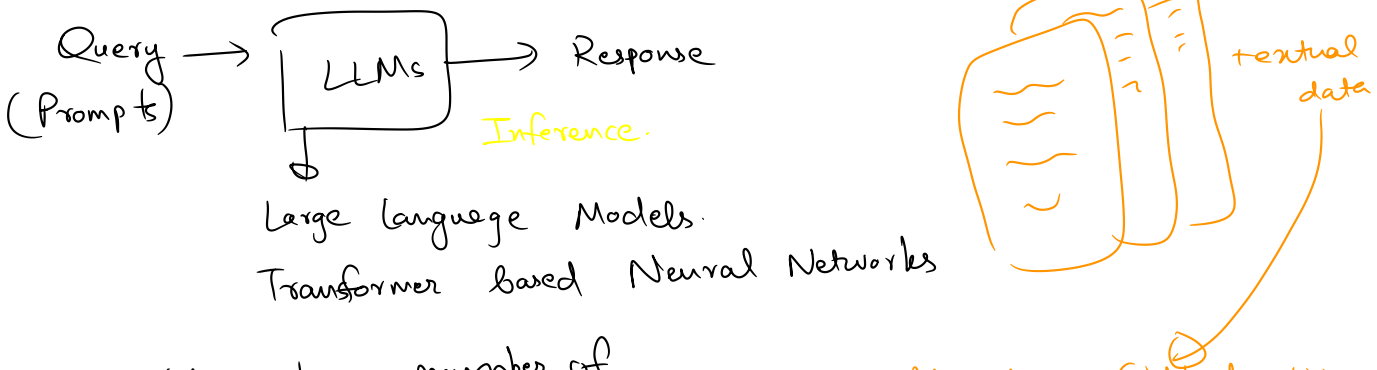
Retrieval Augmented Generation (RAG)



LLMs: huge number of parameters

↓

weights & biases

Parametric knowledge

Challenges (Vanilla LLM):

pretrained on a large Corpus of textual data

1) Prompt " How old is Harry?"

Harry Potter

" What are procedures that are followed in Amazon? "

2) knowledge cut off data

Pretraining LLM : GPT-4 GPT-3

Oct 2024
2025
} known

LLM has no access to the data

13th Jan 2026

" What is today's weather in Kanpur "

14th Jan

3) Hallucinations

Response probabilistic convincing answer.

Query: "When did Beam Rule us?" → LLM → "No, they never ruled us."

"Yes, they ruled us b/n 1920–1947."

("Convincing")

Finetuning

1) SFT → Update all our weights & biases

W_0 → labelled dataset

LLM

Prompts → Correct Ans

Errors

2) PEFT : LoRA & QLORA

$W = W_0 + AB$ → Low-ranked matrices for fine tuning

W_0 → Freezed

1) Private data : Amazon

Documents, Files, Legal terms of Amz

2) Knowledge Cut-off date :

Regularly update our LLM.

3) Hallucination

"No, they never Rule"

India's Ind 1000BCE – 2026

Dynasty, Rule, E

Fine-tuning

Challenge

SFT : we are updating all of our weights

1) Computational Cost: LORA & QLORA: A & B

2) Pretraining : English Literature. LLM French Novels

Fine tune

3) Technical Expertize: Labeled data : Prem LLM → +ve sentiment.

" Language Models are few Shot Learners"

Complex ↑ → emergent properties

IN CONTEXT LEARNING

It is the core capability of LLMs, where the model learns (GPT-3/4, Claude, Llama) to solve a task purely by seeing examples in the prompt — without updating the weights.

Ex

Query: Do sentiment Analysis on the following CSV file.

