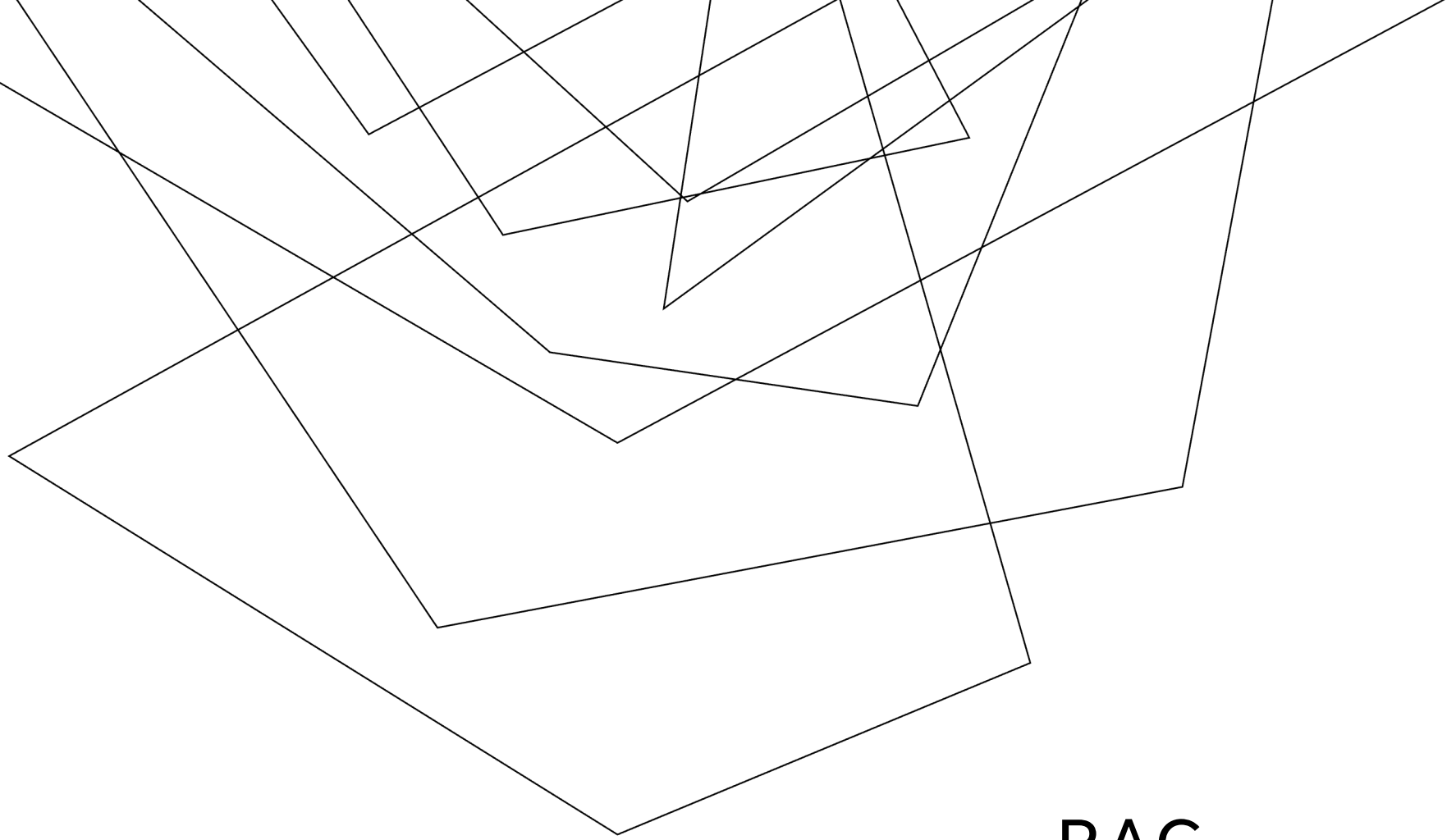
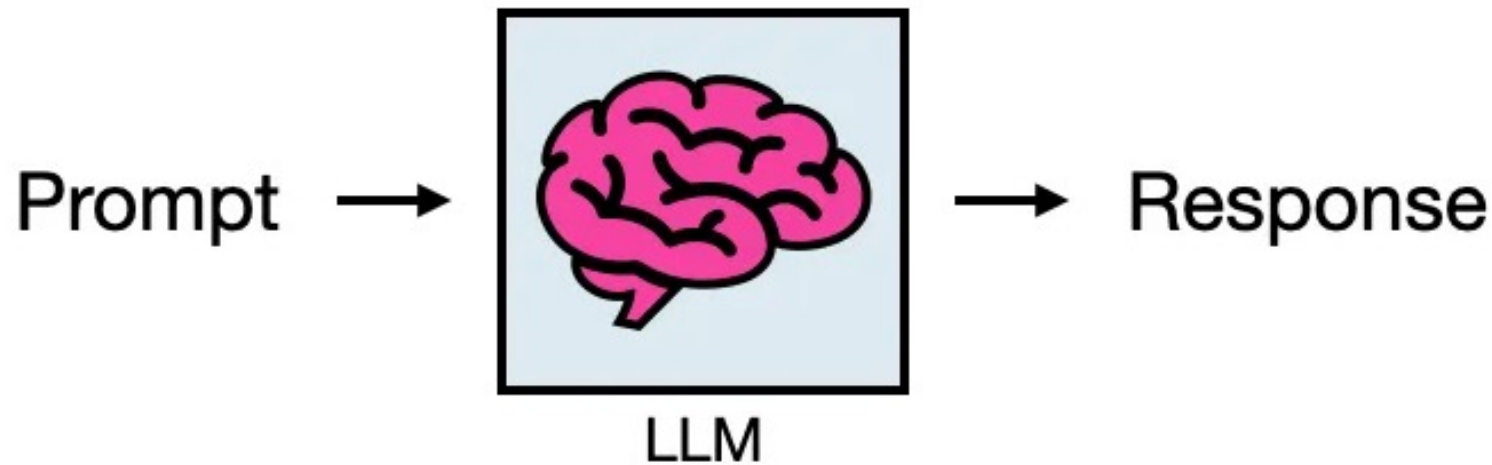
