PROJECT

AI HEALTH ASSISTANT

STUDENT NAMES

Tang Trung Son - *s3978330*
Alex Pham - *s3818206*
Tran Duc Duy - *s3978690*
Nguyen Truong Duong Thinh - *s3914412*
Bui Hoang Quynh Chi - *s3978316*

SUPERVISOR

Nguyen Hieu Thao
Nguyen Quang Minh

COMPANY

FPT

RMIT UNIVERSITY
School of Science, Engineering & Technology

SCAN FOR MORE INFORMATION

PROJECT OVERVIEW

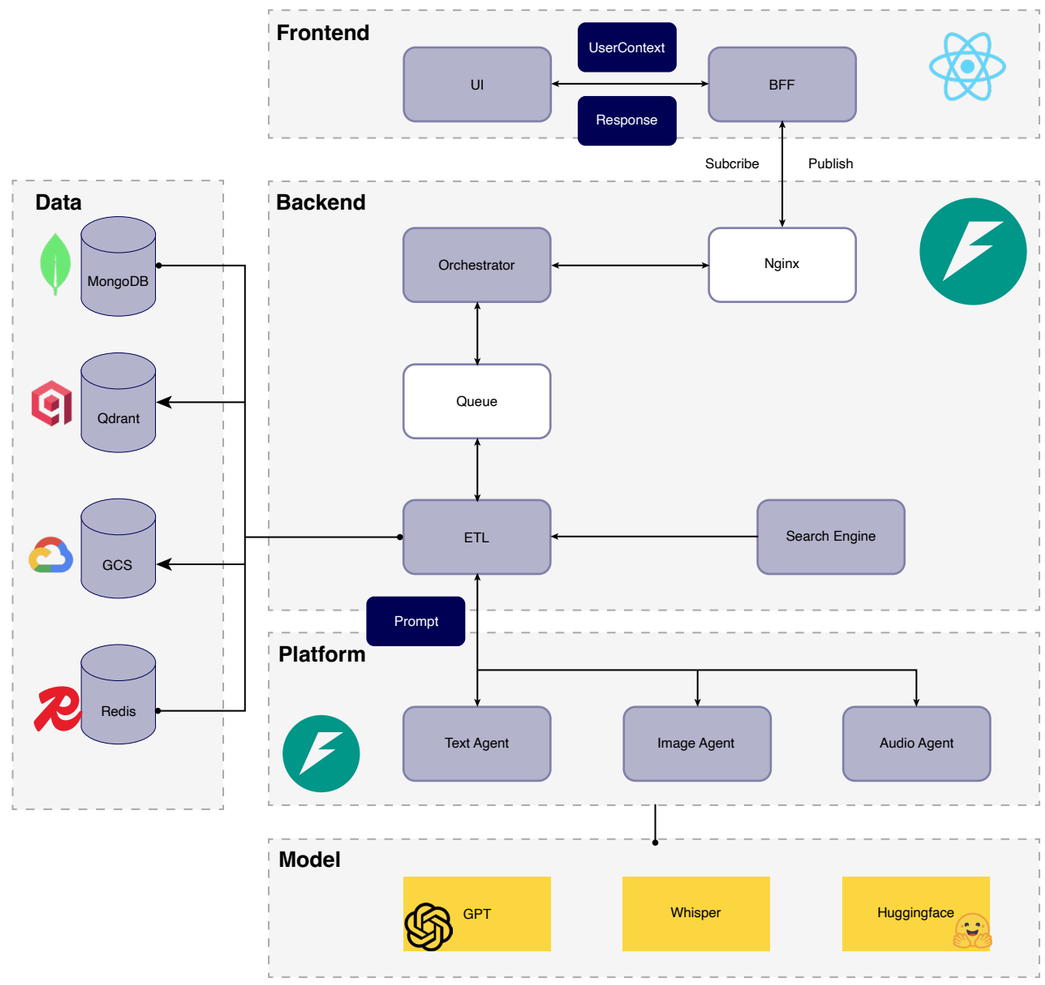
AI Health Assistant is a multimodal system designed to support nurses and healthcare professionals by providing accurate, context-aware assistance across text, images, and audio. The assistant leverages advanced language models, medical search engines, and retrieval-augmented generation (RAG) to deliver reliable responses, summarize patient data, transcribe conversations, and analyze medical documents or images. By combining cutting-edge AI models with healthcare-specific knowledge and compliance mechanisms, the system aims to reduce workload, improve efficiency, and ensure safer clinical decision-making.

