RAG를 활용한 완성도 높은 LLM 서비스 구축

With Langchain & LLamaIndex





RAG with LangChain

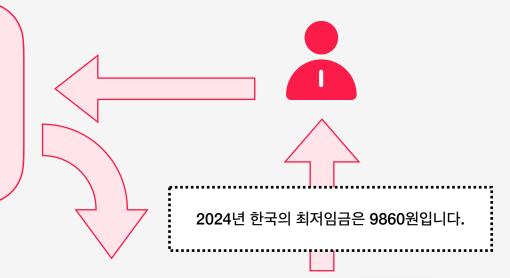
Introduction to LangChain

RAG Pipeline



User query

2024년 최저시급은 얼마야?





documents



chunks embedding



Retrieval

Relevant documents



chunk1: "임금은 2024년 최저 시급이며 교통비가 별도 지급된 다"

chunk2: "2024년 1월부터 최 저시급이 9860원으로 인상되었 다"

Generation

다음 주어지는 문서를 참고해 user question에 대한 답변을 생성하시오

> chunk1: chunk2:

user question:





RAG Pipeline

User query

2024년 최저시급은 얼마야?



Indexing

documents



Llamalndex 🦙

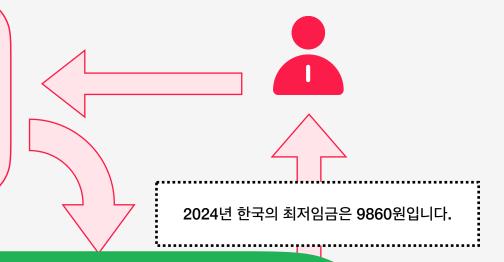




RAG Pipeline

User query

2024년 최저시급은 얼마야?



