## NVIDIA AI-AGENT夏季训练营 RAG智能对话机器人**实现本地文档内容问答**

项目名称：AI-AGENT夏季训练营 — RAG智能对话机器人实现本地内容问答

报告日期：2024年8月18日

项目负责人：刘建辉

#### 项目概述：

1最小化步骤，跑通RAG基本流程

2 可以在zh\_data目录下放入任意txt文件，用于RAG技术增强搜索

3 可以通过文字及语音进行问答

#### 技术方案与实施步骤

* 模型选择：

ai-phi-3-small-128k-instruct 轻量化模型

NV-Embed-QA 嵌入模型

whisper 语音转文字

edge-tts 文字转语音

* 数据的构建：

RAG数据以文本方式存放在zh\_data目录下，程序可以从该目录获取所有内容，并建立对应向量库

* **功能整合**：

使用whisper进行语音转文字，发送给模型，模型返回的文本由edge-tts进行声音构建，并返回对应的文件按钮，可以点击回放

#### 实施步骤：

* 环境搭建：

1 使用conda建立隔离的python运行环境

2 安装langchain-nvidia-ai-endpoints，使用NVIDIA的模型能力

3 安装其他对应的工具库

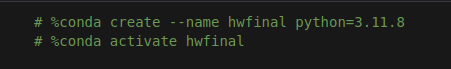
4 建立向量库

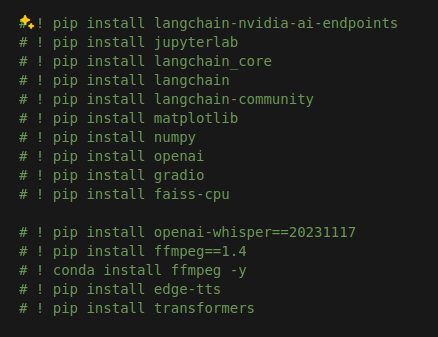
5 实现工具类

6 gradio实现一个web页面

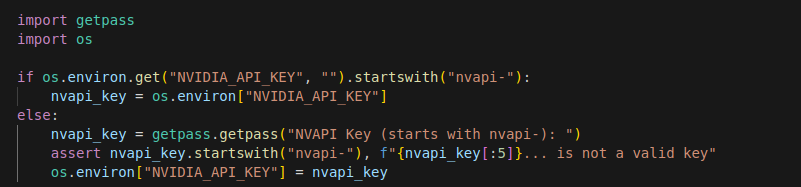
* 代码实现：

1 环境安装

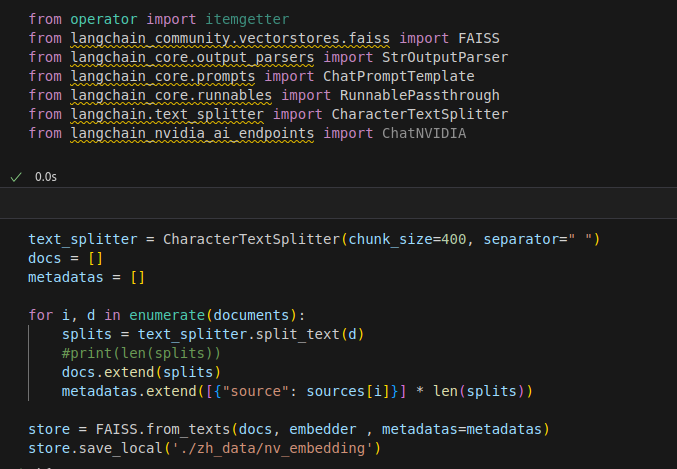
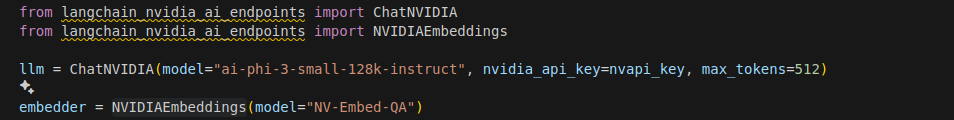




