Boas vindas

DANIEL SÓRIA

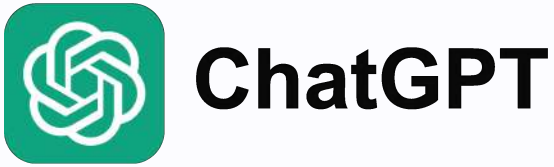
Otimizando LLM’s com RAG

DANIEL SÓRIA

O que são LLM’s?

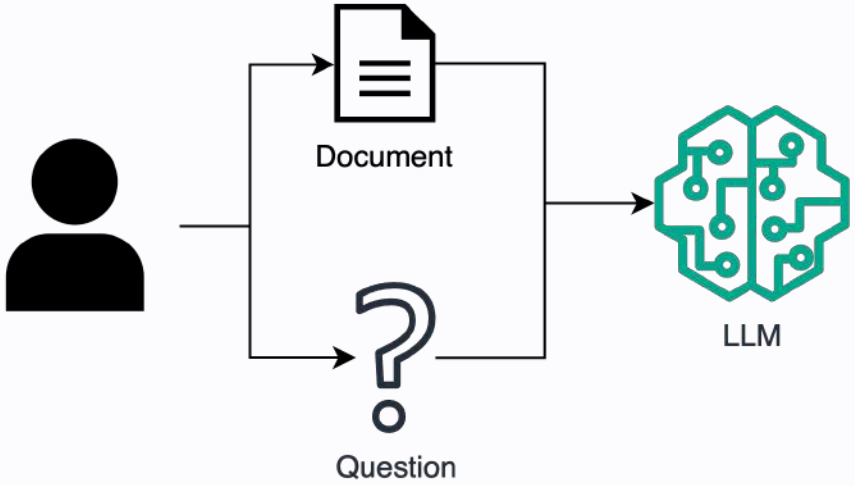
LLM’s - Large Language Models

- Modelos de gerais treinados com uma vasta quantidade de dados. (Large)
- Aprender padrão linguísticos (Language)
- Realizar várias tarefas como tradução, classificação, geração de texto.

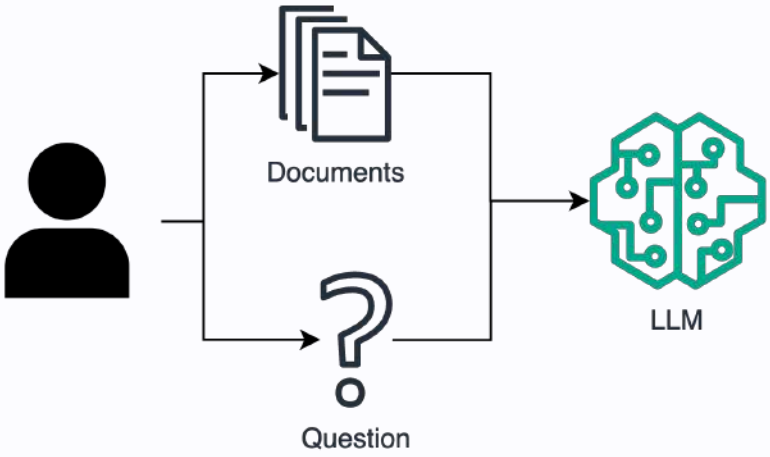


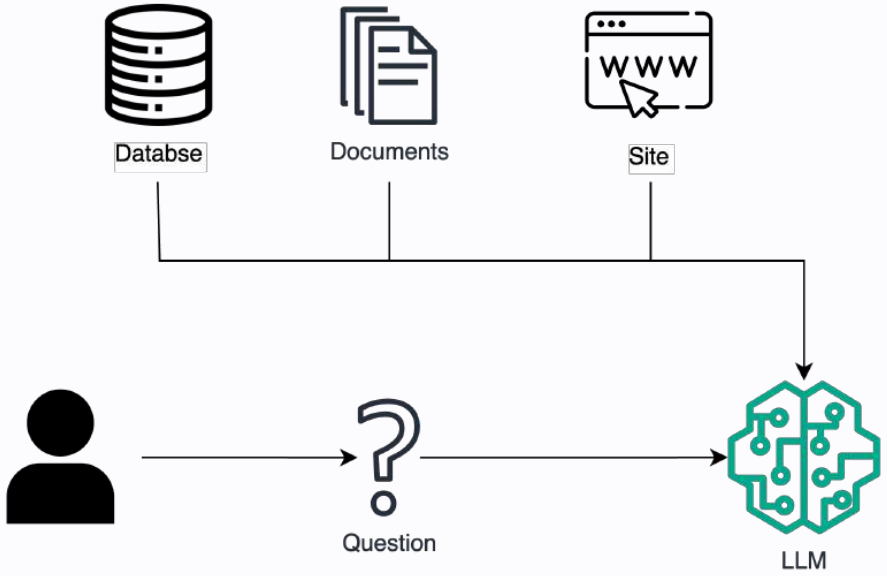
LLM’s - Large Language Models

- Limitados a base (gigantesca) de treinamento
- Dados públicos (ou não) gerais
- Não atendem necessidades específicas de cada empresa/ pessoa



