Retrieval-Augmented Generation (RAG) – Hands-On Practical Document

This document is designed to be used as a primary knowledge source for a Retrieval-Augmented Generation (RAG) system. You will load this PDF into your vector database, generate embeddings, and query it using LangChain.

1. What is RAG?

Retrieval-Augmented Generation (RAG) is an architecture that enhances Large Language Models by allowing them to retrieve relevant information from external documents before generating an answer. Instead of relying only on parametric memory, RAG grounds responses in real documents.

2. Core Components of a RAG System

1. Document Loader: Loads files such as PDFs, text files, or web pages.
2. Text Splitter: Breaks documents into smaller chunks for better retrieval.
3. Embedding Model: Converts text chunks into numerical vectors.
4. Vector Store: Stores embeddings and enables similarity search.
5. Retriever: Fetches the most relevant chunks based on a query.
6. LLM: Uses retrieved context to generate grounded answers.

3. Why RAG is Important

RAG reduces hallucinations, enables access to private or domain-specific data, and allows models to stay up to date without retraining.

4. Typical RAG Workflow

1. Load documents into memory.
2. Split documents into chunks.
3. Generate embeddings for each chunk.
4. Store embeddings in a vector database.
5. Retrieve relevant chunks for a user query.
6. Pass retrieved context to the LLM.

5. Example Questions for RAG Testing

1. What is Retrieval-Augmented Generation?
2. List the core components of a RAG system.
3. Why is RAG preferred over pure LLM generation?
4. Explain the role of embeddings in RAG.

6. Practical Notes

When using this document in LangChain, ensure proper chunk size and overlap to maximize retrieval quality. Smaller chunks improve precision, while overlap preserves context.

7. Final Reminder

This PDF is intentionally structured with clear sections to help your retriever return accurate context. Use it to test, debug, and understand how RAG systems behave during real-world usage.