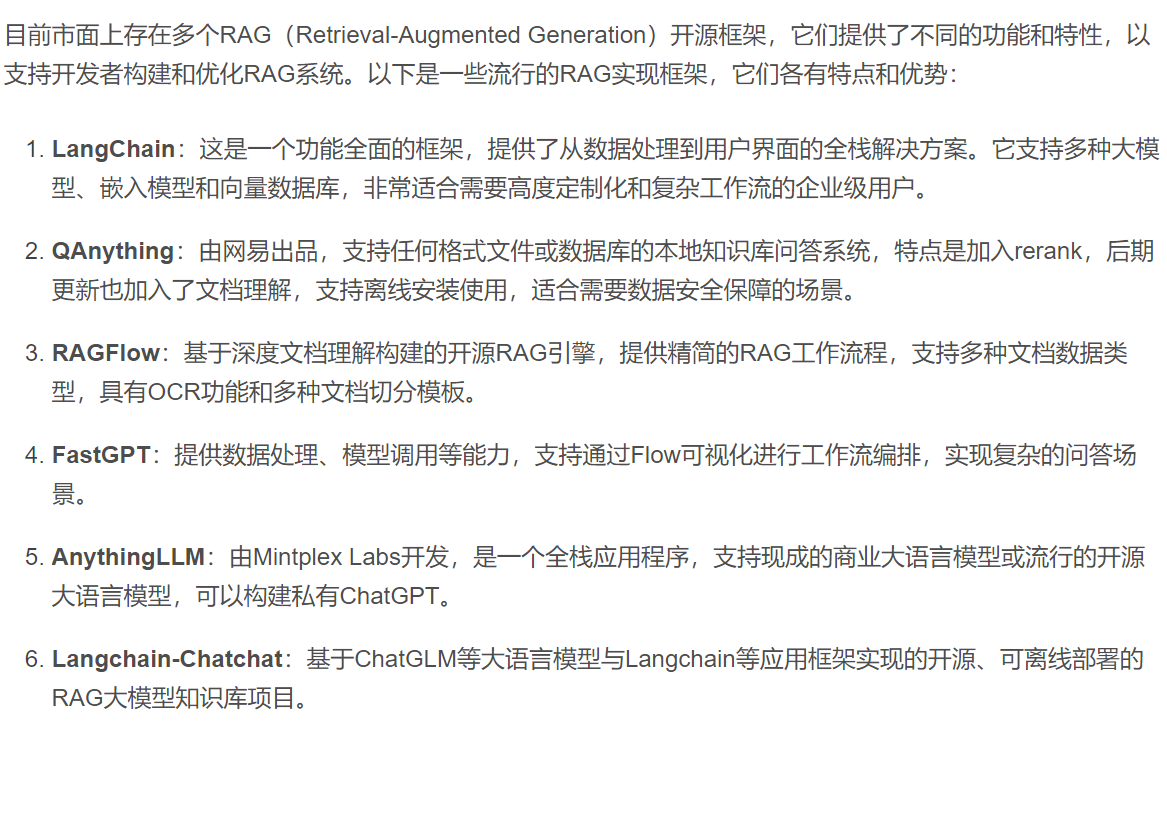
2025/2/17

基于RAG生成报告的基本流程

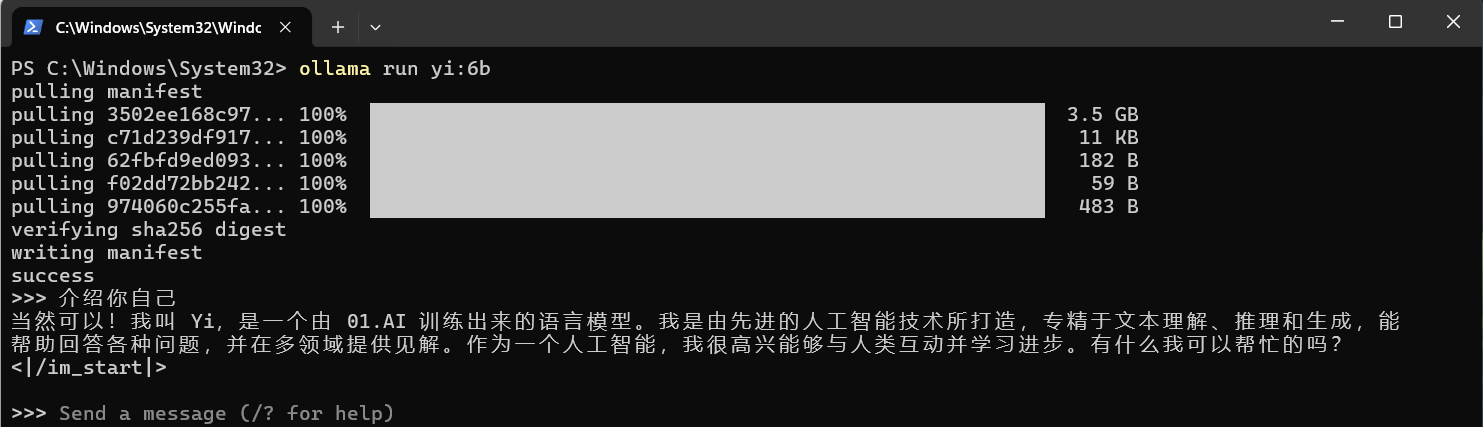
在搭建好相匹配环境后，通过加载器导入并完成分块，利用嵌入模型将文档内容向量化，并存储至向量数据库；配置好检索器，选定语言模型并设计提示模板，组装成 RAG 处理链，可依据用户输入生成报告内容。最后，运用评估工具对报告质量进行初步评估，后续将根据结果持续优化，同时可计划部署 Web 界面以提升用户体验。