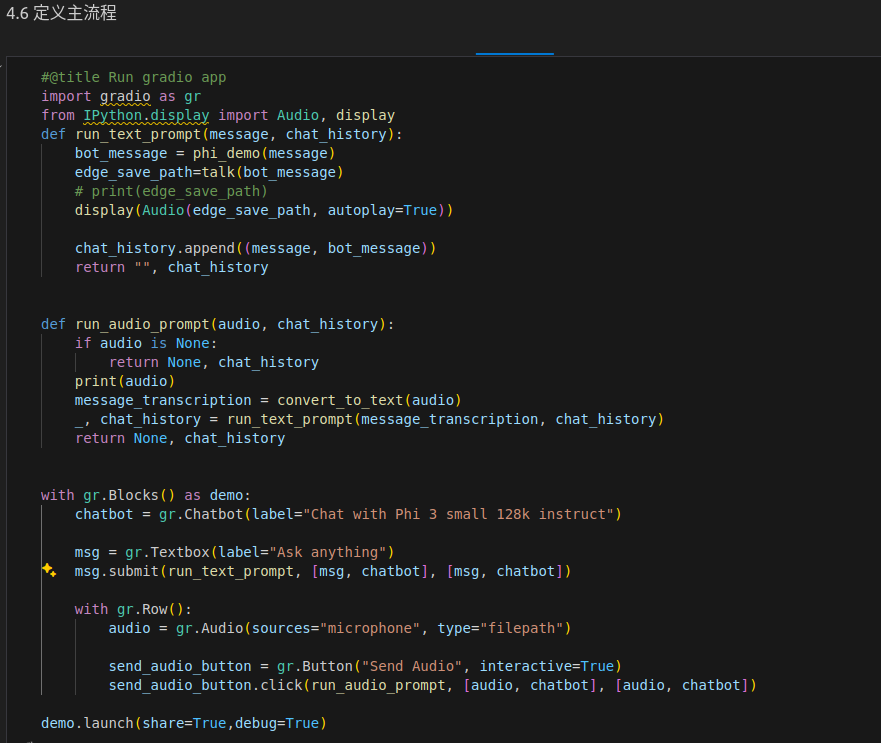
2 设置APIKEY



3 建立知识库



4 主流程



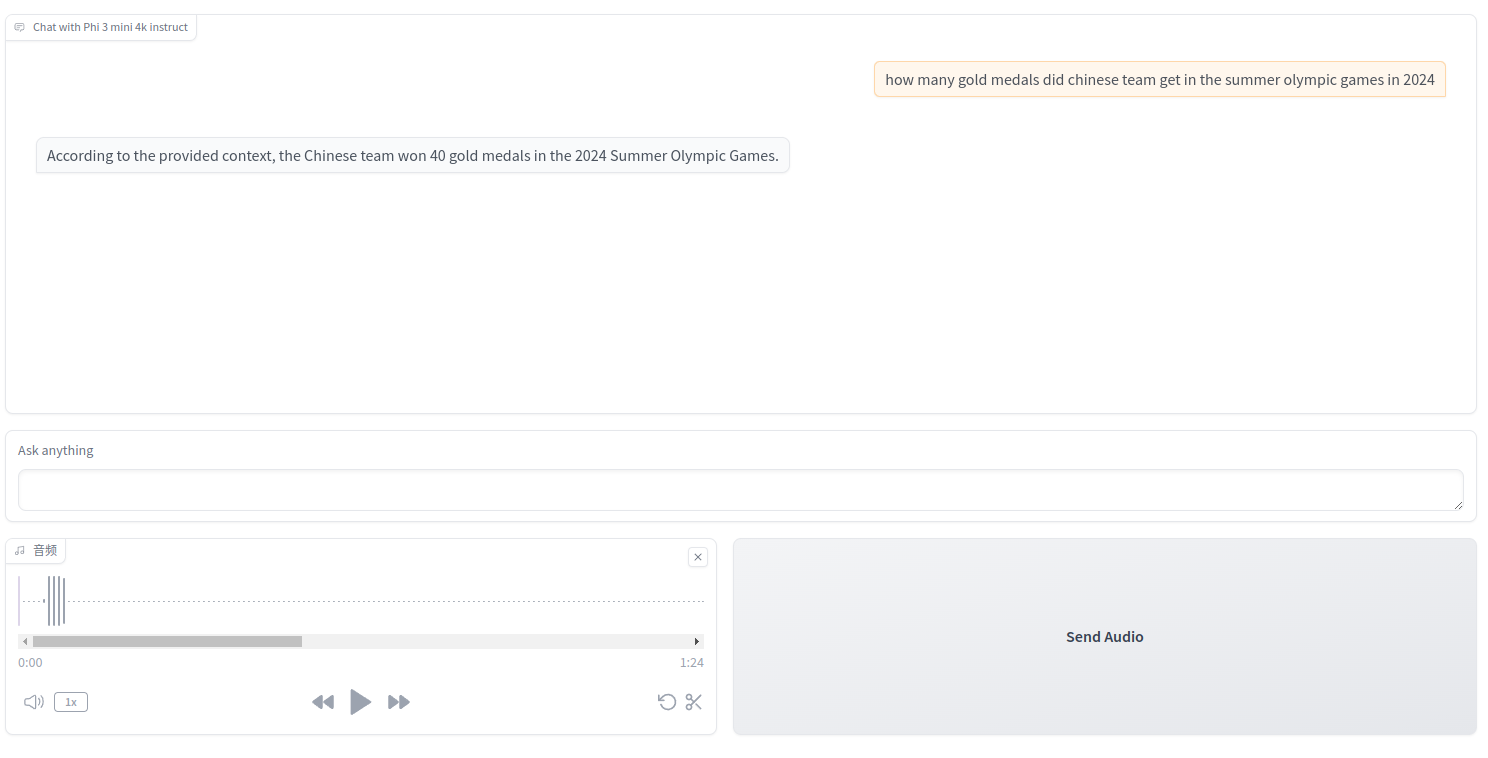
* 测试与调优：

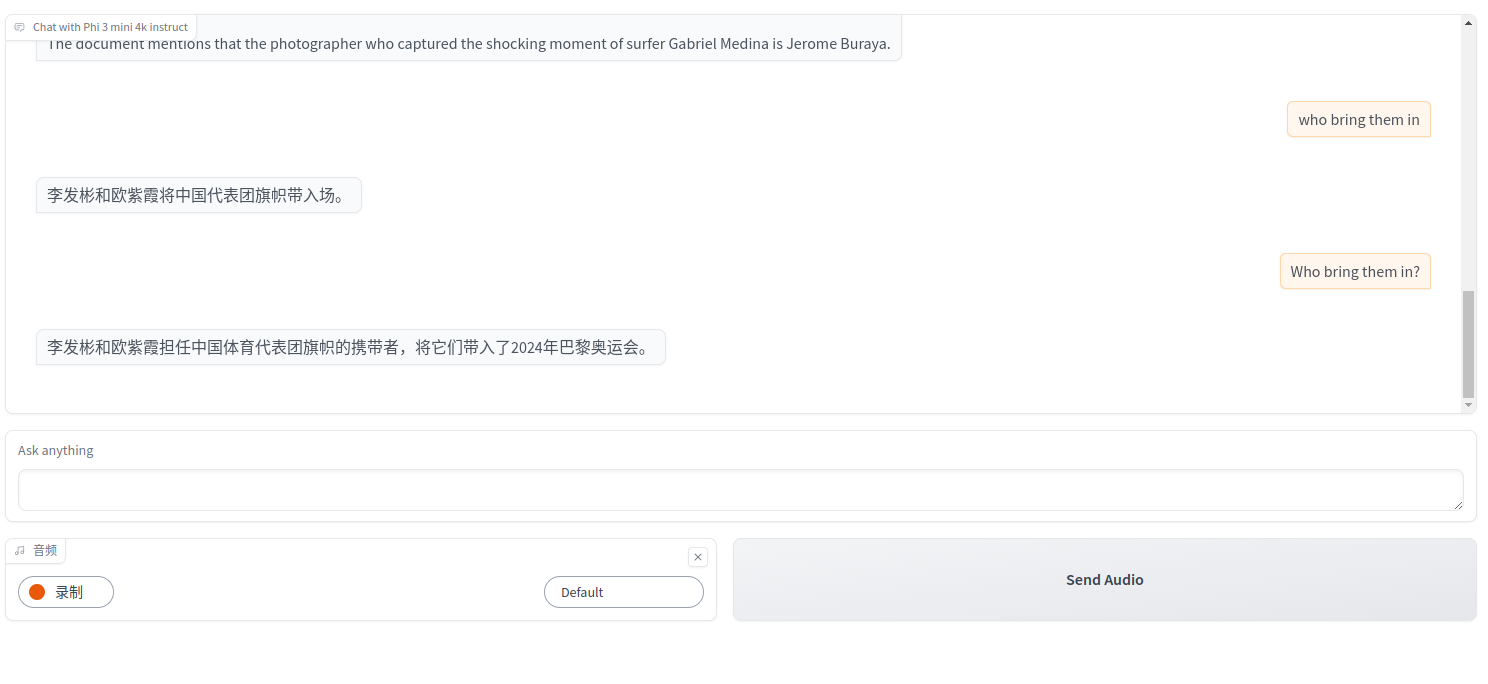
1 弹出的gradio界面，可以用文字进行对话，如输入：

how many gold medals did chinese team get in the summer olympic games in 2024

得到以下回答，是预计zh\_data目录下提供的文本内容进行回答：

According to the provided context, the Chinese team won 40 gold medals in the 2024 Summer Olympic Games.

2 也可以本地录制声音，通过Send Audio按钮发送后，如“who bring them in”模型也会给予对应输出，截图中的重复两次问答分别是文字输入和语音输入，如下：