Examples: Pleam go through the below examples & learn how to judge Sentiments

1) I love my singing → Positive Sentiment
2) I hate my college → Negative Sentiment
3) My life sucks → Negative

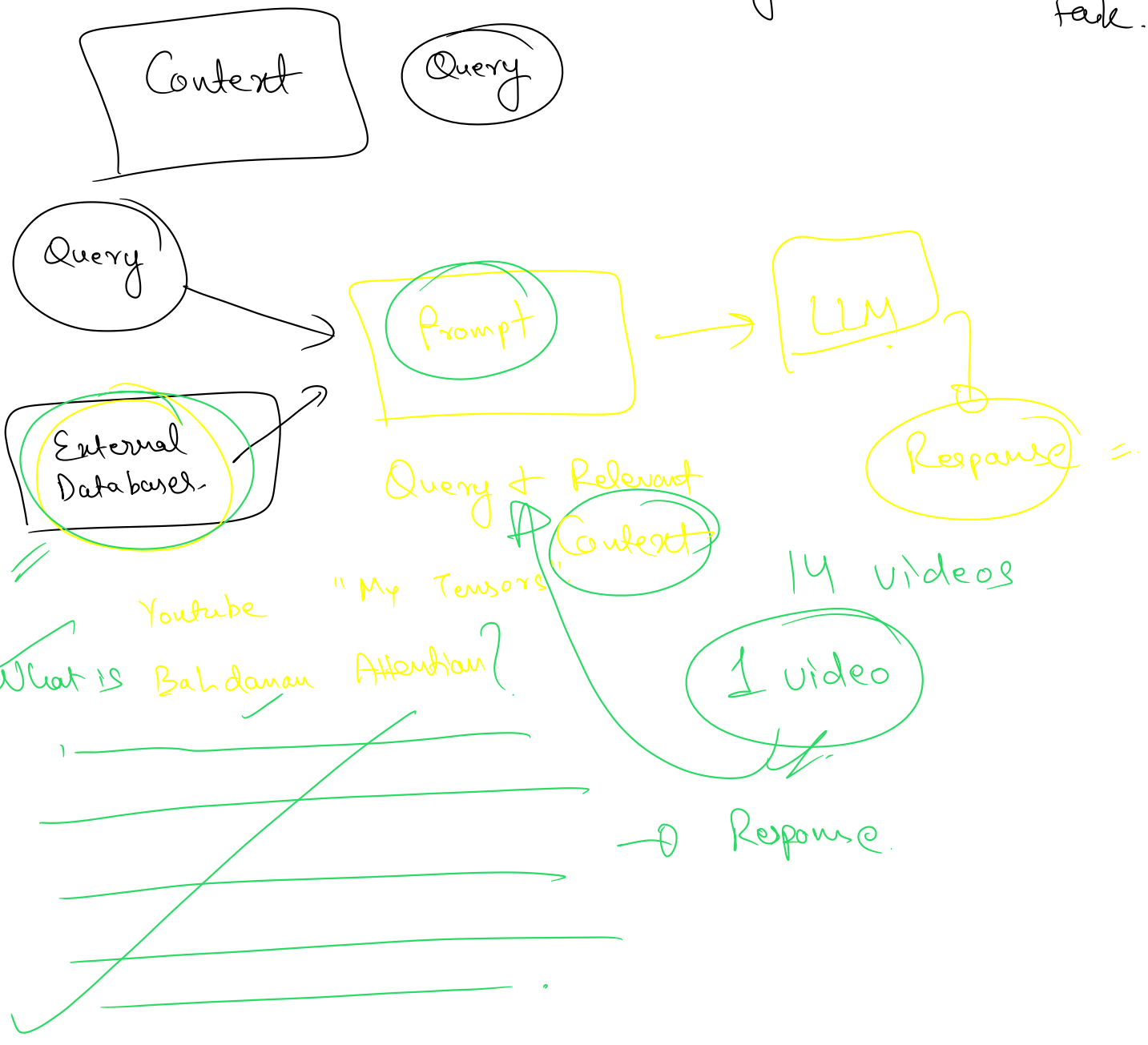
CSV contains: "I enjoy fishing" → ??

