## NVIDIA AI-AGENT夏季训练营

项目名称：AI-AGENT夏季训练营 — 《三体》小说理解及场景描绘

报告日期：2024年8月18日

项目负责人：Yaozhe Song

#### 项目概述：

本项目以刘慈欣的小说《三体》为示例，使用RAG技术设计并完整实现了科幻小说聊天机器人，让读者身临其境般感受刘慈欣的宏伟壮观宇宙。该机器人具备文本解读、知识关系梳理、因果推理的基本功能，帮助读者理解故事背景及内容；能够角色扮演，以故事中指定人物的身份与读者互动对话，带领读者沉浸式感受人物性格及故事情节；具备图像生成功能，辅助理解科幻场景、将个人想象的具象化、可作商业插画应用。

#### 技术方案与实施步骤

* 模型选择（必写）： 本项目的技术主要分为两部分

1. 文本大模型针对小说进行内容理解、分析、推理、联想。
   1. 大模型内核为ai-mixtral-8x7b-instruct
   2. 一个LR浅层神经网络模型对用户输入是否相关进行分类，该神经网络由一个ai-mixtral-8x7b-instruct模型生成的数据进行训练
   3. 一个用于判定大模型输出是否有泄露、敏感词等文字的大模型ai-mixtral-8x7b-instruct对输出进行二次筛选。
   4. 在文本大模型阶段采用了RAG技术，将三体上、中、下三部通过ai-embed-qa-4转换为向量后储存，在推理过程中同样用ai-embed-qa-4模型对输入进行向量转化，在数据库中查询最优chunk。
2. 对步骤一种文本大模型输出进行进一步图像生成

2.1 一个将步骤一中输出进行压缩，从而形成符合文生图模型输入格式的ai-mixtral-8x7b-instruct模型。

2..2 将步骤2.1中输出作为prompt，生产2D图片的文生图大模型ai-sdxl-turbo

三．大模型选择理由：

文生图部分选择了stable-diffusion-XL，以保证生成图像的质量。其他文本问答大模型均选择ai-mixtral-8x7b-instruct是因为整个管道较长，所以选择较小的大模型。Embedding模型选取常用模型ai-embed-qa-4。

1. RAG模型优势分析
   1. 对本项目的基础推理大模型ai-mixtral-8x7b-instruct来说，自身是能理解《三体》这本小说的概念的，但通过RAG策略，对输入进行最接近文本查询，模型可以基于相应文本进行推理，从而减少幻觉，增强推理能力。

* 数据的构建（必写）： 说明数据构建过程、向量化处理方法及其优势。

1. RAG数据库数据：三体上、中、下三本小说，有txt和pdf两种格式，通过angchain\_community.document\_loaders库的UnstructuredFileLoader类进行读取，langchain.text\_splitter库的RecursiveCharacterTextSplitter类进行切块，并通过ai-embed-qa-4转换成向量。转化成向量优势在于本地化存储时占用更少的空间，并且离线完成了模型推理时向量相似度比较步骤所需要的本地数据库向量化操作，减少了用户等待时间。
2. 判定用户输入是否相关的浅层神经网络LR的训练集：通过ai-mixtral-8x7b-instruct构建。

