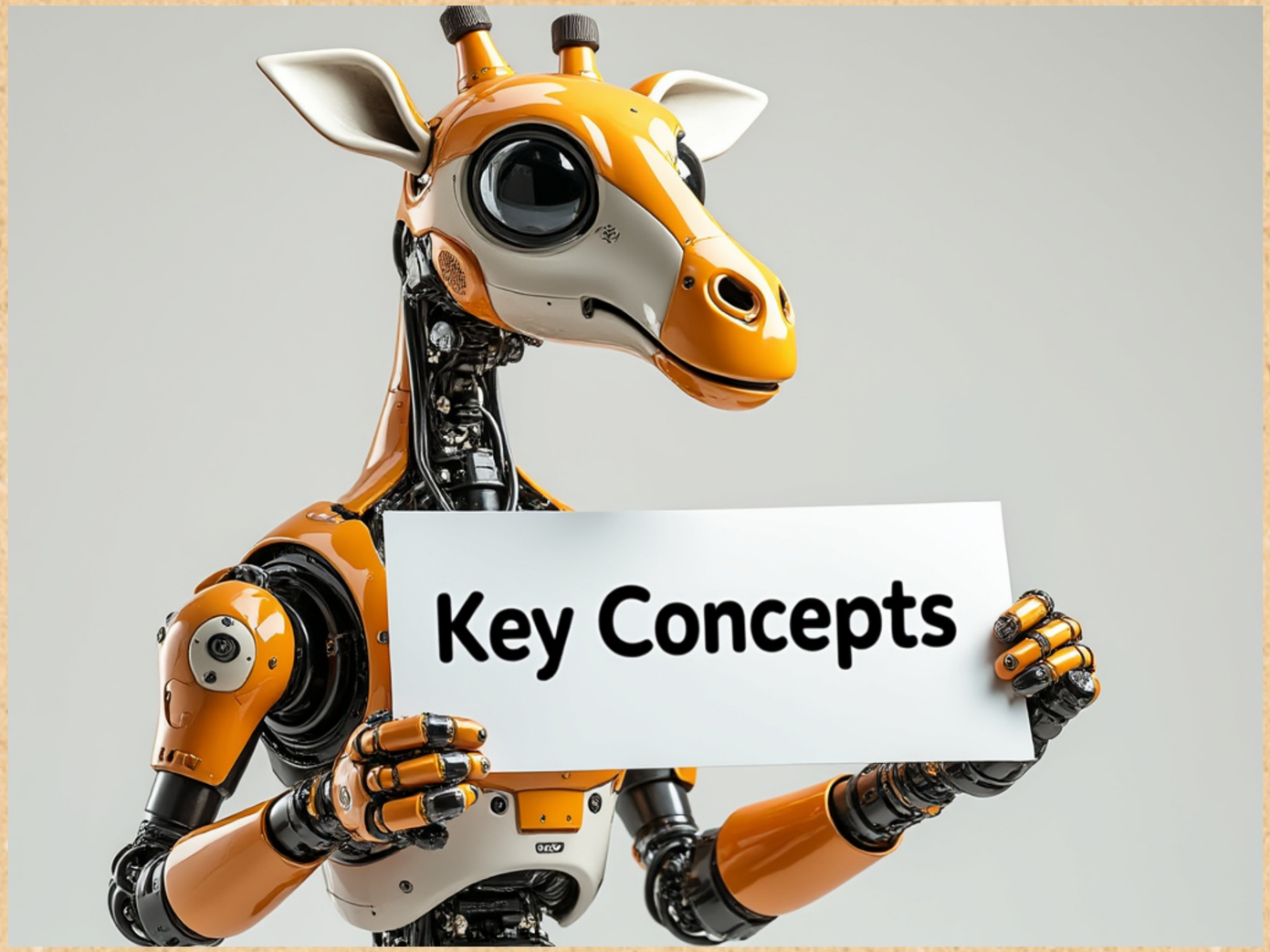
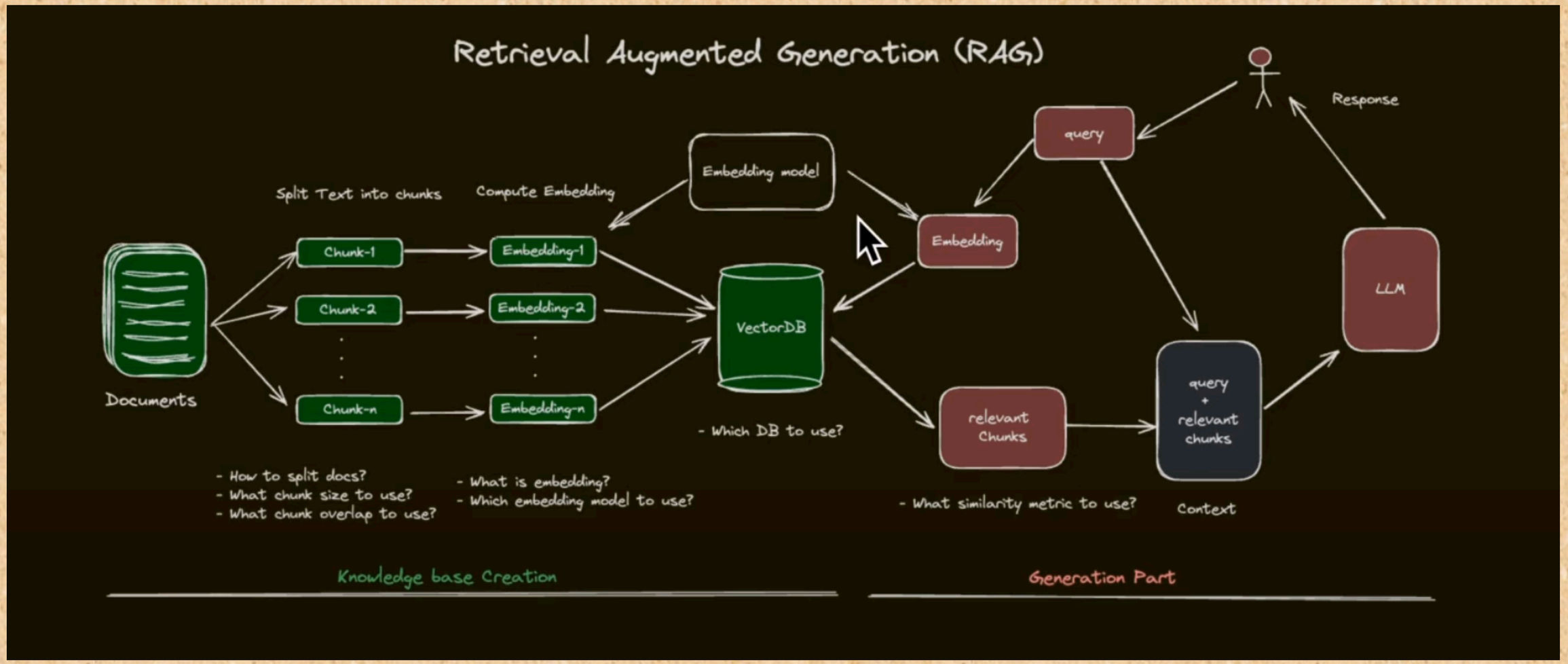
Module 2 - Topic 3

