07 · GRADUATION PROJECT · 2025

Medicaa

A medical question-answering chatbot grounded in a trusted reference — the Gale Encyclopedia of Medicine — instead of general web data.

ROLE AI engineer — data, RAG pipeline, model selection, evaluation and web app

STACK LLaMA models · MiniLM embeddings · Pinecone · LangChain · Python · Flask

84% accuracy, larger LLaMA model

72% accuracy, small LLaMA model

12 s small-model response

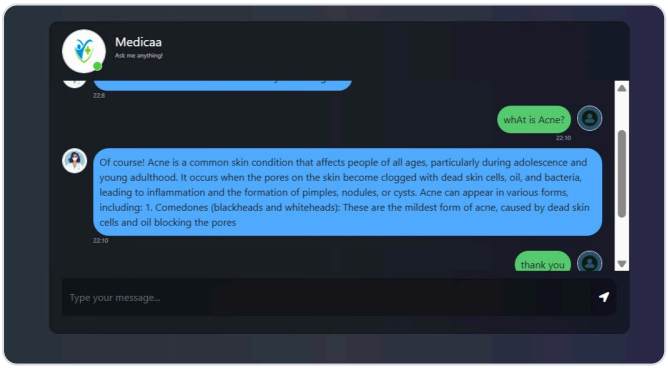
1–1.5 min larger model on CPU

THE PROBLEM

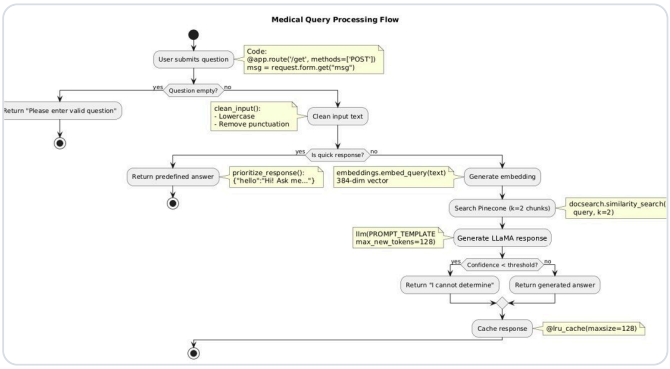
People search the web for medical answers and land on unverified content. Medicaa answers only from a trusted medical reference and runs the language model locally so questions stay private.

WHAT I BUILT

- **RAG pipeline.** Text cleaning, MiniLM embeddings in Pinecone and LLaMA-model generation grounded in retrieved passages, behind a Flask chat app with caching.
- **Model selection.** Compared a larger and a smaller LLaMA-family model on the same questions: the larger model gave complete, accurate answers (84%) but slower; the smaller one was 5–7× faster but cut answers off (72%).
- **Evaluation with domain experts.** 4th- and 5th-year medical students asked specific questions and reviewed answers; robustness tested with invalid queries.



Chat answer grounded in the reference



Query processing flow