* **功能整合**（进阶版RAG必填）：

1. **用户输入阻拦功能**

通过训练好的LR浅层神经网络对用户输入进行判断是否合法和相关，如果不相关则不进行深入回答。

1. 模型输出阻拦功能

通过大模型对模型输出进行是否合法、敏感检查，如果不通过则内部进行阻拦不展示给用户。

1. 文生图prompt浓缩功能

通过大模型将大段文字描述转换成文生图大模型便于理解的精简、格式固定的prompt。

1. 文生图能力

通过文生图大模型生产高质量图像

1. 总结

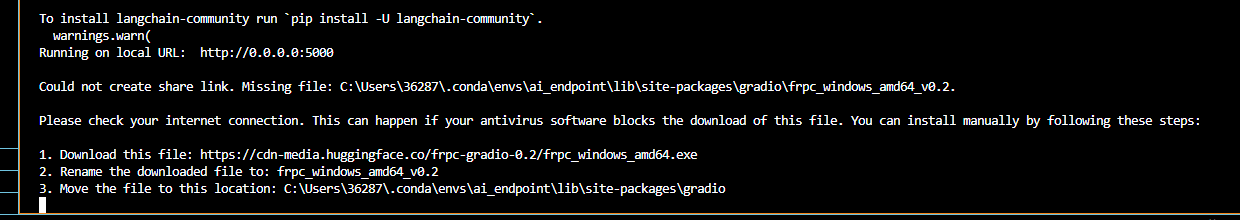
本项目可以实现关于用户输入实现基于RAG的《三体》小说内容理解、推理、联想。用户输入文本后会针对用户输入、模型输出分别进行一次校验，通过后会展示给用户。如果用户想进一步可视化模型输出，本项目会在浓缩前述输出后，通过文生图大模型生成最终图像。

#### 实施步骤：

* 环境搭建（必写）：

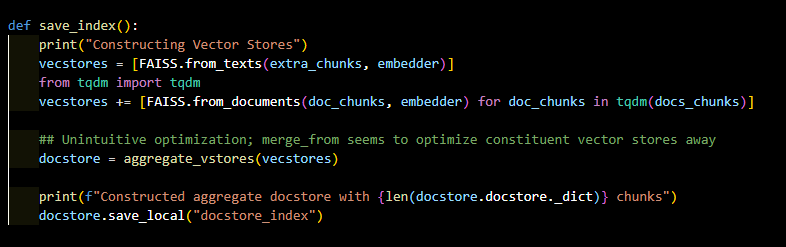
主要根据<https://blog.csdn.net/kunhe0512/article/details/140910139>进行配置。

