Indexes:

Indexes

Indexes in LangChain organize and structure your documents for efficient retrieval.They help LLMs quickly find the most relevant chunks of information to answer queries accurately.

01

Indexes

Definition

An **Index** in LangChain refers to a structured **data representation** (usually of unstructured documents) that enables **efficient retrieval** of relevant information using **embeddings**

In simple terms, indexes help your LLM **remember** where to look when answering a query from a large set of documents.

02

Why Indexes Matter

- ❑ Allow **fast and relevant document retrieval**
- ❑ Power the **Retrieval-Augmented Generation (RAG)** pattern
- ❑ Help LLMs work with **external knowledge** beyond their training data
- ❑ Enable **semantic search** instead of simple keyword search

Components of an Index

Component	Description
Text Splitter	Breaks documents into smaller, manageable chunks
Embedding Model	Converts text into high-dimensional vector representations
Vector Store	Stores and indexes embeddings for similarity search
Retriever	Fetches top relevant documents based on user queries and similarity matching

Workflow of Indexing

Raw Documents

↓

Text Splitter (e.g., RecursiveCharacterTextSplitter)

↓

Embedding Model (e.g., OpenAI, HuggingFace)

↓

Vector Store (e.g., FAISS, Chroma, Pinecone)

↓

Retriever (used in RAG chain)