Section 1: Core Concepts of RAG

Retrieval: Fetching relevant chunks from a knowledge base.

Augmentation: Adding retrieved text to the model prompt.

Generation: LLM produces answers based on augmented context.

Knowledge Base: A store of documents used for retrieval.

Context Window: The amount of text an LLM can consider at once.

Section 2: Embedding & Vector Search

Embedding: Turning text into numerical vectors.

Vector Store: Database storing embeddings for quick search.

Similarity Search: Finding related vectors based on distance.

FAISS: A fast library for vector similarity search.

Cosine Similarity: A measure of how close two vectors are.

Section 3: Chunking & Preprocessing

Chunking: Splitting long text into smaller meaningful units.

Overlap: Extra repeated text to preserve context between chunks.

Metadata: Extra info about chunks like page or section.

OCR: Extracting text from scanned documents.

Tokenization: Breaking text into units an LLM processes.

Section 4: RAG Pipelines & Evaluation

Pipeline: A sequence linking retrieval and generation.

Latency: Time taken to retrieve and answer.

Grounding: Ensuring model answers come from retrieved text.

Hallucination: When an LLM produces incorrect info.

Evaluation: Checking relevance and accuracy of answers.