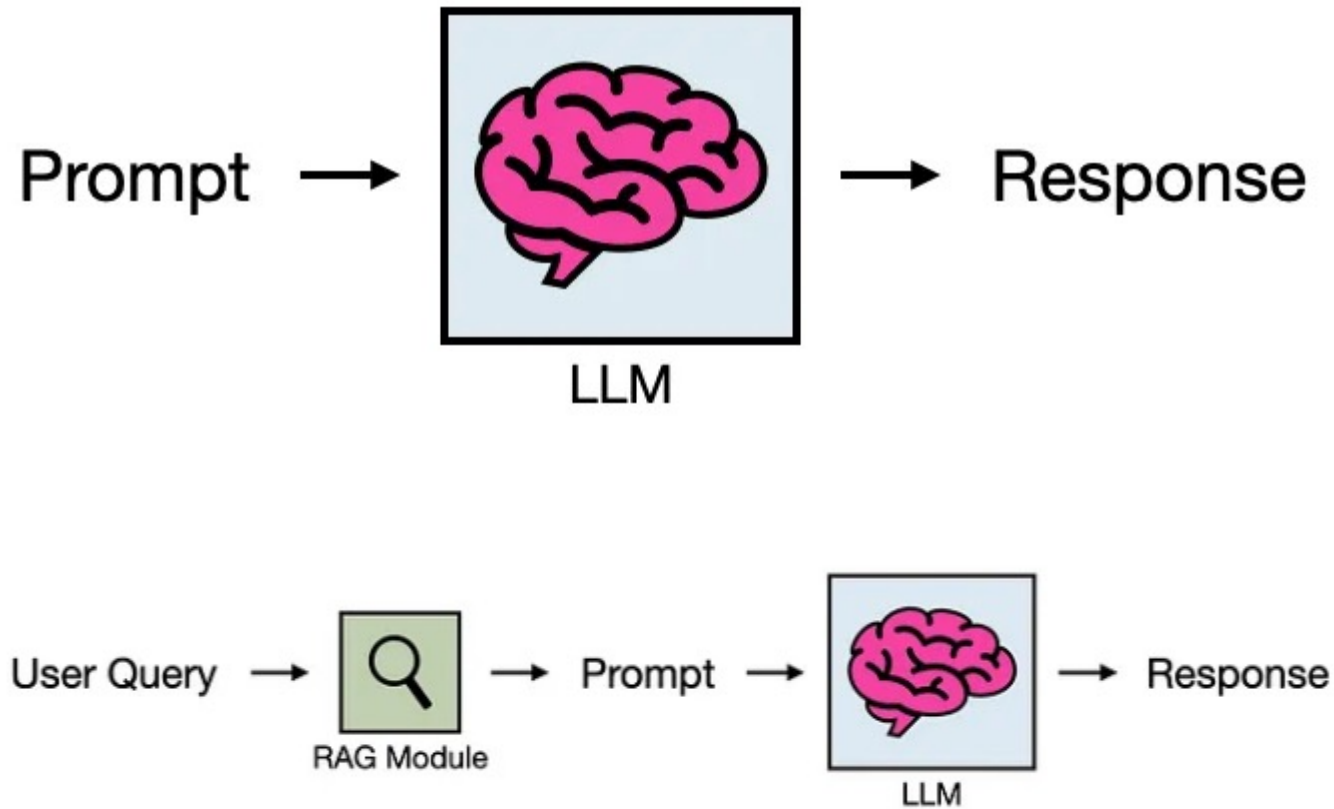