LLM's - Large Language Models





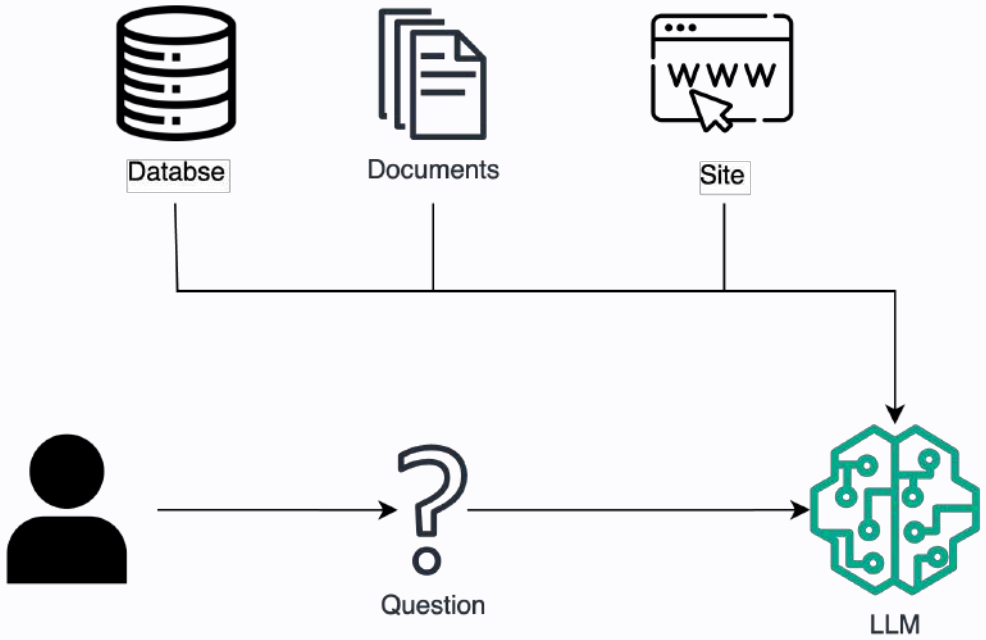
Otimizando LLM’s com RAG

Retrieval

Augmented

Generation

Retrieval-Augmented Generation



Componentes do RAG

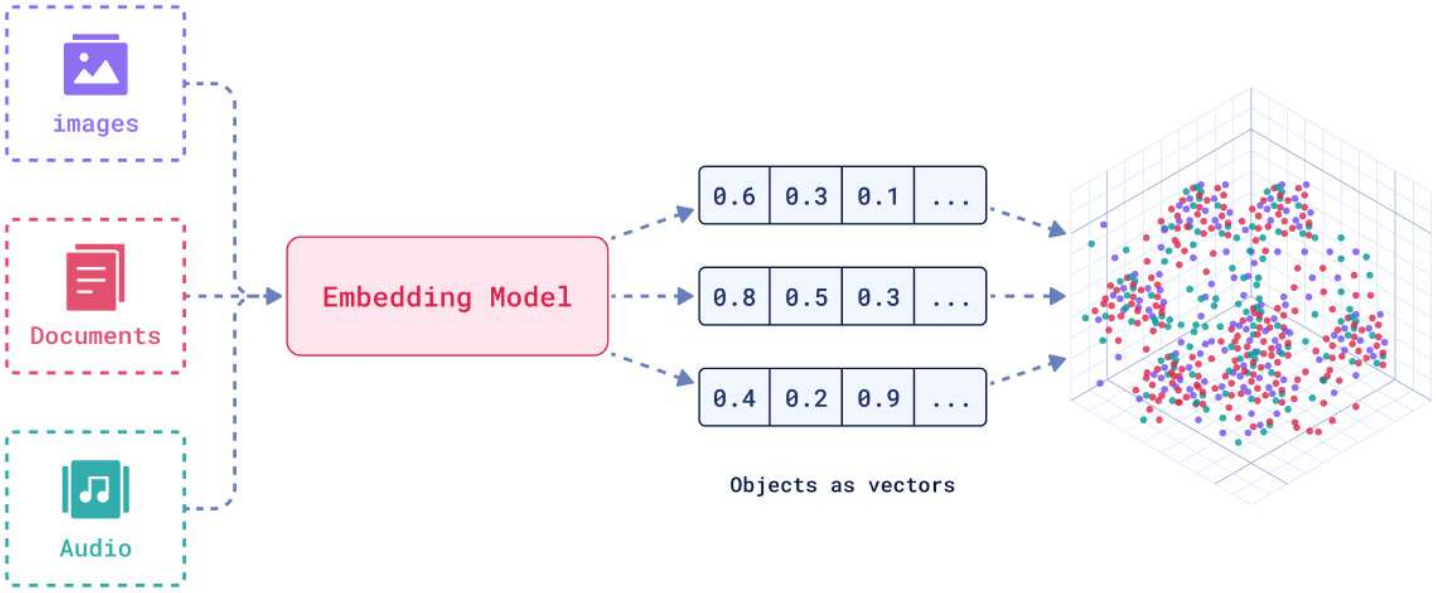
Retrieval-Augmented Generation

Componentes do RAG

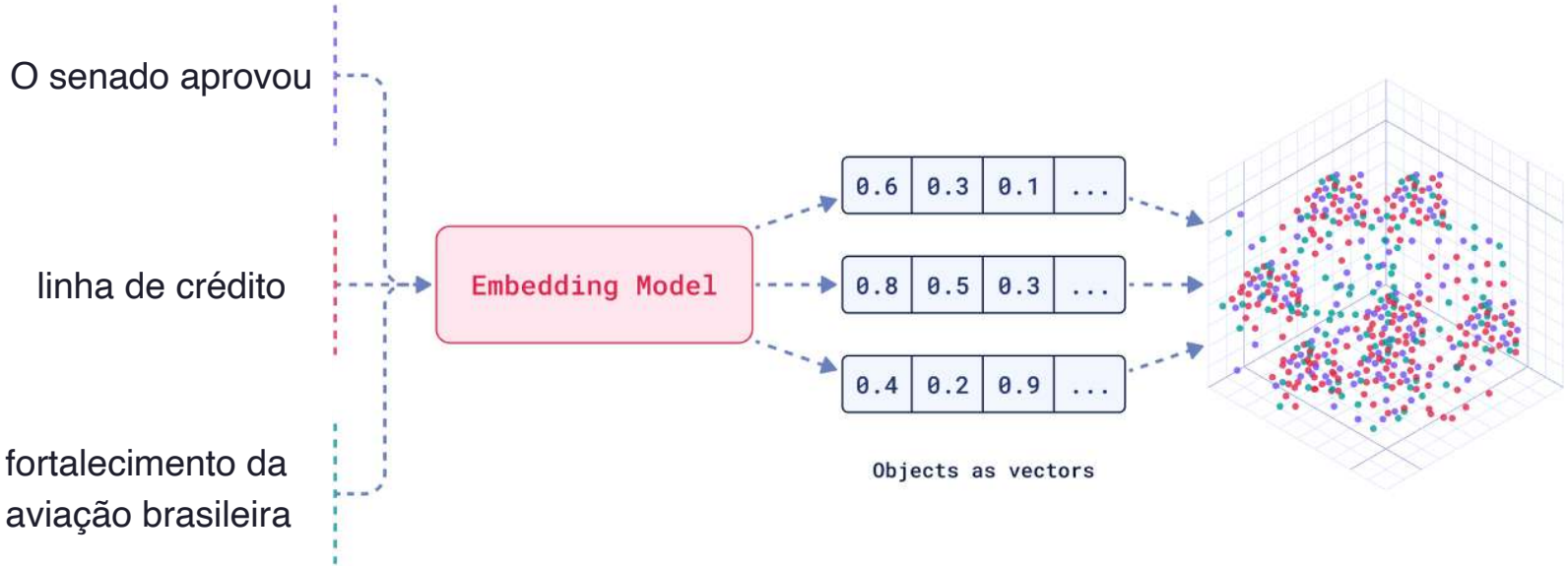
