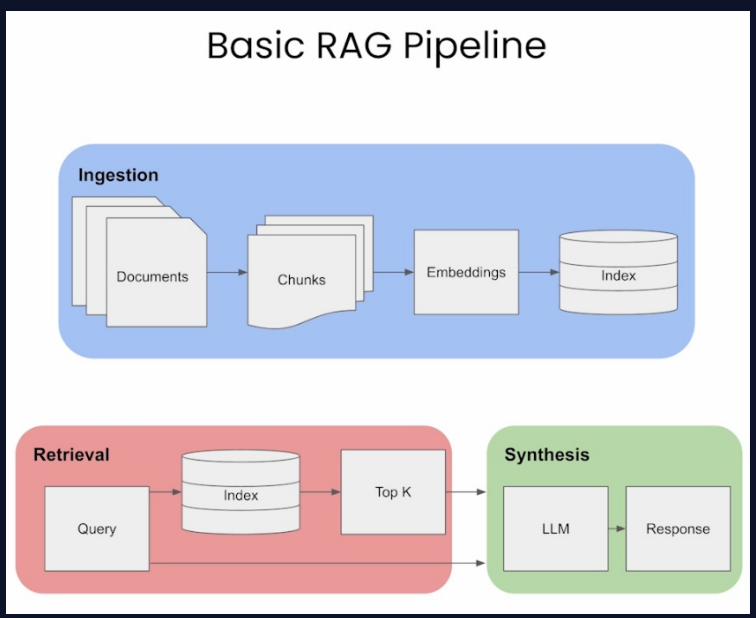
What is RAG:
Understanding Retrieval-Augmented Generation



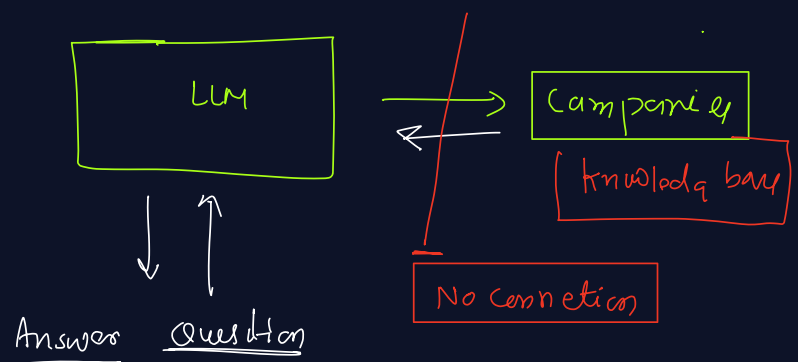
Agenda: What is RAG ?
Why ?
Importance ?
Use case ?
Benefit ?
} 1

Core concept ? } - 2
Components } 3

Challenges .
Limitations -
Future Trends
} 4

Implemation — 5

(1). INey : 1)



Personal Database.

(2) → LLM → India Vs Aus → Who?

↑↑

X (Web search)