* 描述测试过程，包括测试用例的设计、执行及性能调优。
* 集成与部署： 说明各模块集成方法及最终部署到实际运行环境的步骤。

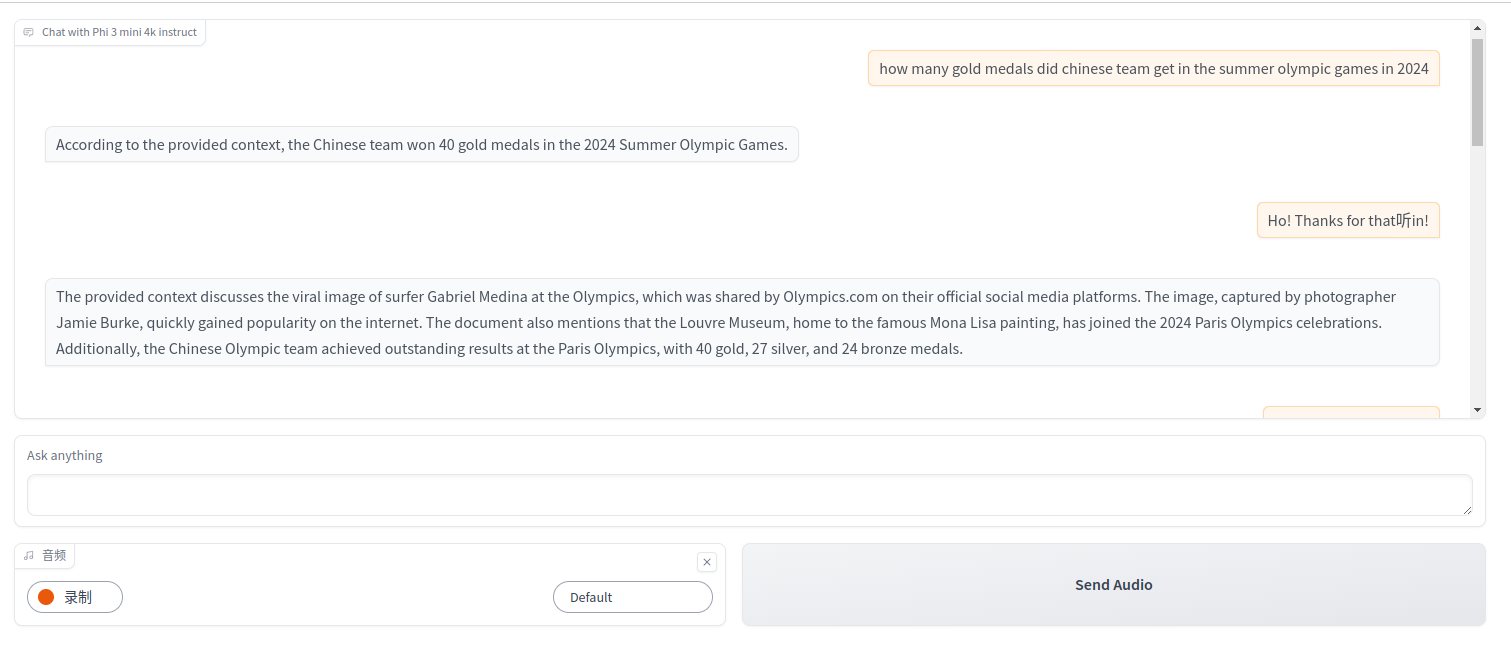
