① RAG
② Finetuning
③ Langchain Advance (LangServe, Langsmith)
④ Lhama-index.
⑤ Project.

③ Real time usecase

RAG

① RAG Architecture
② Pipeline
③ evalute

    - individual
    - complete

④ RAG v/s finetune

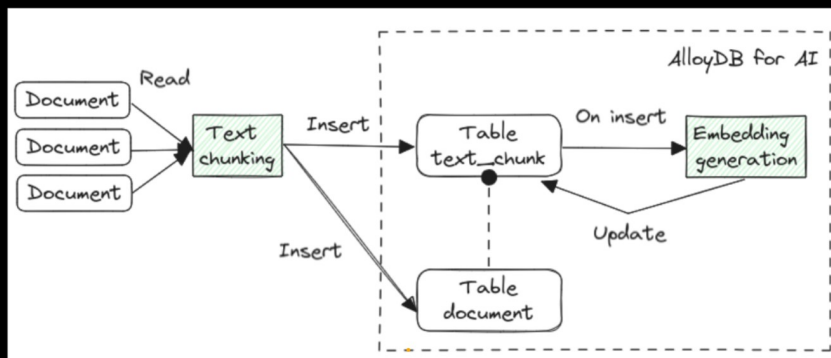⑤ Langchain

⑥ Llamaindex

⑦ vector/Nosq2 (DB)

{
RAG Architecture
Pipeline

RAG from scratch

RAG with Langchain
}

RAG Architecture

1. Ingestion
2. Retrival
3. generation

AlloyDB for AI

Ingestion ⇒ Data ⟶ transformation ⟶ DB

Data ⟶ ⇒ Image / text
Both

Embedding

Storing | Indexing |
Searching

GPT4.4v
Gemini
Gemma
} multimodel RAG

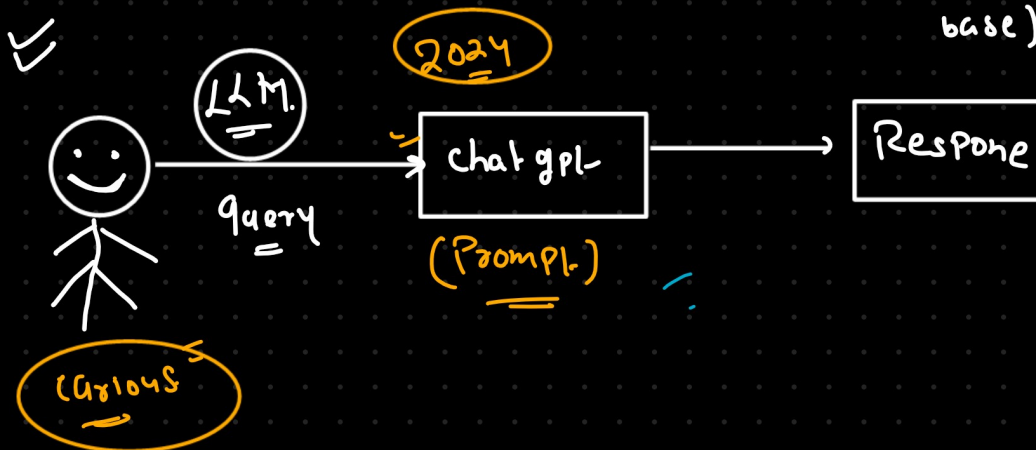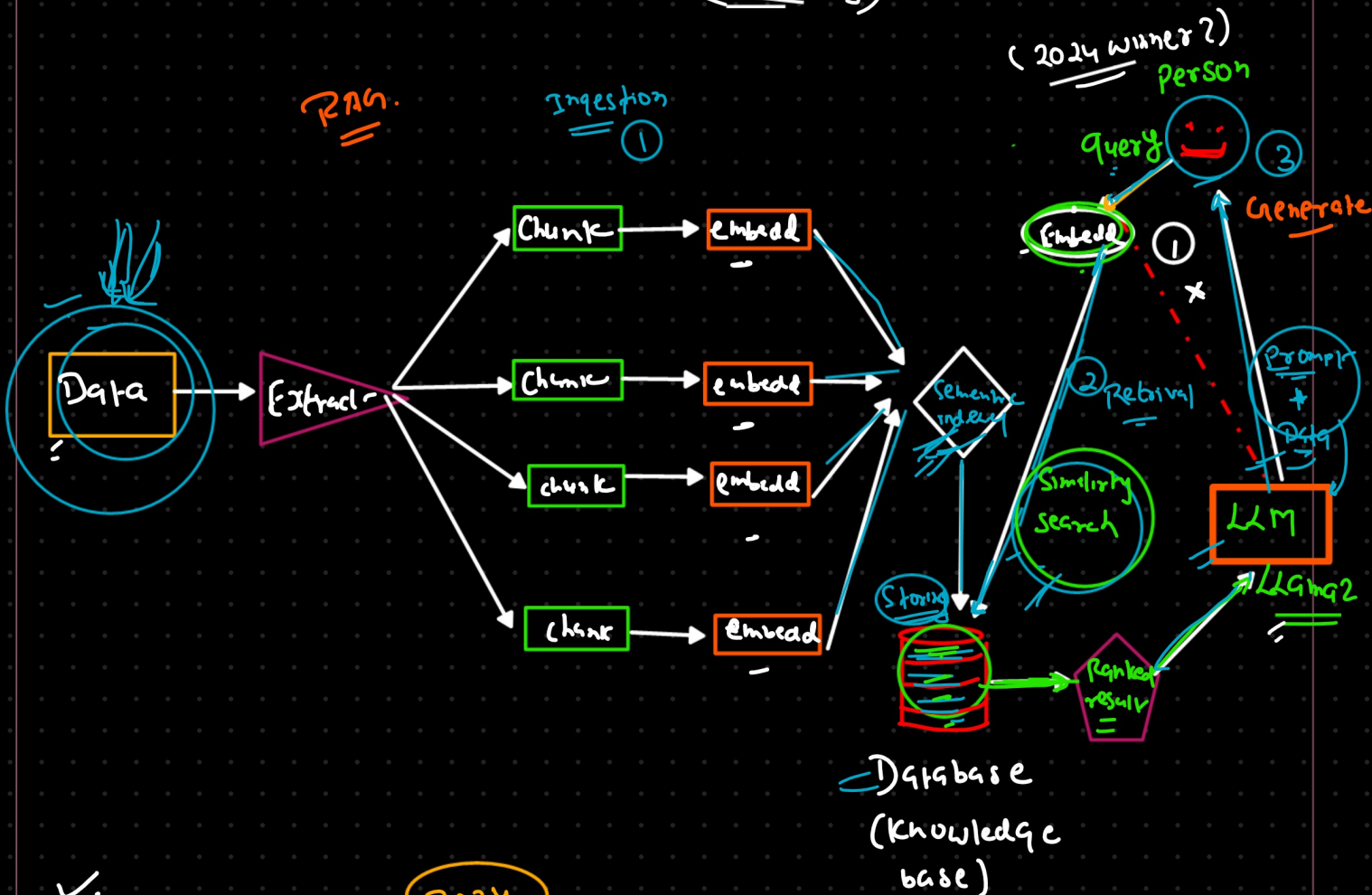## Retrival