X Updated

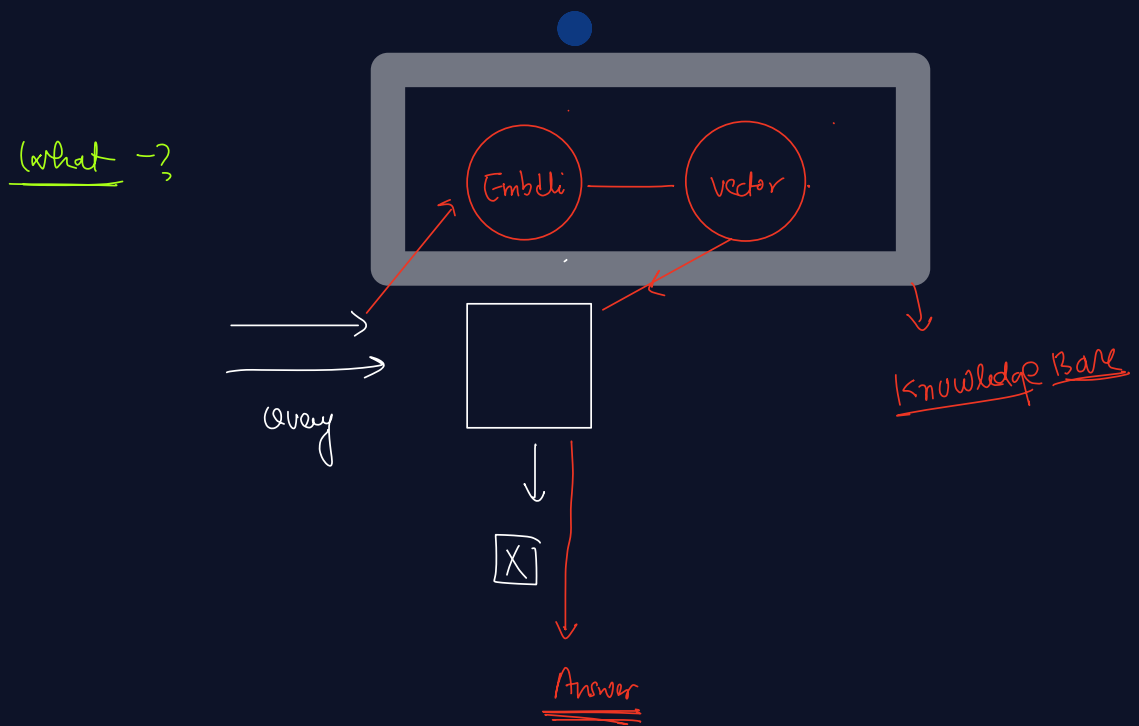
(3). Real time Data ⇒ Live - community X

(4) → → Hallucination (Faulty Not correct)

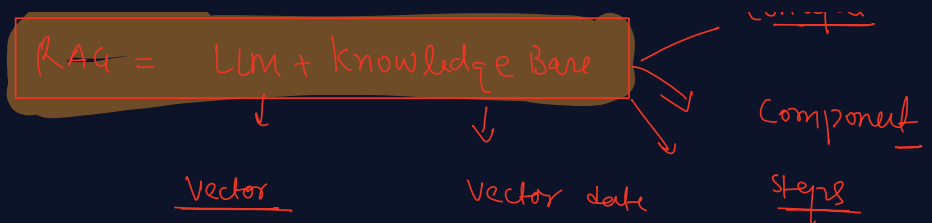
↓

Personal

(5) → Domain specific knowledge.



Consult



(1) → Benefit & Use Case ⇒

1) Domain specific Model
2) Real time data Access
3) Question - Answering.
4) Chatbots for your Database

RAG ⇒ : Concepts :

1) Query ⇒ Question
2) Retrival →
3) Embedding.
3) Vector Database
4) Data Loaders.
5) Splitter & Chunks
6) Augmentation

7) Fusion Logic

8) Evalution Matrix

Data → Knowledge Base → (LLM) + KB

1) Loader →

2) Split, Chunks →

3) Embedding → Text → Numerical (Meaning)

4) Vector Data → place to store data

5) Knowledge Base → Retrivalful process

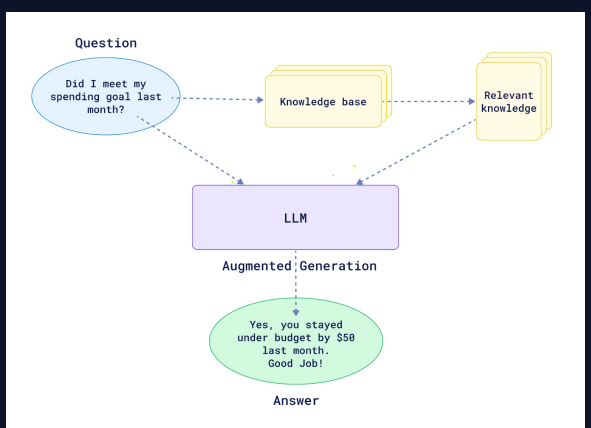
6) Augmented → Combine ct Answer.

