Campus → Company policy → docs   ] → 300 pages

↓
LLM [x]

Train [x]

Database

↓ Pass

LLM → [Language Ans.]

Langu
Info

LLM + Database → O/P → LLM → Represent

RAG

user → Query

Employment Benefit

It's an approach in AI/LLMs where instead of only relying on the model's pre-trained knowledge, the system **retrieves relevant external information (documents, facts, databases, APIs, etc.)** and then **uses that retrieved content to generate a better, more accurate, and contextual response**.

Flow of RAG Based Applications

doc → Company Policy
       300

① Get the data for which you create database.

② creating chunks. → 600 Chunks

2.5   creating Embedding

③ vector Store [database] → → slow.
                                → Inefficient   Textual Comparison

④ Retrieval [Query — Database → Retrieve]

⑤ LLM [Query , Retrieve]

→ Numerical Rep. of Data.
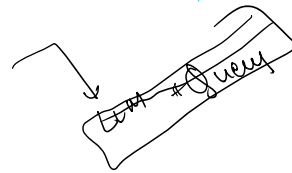
Embedding → Numerical Rep. of Data.

[1 have a cat]

[0.986, 0.782, 0.999....]

Vector

Actual → Chunks ⟶ [Embedding] → Searching.

Vector Database

v_3
v_1 v_4
v_{10}
v_Q
v_{11}
v_2
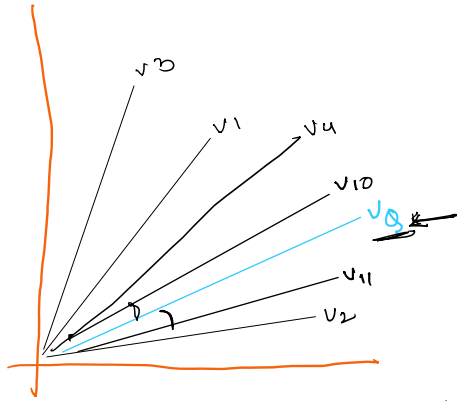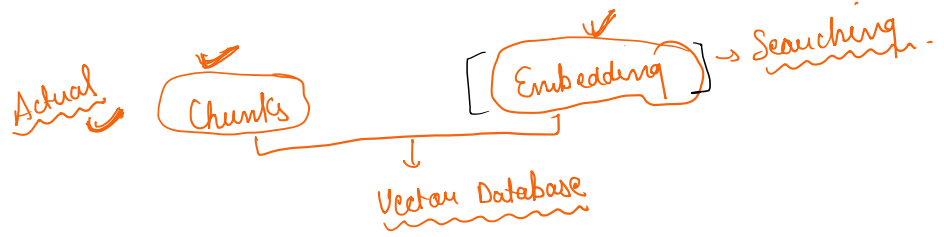
$| v_{10}, v_{11} | ✓$

↓

Text Extract

Query → Vector.

Text + Query

→ Speedy
→ Efficient }