### 项目报告模板

**第十届Sky Hackathon**

**参赛项目书**

参赛学校（公司）：\_\_\_南加利福尼亚大学-波士顿大学 \_\_\_\_\_\_\_\_

参赛队名：\_\_\_\_\_\_\_快乐一家亲\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

指导老师（如果没有可以不填写）：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

团队成员：刘天泽（队长） – 南加利福尼亚大学；

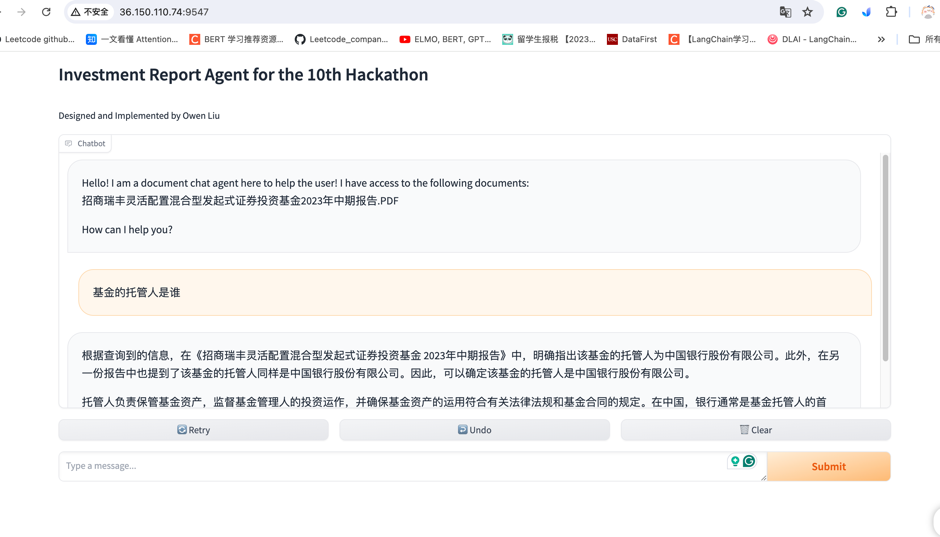
黄安琪 – 波士顿大学

于兆轩 – 海南科技职业大学

项目说明：

（请附上你项目的截图）此处需附加上你们队伍的端口号， 醒目一点。

端口号：**http://36.150.110.74:9947**



****

从图中可以发现，机器人对于关键信息的检索非常准确，示例文档长度60页pdf而目标信息简短且处在格式化数据之中。

### 描述一下你们的作品，包括亮点/特点，作品的功能和初衷

本次参赛作品基于Nvidia nomentron-340b框架和RAG技术构建投资研究智能体Agent，并使用该Agent完成对基金公告数据的解析和理解，回答用户提问以及分析需求等内容。

作品的亮点：区别于传统的RAG技术，本次作品在RAG技术上实现多重优化使其更加符合行业需求。

**在数据嵌入阶段**，区别于仅文档内容嵌入，首先采用基于知识库的chunks预先提出问题，并同时存储问题和答案（Q&A），在向量检索的时候，同时匹配查询query和预存问题的相似度,。其次存入向量库的文字块预先通过大模型生成总结，在针对总结embedding为对应的检索向量（由于基金报告的信息细节非常多而且相似度很高）

**在检索阶段前**，通过大模型将用户提出的原始查询问题扩展成多样化的查询，从而提升搜索的视野和深度。同时使用大模型规范化用户提出的问题，将用户提出的简短的，上下文信息不全的问题补充完整，提高向量检索的相似度。

**在检索过程中**，构建意图识别模块，根据用户查询的问题路由到某一组知识库，缩小检索范围，增加确定性。用户的问题被路由到不同分支之后，可能对应不同的检索逻辑和数据库

作品的功能和初衷：利用该Agent可以回答用户提出基于基金报告的一些问题，从基金报告中提取，分析，推理信息，为用户提供快速了解基金基本情况以及进阶问题的解决策略。

### 描述一下你们利用AIGC工具来生成的内容， 如：如何利用NIM来生成内容， 利用NIM接口生成内容的形式（如：文字，代码，图片）

利用nomentron-340b 根据提供的各种prompt和文本信息生成各种文字回答。基于时间训练限制，未来还会提升多模态的输出部分，例如：用户想要某一支基金在2023-2024年的收益变化折线图。该功能会基于multi agent智能体协作来进行任务区块的判断，调用相应的模型（funcitonal calling）

