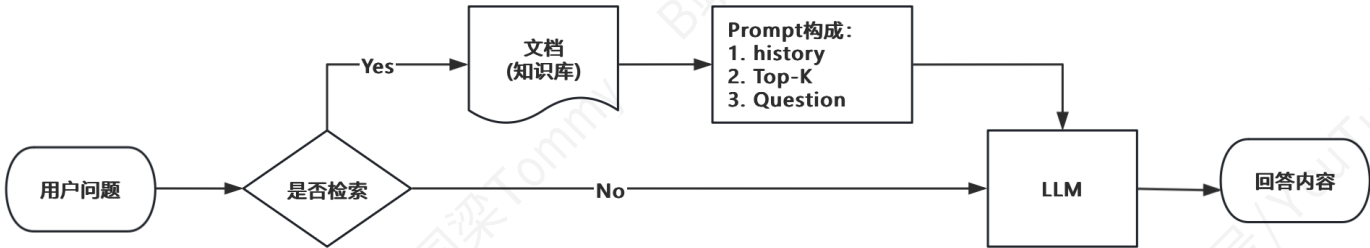


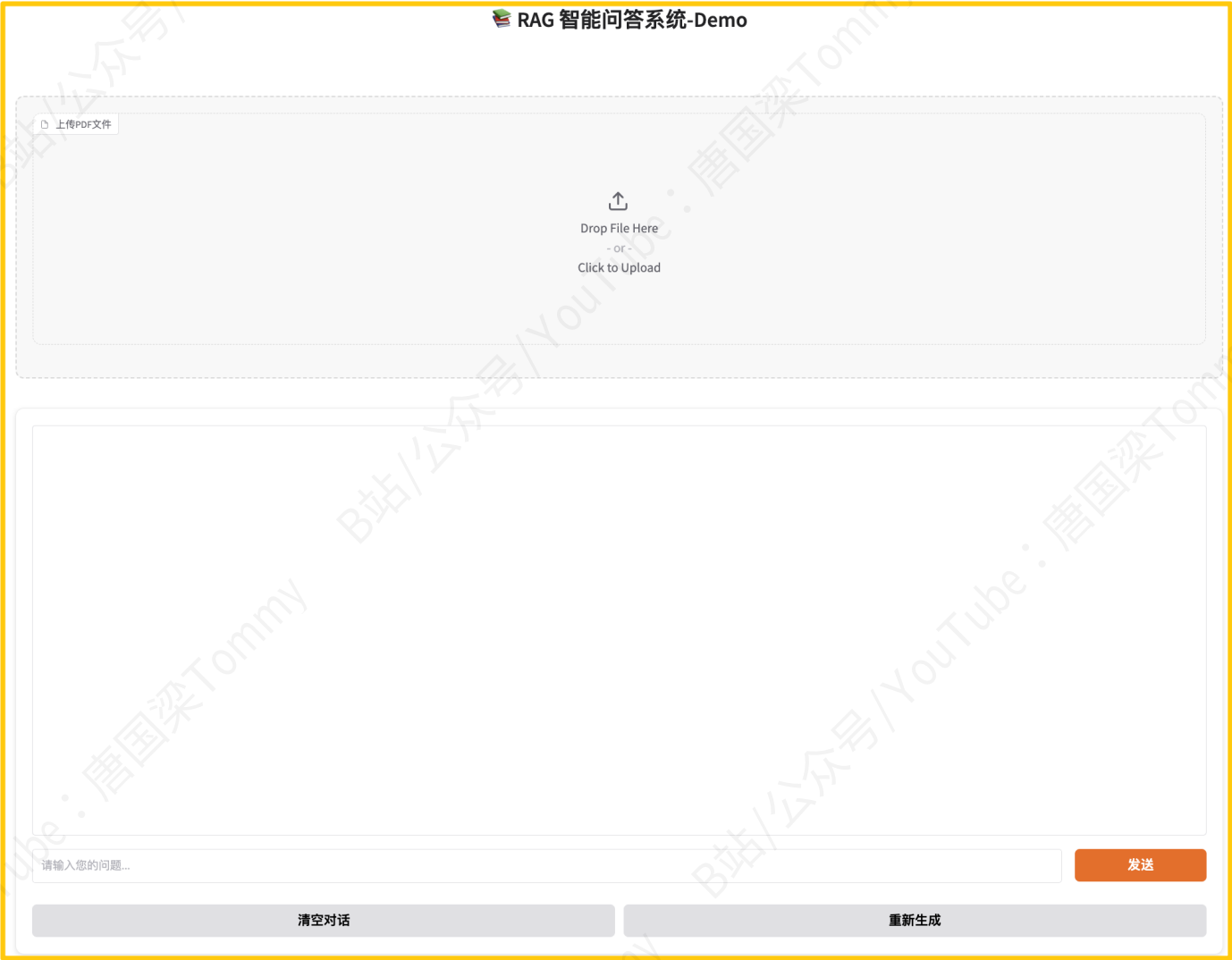
DeepSeek R1 RAG智能问答系统

PDF 智能问答助手 – Demo



基于RAG(检索增强生成)架构的智能问答系统,支持PDF文档的智能问答。本系统使用Ollama作为大语言模型后端,通过向量检索增强文档问答能力。

系统界面预览



1. 功能特点

- 支持PDF文档上传和智能问答

- 🔍 基于FAISS的高效向量检索
- 🗣️ 集成Ollama大语言模型
- 🌐 友好的Web交互界面(基于Gradio)
- 🚀 异步处理提升性能
- 📝 完善的日志记录
- 🐳 Docker容器化部署

2.系统架构

- 前端: Gradio提供Web交互界面
- 后端:
 - Langchain框架处理文档和LLM交互
 - FAISS向量数据库存储文档向量
 - Ollama提供LLM服务

3.环境要求

- Docker & Docker Compose
- Python 3.12+
- 24GB RAM建议

4.项目结构

```
.
├── config/ # 配置文件
├── core/ # 核心功能模块
│   ├── llm_service.py # LLM服务
│   ├── pdf_processor.py # PDF处理
│   └── vector_store.py # 向量存储
├── utils/ # 工具函数
├── fronted/ # Web界面
├── logs/ # 日志文件
├── vector_db/ # 向量数据库
├── Dockerfile # Docker构建文件
├── docker-compose.yml # Docker编排文件
└── requirements.txt # 依赖包
```

5.启动服务

```
# 赋予权限
chmod +x start.sh stop.sh

# 启动所有服务
./start.sh

# 进入ollama容器内, 并下载模型
docker exec -it ollama /bin/bash
ollama run deepseek-r1:7b

# 关闭服务
./stop.sh
```

6. 访问界面

访问 <http://localhost:8804> 使用RAG智能问答系统

