Retrieval-Augmented Generation (RAG) is an architecture that combines the benefits of retrieval-based and generative models.

Instead of generating responses from scratch, RAG retrieves relevant documents from a corpus and then uses a language model to generate answers.

RAG Architecture:

1. Retriever: It fetches relevant documents or passages from a knowledge base using embedding similarity.
2. Generator: A language model (like LLaMA or GPT) conditions on the query and retrieved documents to generate an accurate response.

Applications:

- Question Answering
- Chatbots with domain knowledge
- Document summarization
- Personalized assistants

Advantages:

- Up-to-date responses using dynamic content
- Reduced hallucination by grounding in real data
- Scalable to large knowledge bases

This document is created on July 27, 2025. Use it to test your RAG-based chatbot project.