



法律助手-基于蓝心大模型的RAG应用

Legal Assistant - RAG application based on the Blue Heart Large Model

汇报人：魏苏州 学号：2120230757



CONTENTS

01

软件背景与意义

Project Background and Significance

02

RAG架构

System Architecture

03

技术路线

System Design


04


软件价值


System Implementation




软件背景与意义

 **背景：**大模型浪潮已经席卷了很多行业，但当涉及到行业细分领域时，通用大模型就会面临专业知识不足的问题。相对于成本昂贵的“Post Train”或“SFT”，基于RAG的技术方案往往成为一种更优选择。

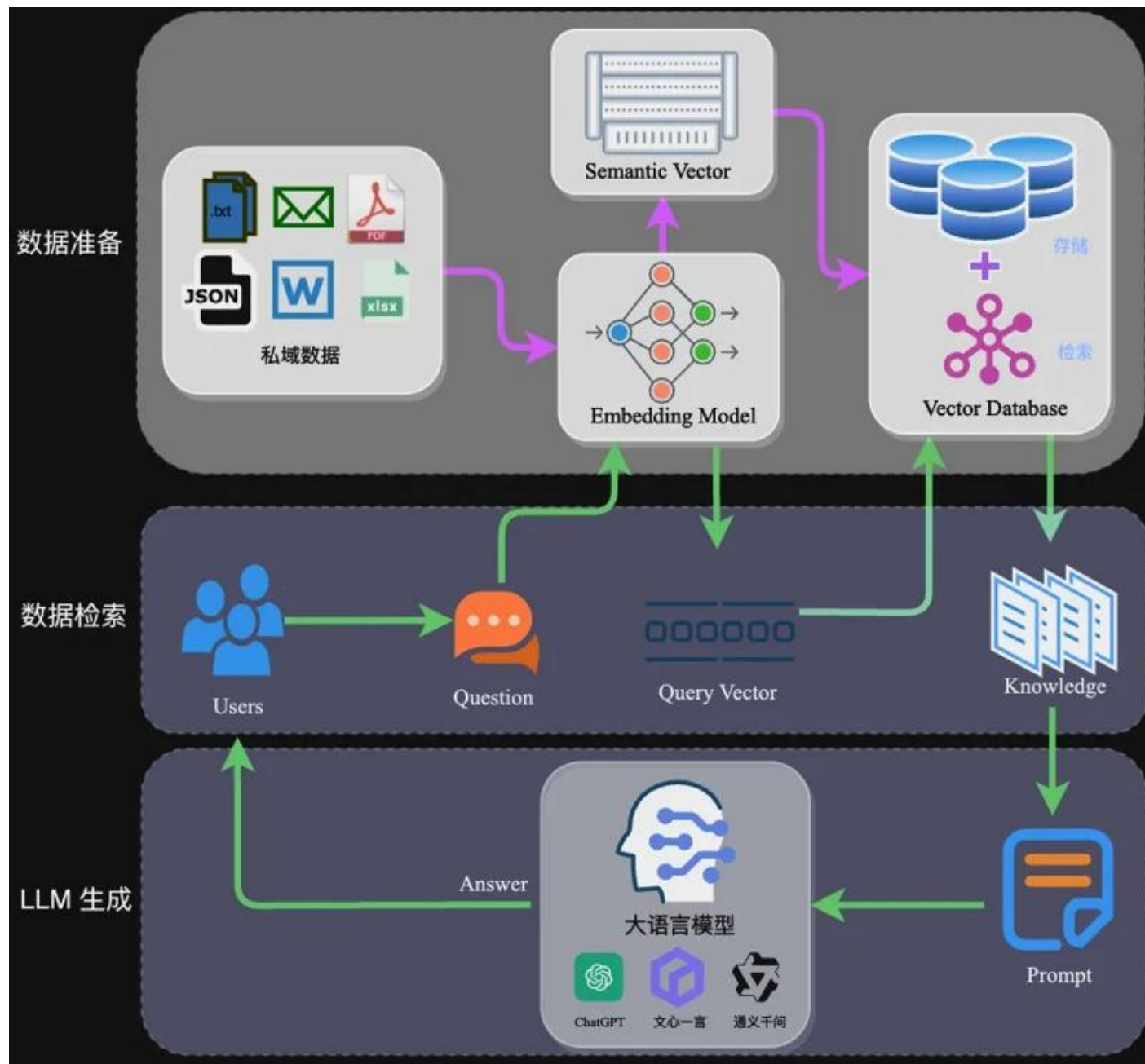
 **知识的局限性：**大模型的训练集基本都是构建于网络公开的数据，对于一些实时性的、非公开的或离线的数据是无法获取到的。