GPT-1, GPT-2 } Basic

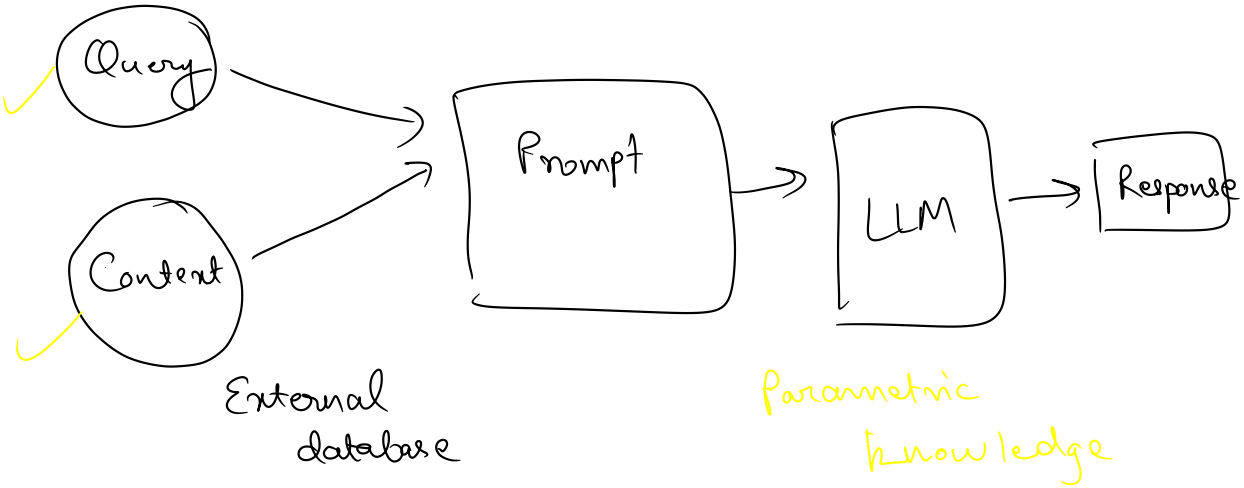
GPT-3 ≤

Few Shot Prompting
One Shot Prompting

Emergent Property: It is a behaviour or ability that suddenly appears in a system when it reaches a certain scale of complexity, even though it was not explicitly programmed on that task.



RAG (Retrieval Augmented Generation)



RAG: Retrieval + Generation

Indexing: It is the process of preparing our knowledge base so that it can be efficiently searched up, at the time of query

Architecting Intelligence

Course Materials → PDFs, PPTs } All the Assignments

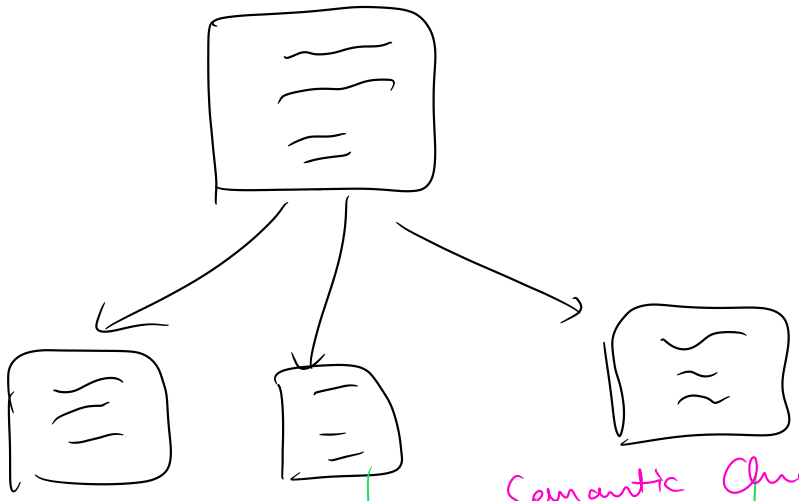
4 sub-steps

1) Document Ingestion: You load your source knowledge into the memory.

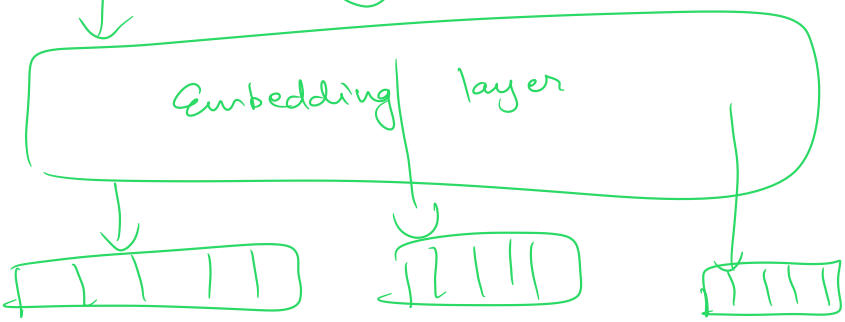
PDF docs
Youtube Transcripts
Github Repos.

