博学谷
www.boxuegu.com
黑马程序员线上品牌

物流行业信息咨询RAG系统

一样的教育，不一样的品质

目录

Contents

01

背景介绍

背景介绍

2023年以来，随着ChatGPT的火爆，使得LLM成为研究和应用的热点，但是市面上大部分LLM都存在一个共同的问题：模型都是基于过去的经验数据进行训练完成，无法获取最新的知识，以及各企业私有的知识。因此很多企业为了处理私有的知识，主要借助一下两种手段来实现：

01

利用企业私有知识，
基于开源大模型进行微调

02

利用企业私有知识，基于大模型搭建RAG系统

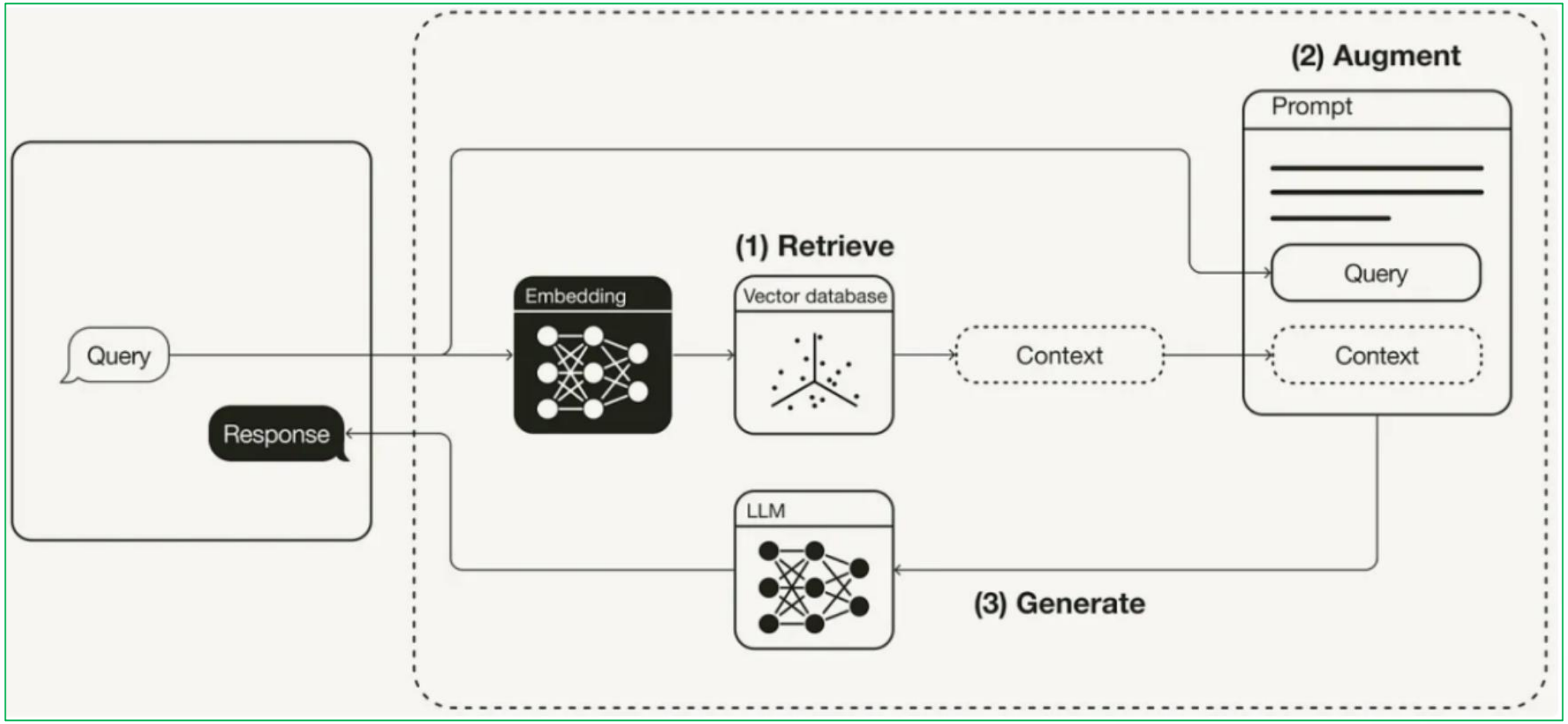
本次项目以"某物流行业"为例，基于物流信息构建RAG系统，测试问答效果。

注意：除物流场景外，使用者可以自由切换其他行业类型知识，实现本地知识库问答的效果。

02

RAG原理

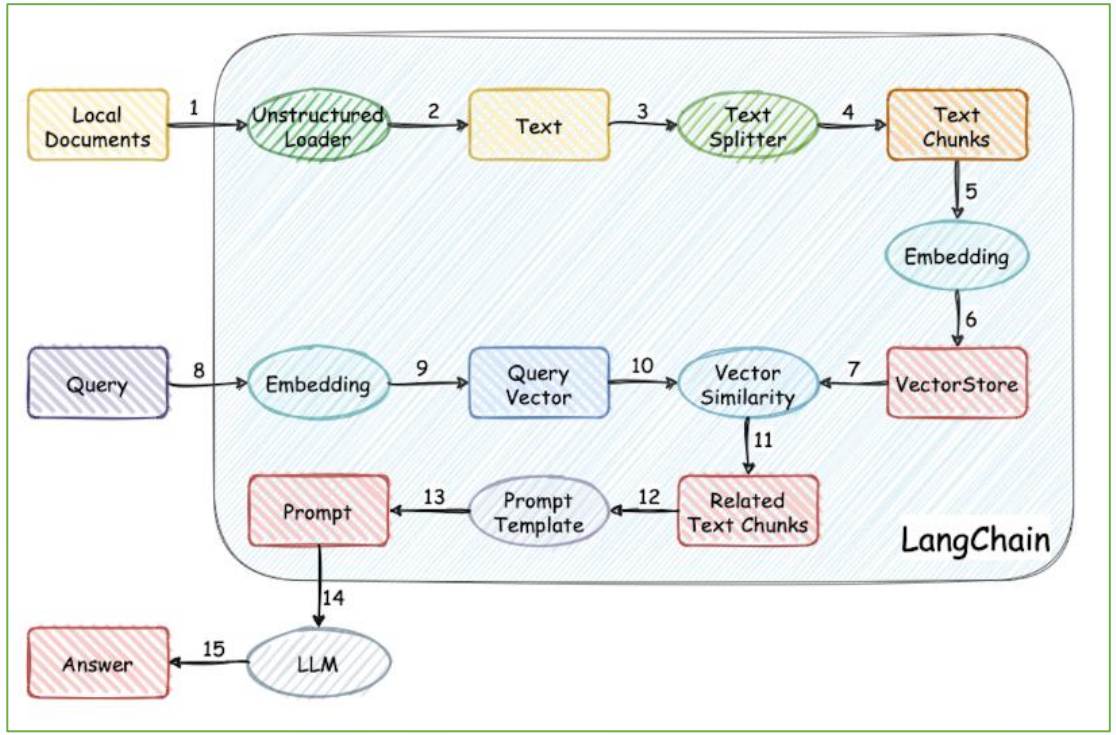
RAG原理



RAG（Retrieval-Augmented Generation）工作原理图

项目流程

原理图



原理描述：

