RAG项目实战（使用llamaindex构建自己的知识库）

1.环境配置

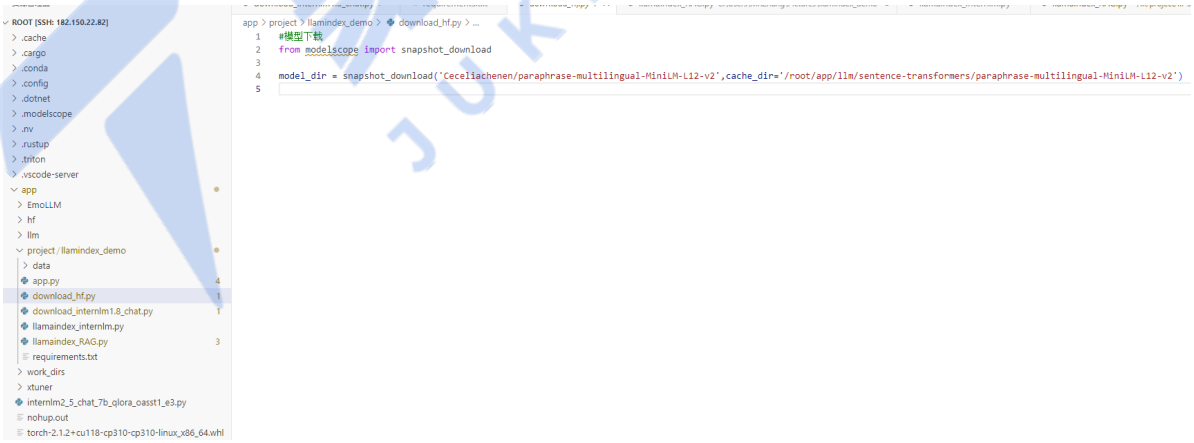
使用 `conda` 配置项目python环境

```
# 创建环境
conda create -n rag python=3.10
# 激活环境
conda activate rag
# 安装一些必要的库(requirements.txt在项目代码包中)
pip install -r requirements.txt
```

2.下载Sentence Transformer 模型

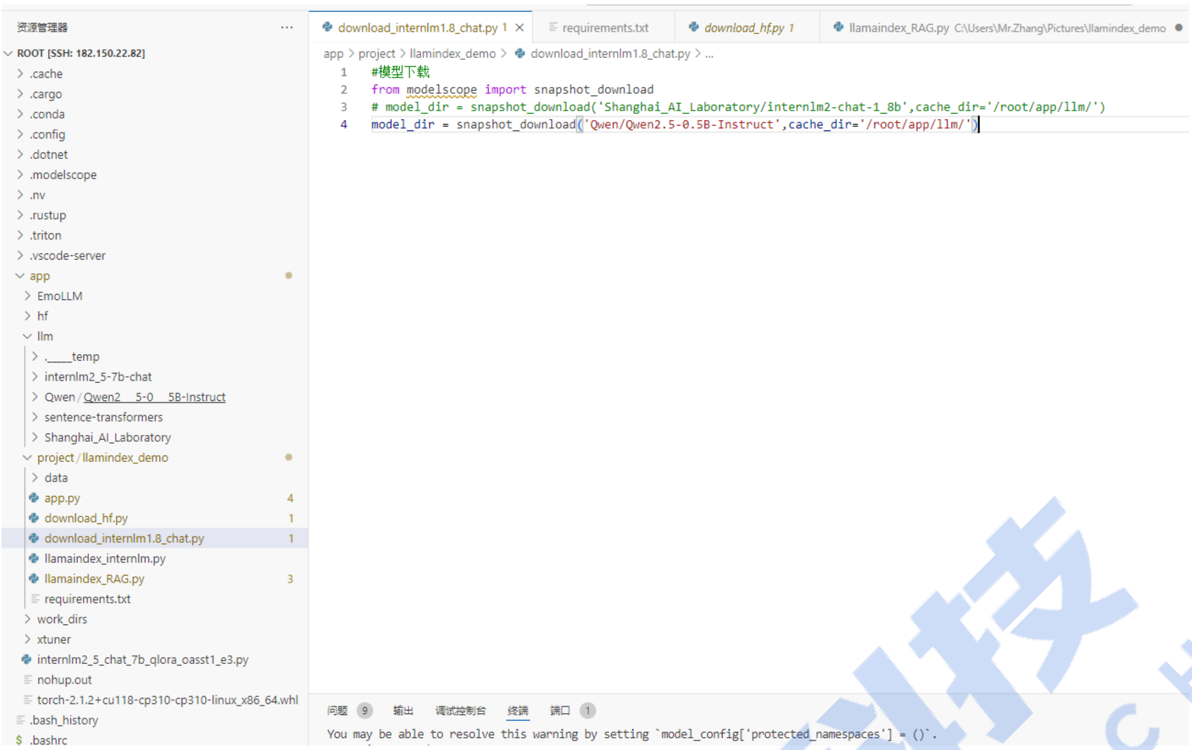
在进行RAG之前，需要使用词向量模型进行Embedding，将文本进行向量化处理，此处选择 Sentence Transformer 模型。执行