INFRASTRUCTURE

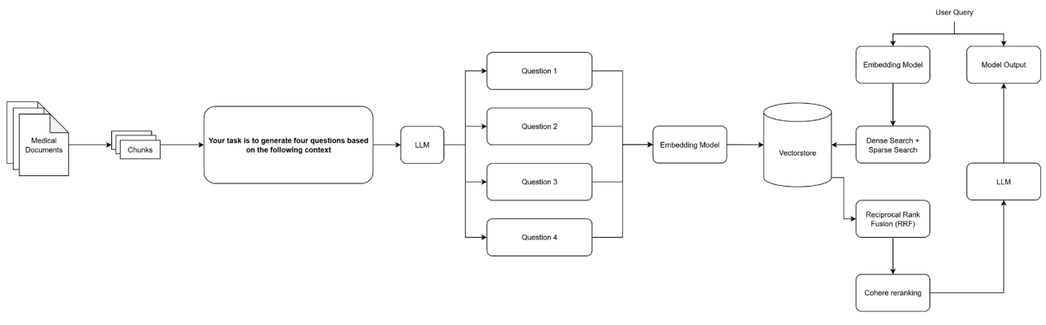
Our AI Health Assistant architecture is fully hosted on Google Cloud Platform (GCP) and integrates multiple layers of data storage, backend orchestration, and AI-powered agents to deliver healthcare support.

- **Frontend (UI & BFF)**: Built with React, this layer provides the user interface and manages session context, responses, and communication with the backend via a Backend-for-Frontend (BFF) service.
- **Backend (Orchestrator, Queue, ETL, Search Engine)**
  - Nginx handles routing and request publishing/subscribing.
  - Orchestrator coordinates tasks between services.
  - Queue manages asynchronous workloads.
  - ETL pipelines process input data and integrate it with downstream agents and the Search Engine for knowledge retrieval.
- **Data Layer (MongoDB, Qdrant, GCS, Redis)**
  - MongoDB stores structured health records.
  - Qdrant provides vector storage for embeddings and semantic search.
  - Google Cloud Storage (GCS) manages large-scale unstructured data like documents and images.
  - Redis supports caching and session management for fast performance.
- **Platform Agents (Text, Image, Audio):** Specialized services handle domain-specific tasks
  - Text Agent for medical Q&A, summarization, and clinical dialogue.
  - Image Agent for analyzing medical images.
  - Audio Agent for transcription and voice-based interactions.
- **Models (GPT, Whisper, Hugging Face):** Foundation AI models provide core intelligence
  - GPT powers natural language understanding and reasoning.
  - Whisper handles medical transcription and speech-to-text.
  - Hugging Face models extend with specialized open-source healthcare NLP capabilities.

By combining these components on GCP, the architecture ensures scalability, reliability, and secure integration of multimodal healthcare data and AI services.

