RAG

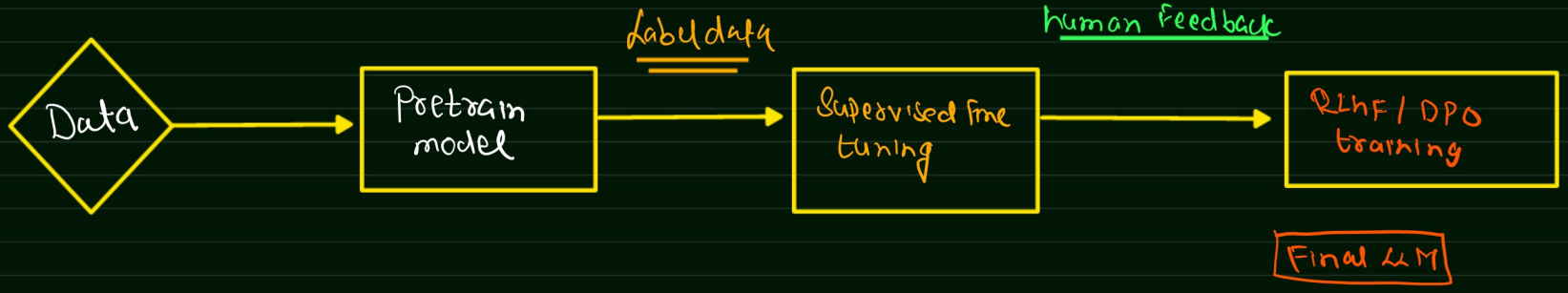
1. What is diff b/w Simple AI Assistant VS RAG
2. Finetuning & RAG
3. Architecture of RAG
4. Implementation of RAG
5. Langchain -
6. Vector Database -
7. Multimodal RAG
8. Graph RAG
9. CAG ⇒ fast retrieval
   - ↑
   - Cache
10. Agentic RAG

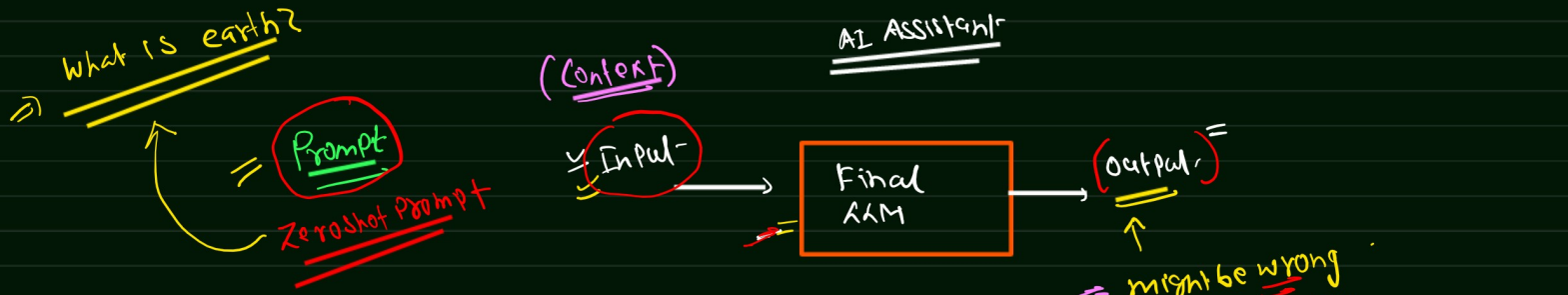
RAG ⇒ look simple
↳ variable

Diff B/w AI Assistant and RAG

Retrieval Augument generation

1. Data ingestion.
2. Data Retriever
3. Data generation





Final LLM ⇒ Problem (hallunication)

- might be wrong
- missinformation
- Wrong information
- No gives any information

Prompt ⇒ 1. Zeroshot Prompt
2. Few Shot Prompt
3. Chain of thought Prompting → Agent ← different instructions / Input — prompting

internet data → LLM

Extracted from data

(Medical domain) → XYZ ⇒ Researching on one specific medicine (latest) (cancer)

Jaspreet ⇒ Company ⇒ (marketing Sales)