7. 性能优化

- 使用异步处理提升并发性能
- 向量检索优化
- 文档分块策略优化
- 多线程处理PDF文档

8. 部署说明

系统使用Docker Compose进行容器化部署:

1. RAGChat服务容器
2. Ollama服务容器

网络配置:

- 内部网络: ragchat-network
- 端口映射: 8804(RAGChat), 11434(Ollama)

AI进阶学习

 B站课堂: <https://space.bilibili.com/3546748815411393/pugv>

 官方网站: <https://www.tgltommy.com/> (国内访问需科学上网)

如果你正在关注大模型 Agent、强化学习后训练、RLHF、DPO、GRPO、RLVR 等前沿方向, 欢迎关注我最新上线的精品课程:
[《大模型 Agent 强化学习实战: 从后训练、奖励设计到工业级对齐系统》](#)

这门课程围绕当前大模型 Agent 强化学习的核心技术路线展开, 不只是介绍零散算法和论文概念, 而是希望帮助学习者建立一套完整的技术地图: 从基础原理、论文脉络、关键算法, 到开源项目源码解析与工业级系统实践, 逐步理解 Agent RL 为什么重要、如何演进, 以及在真实应用中如何落地。

课程配套专属飞书知识库《Agent RL 学习宝典》, 包含学习路线、论文资料、项目说明、源码阅读辅助资料和后续持续更新内容, 方便长期查阅和跟进前沿技术。

Agent RL 变化很快, 但越是变化快, 越需要一套系统的学习框架。如果你希望系统进入 Agent RL 方向, 而不是停留在碎片化了解阶段, 这门课会是一个很好的起点。

课程官网: <https://www.tgltommy.com/p/agent-rl> (国内访问需科学上网)

B站课堂: <https://www.bilibili.com/cheese/play/ss842375604>

更多课程详情, 扫描以下二维码咨询课程助理:

新课首发

大模型 **Agent**
强化学习实战:
| 从后训练、奖励设计到工业级对齐系统

唐国梁Tommy

TGL AI
Tommy

官方微信课程助理



B站/公众号/YouTube: 唐国梁Tommy

**关注微信公众号
获取更多学习资源**



官方网站: TGLTommy.com(需科学上网)