Dados

Tecnologias

Componentes do RAG

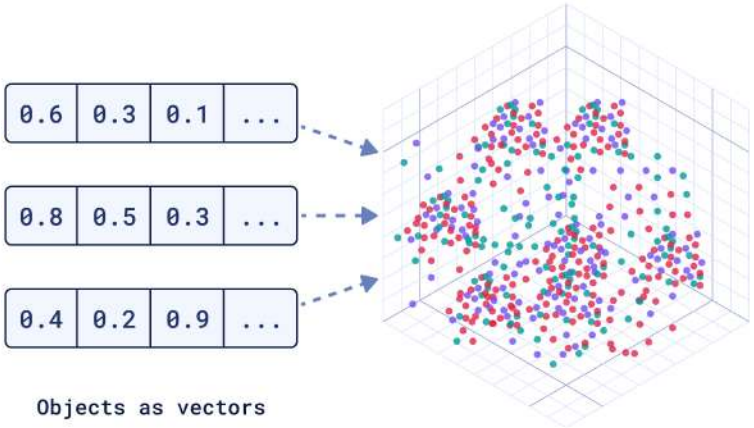


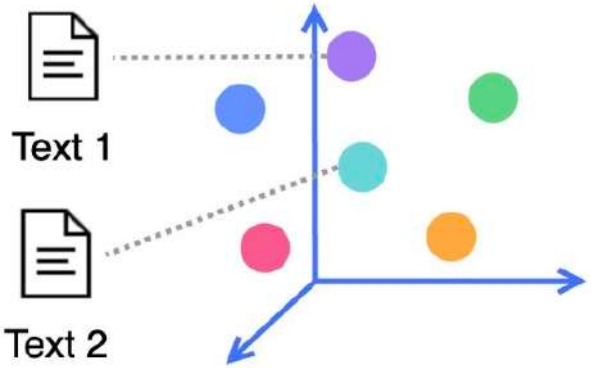
Componentes do RAG



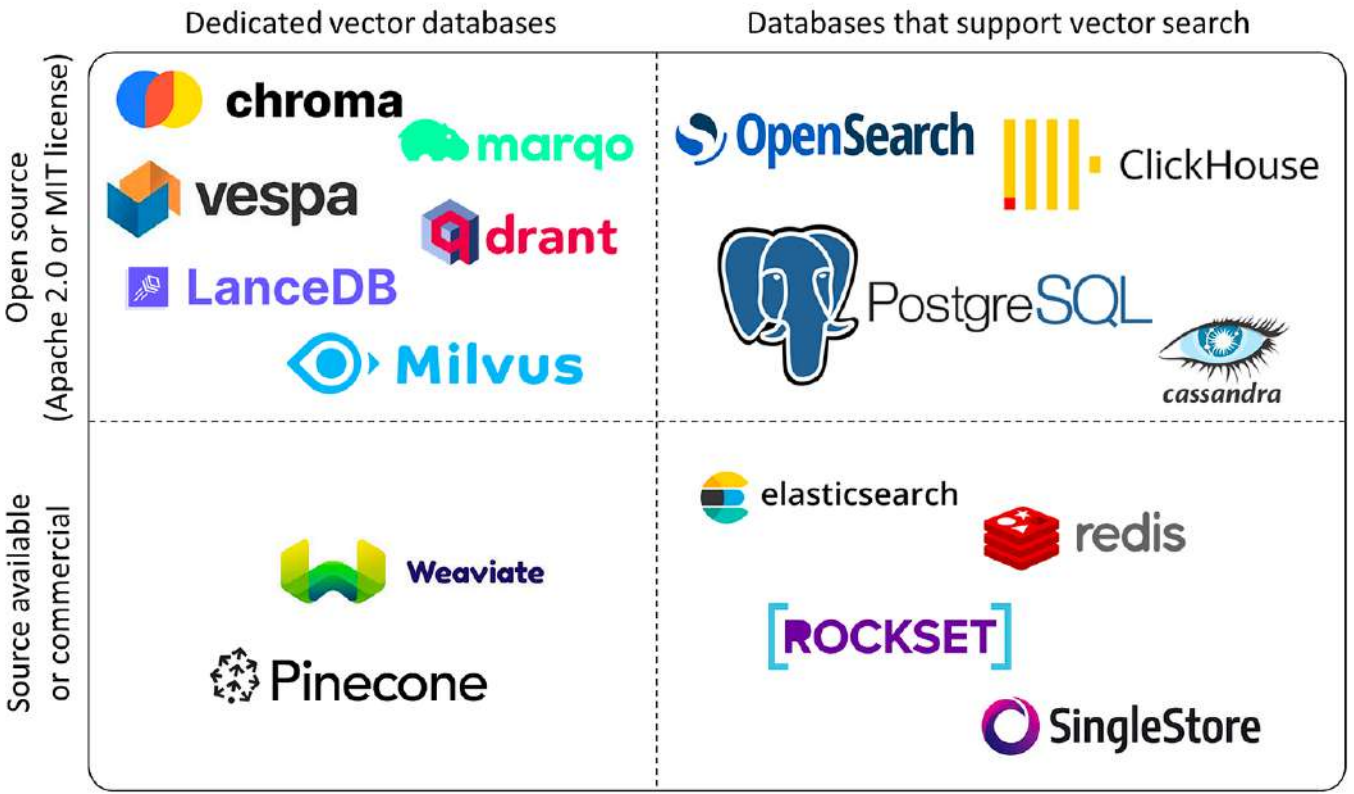
Componentes do RAG

Vector DB





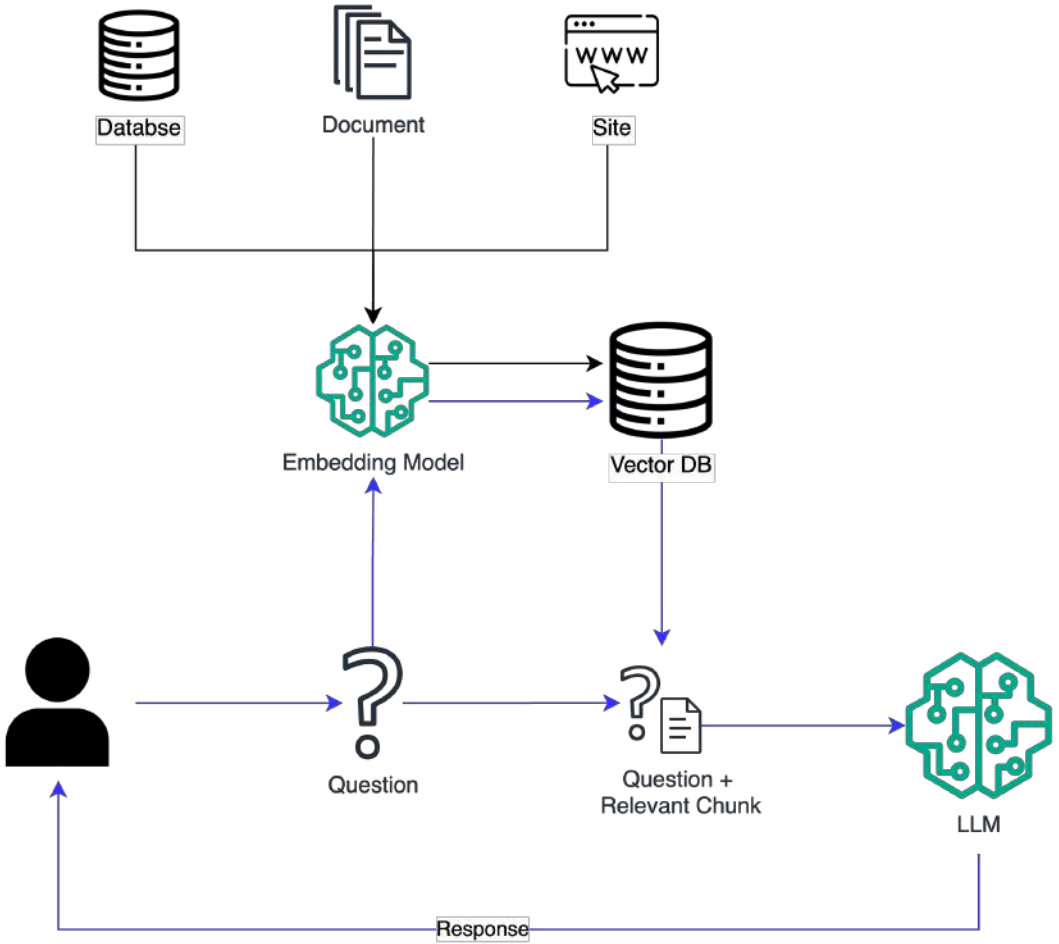
Componentes do RAG



Arquitetura de RAG

