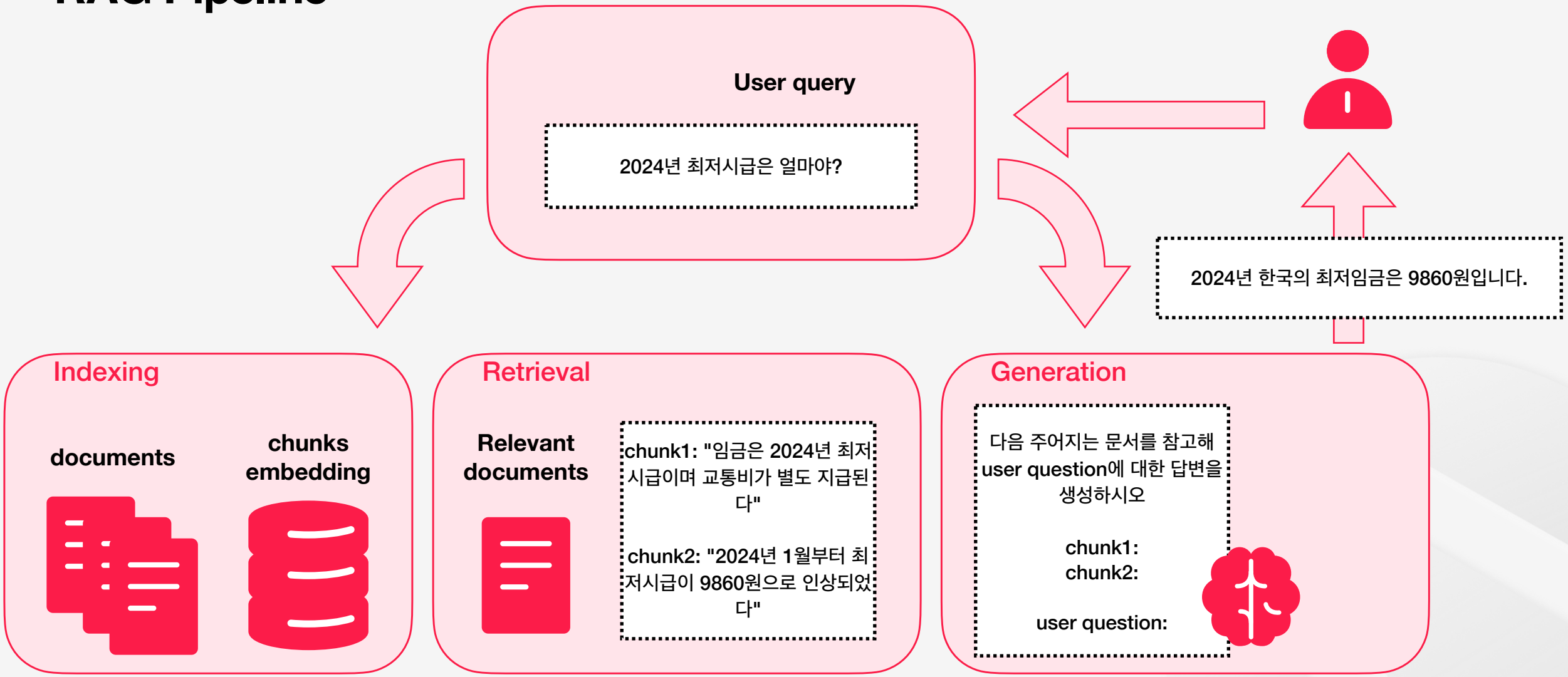
RAG를 활용한 완성도 높은 LLM 서비스 구축

With Langchain & LLamaIndex

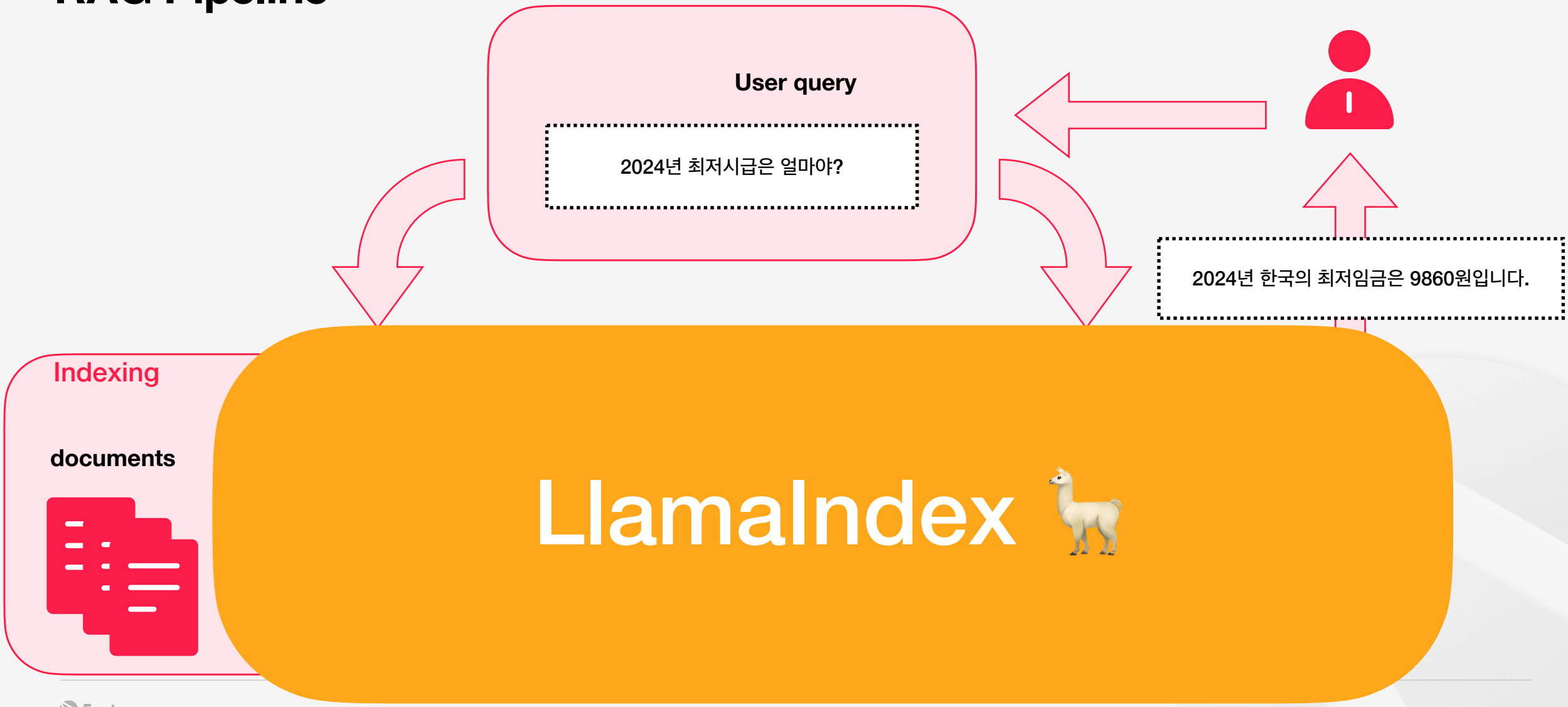
RAG with LangChain

Introduction to LangChain

