1. What is RAG?

RAG is an AI architecture that combines information retrieval with language generation.

Instead of relying only on what the model was trained on (its parametric knowledge), RAG pulls in external knowledge sources (databases, PDFs, APIs, search engines, vector stores, etc.) at runtime.

This helps reduce hallucinations, keeps answers up-to-date, and allows LLMs to work in domain-specific contexts (finance, healthcare, law, etc.) without retraining.

Think of it as giving an LLM "open book access" instead of asking it to rely solely on memory.

2. How RAG Works

(Pipeline)

RAG typically follows two stages:

a) Retrieval

A retriever searches external knowledge bases for relevant chunks of information.

Usually done using:

Vector databases (FAISS, Pinecone, Weaviate, Milvus, Chroma)

Embeddings (e.g., OpenAI, Sentence Transformers, Hugging Face models)

Search engines (Elasticsearch, BM25, Firecrawl, etc.)

Retrieval step gives you the top-k most relevant documents/chunks related to the query.

b)

Generation

The LLM (generator) takes the query + retrieved context and produces a final answer.

The retrieved chunks are stuffed into the prompt context window (via prompt engineering) or used in more advanced architectures (like LangChain's stuffing, map-reduce, refine chains).

4. Why RAG is Important

Up-to-date answers (e.g., news, research papers, company data).

Domain adaptation without retraining (plug in your own docs).

Efficiency: Cheaper than training/fine-tuning massive models.

Hallucination reduction: Model grounds its

answers in real documents.

Explainability: Retrieved docs can be shown as sources.

5. Example Workflow

User asks:

"What are the side effects of drug X?"

Retriever queries a medical knowledge base or PubMed papers.

Top 5 relevant chunks about "drug X side effects" are fetched.

Generator (LLM) combines query + context:

Question: What are the side effects of drug X?
Context:
- Document 1: Drug X may cause nausea and

dizziness.
- Document 2: Some patients report headaches and insomnia.

→ LLM outputs:
"The side effects of drug X include nausea, dizziness, headaches, and insomnia."

Final answer can cite sources.

6. RAG in Practice

Frameworks: LangChain, LlamaIndex, Haystack.

Databases: FAISS, Pinecone, Weariate, Milrus, Chroma.

Models: Any LLM (GPT, LLaMA, Mistral, Claude,

etc.).

Applications:

Chatbots over PDFs/Docs

Customer support AI

Enterprise knowledge assistants

Legal/medical assistants

Personalized study tools