LangChain: PyPDF Loader, GitLoader etc. ⊆

2) Text Chunking: Break our large documents into small, semantically meaning chunks



3) Embedding generation: convert each chunk into a dense vector that captures its semantic meaning

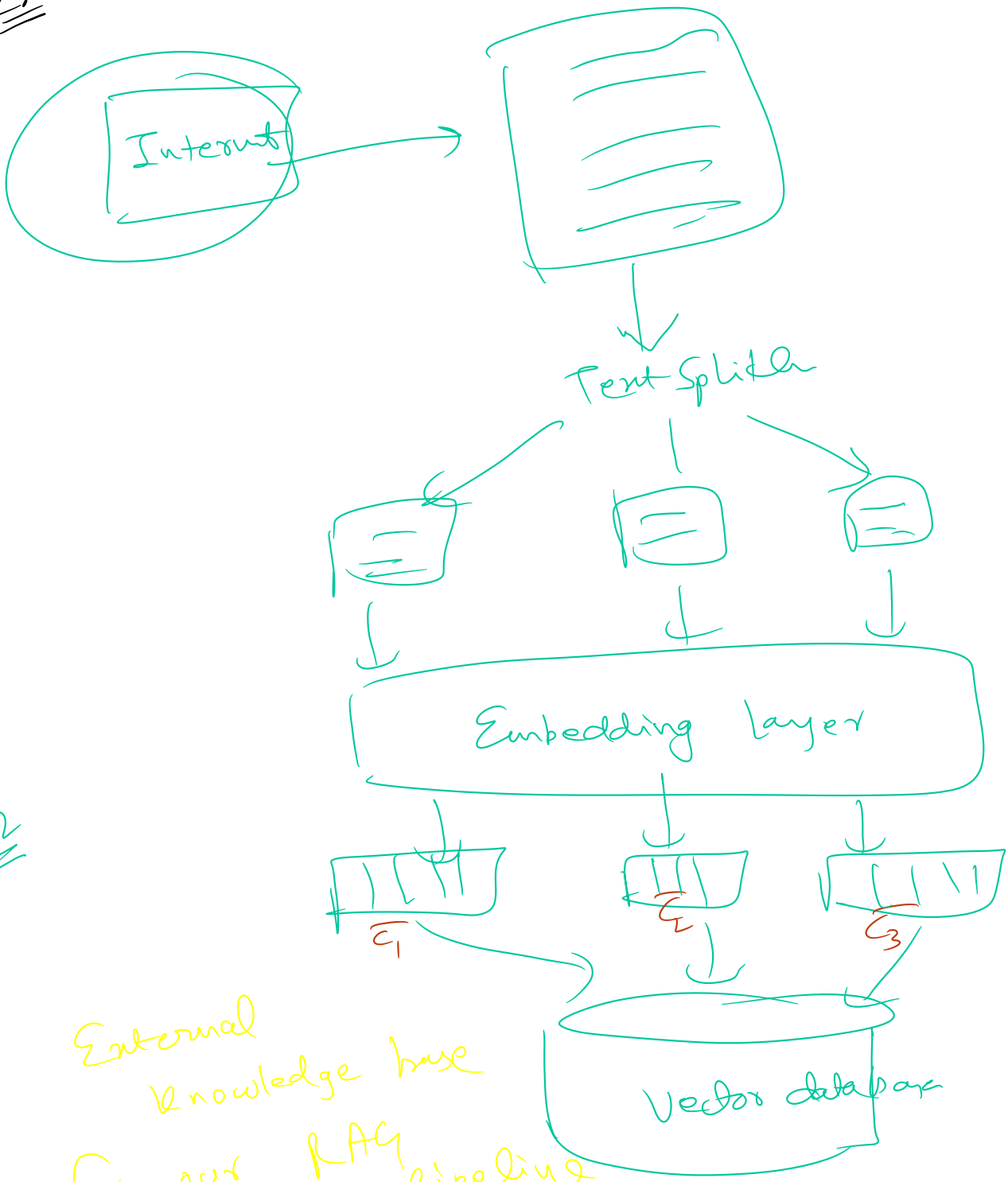


Ex: OpenAI Embedding
Gemini GenAI

4) Store : Original Chunk + Vector embedding

Vector database.