2.3.3 Embeddings and Vector DBs


Key Concepts

How RAG works - Part I



Credits: Prompt Engineering YT

Let’s understand some key concepts!

- Embeddings
- Vector Database

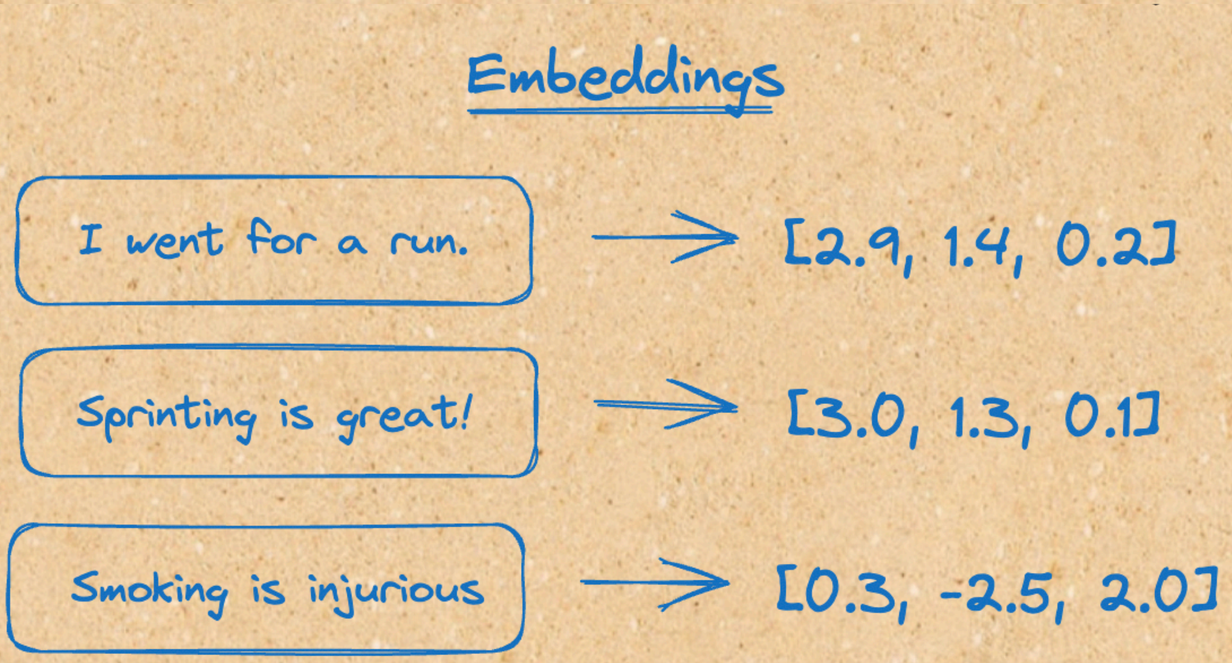
Embeddings

Definition: A technique that converts words, sentences, or other data into numerical vectors, capturing semantic meaning in a machine-readable format.

- Transforms discrete data into continuous vector spaces
- Enables machines to understand and process human language
- Forms the foundation for many NLP tasks and AI applications

How GPT understands English?

Converts data into language understandable by machines



What is happening to the PDF?

Converting PDF to Embeddings

What are Vector DBs?

Definition: A specialized database designed to store, manage, and query high-dimensional vector data efficiently.

- Optimized for similarity search operations on vector embeddings
- Supports fast nearest neighbor searches in large datasets
- Crucial for scaling AI applications like recommendation systems and semantic search

Vector DB vs Traditional DB

	Vector DB	Traditional DB
Data Type	Optimized for high-dimensional vectors	Primarily for structured data (tables, rows, columns)
Query Type	Similarity search (nearest neighbors)	Exact match or range queries
Indexing	ANN algorithms (e.g., HNSW, IVF)	B-trees, hash indexes
Scalability	Designed for high-dimensional data at scale	May struggle with high-dimensional data
Use Case	AI applications, semantic search	CRUD operations, relational data management

Some Popular Vector DBs

- Pinecone
- Chroma
- Qdrant
- FAISS
- Milvus

... many more