#### 项目成果与展示：

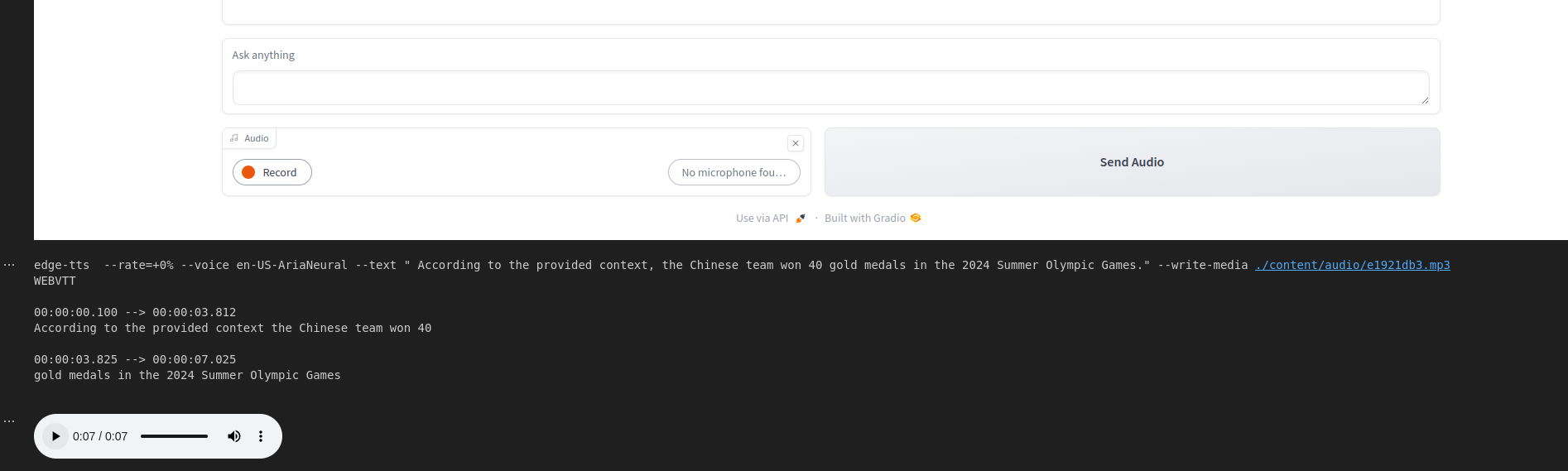
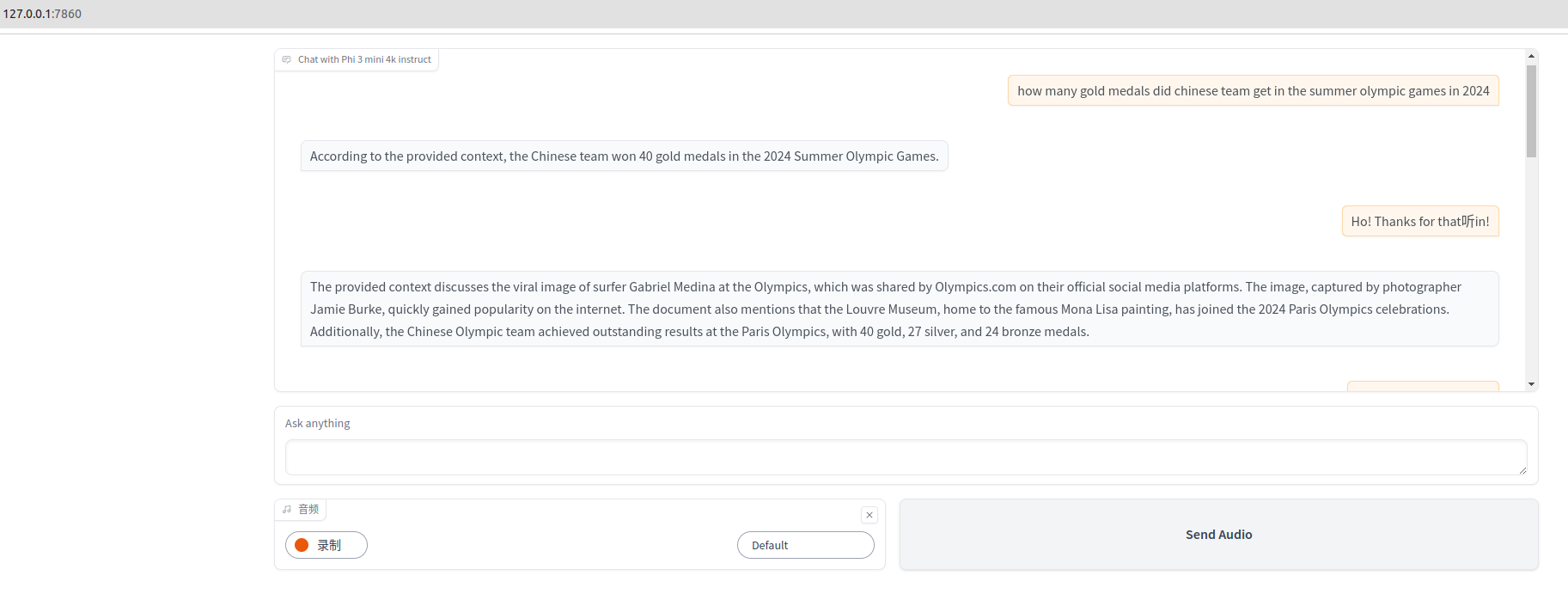
* 应用场景展示(必写)：