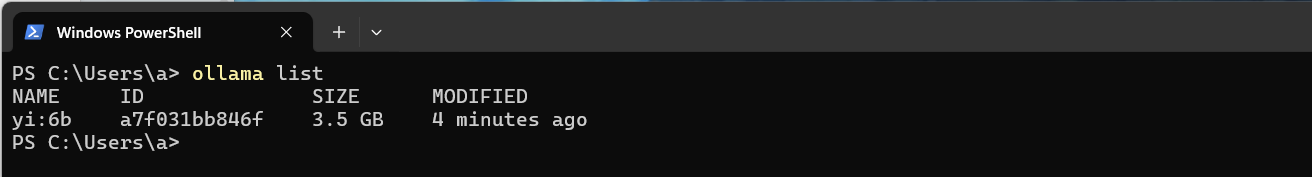
<https://ollama.org.cn/library/yi:6b>

命令：ollama run yi:6b

C:\Users\a\.ollama\models\manifests\registry.ollama.ai\library\yi



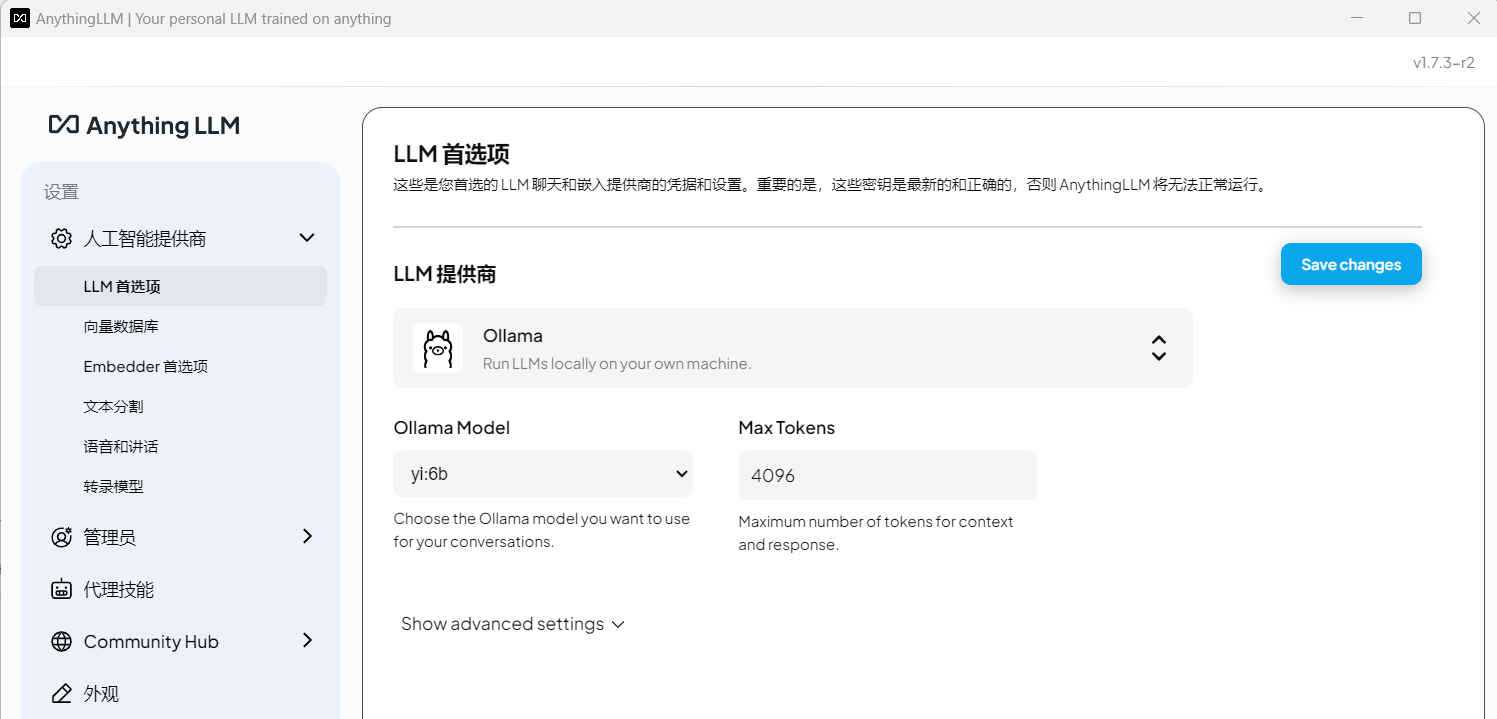
通过ollama在本地加载了零一万物模型yi1.5，选择6b的版本进行部署

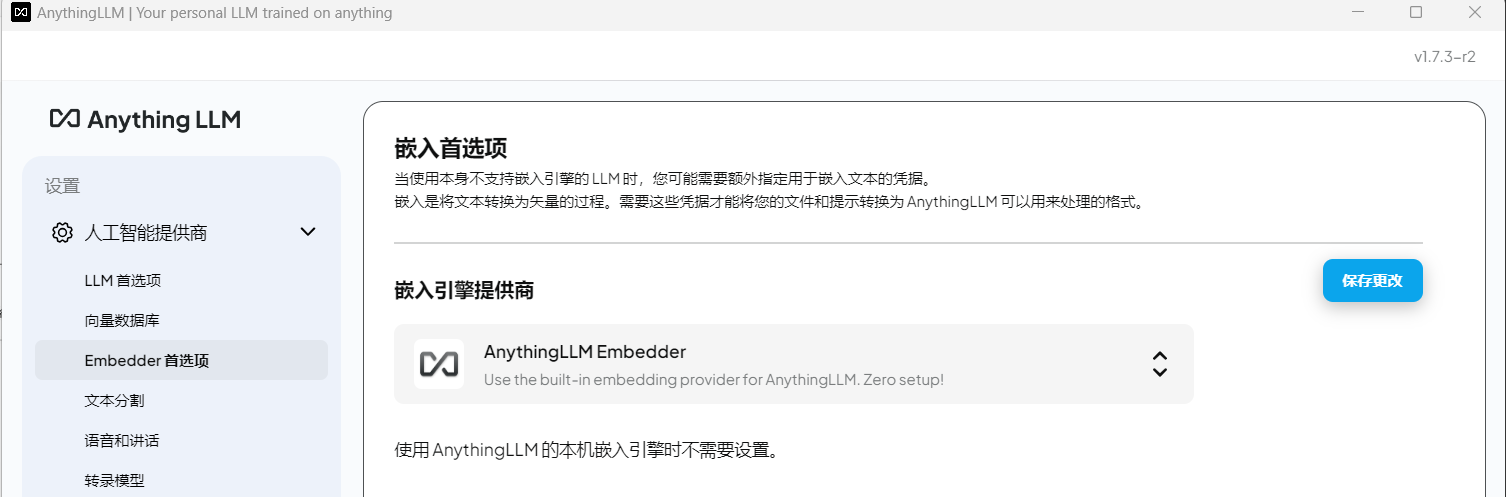


