Jahnavi Pakanati-9013742

Team-6

Sustainable-AI

What Is RAG: - RAG stands for Retrieval-Augmented Generation. It's an AI approach that makes large language models (LLMs) smarter, more accurate, and more reliable by letting them look things up before answering.

What It does:

RAG retrieves relevant information from a knowledge base and includes it in the prompt given to the LLM.

How is it relevant with this project?

Here we were initially doing a search on large document base but then later we dropped the idea of RAG and we are not using it.

How RAG works:

1. **Retrieval Phase:**
   Function used : search_relevant_context(query, max_docs=5)
   Process:
   1. Takes the user's medical query.
   2. Converts it into an embedding using the same MedEmbed model used for documents.
   3. Searches the persistent ChromaDB collection for the most similar stored chunk embeddings.
   4. Returns the top matching chunks along with their metadata (file name, page, etc.).
   5. If ChromaDB isn't available, uses a medical keyword-based fallback search.

2. **Augmentation Phase**
   The returned chunks are not directly used in this code to prompt an LLM, but in a complete RAG system, these chunks would be inserted into the LLM's prompt alongside the question.

   Summary of how it works here
   - This class handles the "R" (Retrieve) part of RAG completely.
   - The "A" (Augment) step is partially done — you get structured results ready to insert into a prompt.
   - The "G" (Generate) step would be done by another component that calls an LLM with the retrieved text