### 如果有， 描述一下你们是如何进行模型训练和推理（比如训练的时间，训练的结果， 优化过程等）

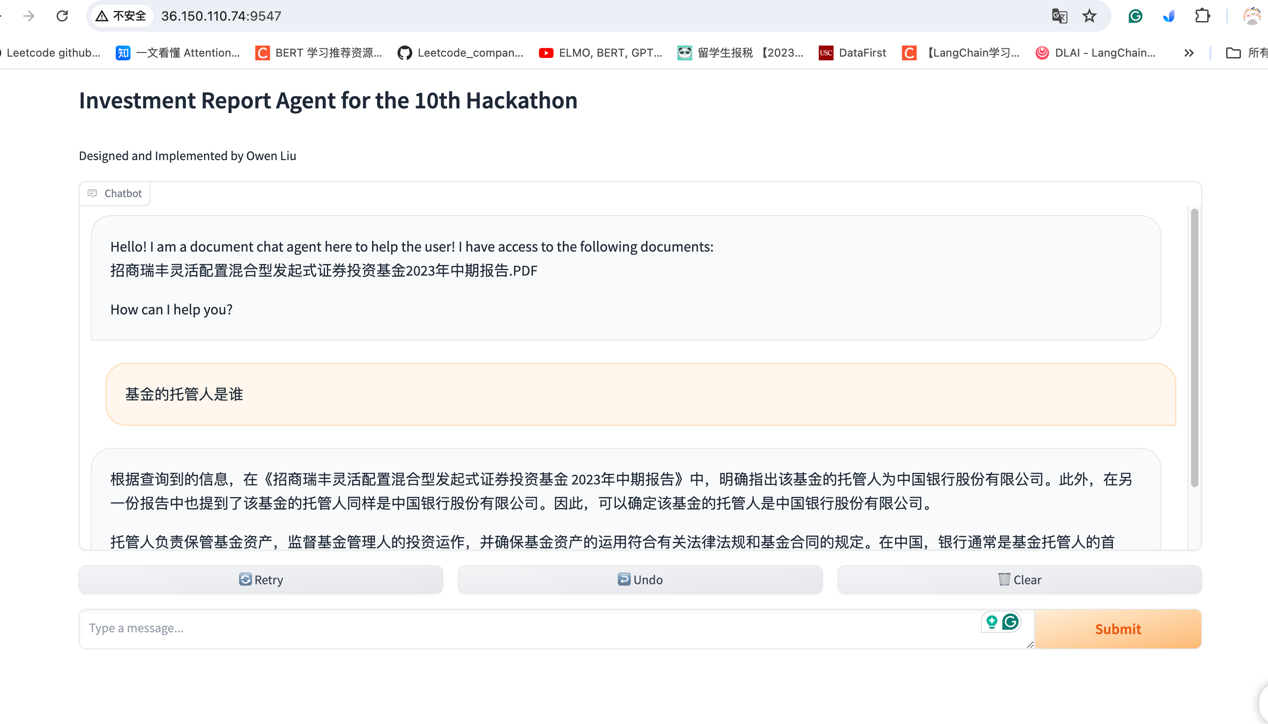
训练时间不足，在文本总结进行高精度索引的部分时，对于166chunks的总结消耗了3个小时左右，未来应重新选用高速的模型进行总结。在多智能体框架中，应该更加注重时间墙的限制，提高响应速度从而提高用户体验。训练结果良好，nomentron-340b的强大功能保证了中间结果的可靠性，但是消耗了更多的时间。优化部分未来可以采用小参数的模型来进行嵌入（不推荐于此次比赛，因为基金报告任务中，信息的准确性和细致性是任务的主要能力要求，而不是快速响应能力）。

### 描述一下你们内为的技术创新点，包括但不限于在组委会提供的技术点

技术创新点同一，采用大量RAG技术优化模型返回的答案，包括但不限于多个阶段的优化，在数据嵌入，检索前，检索中，检索后都有各种方案提供解决问题，由于时间限制，优化的空间还很大。