Ex Chroma, FAISS, etc.



Ex 2

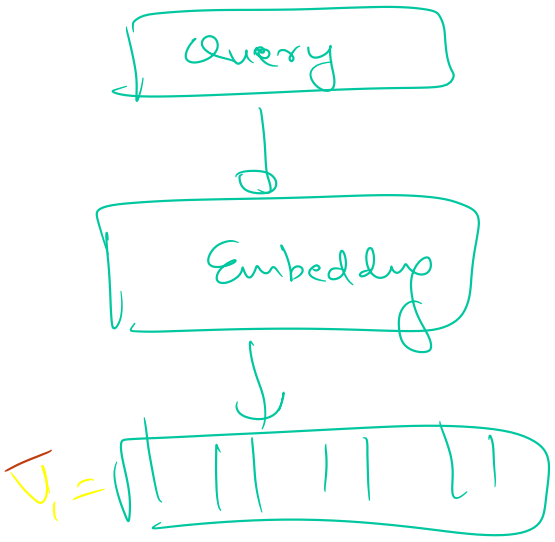
External knowledge base for our RAG pipeline

(2) Retrieval:

finding the most relevant pieces of information from a pre-built index

Cosine Similarity, Euclidean distance, vector database, BM25

$$\frac{\vec{V_1} \cdot \vec{C_1}}{|V_1||C_1|} = S_1, \quad \frac{\vec{V_1} \cdot \vec{C_2}}{|V_2||C_2|} = S_2, \quad \frac{\vec{V_1} \cdot \vec{C_3}}{|V_3||C_1|} = S_3$$



512dim

Rank: S_2, S_3, S_1 → Top (k) chunks

Step-1 : embedding of our query vector

Step-2 : Semantic Search

Step-3 : Ranking (Top-k)

Step-4 : Most relevant

(3) Augmentation

Where the retrieval docunts (chunks of relevant context) are combined with the user's query to form a new, enriched prompt for the LLM

