本地部署全流程

1.下载ollama，通过ollama将模型下载到本地运行；

2.下载RAGflow源代码和Docker，通过Docker来本地部署RAGflow；

3.在RAGflow中构建个人知识库并实现基于个人知识库的对话问答。

一、

1.下载ollama平台

ollama是一个用于本地运行和管理大语言模型（LLM）的工具。

2.配置环境变量

·OLLAMA HOST - 0.0.0.0:11434

作用：让虚拟机里的RAGFlow能够访问到本机上的Ollama

如果配置后虚拟机无法访问，可能是本机防火墙拦截了端口11434

不想直接暴露11434端口：SSH端口转发来实现

更新完两个环境变量记得重启

·OLLAMA MODELS - 自定义位置

作用：ollama默认会把模型下载到C盘，如果希望下载到其他盘需要进行配置。

3.通过ollama下载模型deepseek-r1:1.5b（例子）

· ollama run deepseek-r1:1.5b

二、

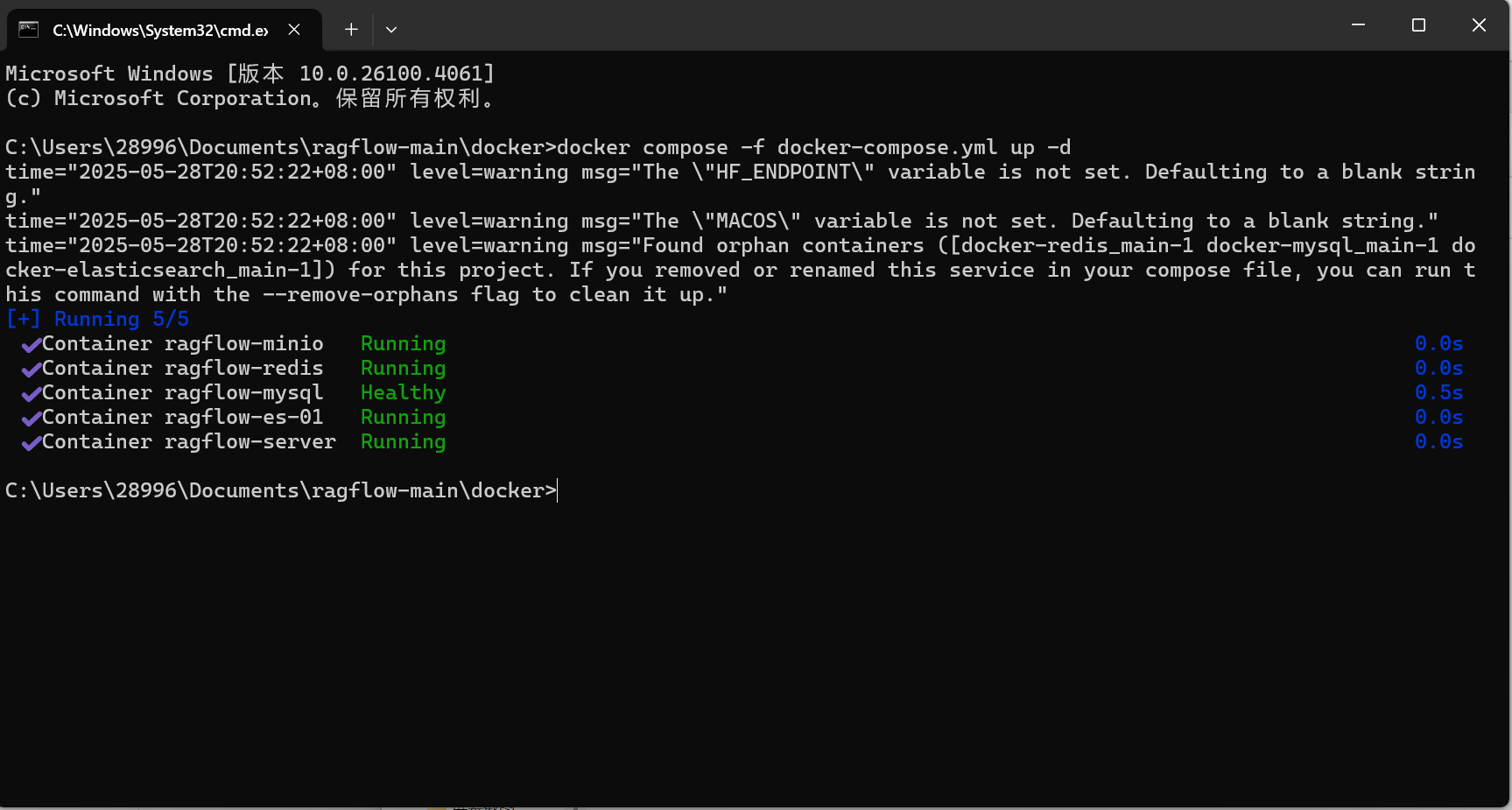
1.下载RAGFlow源代码

2.下载Docker

·Docker镜像是一个封装好的环境，包含了所有运行RAGFIow所需的依赖、库和配置。

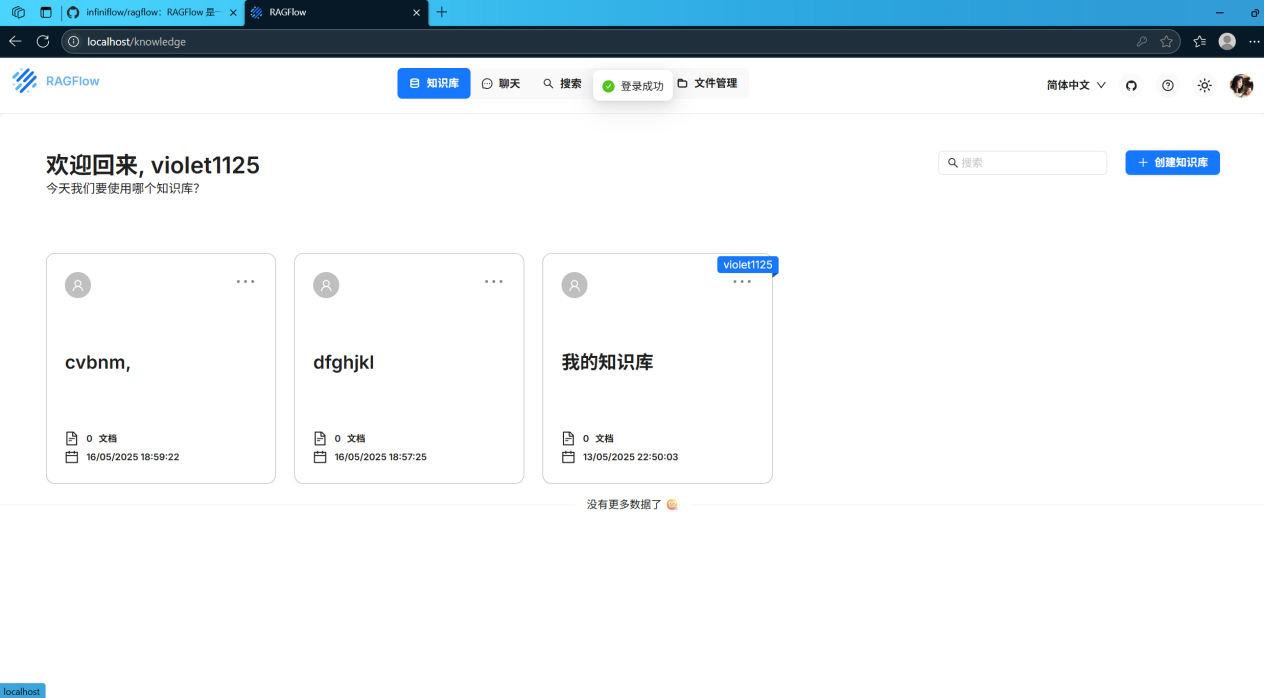
·如果安装遇到踩坑，可以自行搜索一下报错或者问gpt。

·如果镜像拉不下来，试试修改docker的镜像源。



三、

1. docker成功启动后，浏览器输入localhost:80来访问RAGFlow；



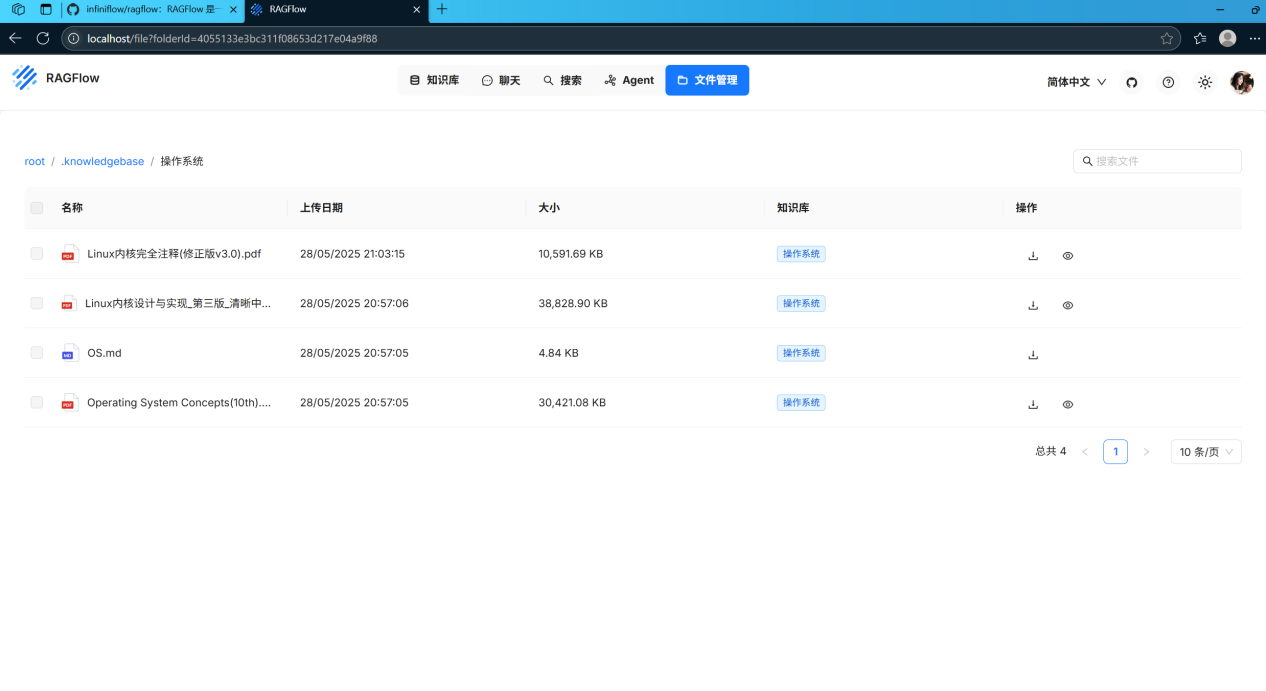
2.在“模型提供商”中添加本地部署的模型；