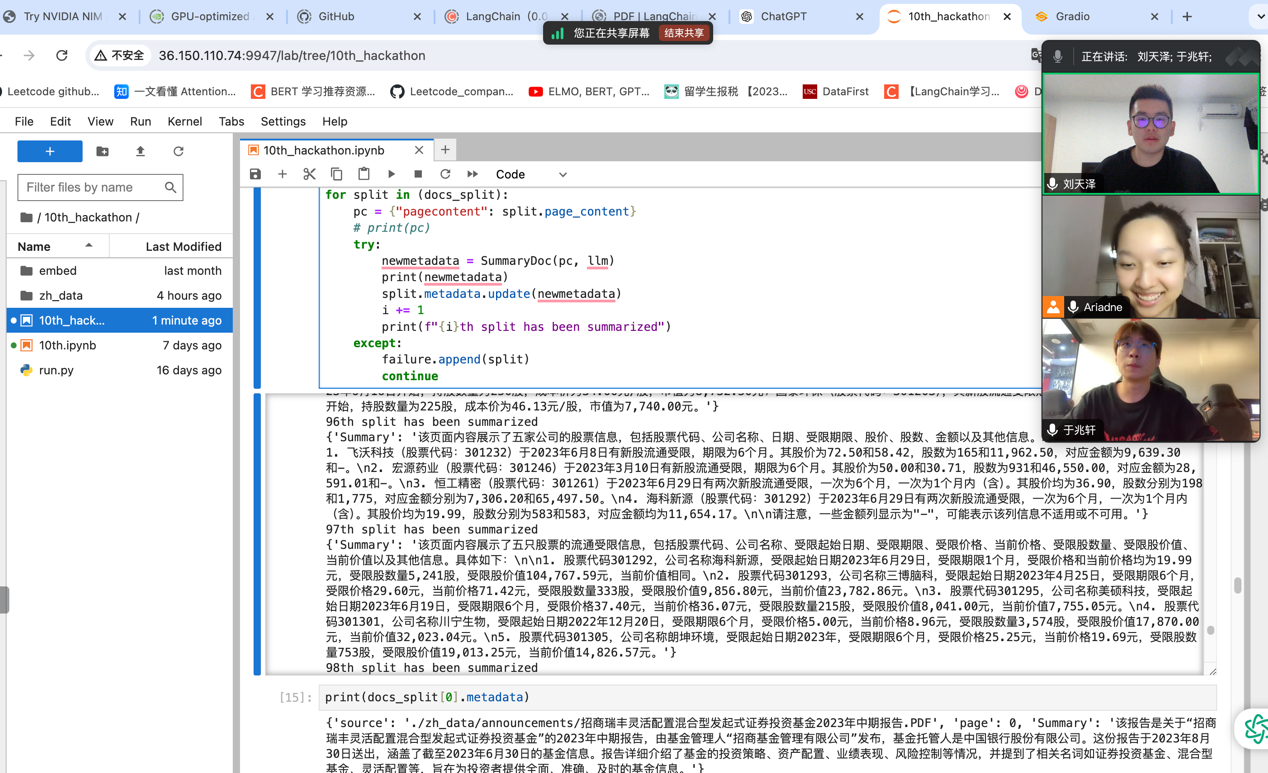
### 描述一下您们是如何优化UI页面的

UI希望页面简洁并具有功能性，能够给用户提供很好的使用体验。时间有限还有很大的提升空间。



## 介绍一下团队（团队合影）

团队由三名队员组成，来自国内外知名高校。



### 总结（团队收获）

1. 收获

非常感谢Nvidia开发者社区提供的这次宝贵的实践机会，从本次比赛中包括课程中学会了很多基于langchain架构的个性化AIGC框架的搭建方式，利用模型搭建属于自己的AIGC Agent并且在实际情境中应用并且基于应用中出现的问题重新调试框架。

1. 遗憾

由于时间关系，本次作品还有大量可以优化的模块和flow parts等待未来完成。例如重排序的方法和模型，向量库的索引优化，知识图谱graphRAG的实现，以及各种丰富知识库内容和结构的方法可以实现。

1. 希望

希望我们Nvidia Hackathon越办越好！下次比赛有更多的时间参加我们的比赛，希望在下次比赛中能够拿出更好的作品和更加创新的ideas和各位优秀的开发者前辈们交流提升！