Example

""" You are an helpful assistant

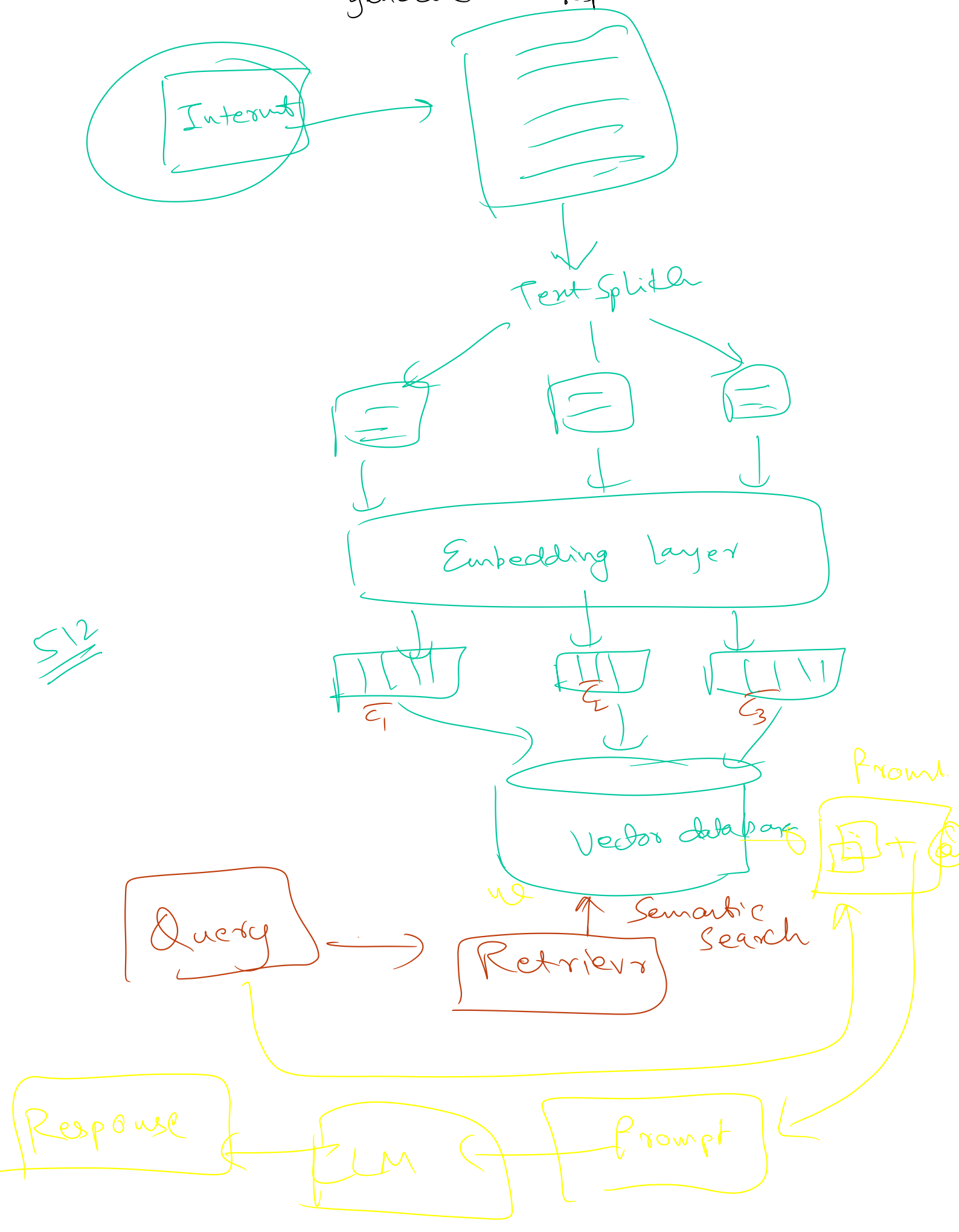
Answer the question ONLY from the provided context. If the Context is insufficient, just say you don't know.

{context}

Question: {question}

""" """

(4) Generation — LLM uses the user's query & retrieved & augmented context to generate a response



Vanilla LLM

RAG

1) Private data

"Amazon"

"Text Ingestion"

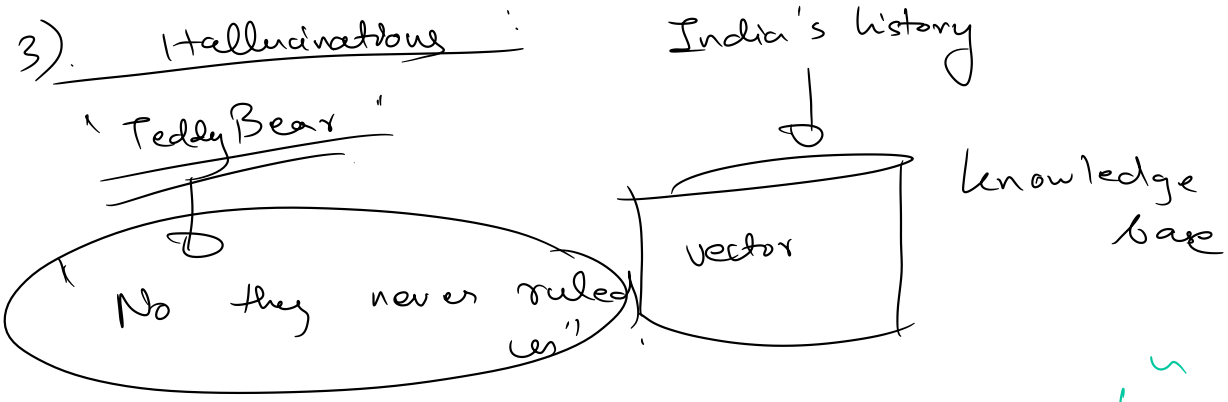
2) Knowledge Cut off date

LLM : 1st Jan 2026

14th Jan 2026

Weather

3) Hallucinations



Legal : - - - - -

Medical field : - - - -

App^n RAG