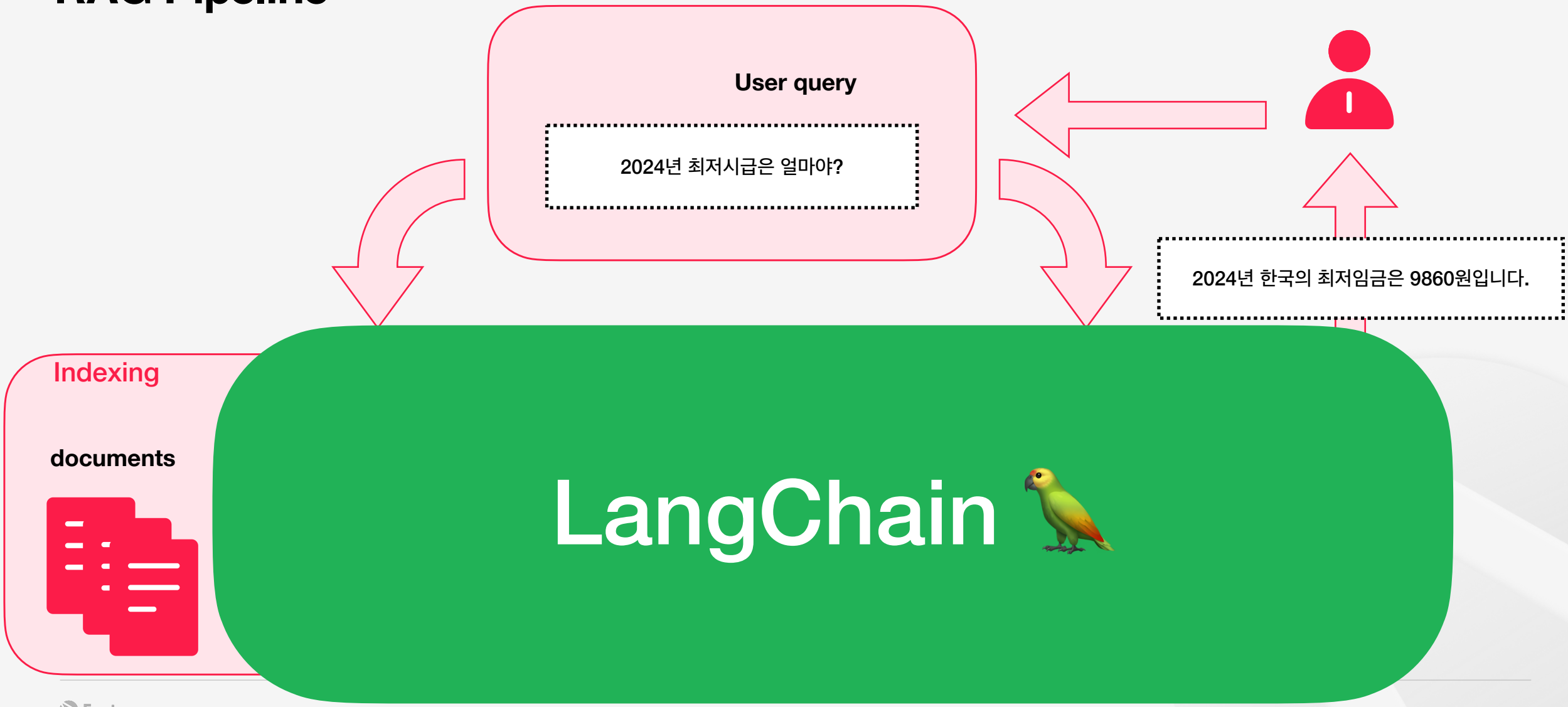
RAG Pipeline

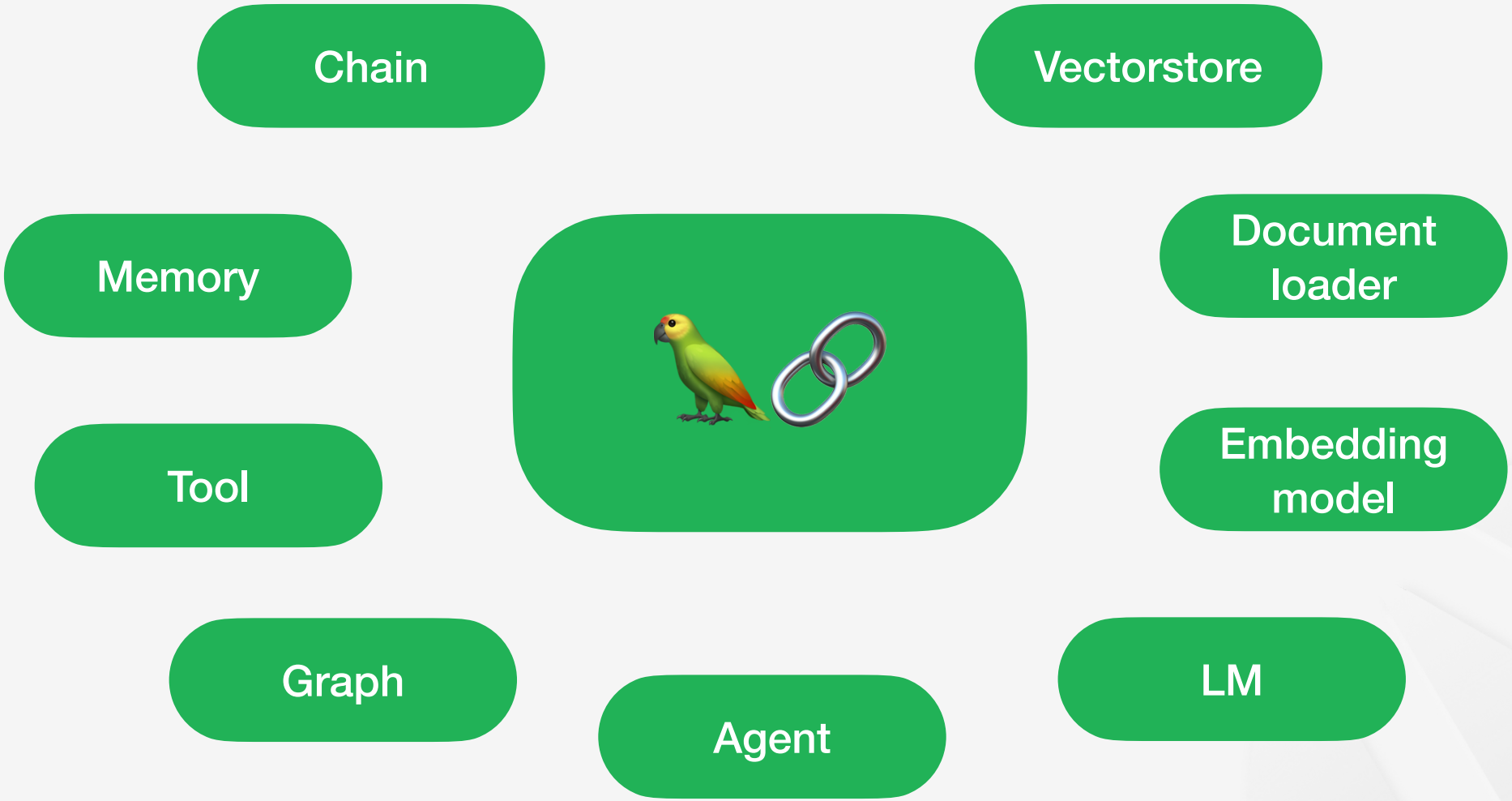


RAG Pipeline



RAG Pipeline




Chain
Vectorstore
Memory
Document loader
Tool
Embedding model
Graph