RAG

What is RAG?

The basic usage of an LLM consists of giving it a prompt and getting back a response.

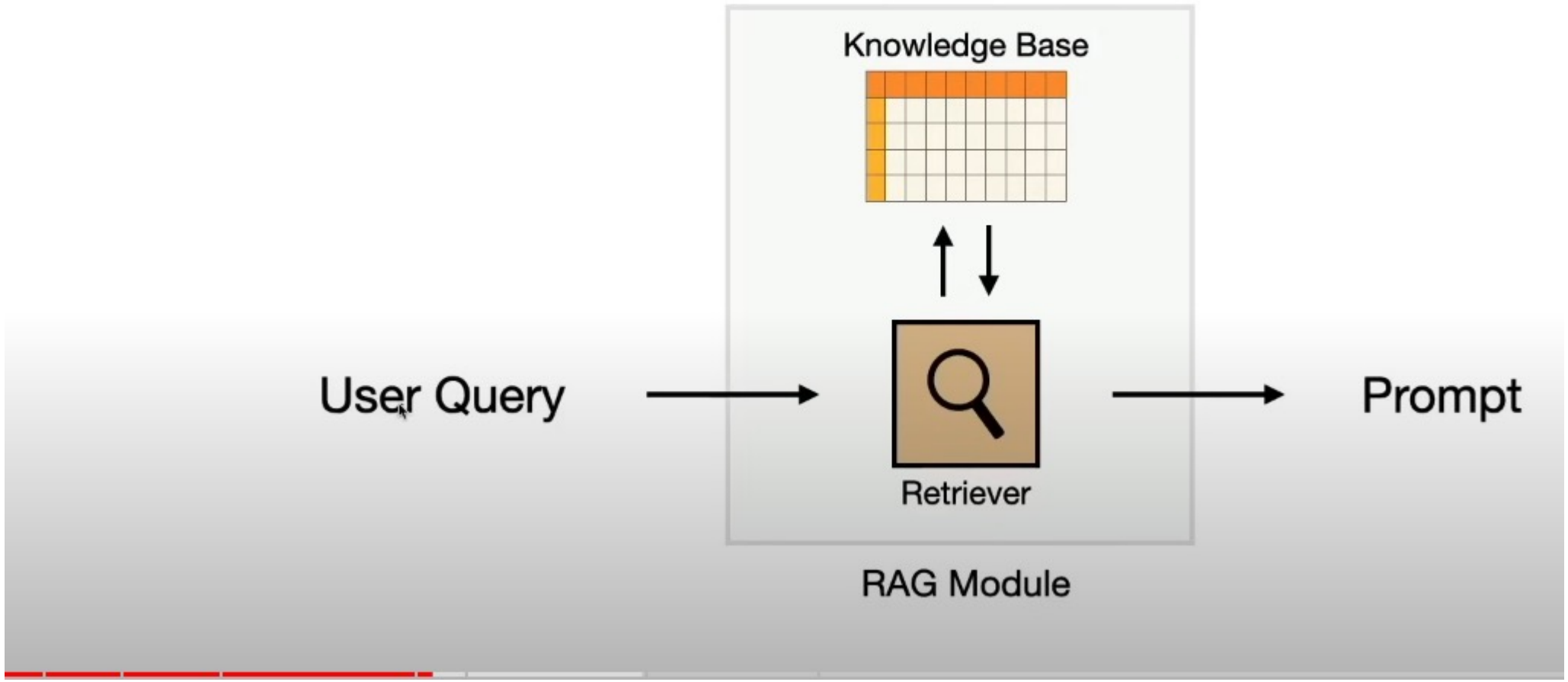




Overview of RAG system. Image by author.

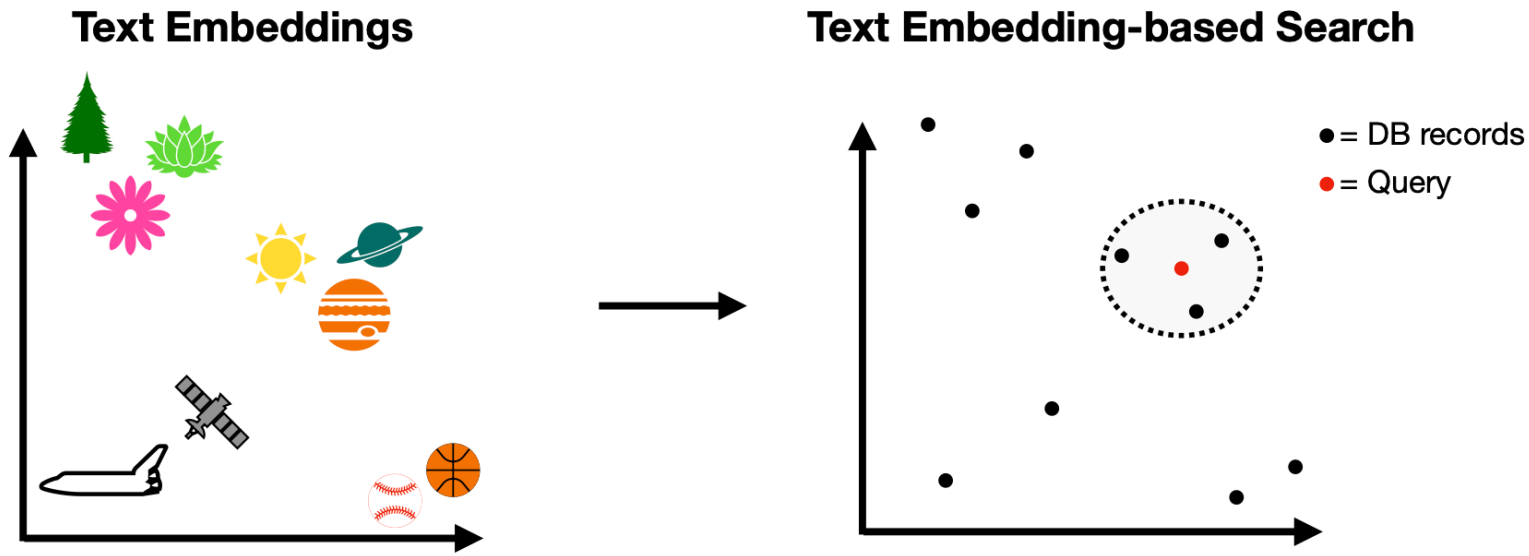
How it works?

