LLM Health Assistant

Yuhang Jiang, Ebuka Nwafornso
University of Trento

Abstract
The LLM Health Assistant is a health consultation platform based on a large language model (LLM), leveraging generative AI and retrieval-augmented generation (RAG) technologies to provide users with personalized and intelligent health Q&A services. The system integrates multiple functional modules, including text interaction, voice interaction, PubMed paper retrieval, user information management, and conversation storage.

1 Technology Overview

This project prioritizes technologies that offer high response speed, stability, and cost-effectiveness.

1.1 Backend Framework

The backend uses **FastAPI** with an **SQLite** database for user management and **Pinecone** for conversation storage. FastAPI ensures efficient asynchronous processing, strong type safety, and seamless **JWT-based authentication** via **RESTful APIs**.

1.2 Frontend Technology

The frontend is built with **HTML, CSS, and JavaScript**, integrating **Fetch API** for asynchronous data exchange. This approach enhances customization, UI flexibility, and user experience.

1.3 Large Language Model (LLM)

The system employs **GLM-4-Plus** for text-based processing and **GLM-4-Voice** for real-time speech interaction. These models ensure high accuracy, long-context understanding, and adaptable responses.

1.4 Database Selection

SQLite stores user data, while **Pinecone** enables efficient semantic search for personalized responses. Data access is managed securely via **FastAPI**.

1.5 Security Authentication

The system implements **OAuth2.0 + JWT** for user authentication and **bcrypt** encryption for password protection. Tokens have expiration control to mitigate security risks.

1.6 External Data Augmentation (PubMed API)

The **PubMed API** retrieves up-to-date medical literature, enhancing response accuracy with real-world scientific data.

2 System Architecture

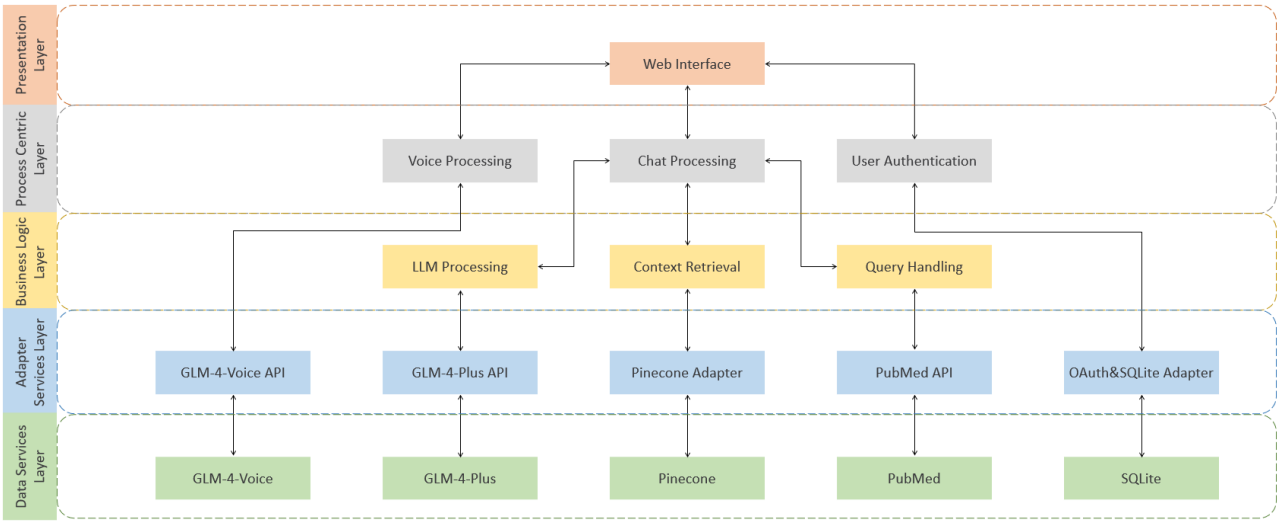

Figure 1: System Architecture

This system adopts a **Layered Architecture** design, which consists of five layers: **Presentation Layer**, **Process Centric Layer**, **Business Logic Layer**, **Adapter Services Layer**, and **Data Services Layer**. This architecture ensures modular decoupling, scalability, and efficiency.

2.1 Presentation Layer

Web Interface (User Interface):

- Serves as the interaction interface for users, supporting text input and voice interaction.
- Sends user requests to **Chat Processing** and **Voice Processing** and receives the final processed results.

2.2 Process Centric Layer

Voice Processing:

- Processes voice input and calls the **GLM-4-Voice API** for speech recognition and synthesis.
- Voice data is not stored and is only used for real-time interaction.

Chat Processing (Text Chat Processing):

- Handles user text input and interacts with the **Business Logic Layer** to obtain responses.
- Manages the conversation flow and works with **Context Retrieval** to provide more intelligent replies.

User Authentication:

- Ensures user authentication and API access security.

2.3 Business Logic Layer

LLM Processing:

- Calls the **GLM-4-Plus API** to generate appropriate text responses based on user input.

Context Retrieval:

- Queries user chat history to provide context-aware intelligent responses.
- Connects to the **Pinecone Adapter** for vectorized storage and retrieval of historical conversation data.

Query Handling:

- Parses user queries and calls the **PubMed API** for medical literature retrieval based on query type.

2.4 Adapter Services Layer

GLM-4-Voice API:

- Processes voice input and returns both text and speech output.
- Connects to **GLM-4-Voice** to ensure high-quality voice synthesis.

GLM-4-Plus API:

- Handles AI-powered text conversations by calling **GLM-4-Plus** for natural language understanding and response generation.

Pinecone Adapter (Vector Database Adapter):

- Connects to the **Pinecone** vector database for storing and retrieving user conversation history.

PubMed API:

- Retrieves medical literature and provides professional health consultation based on the **PubMed** database.

OAuth & SQLite Adapter (Authentication and Database Adapter):

- Manages user authentication using **OAuth**.
- Interacts with the **SQLite** database to store and manage user information.

2.5 Data Services Layer

GLM-4-Voice:

- A cloud-based large model providing speech recognition and synthesis capabilities.

GLM-4-Plus:

- Mainly used for text processing, supporting intelligent Q&A and conversation generation.

Pinecone:

- Stores user chat history and provides efficient context retrieval functions.

PubMed:

- A medical literature database that provides health information based on the latest research.

SQLite:

- Stores user basic information and authentication data.

---

3 APIs Used

The LLM Health Assistant integrates several APIs to provide intelligent health consultations and voice interaction services. Below are the external APIs used in the system:

- **GLM-4-Plus API**: Processes user health queries and generates intelligent responses.
- **GLM-4-Voice API**: Supports voice-based interaction by converting speech to text and generating spoken responses.
- **Pinecone**: Stores and retrieves past conversation history for context-aware interactions.
- **PubMed API**: Retrieves up-to-date medical literature for evidence-based responses.