Retrieval-Augmented Generation

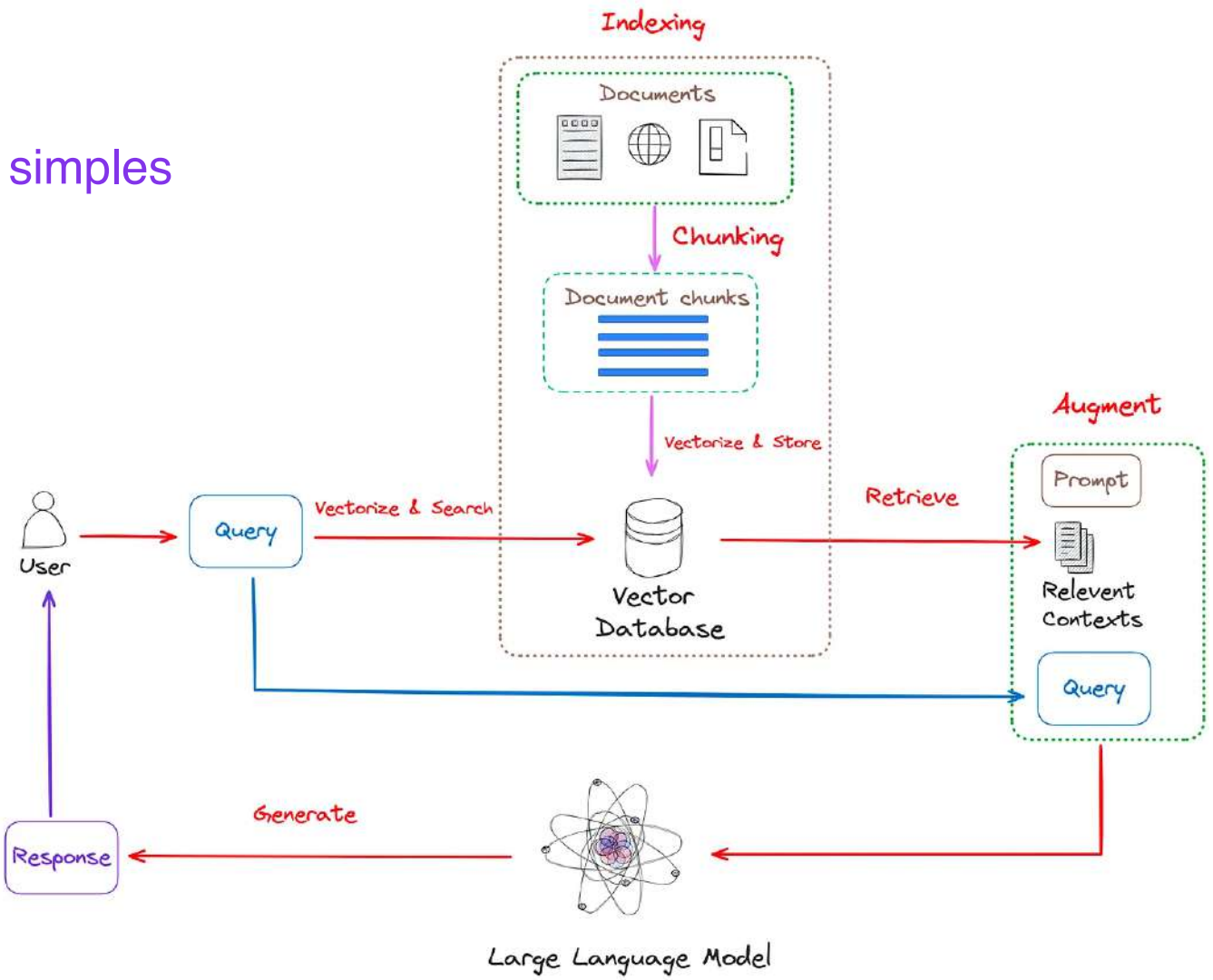
Arquitetura simples de RAG



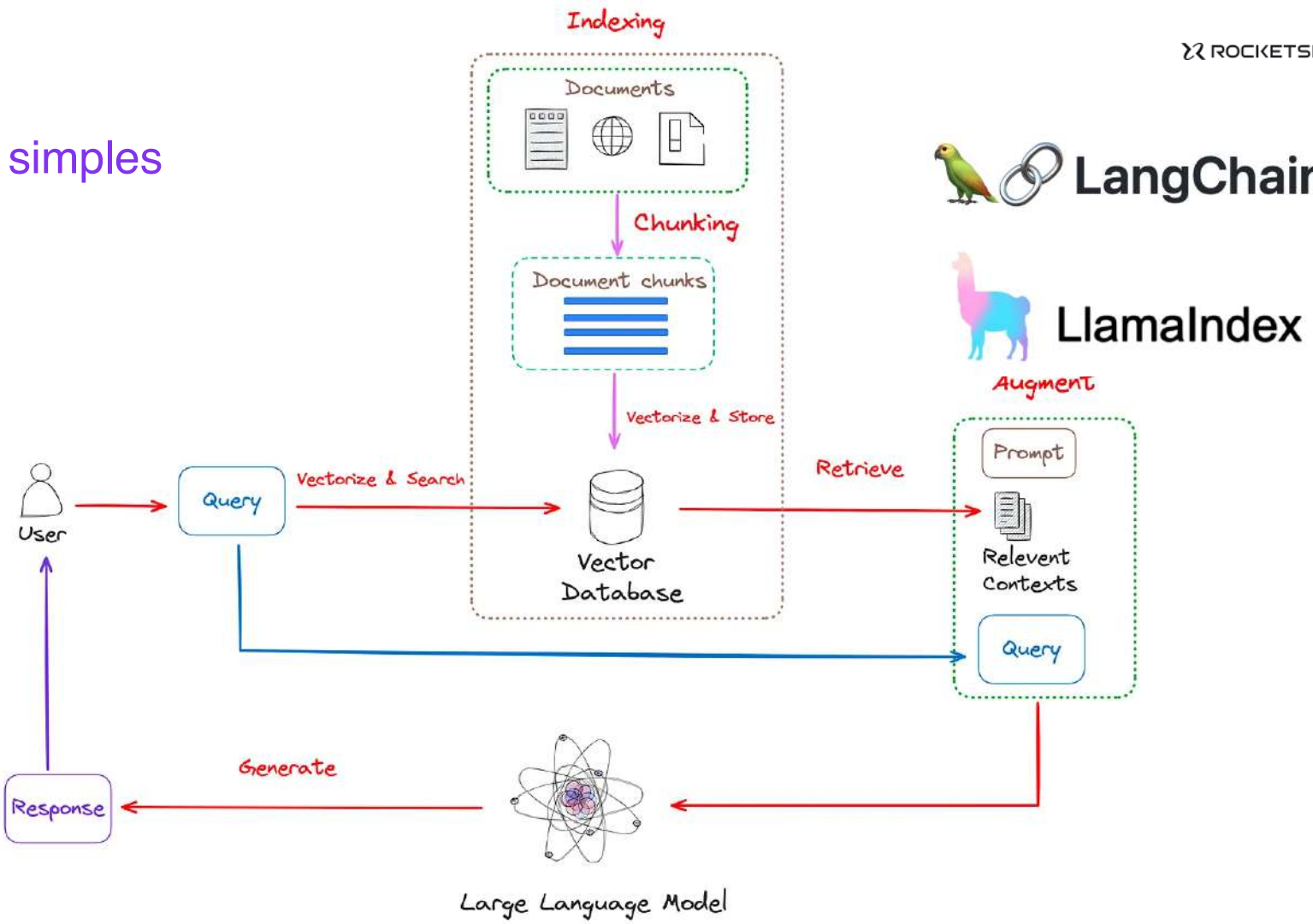
Solução simples de RAG

Ask Your PDF

Arquitetura simples de RAG



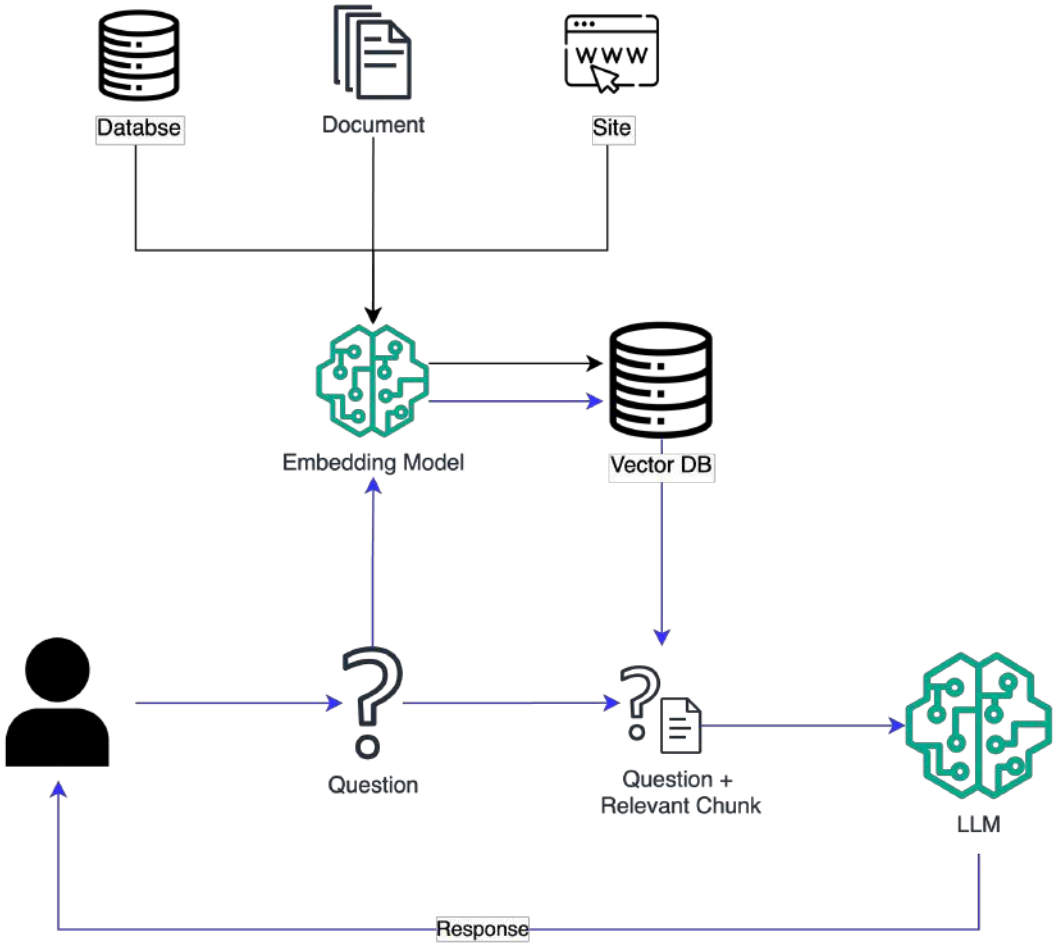
Arquitetura simples de RAG



LangChain

