Development of LLM based Chatbot

Frameworks & platforms to be used

- Django – Backend (Auth, dashboard, metrics, chat session history, etc.)
- OpenAI – Fine tuned LLM
- Langchain – chatbot (add context, RAG, prompt templates, etc.)
- AWS EC2 – deployment (other providers can be considered)
- Nginx – https (required for login with google)
- Android/iOS – PWA for time being

Fine Tuning

- OpenAI allows fine tuning
- Need to try different adjustments in batch size, epoch, learning rates, etc.
- The LLM can be fine tuned on Red flag detection & patient education (points from project proposal pdf). Few shot examples?
- Can take a long time to complete depending on dataset size

Passing Context

- Context regarding patient history be passed using prompt templates
- Instructions on Mental health assessment using PHQ-9, etc. can be included here.
- Generating summary in the dashboard

Latency

- There's not much we can do here
- There will be many calls to the llm. Paid plans might have lower **latancy** and **rate limiting**. (have to check)
- Apart from LLM calls, database can become bottleneck (later after production) - may need to host separate database server.

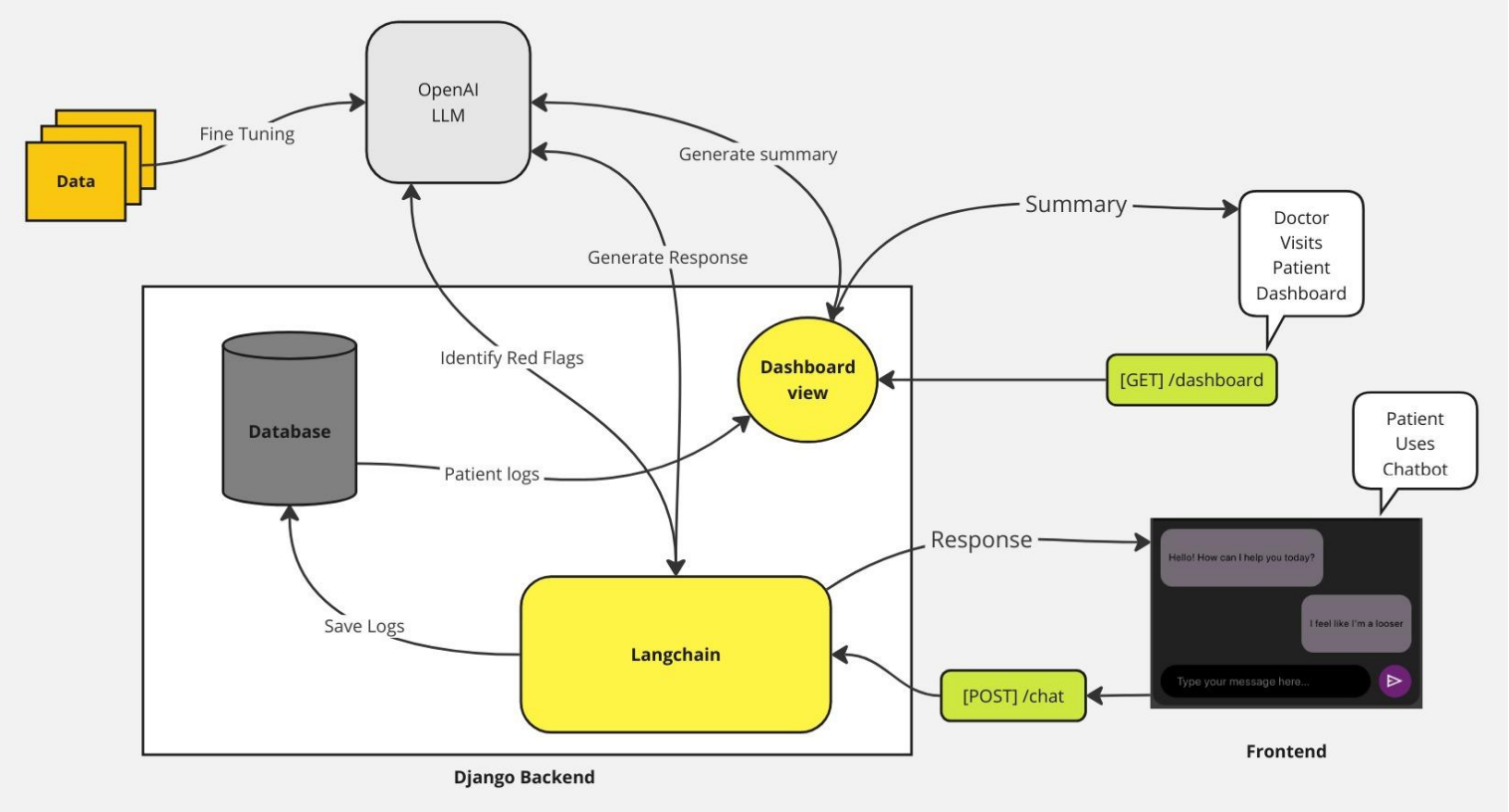
RAG vs Agents

- RAG info taken from similarity match of patient input with RAG context data.
- The patient is likely never going to use any medical jargon in the prompt. Thus RAG is unlikely to pick up any useful context.
- Agents can dynamically decide whether to fetch a factual description of any term using any ***tools*** it has.

Doubts

- Specifics of Red flag detection and patient diagnosis. Be given by the dataset for fine tuning.
- Mental health assessment how frequently? On every chat message of patient?
- Voice to text quality problems. Is the fine tuning data (patient interactions) speech to text generated? (quality questionable)

Project Architecture



Thank You!