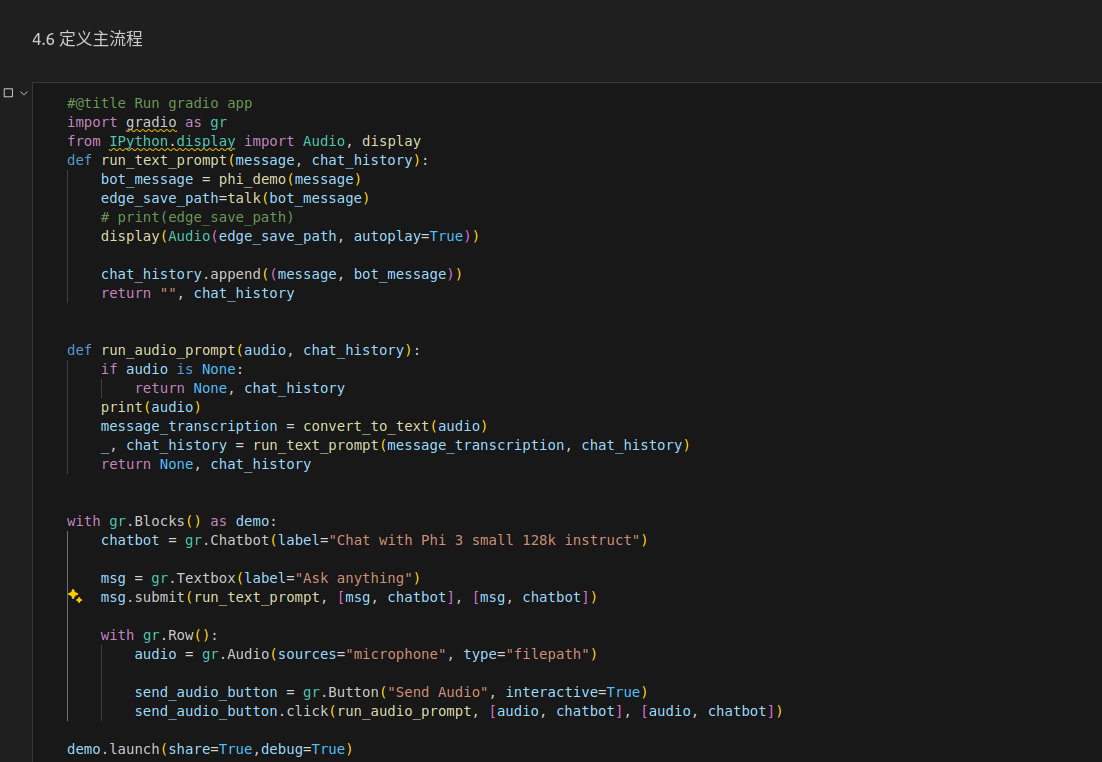
用于会议记录问答，把会议记录放入zh\_data对应目录，就可以回答会议记录上提到的内容。

* 功能演示（必写）：

主界面：

输出TTS输出的语音：

主要代码流程：



#### 问题与解决方案：

* 问题分析：

1 调用嵌入模型NVIDIAEmbeddings模型调用时，出现401错误

* 解决措施：

调用ChatNVIDIA时，llm = ChatNVIDIA(model="ai-phi-3-small-128k-instruct", nvidia\_api\_key=nvapi\_key, max\_tokens=512)

有传入nvapi\_key，但调用NVIDIAEmbeddings没有传入对于的APIKEY，可能会出现401错误，以为是上面的KEY过期了，实际上是内部也需要一个KEY，所以运行时最好在环境变量中设置NGC\_API\_KEY=”nvapi-xxxx”，以免出现返回401的权限问题

#### 项目总结与展望：

* 项目评估：

目前只是跑通RAG基本流程，以及尝试进行语音转换，实现基本的多模态。

* 未来方向：

1 更多文档类型支持

2 考虑加入AUDIO2FACE，进行数字人语音驱动。

#### 附件与参考资料

https://blog.csdn.net/babylon3000/article/details/141217141?csdn\_share\_tail=%7B%22type%22%3A%22blog%22%2C%22rType%22%3A%22article%22%2C%22rId%22%3A%22141217141%22%2C%22source%22%3A%22babylon3000%22%7D

# **2024 NVIDIA开发者社区夏令营环境配置指南(Win & Mac)的补充说明**