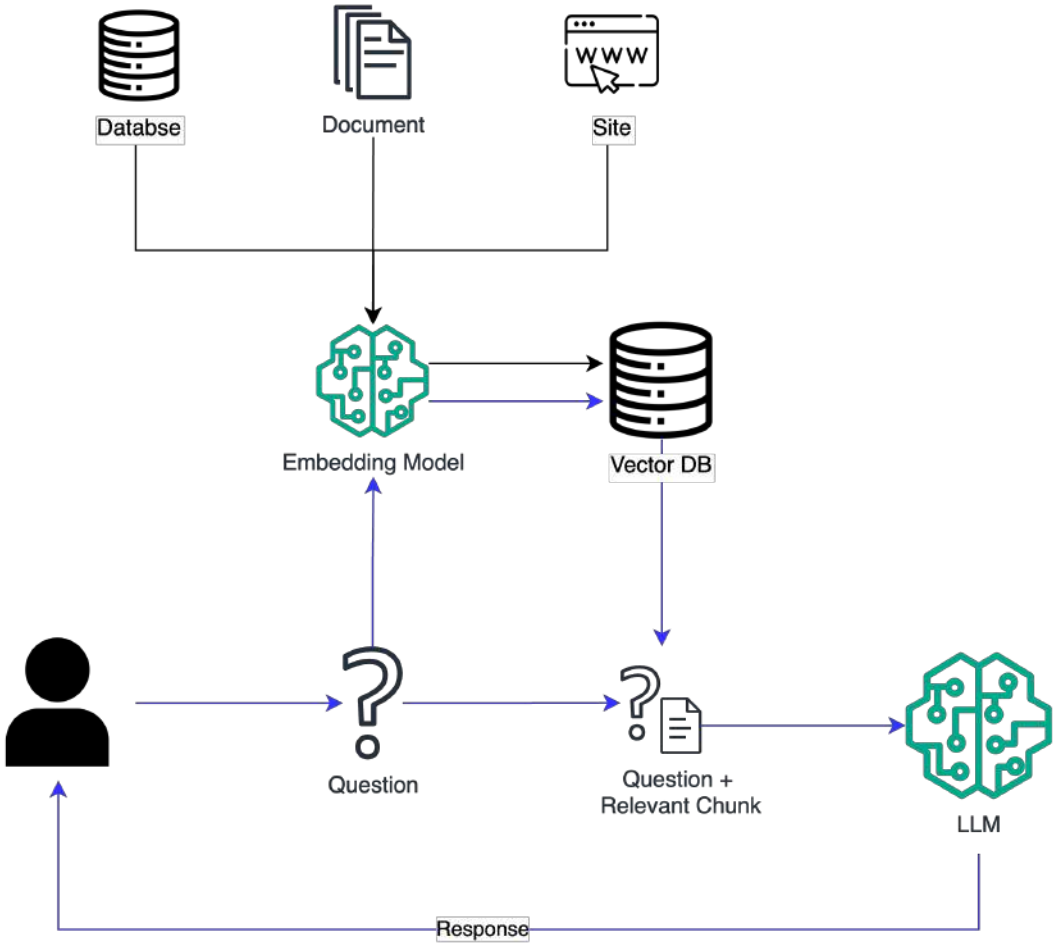
LlamaIndex

Arquitetura simples de RAG

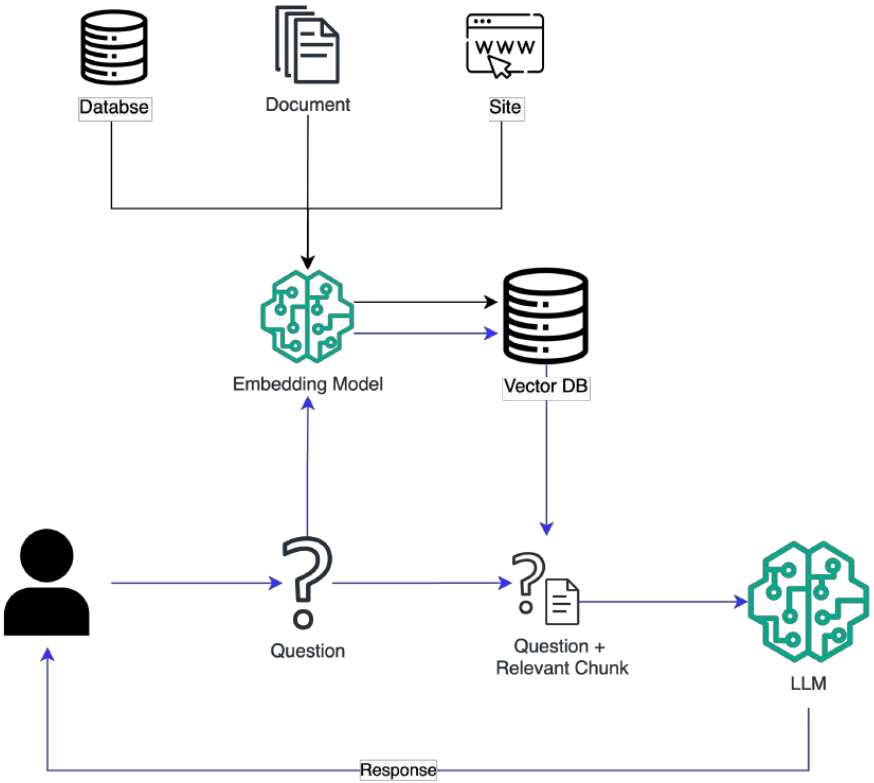


Desafios e limitações do RAG

Arquitetura simples de RAG



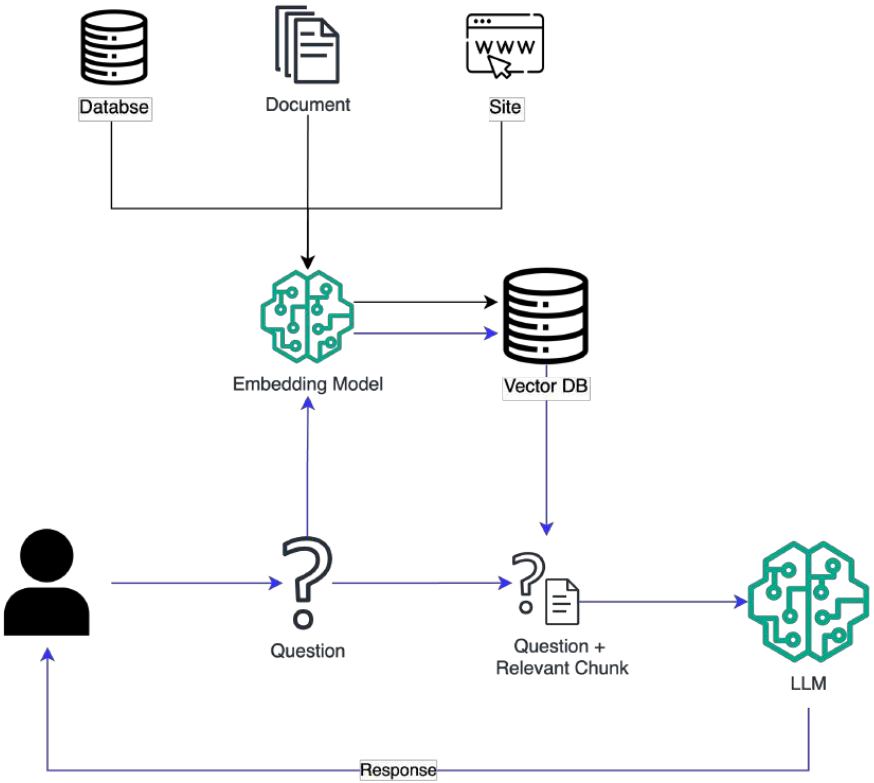
Arquitetura simples de RAG



1. Chunk size e Top-K;
2. Conhecimento de mundo;
3. Perda de informação;

