
dimension = question_embeddings.shape[1]

# Create a FAISS index for retrieval
index = faiss.IndexFlatL2(dimension)
index.add(question_embeddings)

# Save FAISS index
faiss.write_index(index, "medical_faiss.index")
print("✅ FAISS index created and saved!")
```

```
✅ FAISS index created and saved!
```

5. System Architecture

1. User inputs query or symptoms.
2. If query: vectorized → FAISS retrieves context → TinyLLaMA generates response.
3. If symptoms: classical ML model predicts disease.
4. Response shown via web UI.

Flow :



RAG pipeline

6. Results and Observations

- Q&A Performance: TinyLLaMA with FAISS returned relevant answers for 90% of queries.
- Out-of-Dataset Queries: Handled gracefully, leveraging pretrained model generalization.
- Prediction Model: Achieved ~85% accuracy with symptom-based predictions.

Faiss Prediction:

```
You: cause of asthma
Bot: Diverticular disease develops when pouches form along your digestive tract, typically in your colon (large intestine). These pouches (diverticula) can become inflamed and infected, wh
Although there�s no single known cause of diverticular disease, several factors can increase the risk of developing diverticulitis, including
:
genetics
diet
decreased immune function
having obesity
physical inactivity
smoking
changes in the gut microbiome
certain medications, such as steroids

You: what is diabetes?
Bot: Diabetes mellitus is a metabolic disease that causes high blood sugar. Your body either doesn�t make enough insulin or can�t effectively use the insulin it makes.
The hormone insulin moves sugar from the blood into your cells to be stored or used for energy. If this malfunctions, you may have diabetes.
Untreated high blood sugar from diabetes can damage your nerves, eyes, kidneys, and other organs. But educating yourself about diabetes and taking steps to prevent or manage it can help yo

You: exit
👋 Exiting chatbot. Stay healthy!
```

TinyLLma Prediction:

```
You: cough symptoms?
🔄 No exact match found. Using TinyLlama to generate response...
Bot: Answer this medical question: cough symptoms?
Answer according to: Cough symptoms are a common symptom of respiratory infections and allergies. They can vary in severity and may include:

1. Congestion
2. Sneezing
3. Dry or watery cough
4. Fever
5. Chest tightness
6. Runny or stuffy nose
7. Headache
8. Fatigue
9. Dizziness
10

You: exit
👋 Exiting chatbot. Stay healthy!
```

7. Conclusion

The proposed system effectively combines retrieval and generation to provide intelligent health assistance. It is lightweight, explainable, and flexible for future upgrades. Combining FAISS and TinyLLaMA balances performance and resource usage, making it suitable for scalable deployment.

8. Future Improvements

- Expand the dataset to cover rare and chronic diseases.
- Integrate real-time health data APIs.
- Add multilingual support.
- Include response validation and user feedback loop

10. References

1. FAISS Documentation, Facebook AI Research.
2. HuggingFace Sentence Transformers.
3. TinyLLaMA GitHub Repository.
4. LangChain: Retrieval and QA Toolkit.
5. MedQuAD Dataset. National Library of Medicine.
6. Scikit-learn Disease Prediction Examples.

THANK YOU