3.在“系统模型设置”中配置Chat模型和Embedding模型（用RAGFlow自带的即可）；



4.创建知识库，上传文件，解析文件；

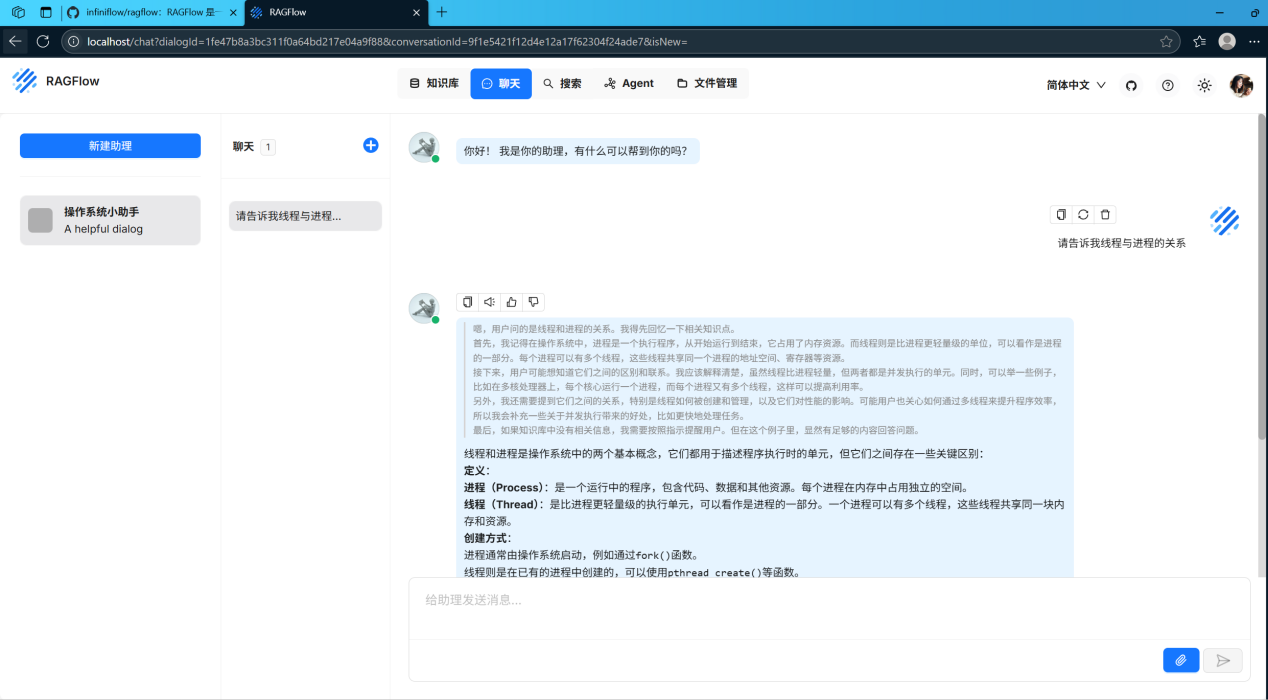




5.创建聊天助手（注意prompt和tokens的配置）；



6.开始对话。



Q：只想迅速搭建个人知识库，可以不本地部署吗？

A：当然可以！这样实现起来更简单，而且效果一般来说会更好。

具体步骤:

1.下载RAGFlow源代码和docker，通过docker本地部署RAGFlow（RAGFlow目前没有官方的网页版）（直接从第二步开始）；

2.在RAGFlow中配置任意的Chat模型和Embedding模型（需要到这些模型对应的官网去付费申请apiKey)；

优点：

不用本地部署大模型，简单易操作;

企业大模型性能一般更优越;

缺点：

隐私安全问题;

调用企业大模型API的成本;