# RAG [Retrieval Augmented Generation]

## Why we need RAG?

① Hallucination ← model being over-creative and not factual

dealing with

① temperature

② Fine-tune a LLM (Creating Distilled models) (Peft, LoRA)

③ Adding Context
- Using Prompt Template (FewShotTemplate)
- Using memory data

② Stale Knowledge

③ Size Limits

④ Specific Context Awareness.

## RAG helps us:

① Make context RECENT

② more accurate and on-point answers.

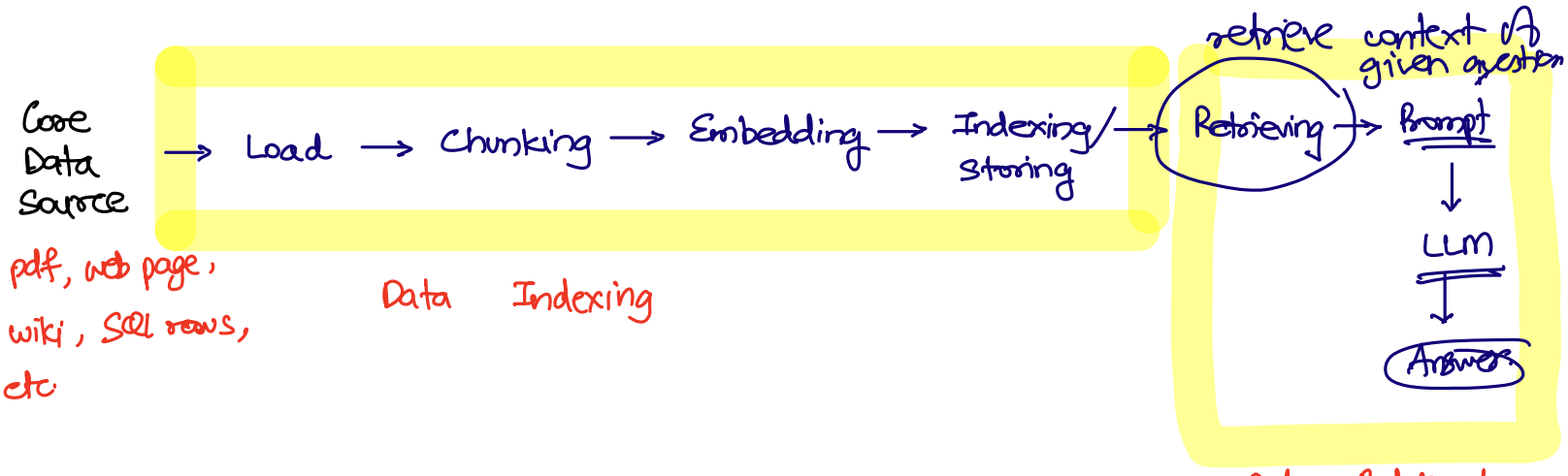
③ Uses your own data to get answers.

## What is RAG?

RAG is a pattern where LLM answers a question using

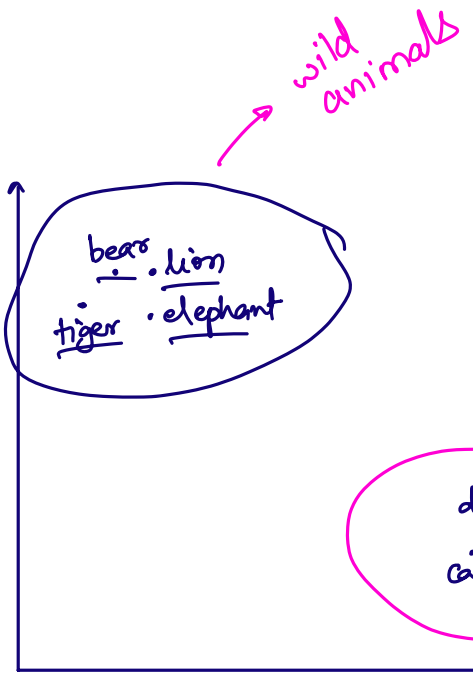
EXTERNAL UP-TO-DATE PROPRIETORY DOCUMENTS

that are retrieved at query-time.




## Embeddings

embedDim = 2




Similarity search

vector form