基于个人知识库与检索增强生成机制的导师agent

|  |  |
| --- | --- |
| 项 | 说明 |
| 名称 | 基于个人知识库与检索增强生成机制的导师agent（简称导师agent） |
| 行业 | 计算机，教育，学生 |
| 技术 | 1. 智谱AI大模型API (GLM-4, Embedding-2)； 2. 检索增强生成（RAG）技术； 3. 向量数据库（ChromaDB）；  4. Python后端开发（Flask框架）；  5. 前端交互技术（HTML/CSS/JavaScript）；  6. 文档处理与文本分割技术；  7. API接口设计与调试； |
| 其他 | 本项目聚焦于实现大模型应用的核心逻辑，提供清晰的API接口，学员可在此基础上开发不同风格的前端界面。 | |

# 一． 项目摘要

“导师agent”是一款基于检索增强生成（RAG）技术的垂直领域大模型应用，旨在为计算机专业的学生、教师及爱好者提供一个精准、可靠、专业的智能问答助手。项目通过整合智谱AI的GLM-4大语言模型和Embedding-2文本向量模型，结合自建的计算机专业知识库（涵盖算法、编程、网络、系统等方向），实现了问答内容的有据可依。系统首先将用户问题在本地向量知识库中进行语义检索，找到最相关的资料片段，再将这些片段作为上下文提供给大模型，最终生成专业、准确且不易产生幻觉的答案。本项目有效降低了使用大模型的技术门槛和成本，为教育领域提供了一个可私有化部署的智能辅导解决方案。

# 二． 项目需求

1. **核心需求**：用户能够通过自然语言提问，并获得基于权威计算机专业知识库的精准答案，而非大模型的通用回答。
2. **知识库管理需求**：支持管理员上传、更新和管理专业知识文档（如PDF、Word、TXT格式的教材、文档），系统能自动处理并存入向量数据库。
3. **问答性能需求**：响应速度需在可接受范围内（通常数秒内），答案需结构清晰、专业性强，并可追溯答案来源。
4. **系统扩展需求**：架构设计应允许轻松扩展知识库领域（如从计算机扩展至数学、物理），或更换底层大模型供应商。
5. **用户体验需求**：提供简洁友好的Web界面，支持多轮对话，交互过程流畅。

# 三． 项目功能

## 1. 功能结构

* **知识库处理模块**：负责文档的加载、文本分割、向量化与存储。
* **语义检索模块**：接收用户问题，将其向量化并从知识库中检索最相关的片段。
* **RAG生成模块**：构建Prompt，协调检索结果与大模型生成答案。
* **Web交互模块**：提供用户界面和API接口，实现用户与Agent的交互。

## 2. 功能介绍

* **智能专业问答**：解答计算机科学领域的复杂问题，如解释概念、分析算法、调试代码思路等。
* **知识库检索与引用**：所有回答均基于本地知识库生成，并可展示答案引用的来源片段，增强可信度。
* **多格式文档支持**：支持将PDF、Word、TXT等多种格式的教材和资料一键录入知识库。
* **Web交互界面**：提供简洁的聊天式界面，用户可直观地提问和查看回答。
* **管理员后台**（可选）：允许管理员监控系统状态、管理知识库文档、查看问答日志。

# 四．项目使用的技术与学习建议大纲

|  |  |
| --- | --- |
| 技术名称 | 推荐学习资料 |
| Python编程 | 《Python编程：从入门到实践》 |
| 大模型与RAG技术 | 智谱AI官方文档、OpenAI API文档、Postman接口测试工具 |
| 检索增强生成（RAG） | 《RAG Survey》论文、LangChain中文文档、LlamaIndex官方教程 |
| 向量数据库 | Chroma官方文档、Milvus入门教程、向量数据库技术白皮书 |
| Flask Web开发 | Flask官方文档、《Flask Web开发实战》、CSDN Flask实战博客 |
| 前端基础 | MDN Web文档、W3School前端教程、Bootstrap框架文档 |

# 五．附录