Agent
LM

Comparative Analysis of RAG Frameworks

Chain

Memory

Tool

Graph

Agent

Vectorstore

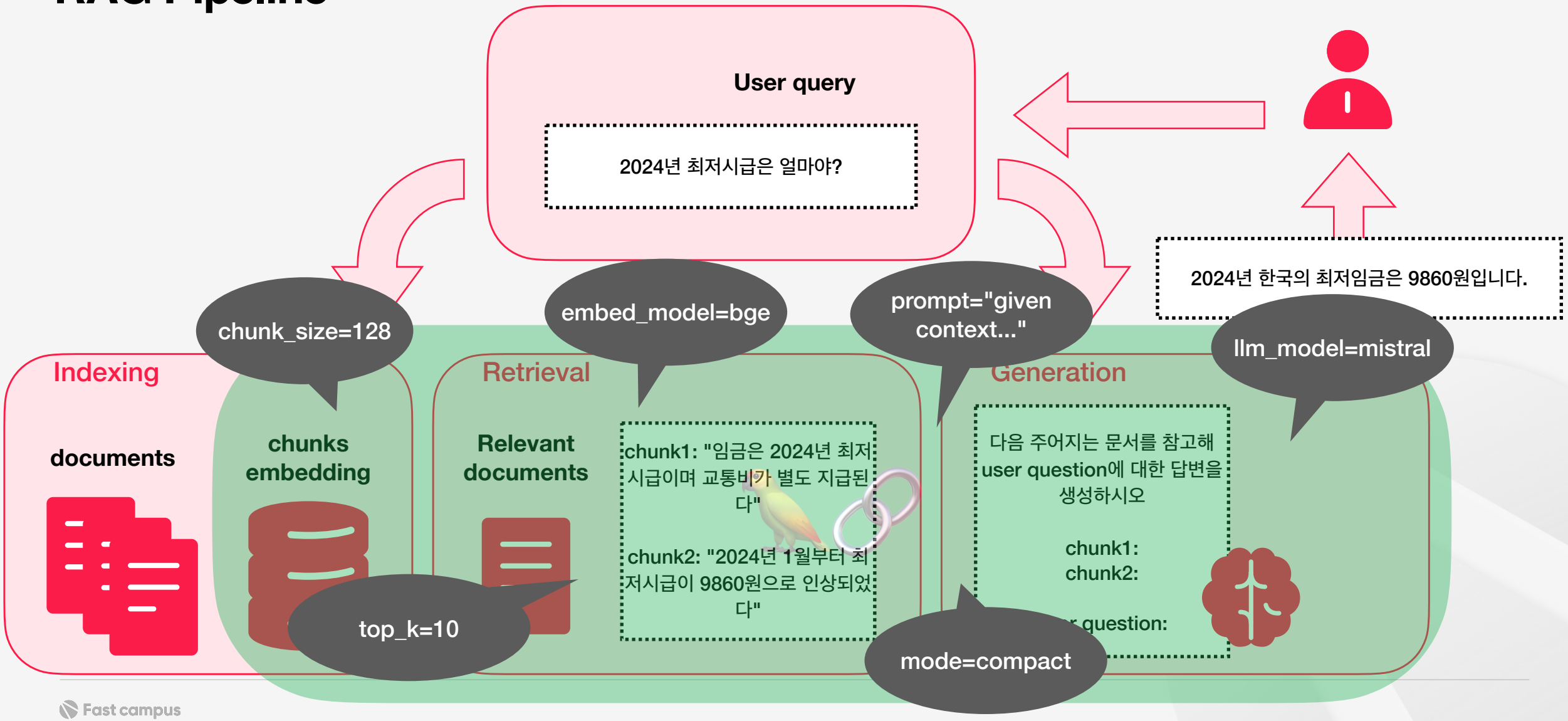
Document loader

Embedding model

LM

Recap