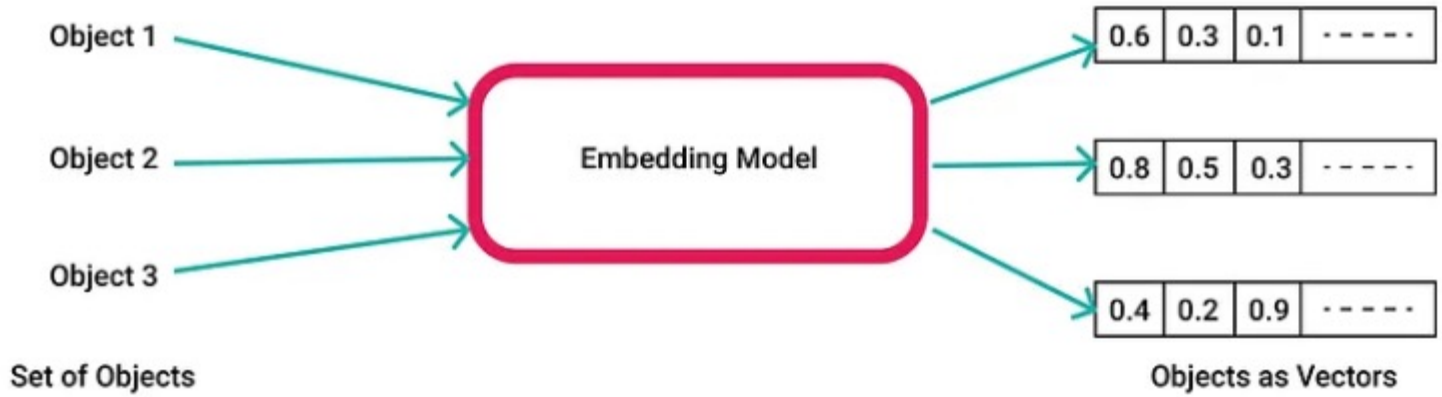
2 key elements: retriever and knowledge base



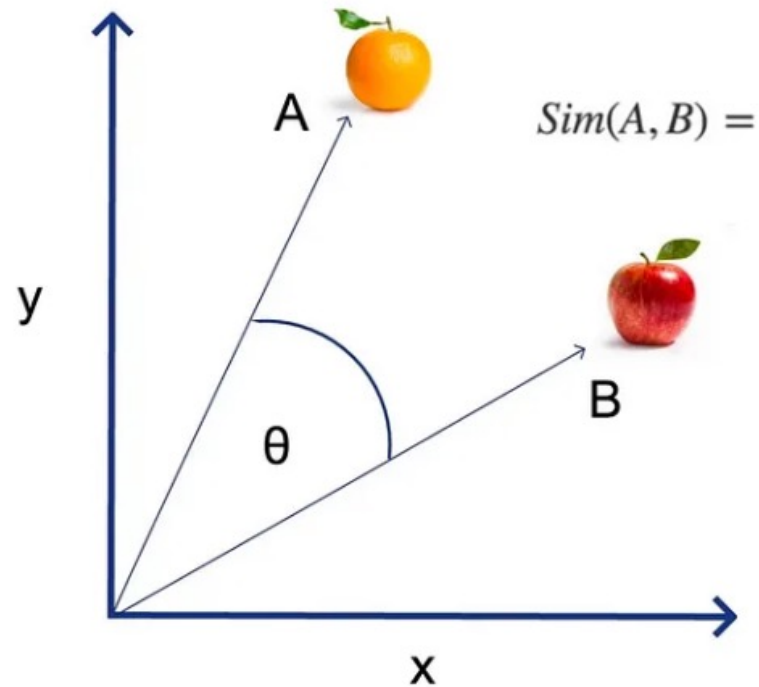
How it works?

2 key elements: **retriever** and knowledge base



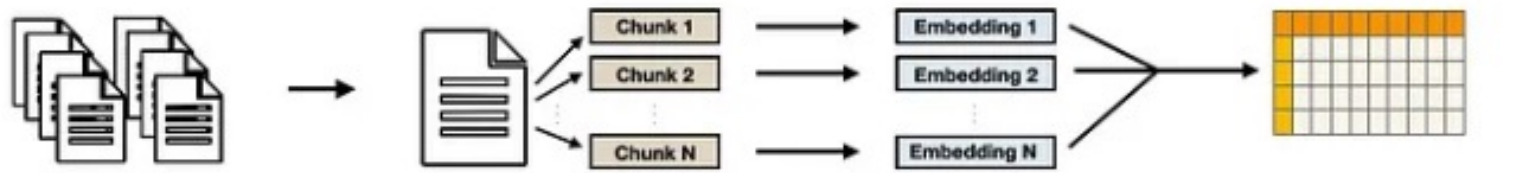


vector embedding



$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|}$$

Cosine similarity



1) Load Docs 2) Chunk Docs 3) Embed Chunks 4) Load into VDB