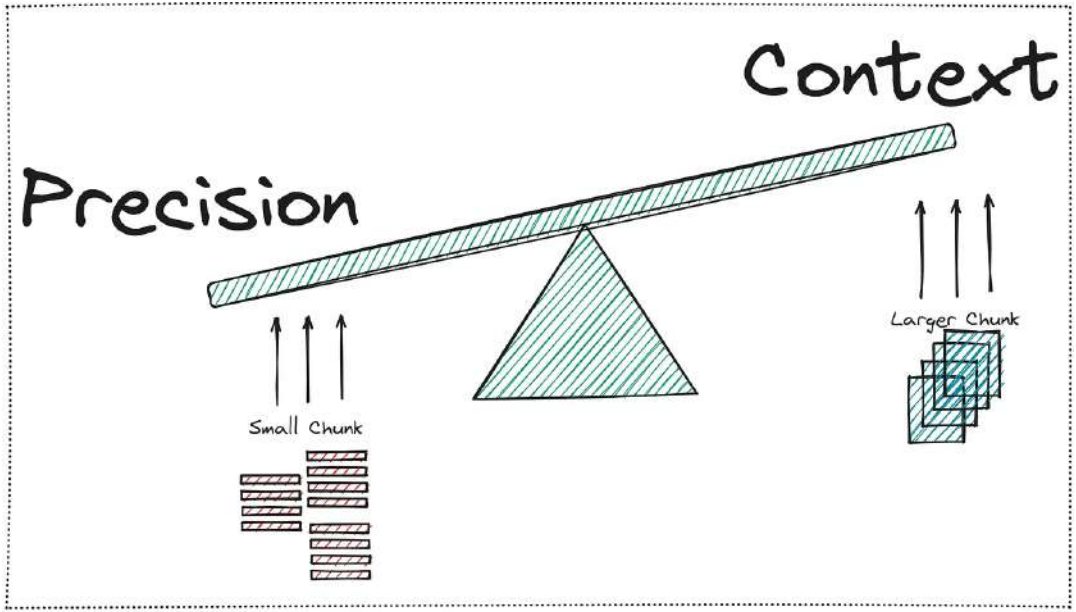
Advanced RAG

Naive RAG



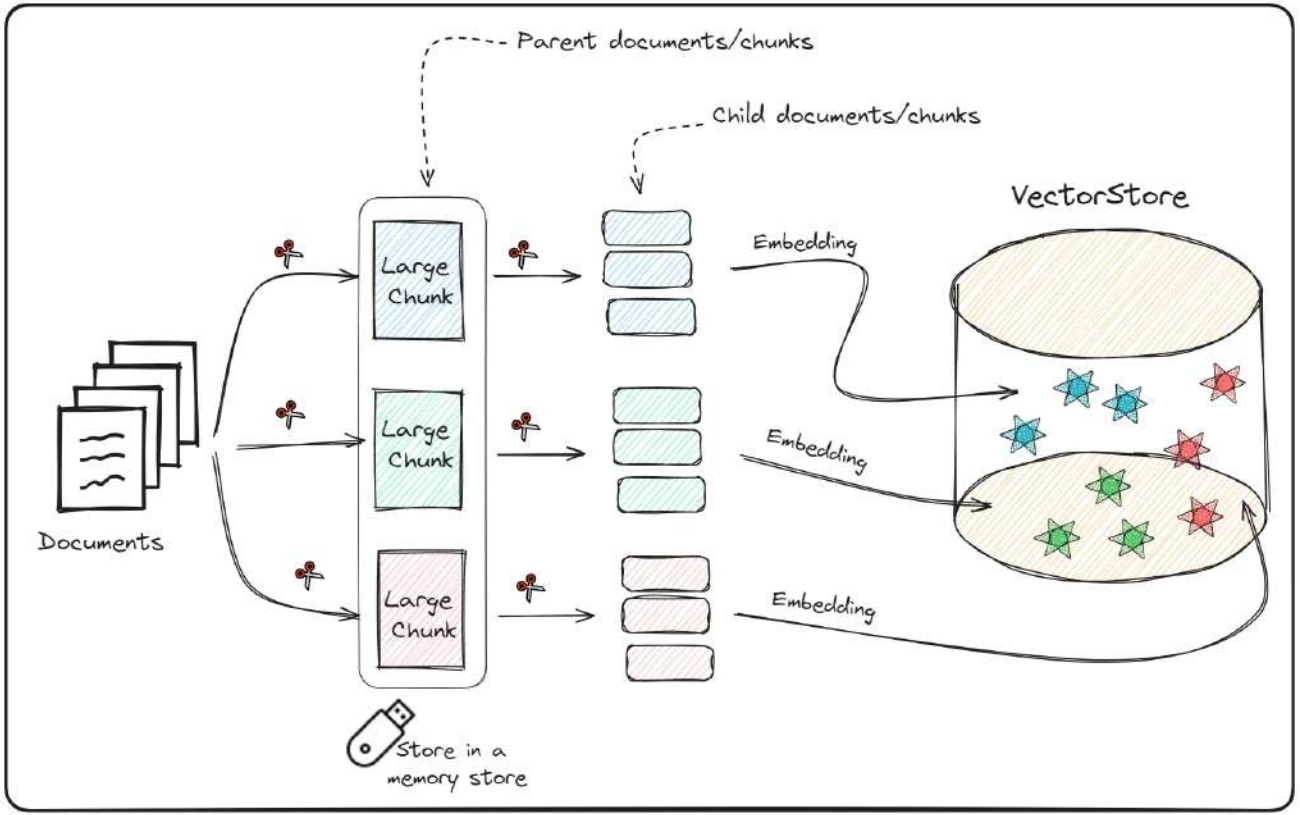
1. Chunk size e Top-K;
2. Conhecimento de mundo;
3. Perda de informação;

Naive RAG



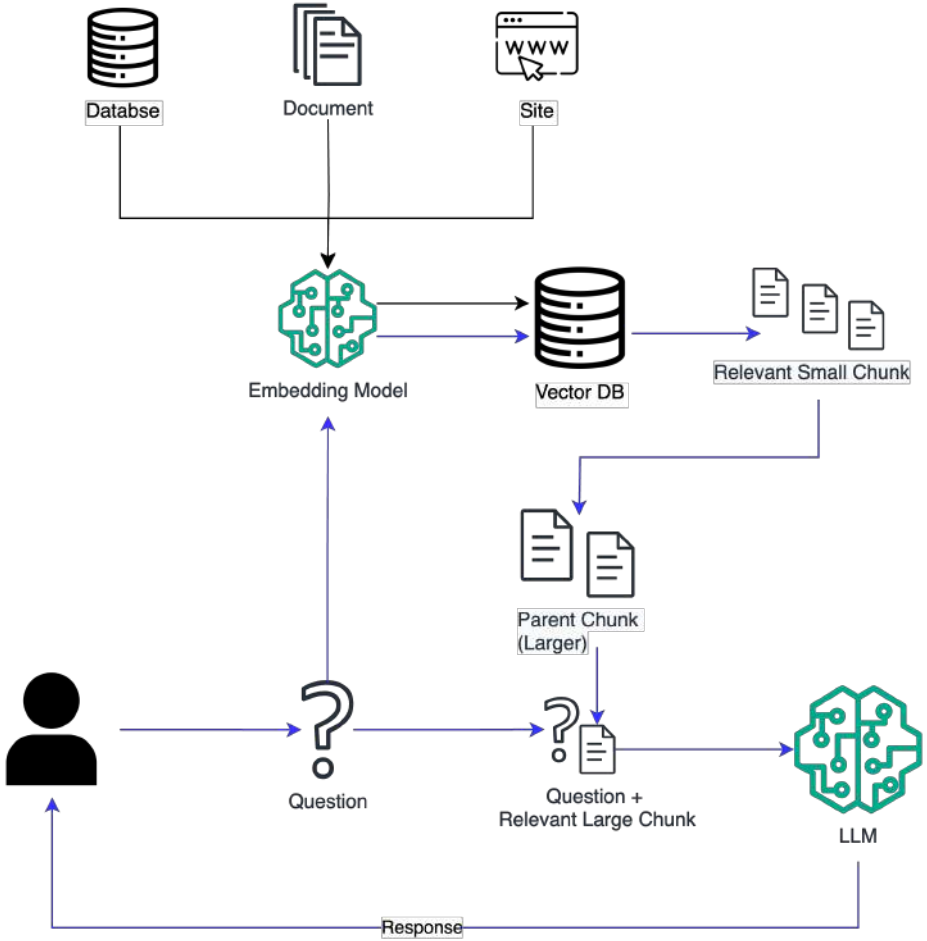
1. Chunk size e Top-K;
2. Conhecimento de mundo;
3. Perda de informação;

