RAG System Test Document

Section 1: Introduction to Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a framework that combines information retrieval with natural language generation. Instead of relying solely on a pre-trained model's internal knowledge, a RAG system retrieves relevant documents from an external knowledge base and uses them to generate more accurate and context-aware responses.

Section 2: Key Components of a RAG System

1 1. Document Loader – Loads documents from PDF, TXT, HTML, or other sources.

2 2. Text Splitter – Breaks documents into smaller chunks for embedding.

3 3. Embedding Model – Converts text chunks into vector representations.

4 4. Vector Store – Stores embeddings for efficient similarity search.

5 5. Retriever – Finds the most relevant chunks based on user queries.

6 6. Generator (LLM) – Produces the final answer using retrieved context.

Section 3: Example Knowledge Base Content

Python is a high-level programming language known for its simplicity and readability. It is widely used in web development, data science, artificial intelligence, and automation. The capital of Bangladesh is Dhaka. The capital of France is Paris. LangChain is a framework designed to simplify building applications powered by large language models.

Section 4: Sample Questions for Testing

1 • What is Retrieval-Augmented Generation?

2 • What is the role of a vector store in RAG?

3 • What is the capital of Bangladesh?

4 • What is LangChain used for?

5 • Name two common use cases of Python.