⇨ What is RAG?

RAG = Retrieval + Augmentation + Generation

It is a technique that retrieves relevant external information and uses it to generate accurate, grounded answers through a language model.

=> Why is RAG Used?

- Solves limited knowledge of LLMs
- Reduces hallucinations
- Improves traceability and explainability
- Powers domain-specific Q&A systems

=> 6 Important Stages of a RAG System:

1. Ingestion – Load documents (PDFs, Word, etc.)

2. Chunking – Split long documents into smaller parts

3. Embedding – Convert text into vectors using an embedding model

4. Indexing – Store vectors in a vector database like FAISS

5. User Query – Convert the user query to an embedding

6. Retrieval + Generation – Retrieve similar chunks, generate final response

=> Explanation of Each Stage:

1. Ingestion: Load input documents into the system.

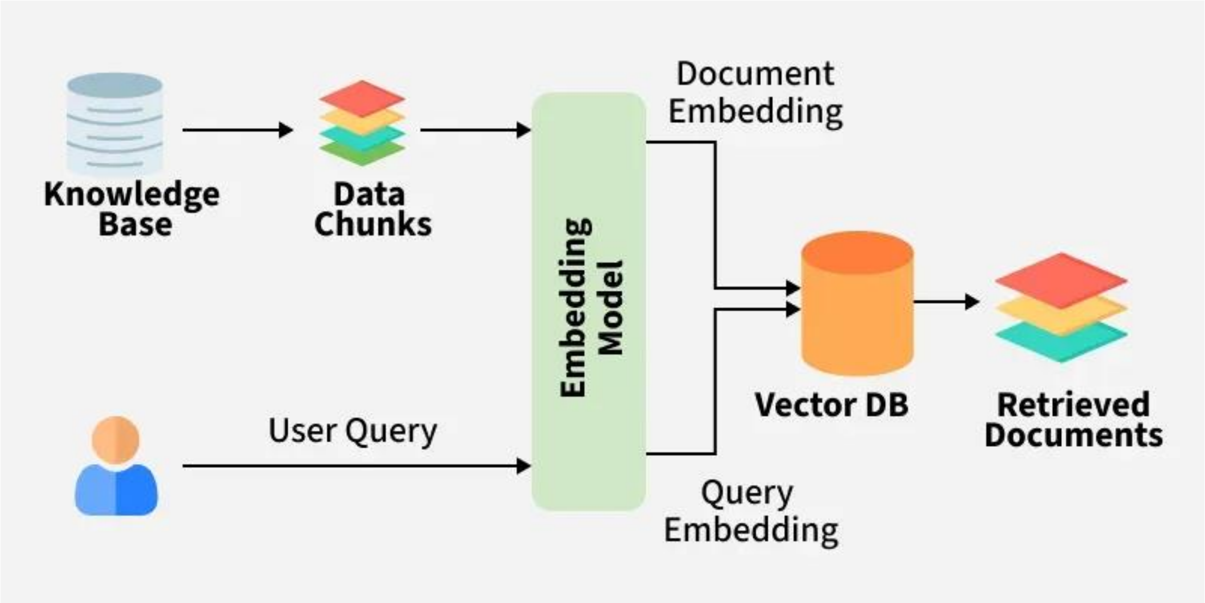
2. Chunking: Break text into smaller units like paragraphs or blocks.

3. Embedding: Represent each chunk numerically using models like MiniLM.

4. Indexing: Store embeddings in a fast-searchable vector database.

5. Query Processing: Convert user query to a vector.

6. Retrieval & Generation: Match query vector to chunks, generate an answer.


Knowledge Base
Data Chunks
Embedding Model
Document Embedding
User Query
Query Embedding
Vector DB
Retrieved Documents

Flowchart of RAG Stages:

```
[Documents]

    |

[Ingestion]

    |

[Chunking]

    |

[Embedding Model]

    |

[Vector DB (Indexing)]

    |

[User Query]

    |

[Query Embedding]

    |

[Retrieve Top-K Chunks]

    |

[Language Model]

    |

[Answer Generation]

    |

[Final Response]
```

Importance of RAG in GenAI:

- Ensures factual accuracy
- Allows real-time access to new knowledge
- Easy domain adaptation
- Reduces model retraining needs
- Enables enterprise-level AI applications

5 Real-World Applications of RAG:

1. Enterprise Chatbots
2. Legal Document Analysis
3. Healthcare Support Tools
4. Academic Research Assistants
5. Intelligent E-commerce Search