query ☺

Data

(Indexing)

(Similirty score)

Relevant Data

Ranking

Generation

Prompt

Relevant Data → ◇ LLM → Final output

= (Augument)

= (Genrating)

RAG.

Ingestion ①

(2024 winner ?) person

query 🙂 ③

Generate

Data → Extract →
- Chunk → embedd
- Chunk → embedd
- chunk → Embedd
- Chunk → Embedd

→ Semantic index

Embedd ①

② Retrival

Prompt + Data

Similirty Search

Storing

Ranked result

LLM

Llama2

Database (Knowledge base)

2024

LLM. 🙂 — chat gpt → Respone

query

(Prompt)

Curious
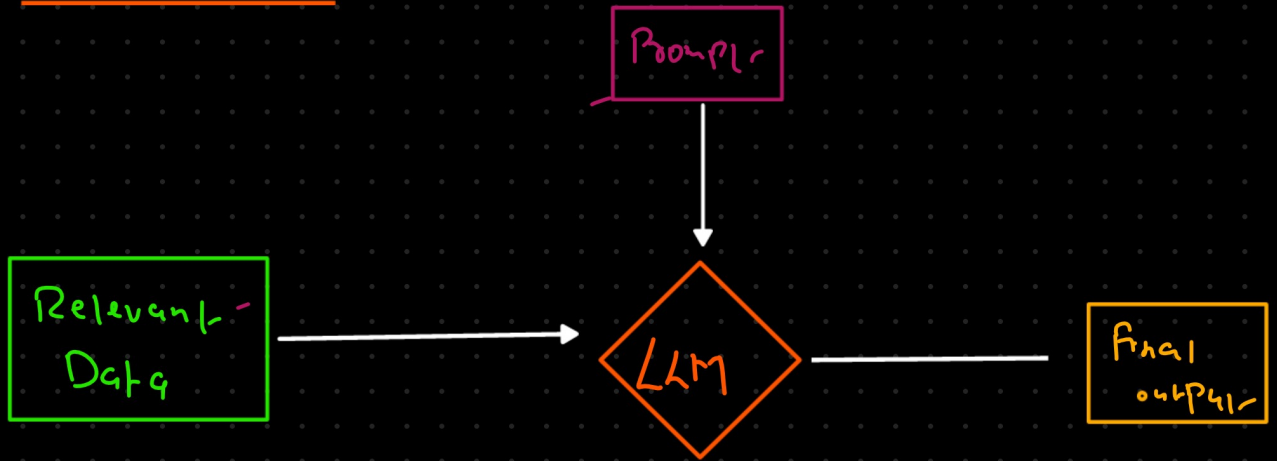
1. Finetuning

2. 3rd Party Agents

3. RLHF