XYZ

If

gpt → LLM

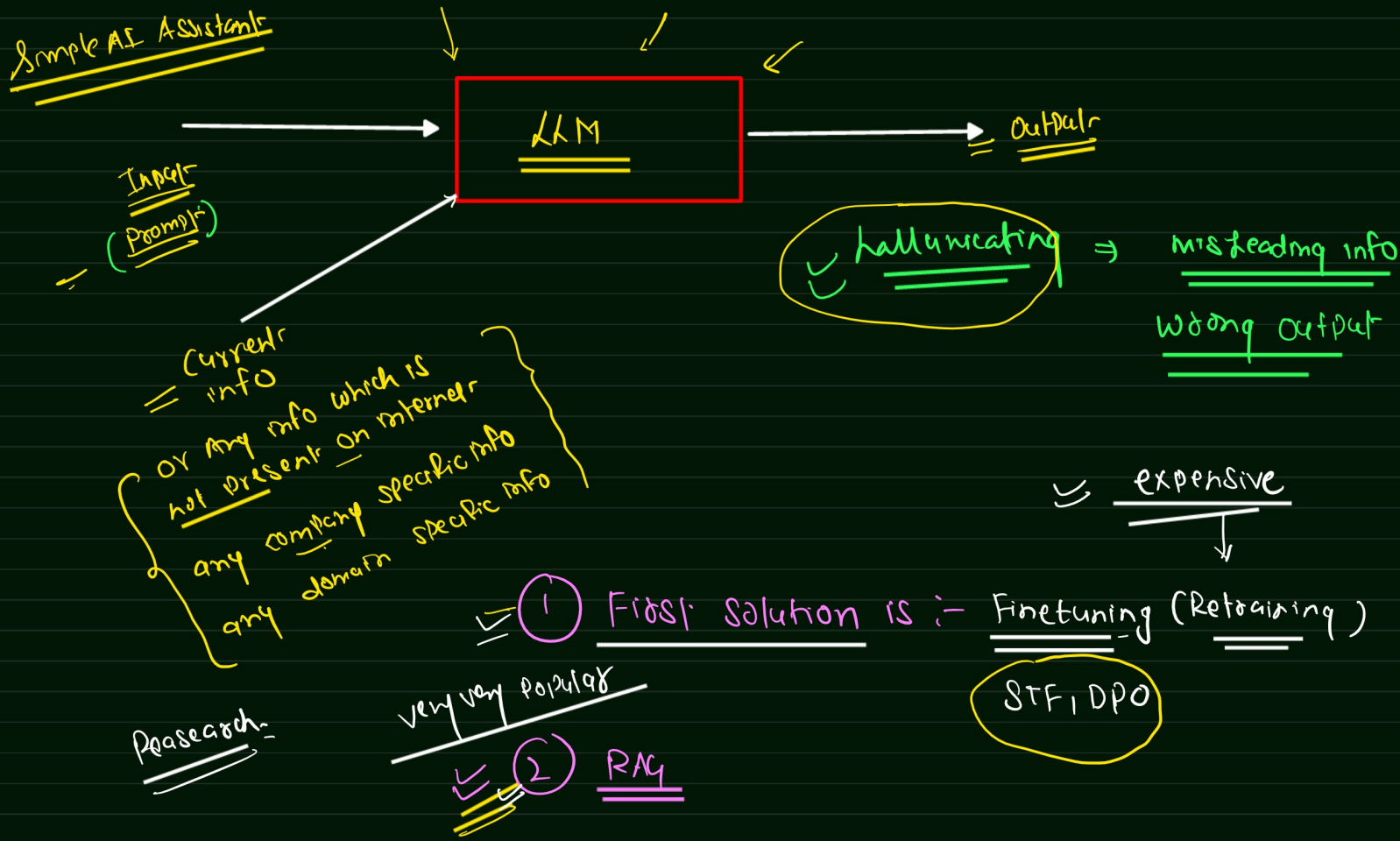
retrain — Finetuning — SFT, DPO

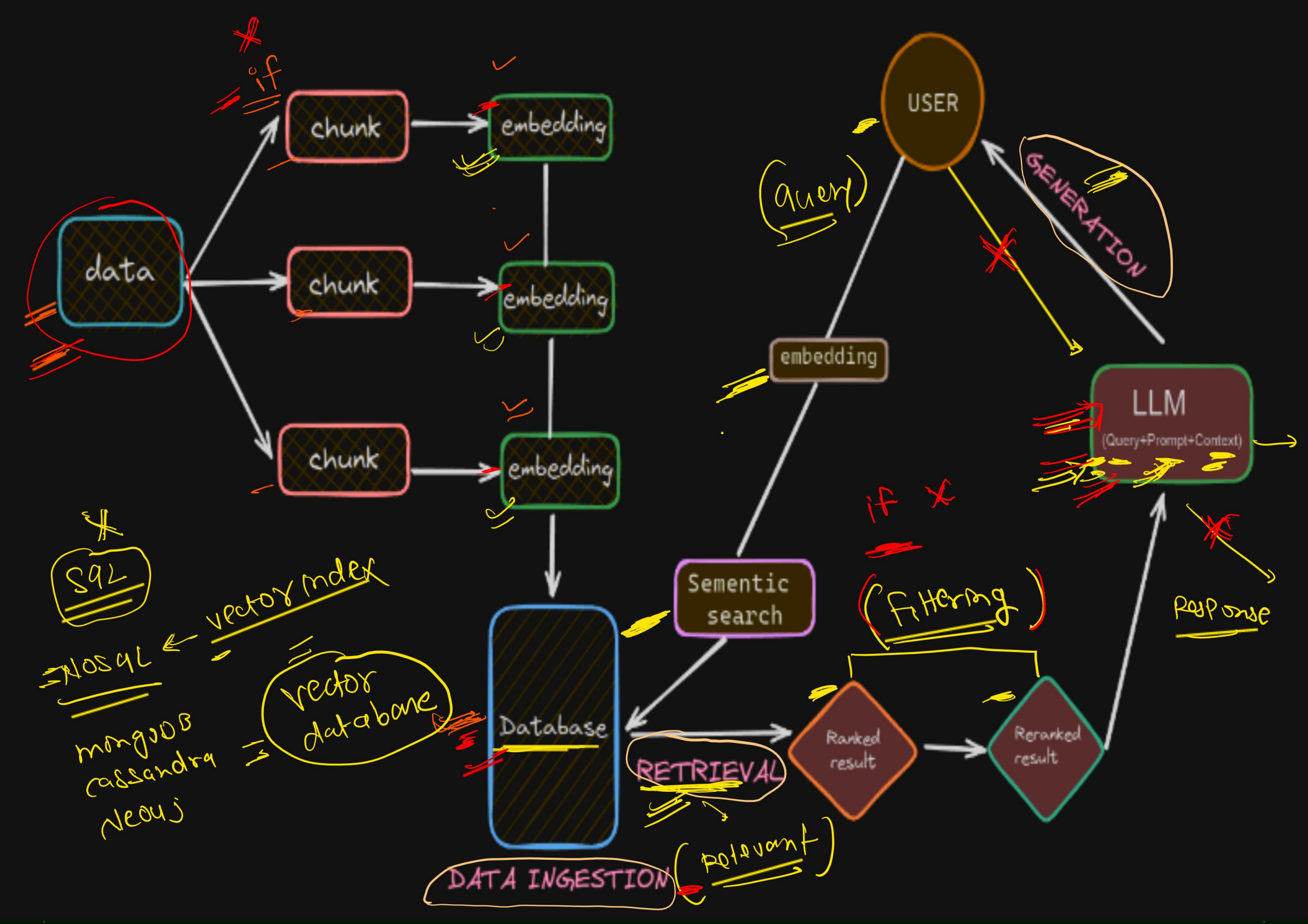
RAG — Context — Few-shot learning

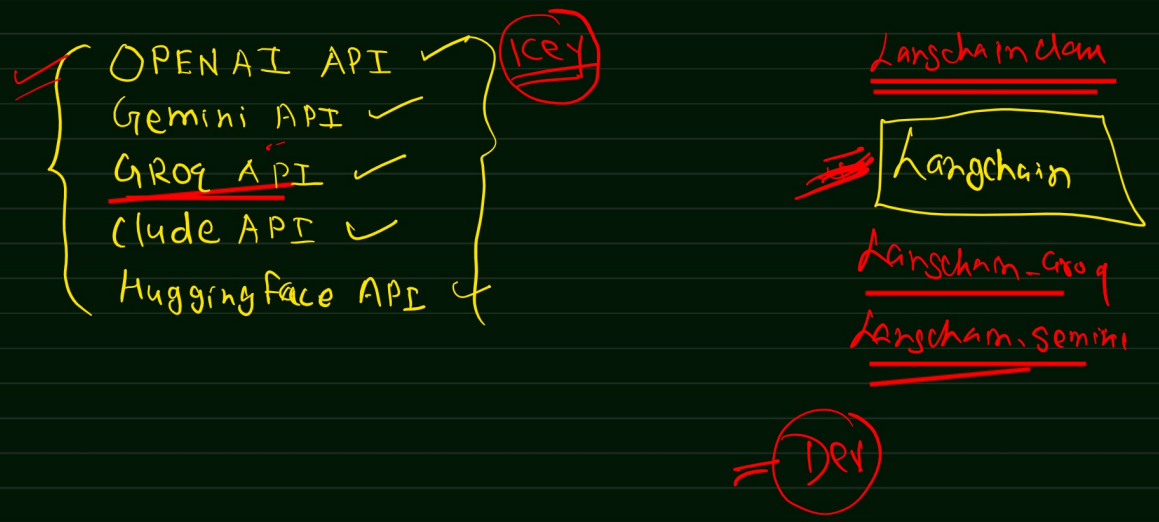
Few-shot Prompting ← RAG

{Artical}

(refer above artical and tell me what is earth?)


Simple AI Assistant
LLM
Output
Input (Prompt)
Current info
or any info which is not present on internet
any company specific info
any domain specific info
Hallucinating ⇒ misleading info
Wrong output
expensive
1 First Solution is :- Finetuning (Retraining)
STF, DPO
Research
very very popular
2 RAG


data
chunk
chunk
chunk
embedding
embedding
embedding
USER
(query)
embedding
GENERATION
LLM
(Query+Prompt+Context)
Response
Sementic search
Database
vector index
SQL
NOSQL
mongoDB
cassandra
neo4j
vector database
RETRIEVAL
Ranked result
Reranked result
(filtering)
DATA INGESTION
(relevant)


OPENAI API
Gemini API
GROq API
Clude API
Huggingface API
Key
Langchain clam
Langchain
Langcham_Groq
Langcham.Semini
DEV