Instruction-Tuned Healthcare Chatbot with RAG Architecture

Motivation

Large language models (LLMs) have significantly improved the capabilities of conversational agents. However, their generic pretraining often results in hallucinations, particularly in sensitive areas like healthcare. This project combines instruction tuning and retrieval-augmented generation (RAG) to ground chatbot responses in reliable, external medical knowledge.

Approach

- **Instruction Tuning:** A supervised fine-tuning objective is applied:

$$\mathcal{L}_{\text{instr}} = -\sum_{i=1}^{N} \log P(y_i|x_i;\theta)$$

  where x_i is the instruction prompt, y_i is the expected output, and θ are model parameters.

- **Retrieval-Augmented Generation (RAG):** Combines dense retrieval with generative modeling:

$$P(y|q) = \sum_{d\in\mathcal{D}} P(y|q,d)\cdot P(d|q)$$

  where q is the query, $\mathcal{D}$ is the document set, and $P(d|q)$ is computed using vector similarity.

Architecture

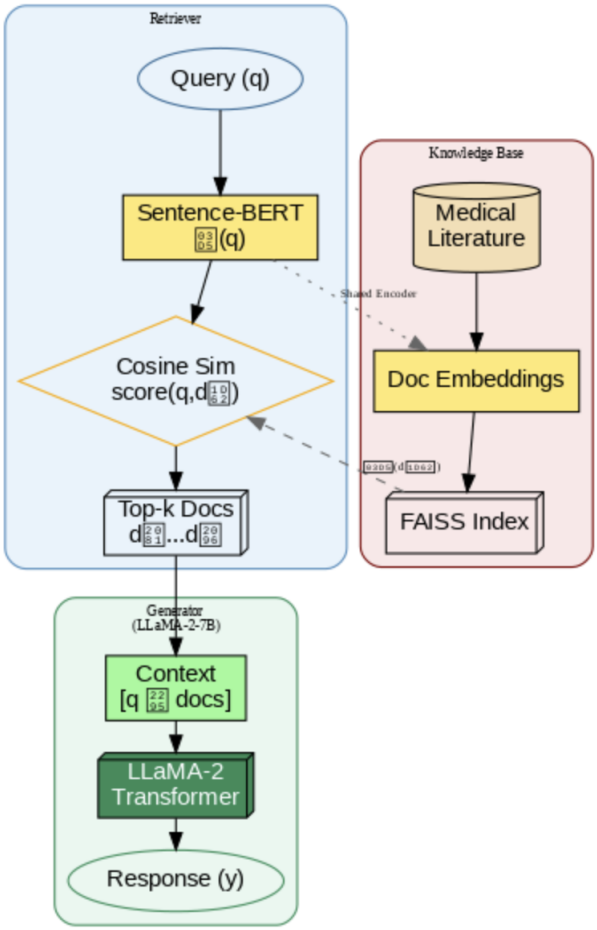
- **Model:** Transformer-based LLaMA 2–7B
- **Retriever:** FAISS with dense vectors from SentenceTransformers
- **Corpus:** Curated medical documents from trusted clinical literature
- **Pipeline Steps:**
  1. Embed query q and compute similarity scores:

$$\text{score}(q,d_i) = \cos(\phi(q),\phi(d_i))$$

  2. Retrieve top-k documents based on similarity.
  3. Concatenate query and context:

$$x = \text{concat}(q,d_1,\ldots,d_k)$$

4. Generate response $y \sim P(y|x;\theta)$



Results

- 25% increase in factual consistency over baseline LLaMA, measured using ROUGE-L and BERTScore
- Achieved sub-second average inference time using optimized batch inference
- Significantly reduced hallucination rate in clinical QA tasks

Future Work

- Integrate Reinforcement Learning with Human Feedback (RLHF) for continual improvement
- Add multilingual support with cross-lingual retrieval embeddings
- Leverage structured ontologies like SNOMED and UMLS for symbolic augmentation
- Experiment with hybrid sparse-dense retrieval to enhance document recall