Indexing

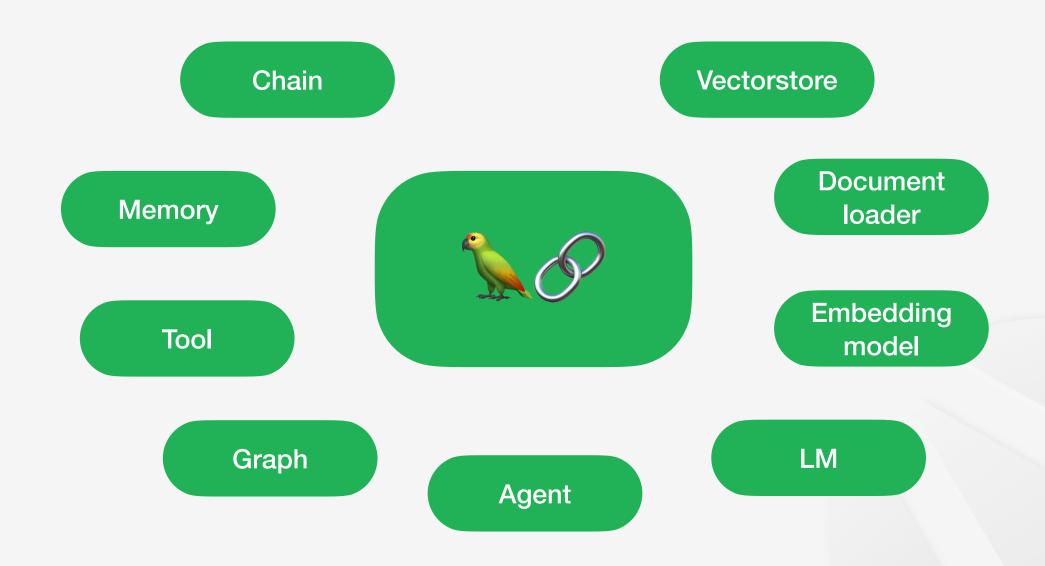
documents



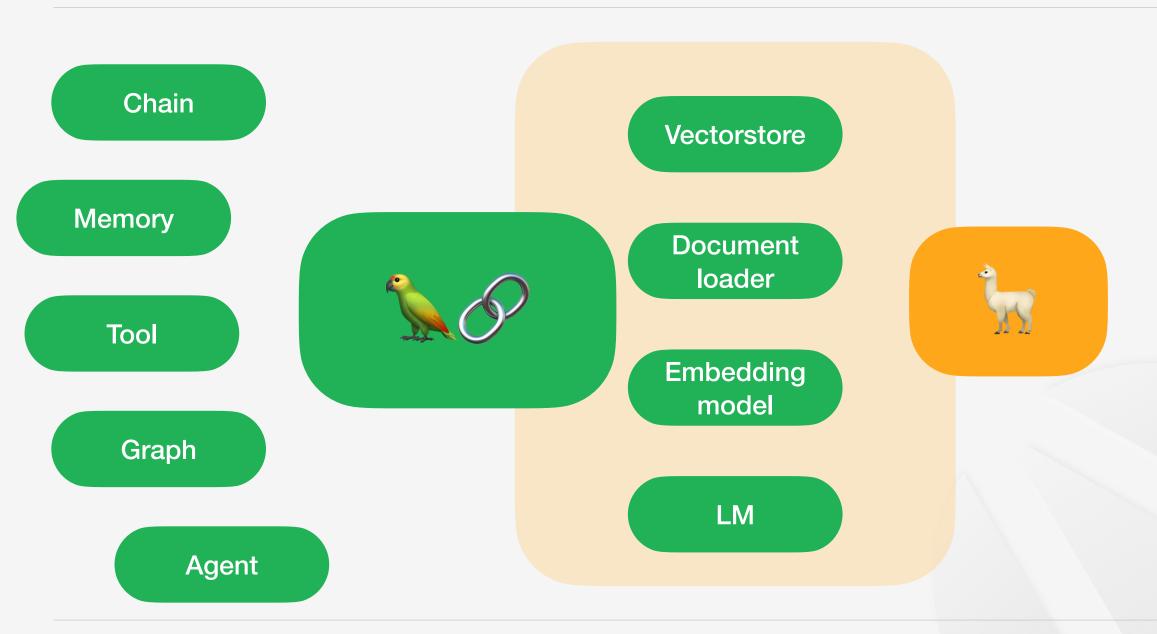
LangChain 🔪



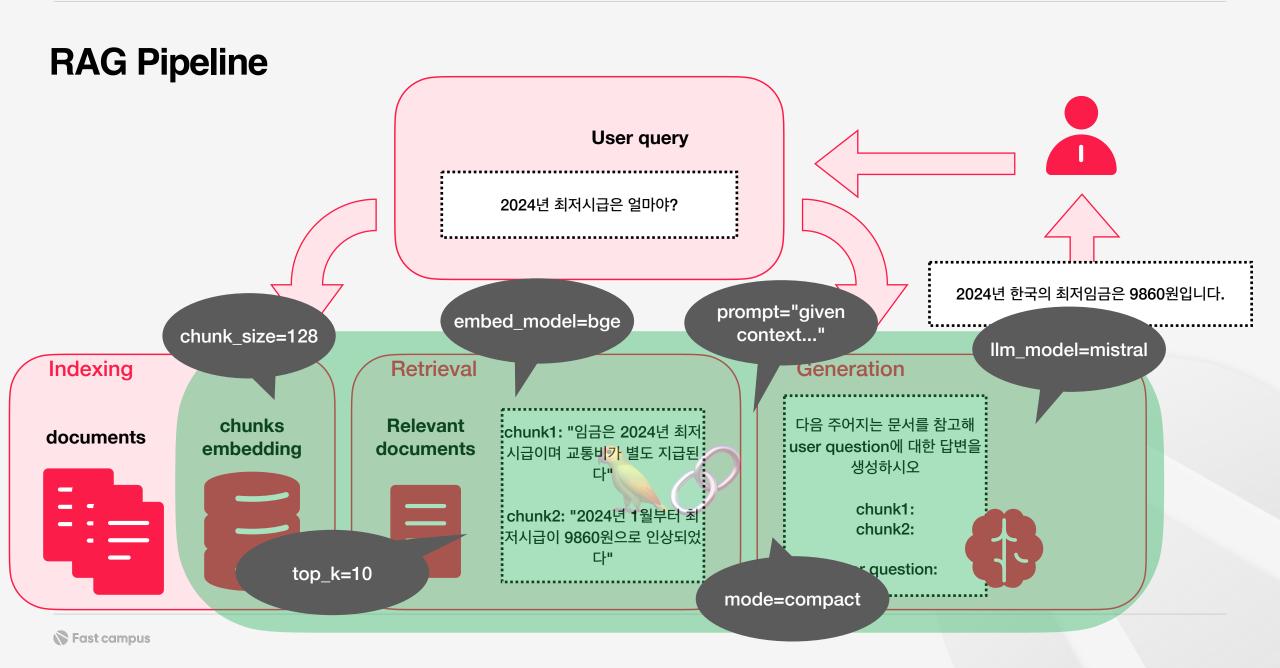


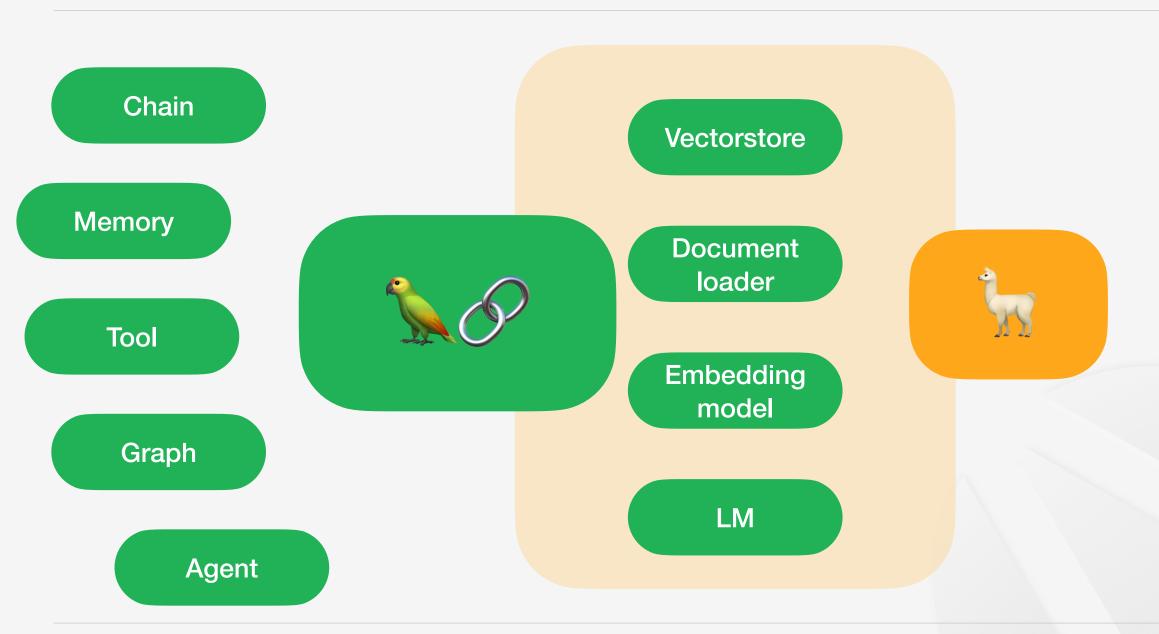


Comparative Analysis of RAG Frameworks



Recap





Ollama

```
ollama run llama2
          curl http://localhost:11434/api/generate -d '{
 "model": "llama2",
 "prompt": "What is Machine Learning?"
  from langchain_community.llms import Ollama
  llm = Ollama(model="teacher")
  query = "What is Machine Learning?"
  result = llm.invoke(query)
   print(result)
 model : [[amaz , created_at : 2024-04-08113:31:08.56/4352 , response :
{"model":"llama2","created_at":"2024-04-08T13:31:08.583829Z","response":"AI","done":false}
{"model":"llama2"."created at":"2024-04-08T13:31:08.600678Z"."response":")"."done":false}
    >>> Send a message (/? for help)
```

LangGraph