加载文件

读取文件

文本分割

文本向量化

问句向量化

文本向量中匹配出与问句向量相似的top_k个

匹配出的文本作为上下文和问题一起添加到prompt中

提交给LLM生成答案

项目原理

主要功能

基于本地知识库的问答

系统可以根据用户的提问，在本地的知识库中进行搜索，并返回相关的答案

多模型支持

项目支持使用不同的语言模型，可以根据需求选择合适的模型进行使用

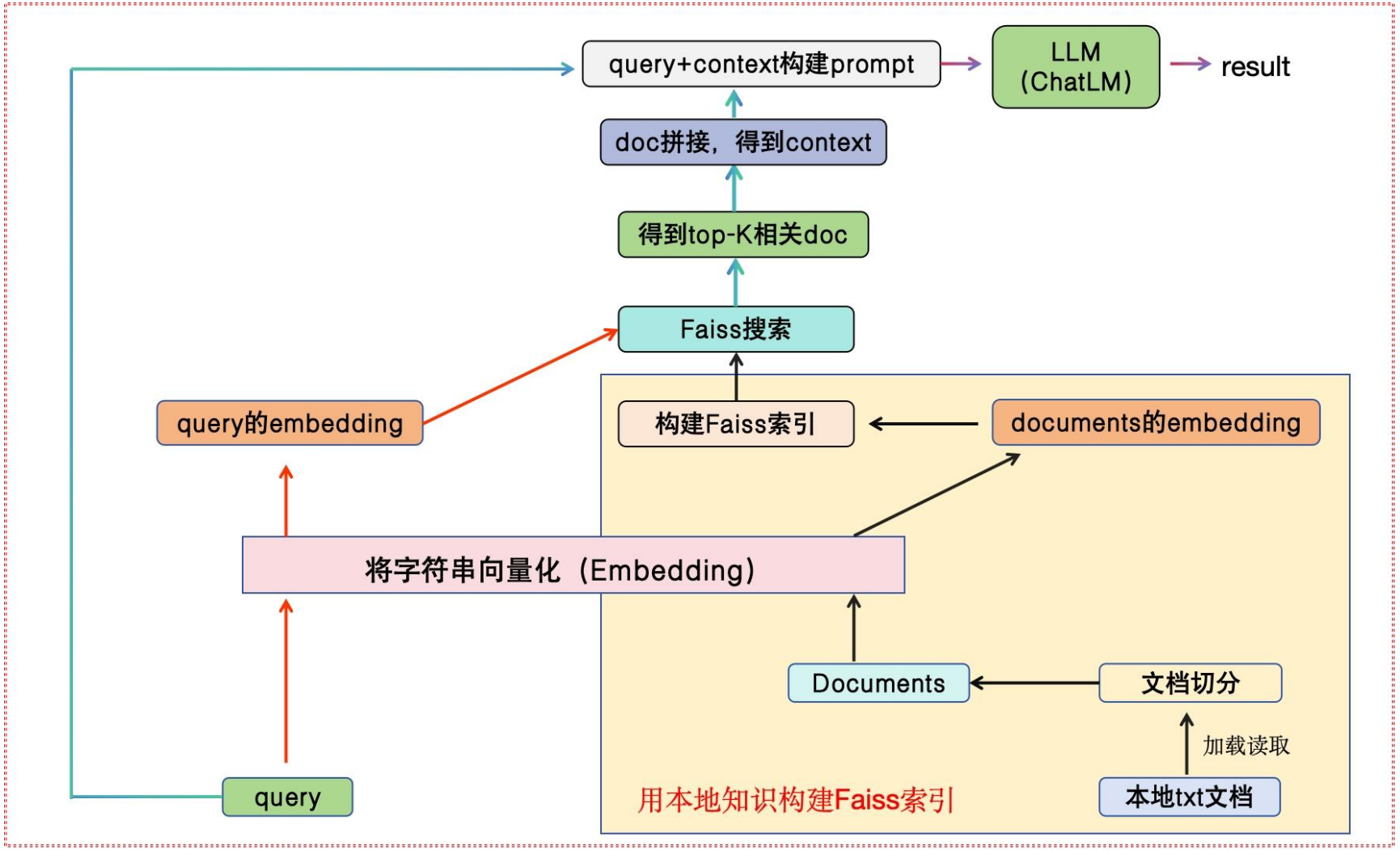
离线私有化

可以将该问答系统部署在本地环境中，确保数据的安全性和隐私性

03

项目流程

项目流程


query+context构建prompt
LLM (ChatLM)
result
doc拼接，得到context
得到top-K相关doc
Faiss搜索
query的embedding
构建Faiss索引
documents的embedding
将字符串向量化（Embedding）
Documents
文档切分
加载读取
query
用本地知识构建Faiss索引
本地txt文档

04

环境配置

环境配置

查看python的版本

确保你的机器安装了Python3.8-Python3.11

```
python --version
```

安装全部依赖

```
pip install faiss-cpu
pip install langchain
pip install ollama
```

模型下载

如需在本地或离线环境下运行本项目，需要首先将项目所需的模型下载至本地，本地将基于Ollama第三方管理框架，实现模型的下载和管理

05

代码实现

代码实现

本地知识库构建

目的

将本地pdf文档信息进行抽取，然后进行分块，最后Embedding存储向量数据库中

代码路径

/Users/**/Knowledge_QA/local_db.py

具体代码：参考pdf附属资料

代码实现

构建
本地问答
RAG系统

目的

基于本地实现RAG系统的检索问答

代码路径

/Users/**/Knowledge_QA/local_qa.py

具体代码：参考pdf附属资料

代码实现

构建web界面实现RAG检索

代码路径

/Users/**/Knowledge_QA/web_qa.py

具体代码：参考pdf附属资料

代码实现—具体代码

◆ 结果展示