llamindex_demo/download_hf.py下载。

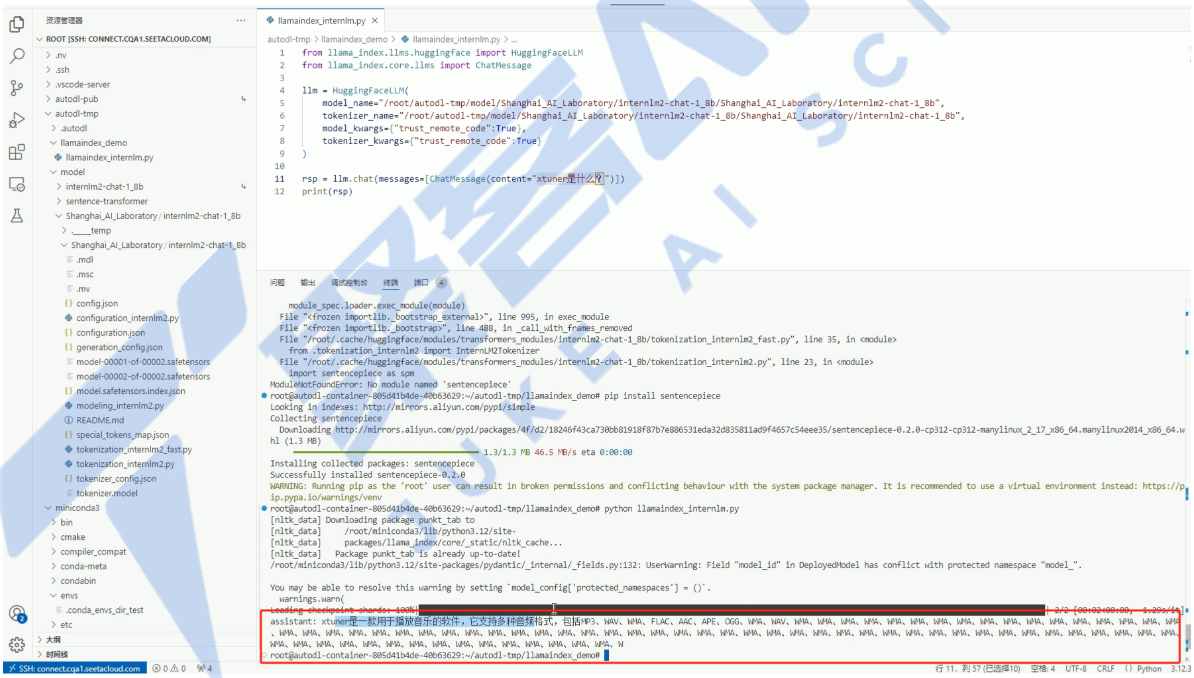


3.下载InternLM2 1.8B/qwen2.5_0.5B 模型

执行llamindex_demo/download_internlm1.8_chat.py下载。

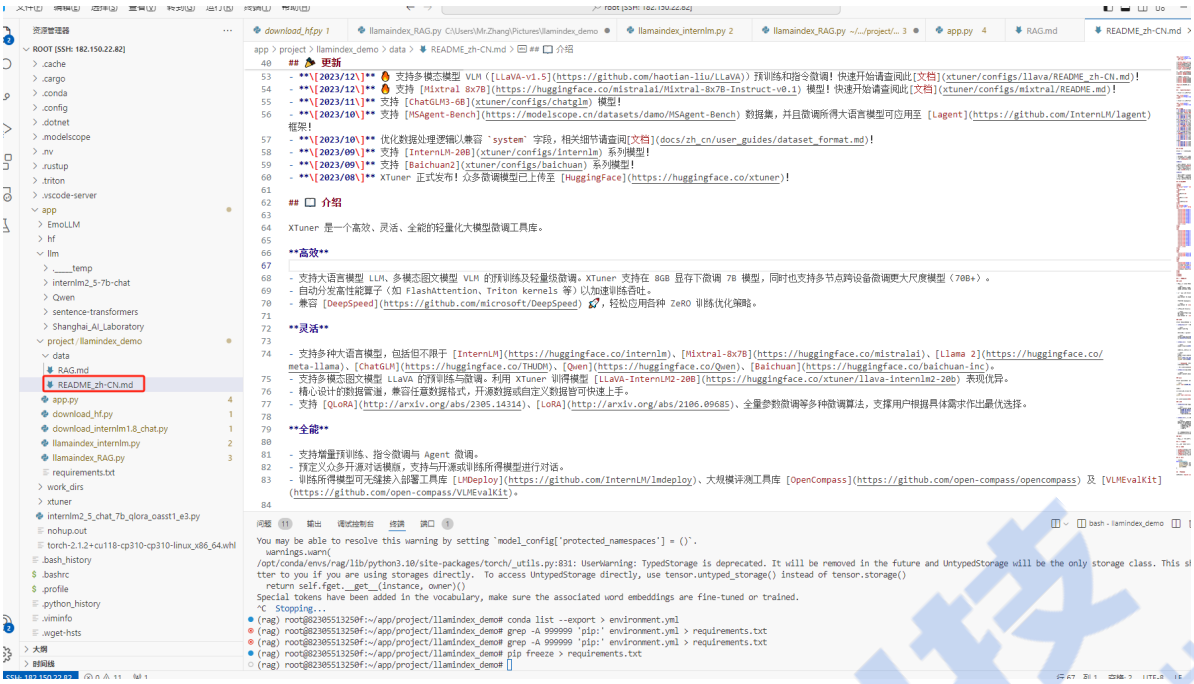


执行llamindex_demo/llamaindex_internlm.py进行提问测试，得到结果如下。可以看出模型本身并不具备关于`xtuner`的相关知识，回复也比较杂乱。



4.创建知识库

在`llamindex_demo`文件夹下，创建`data`文件夹，用于构建知识库。



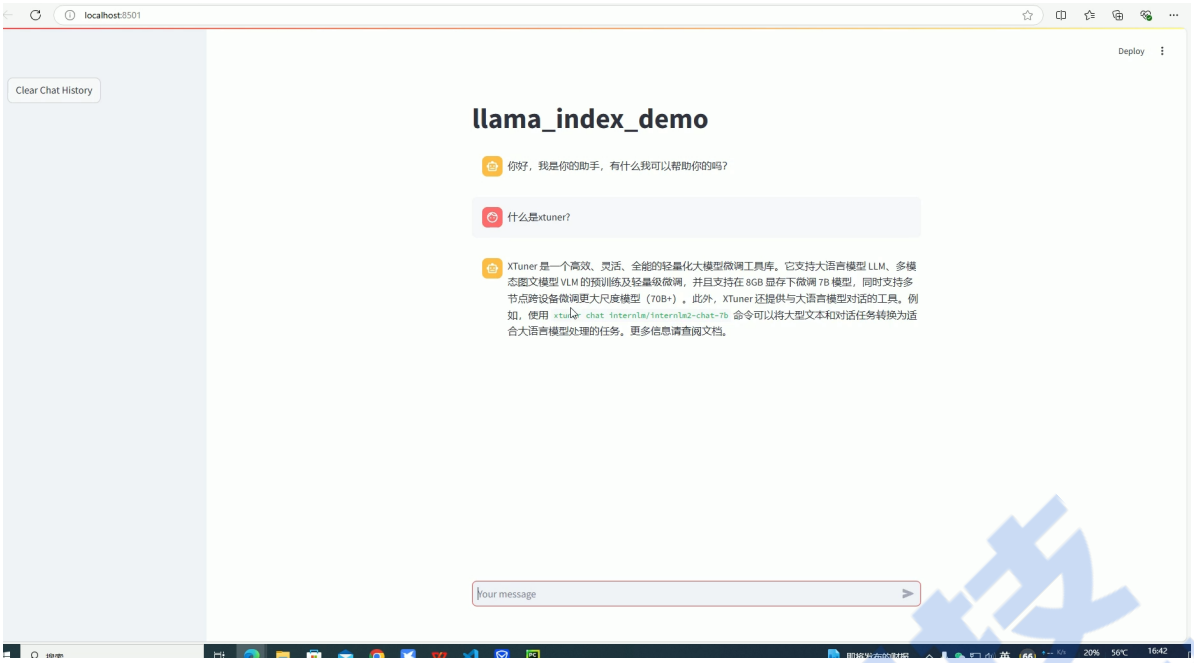
执行llamindex_demo/llamaindex_RAG.py，运行次测试后，可以看到可以正确回答，并且可以给出回答的出处：

5.创建web应用

执行llamindex_demo/app.py，

```
streamlit run app.py
```

运行后可以打开网页端，可以进行提问:


localhost:8501
Deploy
Clear Chat History
llama_index_demo
你好，我是你的助手，有什么我可以帮助你的吗？
什么是xtuner?
XTuner 是一个高效、灵活、全能的轻量化大模型微调工具库。它支持大语言模型 LLM、多模态图文模型 VLM 的预训练及轻量级微调，并且支持在 8GB 显存下微调 7B 模型，同时支持多节点跨设备微调更大尺度模型（70B+）。此外，XTuner 还提供与大语言模型对话的工具。例如，使用 xtuner chat internlm/internlm2-chat-7b 命令可以将大型文本和对话任务转换为适合大语言模型处理的任务。更多信息请查阅文档。
Your message