Retrieval-Augmented Generation (RAG) & LangChain

What is RAG? Retrieval-Augmented Generation (RAG) combines large language models (LLMs) with external knowledge sources. Instead of relying only on the model's internal parameters, RAG retrieves relevant documents from a vector database and injects them into the prompt, improving accuracy and grounding responses in real data.

Core Components of RAG **Document Ingestion:** Loading data (PDFs, text, web pages). **Chunking:** Splitting documents into meaningful pieces for efficient retrieval. **Embedding:** Converting chunks into numerical vectors. **Vector Store:** A database that enables similarity search (FAISS, Chroma, Pinecone). **Retriever:** Finds the most relevant chunks for a user query. **LLM:** Generates output using retrieved context.

What is LangChain? LangChain is a framework designed to simplify the development of LLM■powered applications. It provides modules and abstractions for chaining prompts, LLMs, retrievers, tools, agents, and workflows.

Why Use LangChain for RAG? Easy integration of vector stores Built■in retrievers and query processors Support for tool■using agents Memory and state management for long conversations Production■ready workflows with LangGraph

Typical RAG Pipeline in LangChain Load documents with **DocumentLoaders**. Split using **TextSplitters**. Embed with **Embeddings Models** (OpenAI, HuggingFace, etc.). Store vectors in **Chroma/FAISS/Pinecone**. Create a **Retriever**. Build a **RAG Chain** — Retrieval + LLM. Query and get grounded answers.

RAG Use Cases Chatbots with custom knowledge Enterprise search Automated documentation assistants Legal, medical, and financial knowledge apps Research copilots

LangChain + LangGraph LangGraph enables creation of reliable, stateful, multi■agent workflows. When combined with RAG, it supports: Conditional routing Critic/revision loops Multi■agent collaboration Traceable and debuggable pipelines

Conclusion RAG enhances LLM accuracy by grounding responses in real data, while LangChain provides the tools to build structured, scalable, and production■ready RAG applications.