ArogyaX AI : Medical-Chatbot

1. Problem Formulation

With the increasing pace of digital healthcare platforms, people are increasingly using online platforms for healthcare advice. However, most of the online platforms available today are providing generic, incomplete, or sometimes misleading information. In the healthcare sector, misleading information can result in severe consequences.

Conventional chatbots lack in-depth contextual understanding, while isolated Large Language Models (LLMs) tend to produce hallucinated answers when not backed by authentic medical knowledge. Moreover, most of the current models are restricted to text-based conversation and do not incorporate voice or image analysis.

There is a pressing need for a context-aware and multimodal AI-based healthcare assistant that can provide authentic and grounded answers.

2. Problem

The main research problem being tackled in this research is:

> How can we develop a trustworthy and multimodal AI medical assistant that supports context-aware and grounded medical responses using the Retrieval-Augmented Generation technique for text, voice, and image inputs?

3. Objectives

To address the identified challenges, the objectives of this work are:

1. To design a Retrieval-Augmented Generation (RAG) pipeline for medical question answering.
2. To implement semantic vector search for efficient knowledge retrieval.
3. To integrate speech-to-text and text-to-speech modules for voice interaction.
4. To incorporate image understanding using multimodal AI models.
5. To evaluate the system's reliability and contextual relevance.

4. Methodology

1. System Design

- The system design is modular and comprises the following:
- Data preprocessing module
- Vector embedding and storage module
- Retrieval-Augmented Generation module
- Multimodal processing module
- Deployment and infrastructure module
- Each module is designed to work independently while ensuring seamless integration.

2. Data Collection and Preprocessing

- The medical documents are obtained from structured sources.
- The steps involved are:
- Text preprocessing and formatting
- Document chunking into smaller pieces
- Text chunking to embeddings using an embedding model
- Chunking helps improve the precision of document retrieval and contextual relevance.

3. Vector Embedding and Storage

- The text chunks are represented as numerical vectors using a pre-trained embedding model.
- The vectors are stored in Pinecone, a vector database that enables efficient similarity searches.
- This enables the system to:
- Search for semantically similar documents

- Overcome keyword-based constraints

- Enhance contextual understanding

4. Retrieval-Augmented Generation (RAG)
    - The RAG pipeline is designed as follows:
    - The user query is embedded.
    - The top-k most similar vectors are searched using Pinecone.
    - The retrieved documents are aggregated as contextual input.
    - The context is fed to the LLM.
    - The LLM produces a grounded response.
    - This design enables the system to generate responses based on the retrieved medical knowledge rather than relying on the model's memory.
5. Multimodal Integration

    Speech Processing

    - Speech-to-Text translates voice inputs into text queries.
    - Text-to-Speech translates responses into voice outputs.

    Vision Processing

    - Image inputs are processed by a multimodal AI model.
    - Image interpretations are integrated with text explanations.
    - Final response is generated by LLM with contextual understanding.

This allows interaction with the system beyond text-based interfaces.