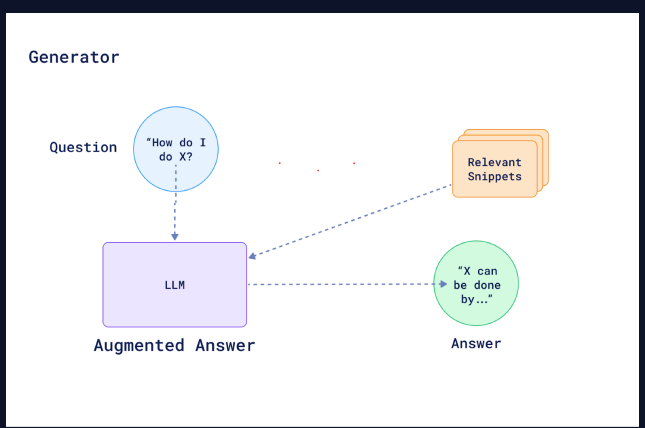
7) Fusion logic → Q + KB →

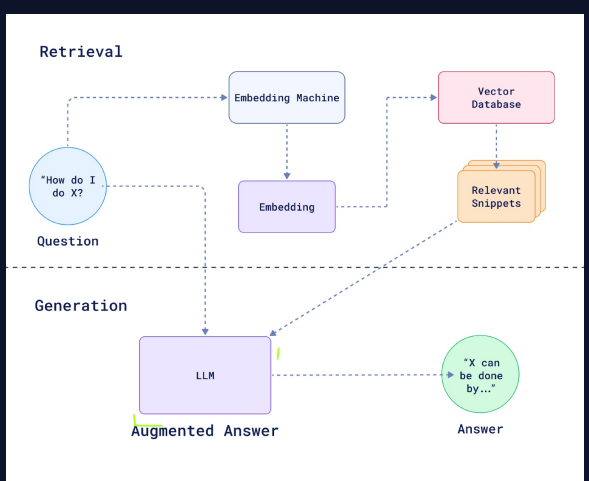
(Q) → Evalution Matrix → 11 Answer

1) Human Answer
2) Confusion
3) Rouge /
2) BLEU

Architecture & Components





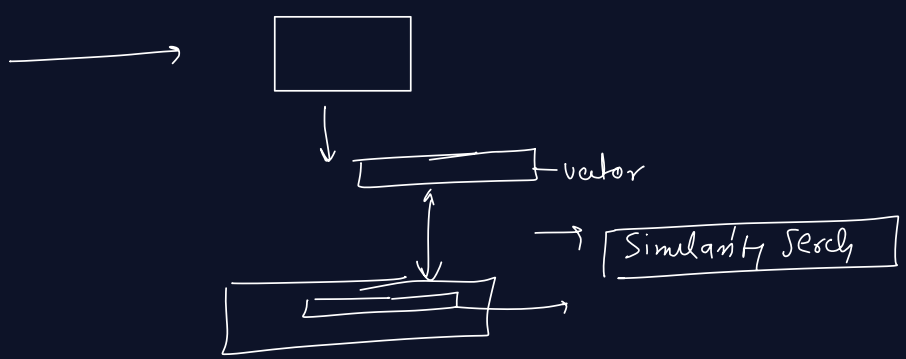


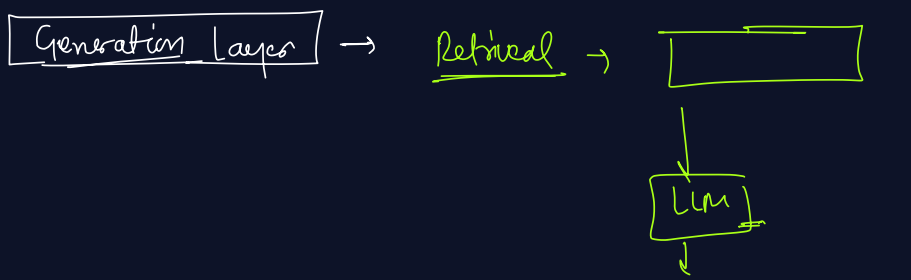
Components: 0) Ingestion →
1) Generation → →
2) Retrieval →
3) Data pipeline
4) Integration

(1) Ingestion → 1) Data Loader
↓
Chunks / Splitter
↓
Embedding
↓
Vector Database
↓
Pinecone
Chromadb
FAISS

- Ingestion

(2) → Retrieval layer
↓
Cosine similarity





Retrival R

↓

Augmented A

↓

Generation G

Challenges:

1) Dependency of High Quality Data

Knowledge Base

(2). Scalibility

(3). Latency (Speed) → Slow. (wait to get Respon)

(4). Storage vectorban → cost

Expensive

(5) → Augmented Answer → (Not correct Answer)