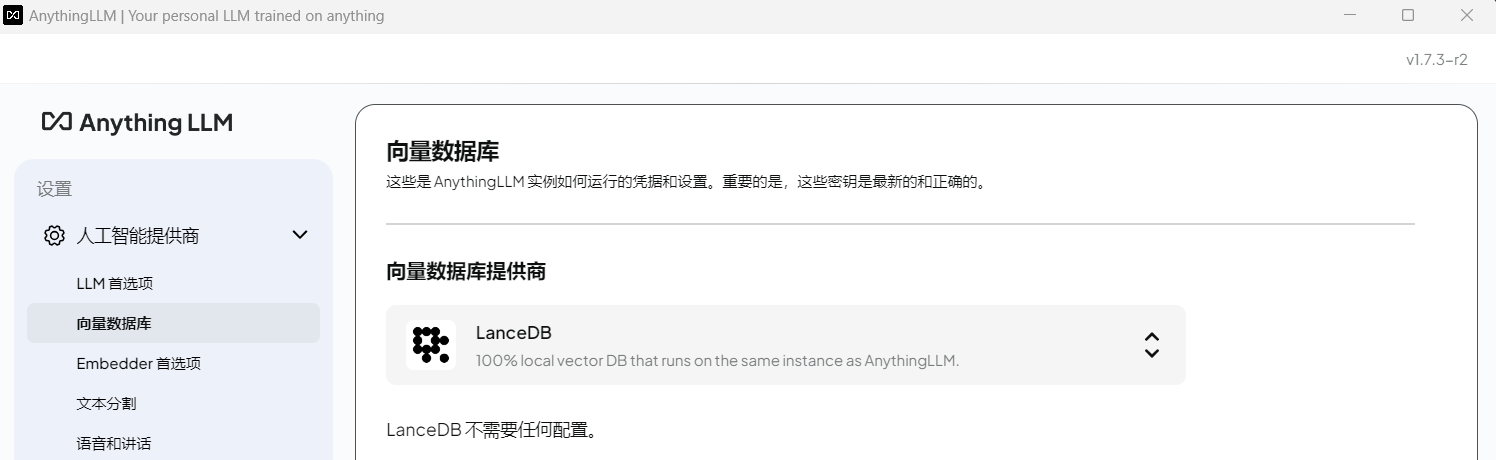
选择anythingLLM框架进行结合部署，后续可根据需要进行更改

<https://zhuanlan.zhihu.com/p/23179498568>

<https://anythingllm.com/desktop>

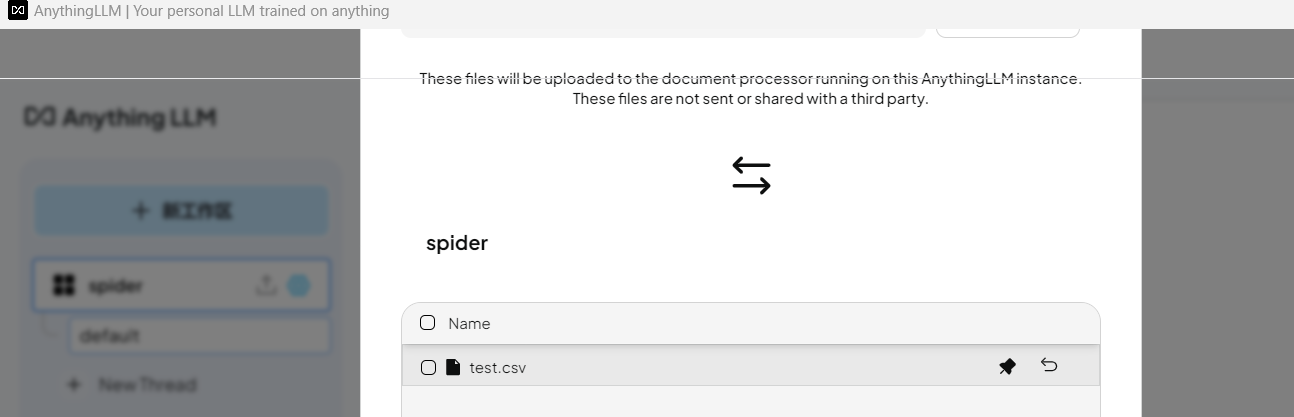




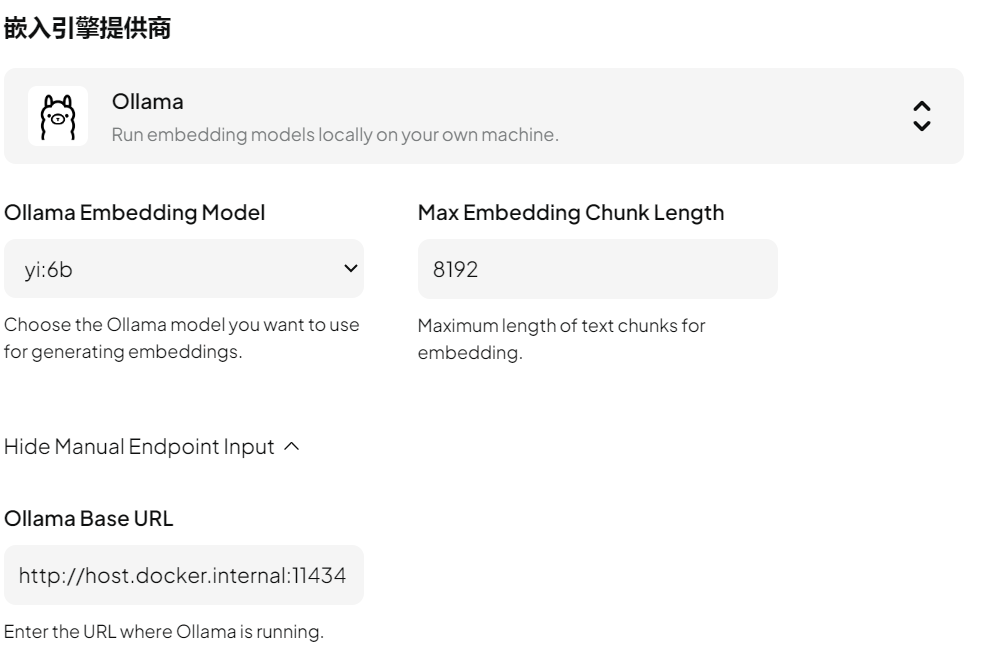




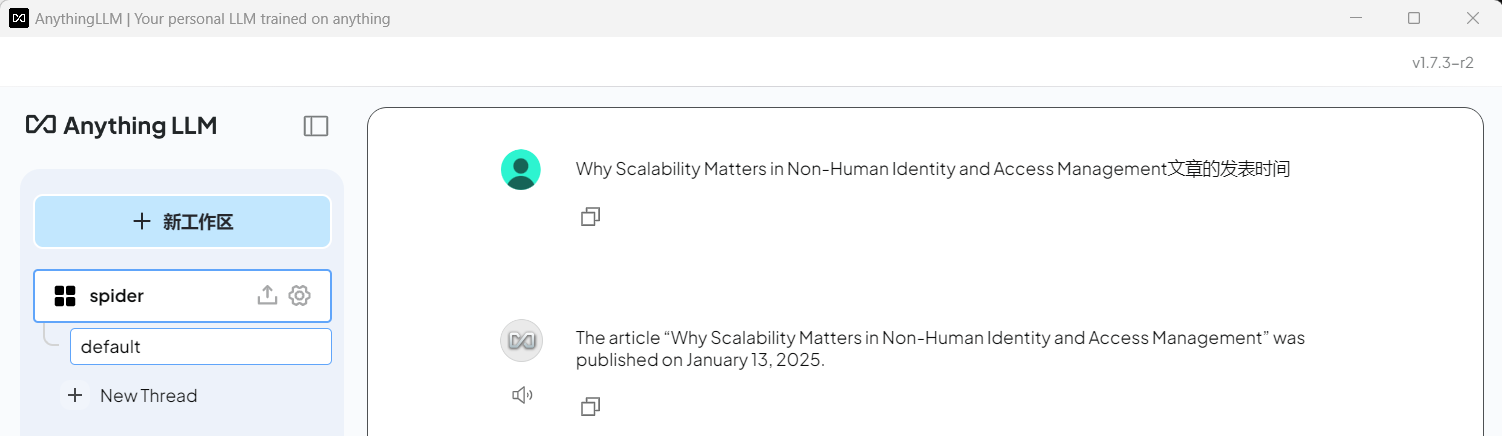
导入了test.csv文件尝试：