 **幻觉问题：**有时候大模型会一本正经地胡说八道，尤其是在大模型自身不具备某一方面的知识或不擅长的场景。而这种幻觉问题的区分是比较困难的，因为它要求使用者自身具备相应领域的知识。

 **意义：**本软件利用RAG架构，通过检索获取相关的知识并将其融入Prompt，让大模型能够参考相应的知识从而给出合理回答。

RAG结构

RAG的核心理解为“检索+生成”，前者主要是利用向量数据库的高效存储和检索能力，召回目标知识；后者则是利用大模型和Prompt工程，将召回的知识合理利用，生成目标答案。





技术路线

数据准备阶段

数据提取

文本分割

向量化

数据入库

应用阶段

1

用户提问

2

数据检索

3


注入Prompt

4

LLM生成答案





技术路线

 **数据提取**：包括多格式数据加载、不同数据源获取等，根据数据自身情况，将数据处理为同一个范式，包括数据过滤，格式化等。

 **文本分割**：文本分割主要考虑两个因素：1) embedding模型的Tokens限制情况；2) 语义完整性对整体的检索效果的影响。

- 句分割：以“句”的粒度进行切分，保留一个句子的完整语义。常见切分符包括：句号、感叹号、问号、换行符等。
- 固定长度分割：根据embedding模型的token长度限制，将文本分割为固定长度（例如256/512个tokens）。

 **向量化 (embedding)**：向量化将文本数据转化为向量矩阵，该过程会直接影响到后续检索的效果。

 **数据入库**：向量化后构建索引，写入数据库，适用于RAG数据库：FAISS、Chromadb、milvus等。



技术路线

✅ **数据检索：**常见的数据检索方法包括：相似性检索、全文检索等，根据检索效果，一般可以选择多种检索方式融合，提升召回率。

- 相似性检索：即计算查询向量与所有存储向量的相似性得分，返回得分高的记录。常见的相似性计算方法包括：余弦相似性、欧氏距离、曼哈顿距离等。
- 全文检索：全文检索是一种比较经典的检索方式，在数据存入时，通过关键词构建索引；在检索时，通过关键词进行全文检索，找到对应的记录。

💡 **注入Prompt：**Prompt作为大模型的直接输入，是影响模型输出准确率的关键因素之一。在RAG场景中，Prompt一般包括任务描述、背景知识（检索得到）、任务指令（一般是用户提问）等，根据任务场景和大模型性能，也可以在Prompt中适当加入其他指令优化大模型的输出。



软件价值



法律助手

国家层面

- 有助于国家法律法规的完善
- 提供法律依据，辅助案件的判决

社会层面

- 有助于全民普法，形成良好的社会风气
- 提高民众司法观念和增加民众法律知识

教育层面

- 有助于法律条文的学习和教育，减轻机构压力
- 提供学习法律的新渠道

个人层面

- 有助于个人法律意识的增强和知识的学习
- 提供咨询服务，保护和维护自身合法权益



感谢观看，恳请指导

— Thanks for listening and please guide me —

汇报人：魏苏州