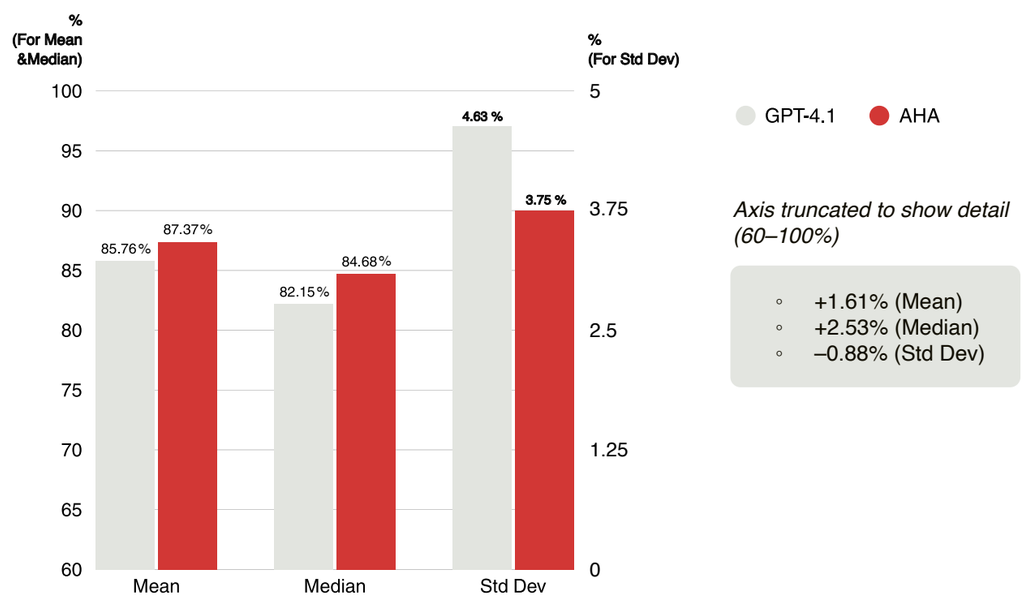


EXPLANATION OF THE WORKFLOW



This pipeline showcases how medical documents are transformed into a question–answer retrieval system. First, documents are chunked and an LLM generates four representative questions per chunk. These questions are embedded and stored in a vector database along with their answers. When a user query comes in, it is embedded and compared against the stored questions (rather than raw context) using cosine similarity. Hybrid dense + sparse search is applied, followed by Reciprocal Rank Fusion (RRF) and an additional reranking step with Cohere to ensure the most accurate matches. Finally, the LLM uses the top results to generate the model output.

CHART 1: ANSWER RELEVANCY BY MODEL

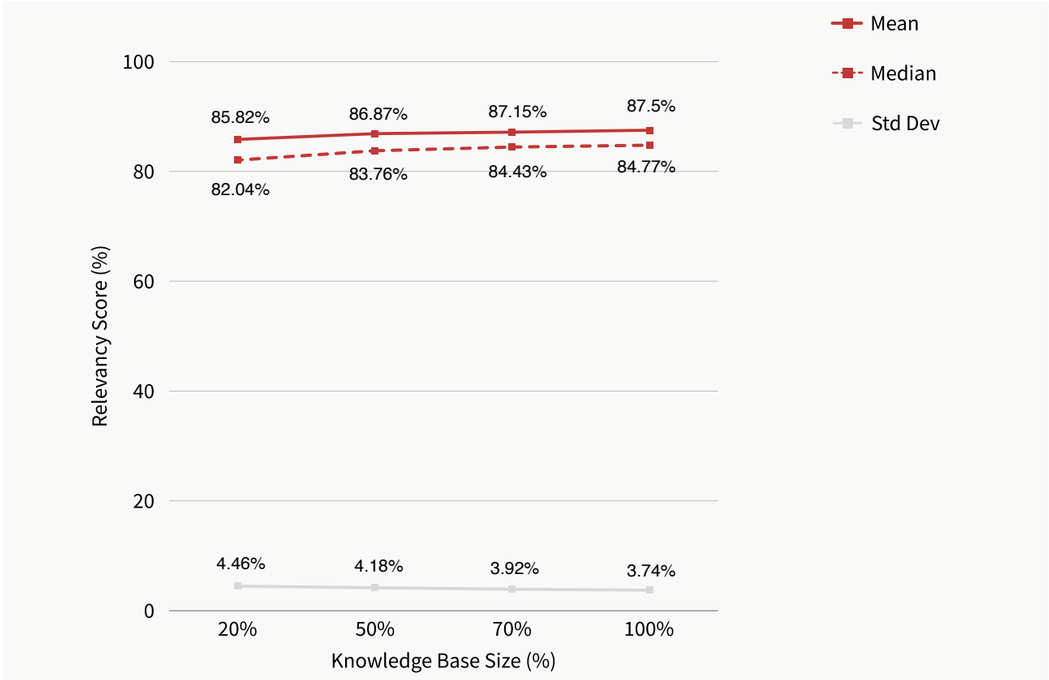


Axis truncated to show detail (60–100%)

- +1.61% (Mean)
- +2.53% (Median)
- –0.88% (Std Dev)

AHA achieves higher mean and median relevancy scores than GPT-4.1, while maintaining lower variability. This shows AHA provides more consistent and reliable answers in health conversations.

CHART 2: ANSWER RELEVANCY BY KNOWLEDGE BASE SIZE



As the knowledge base grows, both mean and median relevancy steadily improve, while variability decreases. This demonstrates that enlarging the knowledge base consistently boosts answer quality and reliability.