可能存在的情况：上传文档存在问题，可尝试下面的解决方法，在设置里更改相关设定



成功检索test.csv文件内容回复：导入文档后实现了检索增强的效果



·查询发现anythingLLM 提供 RESTful 风格的 API 接口，可以通过编写代码实现自动化调用

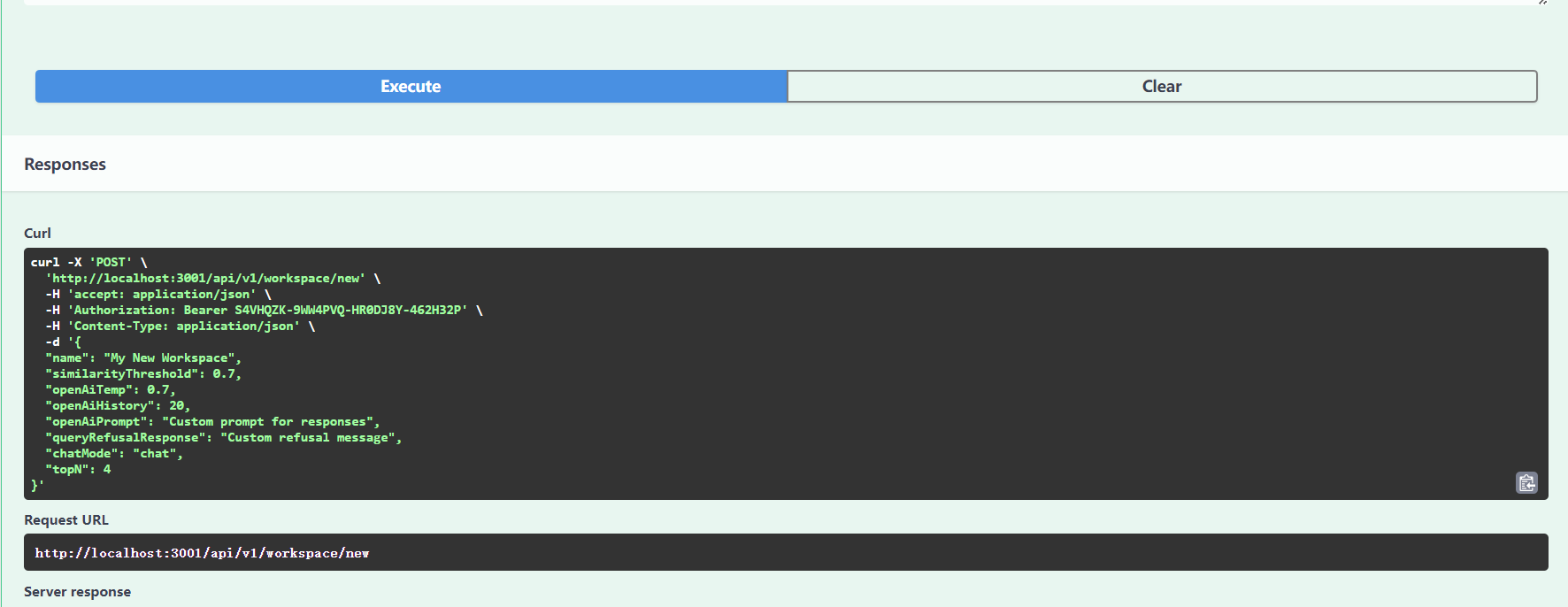
·可以通过调用 AnythingLLM 的 API 将文档上传并进行嵌入处理、实现问答功能等

·我在本地实现了ollama+yi:6b+anythingLLM框架的结合部署，如何编写代码实现对AnythingLLM的API接口实现自动调用，进而实现创建工作区并嵌入文档、以及问答等功能

<https://blog.csdn.net/XZH_iD/article/details/139908980>

API KEY:S4VHQZK-9WW4PVQ-HR0DJ8Y-462H32P

[http://localhost:3001/api/docs/](http://localhost:3001/api/docs/#/)



curl -X 'POST' \

'http://localhost:3001/api/v1/workspace/new' \

-H 'accept: application/json' \

-H 'Authorization: Bearer S4VHQZK-9WW4PVQ-HR0DJ8Y-462H32P' \

-H 'Content-Type: application/json' \

-d '{

"name": "My New Workspace",

"similarityThreshold": 0.7,

"openAiTemp": 0.7,

"openAiHistory": 20,

"openAiPrompt": "Custom prompt for responses",

"queryRefusalResponse": "Custom refusal message",

"chatMode": "chat",

"topN": 4

}'