RAG Pipeline



Chain

Memory

Tool

Graph

Agent

Vectorstore

Document loader

Embedding model

LM

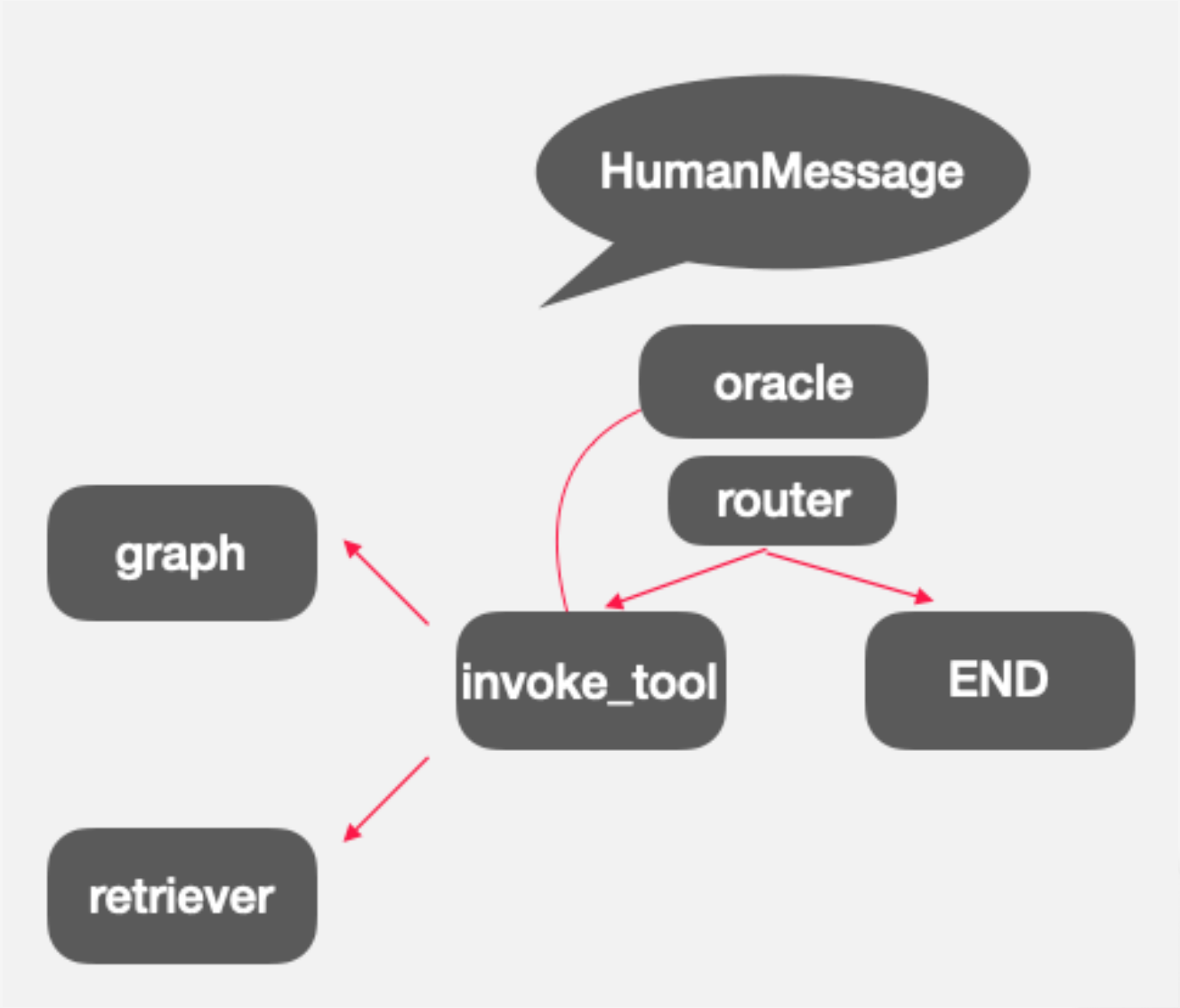
Ollama

```
~ ollama run llama2
~ curl http://localhost:11434/api/generate -d '{
  "model": "llama2",
  "prompt":"What is Machine Learning?"
}'
```

```python
from langchain_community.llms import Ollama

llm = Ollama(model="teacher")

query = "What is Machine Learning?"

result = llm.invoke(query)
print(result)
```

```
{"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


HumanMessage
oracle
router
graph
invoke_tool
END
retriever