Parent Document Retriver

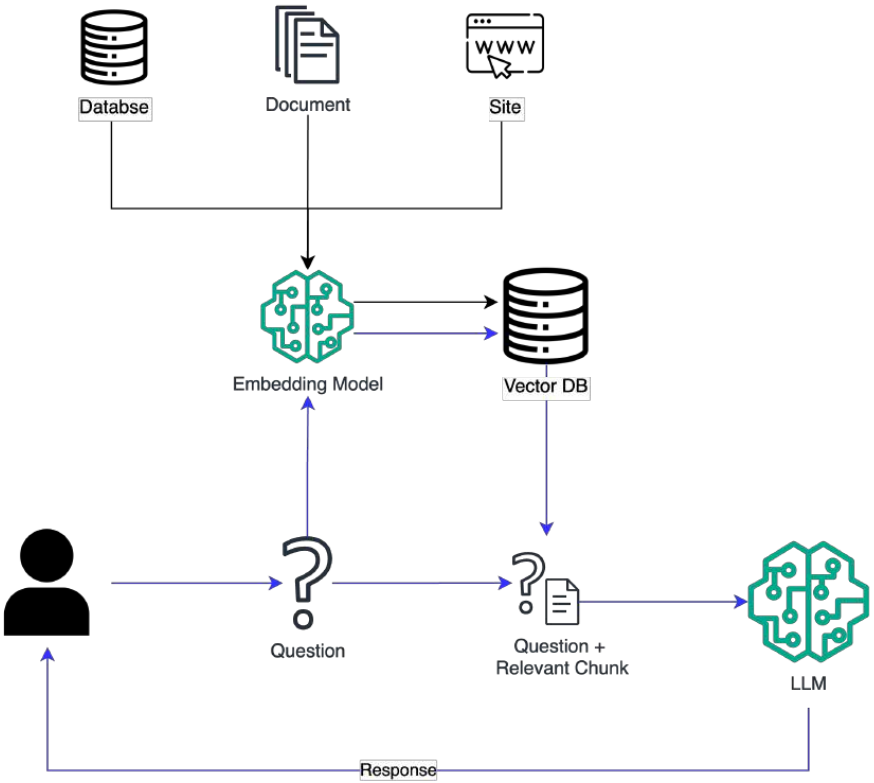


Parent Document RAG

Parent Document Retriever

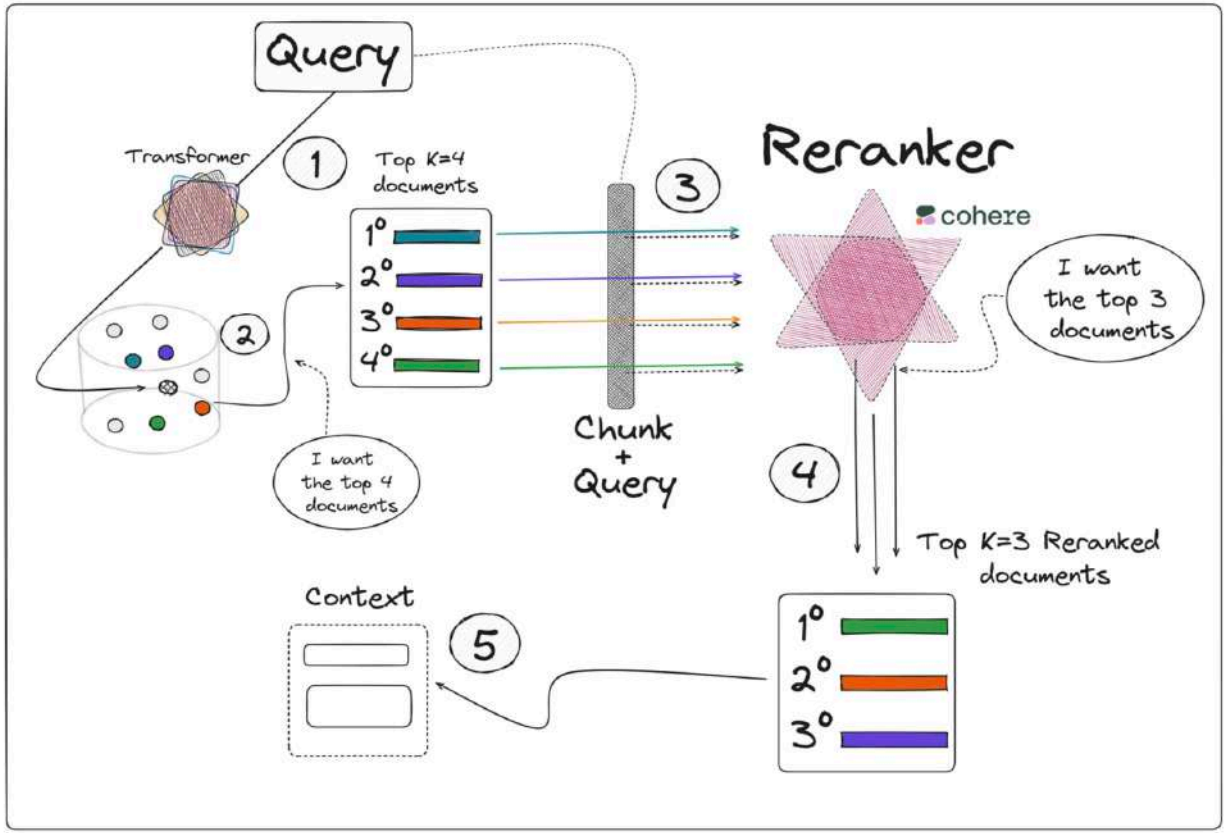


Naive RAG

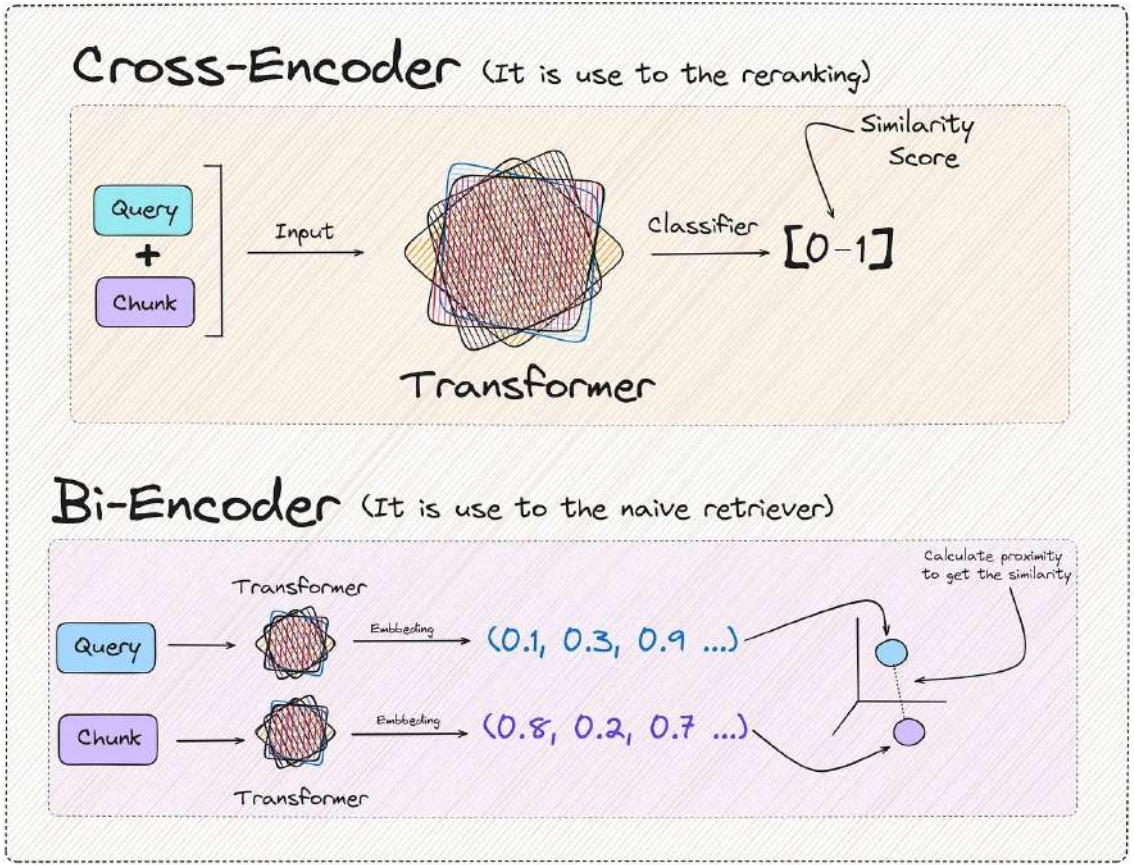


1. Chunk size e Top-K;
2. Conhecimento de mundo;
3. Perda de informação;

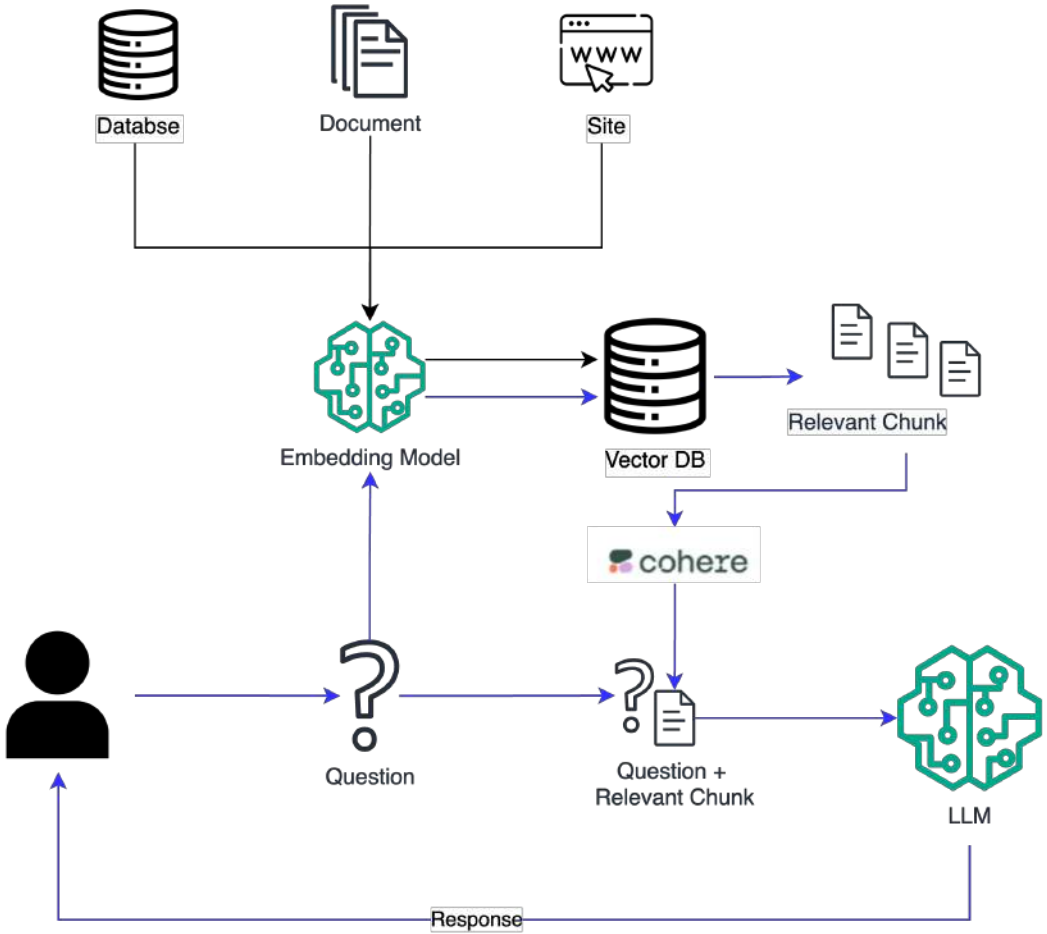