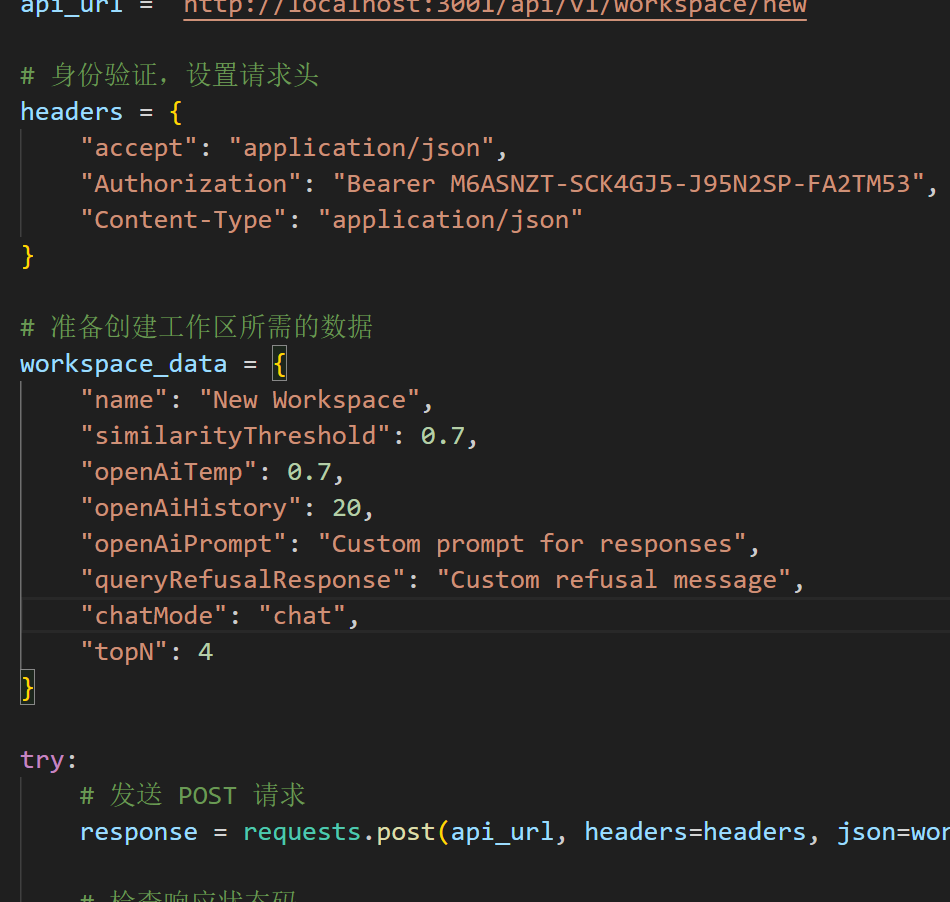
http://localhost:3001/api/v1/workspace/new

1.创建工作区——step1.py

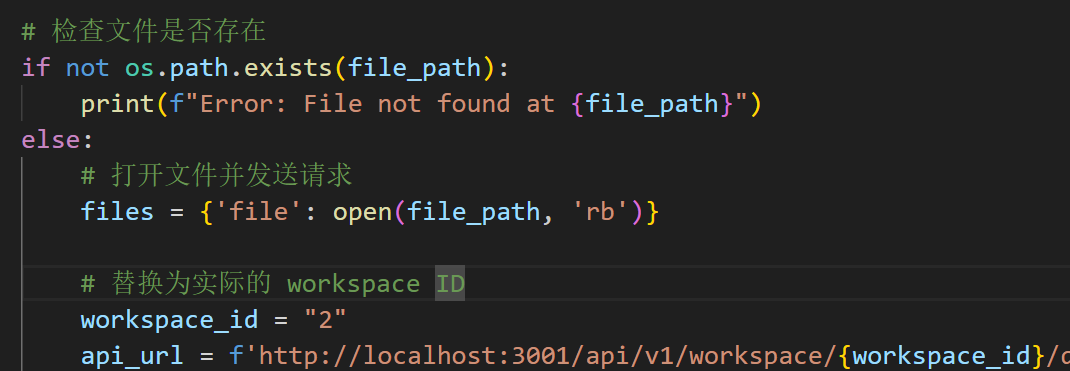


2.





操作后通过api接口嵌入文档的速度比通过ui界面快许多



响应速度慢的问题正在通过查询修改增加Ollama的 num\_ctx 参数解决

目前发现在调用 Ollama 的 API 时，可以通过 options 参数动态设置 num\_ctx