Contextual Compression RAG (Reranker)



Contextual Compression RAG (Reranker)

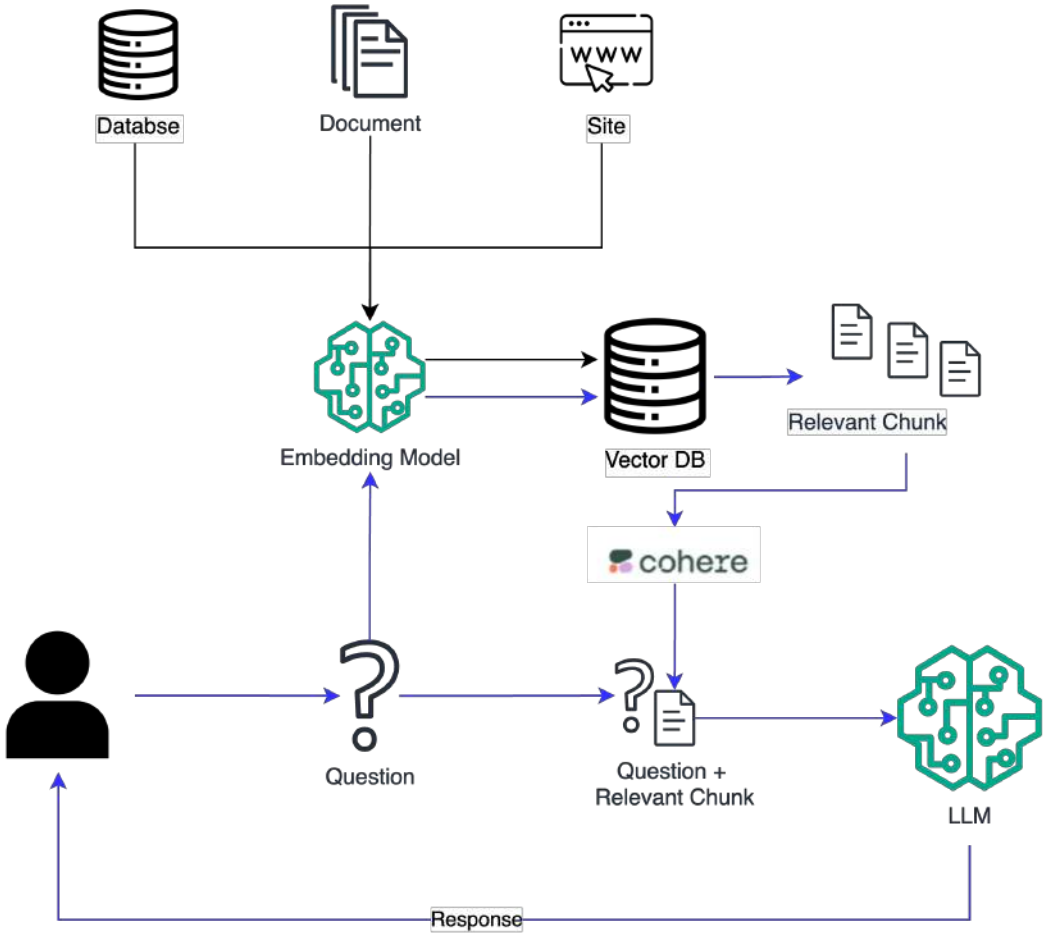


Contextual Compression RAG (Reranker)

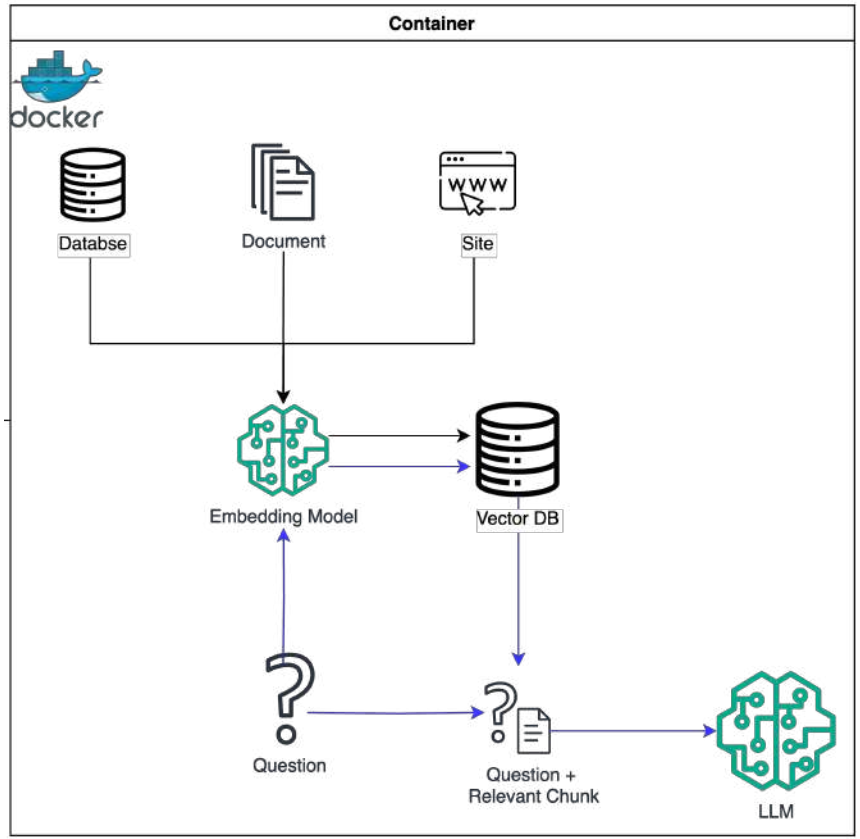


Deploy do RAG na Cloud

Arquitetura simples de RAG



Arquitetura simples de RAG



Arquitetura simples de RAG