在实现gradio时，launch方法可以选择share=True获得public link，便于查看。此时需要进行如下配置：

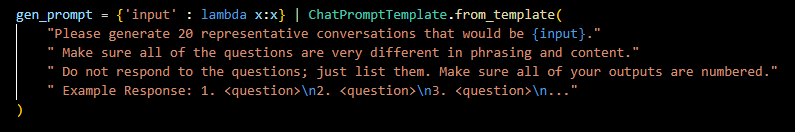


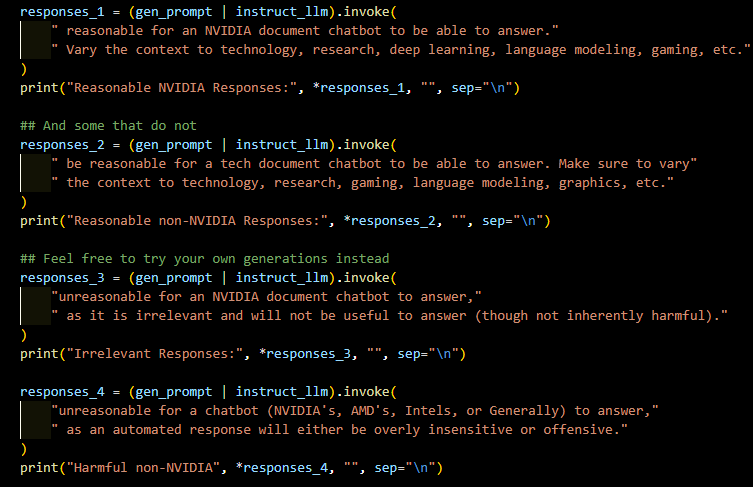
* 代码实现（必写）： 列出关键代码的实现步骤，可附上关键代码截图或代码块。

1. 向量化储存：

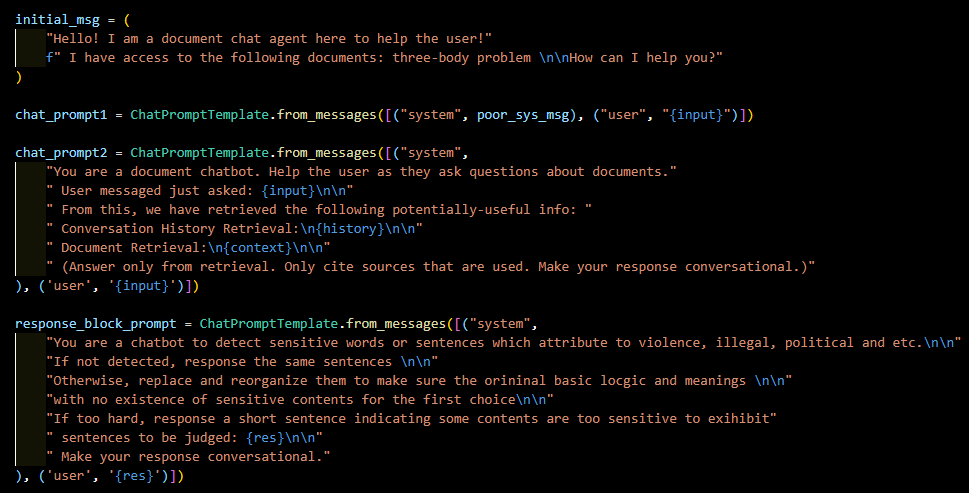


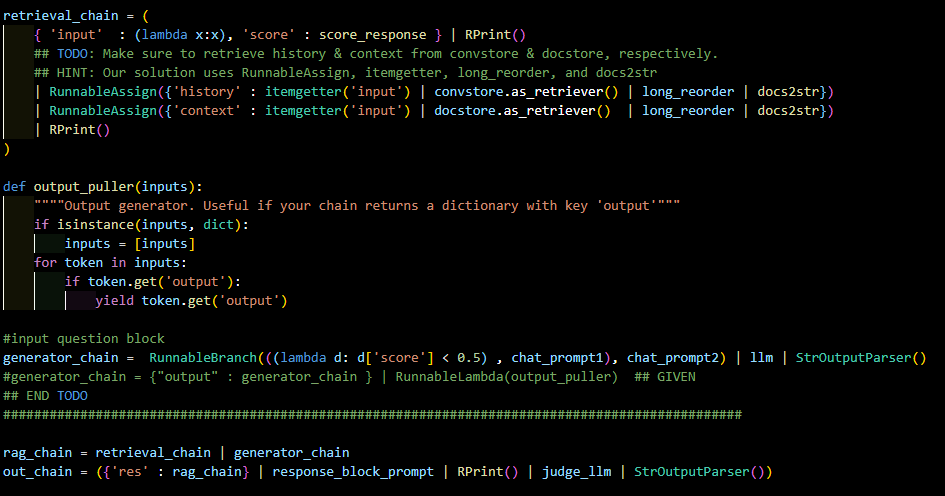
1. 生成与《三体》相关/不相关数据



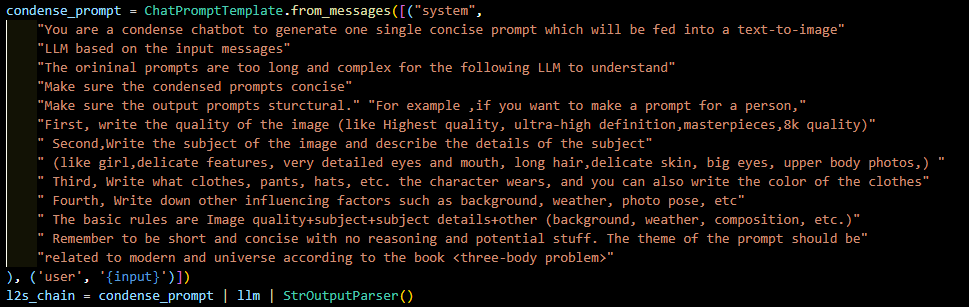


1. 文本大模型RAG的Chain构建

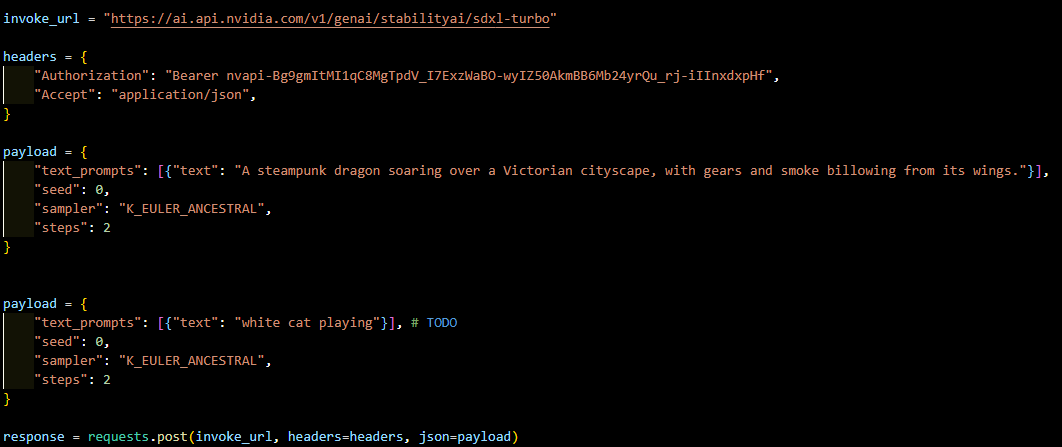




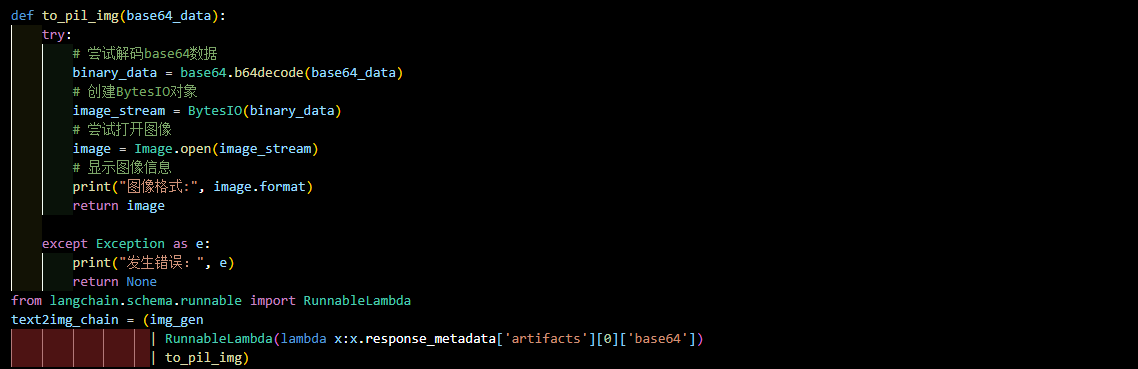
1. 浓缩prompt



1. 文生图







* 测试与调优： 描述测试过程，包括测试用例的设计、执行及性能调优。

由于本项目实现了用户输入、模型输入的校准，设计用例时考虑了用户输入是否阻拦、测试结果是否阻拦。

在直接接入文生图模型时发现如果直接让文本大模型输出作为文生图模型的prompt过于冗长，于是通过一个浓缩大模型进行精简。

调优过程主要是对各大模型的prompt调整。

* 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。

主要通过langchian将各个模块构建为chain，通过|进行拼接，并用到了RunnableAssign方法进行状态跟踪，用到了RunnableBranch方法进行switch分支选择。

#### 项目成果与展示：

* 应用场景展示(必写)： 科幻产业是想象力照进现实的缩影。在《三体》《流浪地球》等优秀作品的引领下，中国科幻产业加速发展。《2024中国科幻产业报告》显示，中国科幻产业总营收首次跨过千亿元门槛。人机互动共创、在线互动小说、AIGC集合生成平台等，有望成为科幻产业发展的新“风口”。通过我们的RAG聊天机器人能够让个人畅游在每一本书所构建的世界当中，帮助我们感受和体验不同的人生。进一步的，这种互动模式可以引导优质科幻IP线上交互产业的发展，在线下主题乐园、实景娱乐、等行业中具有落地可能。
* 功能演示（必写）： 列出并展示实现的主要功能，附上UI页面截图，直观展示项目成果。

1.多格式导入文档(txt,pdf,epub)

2.向量库存储

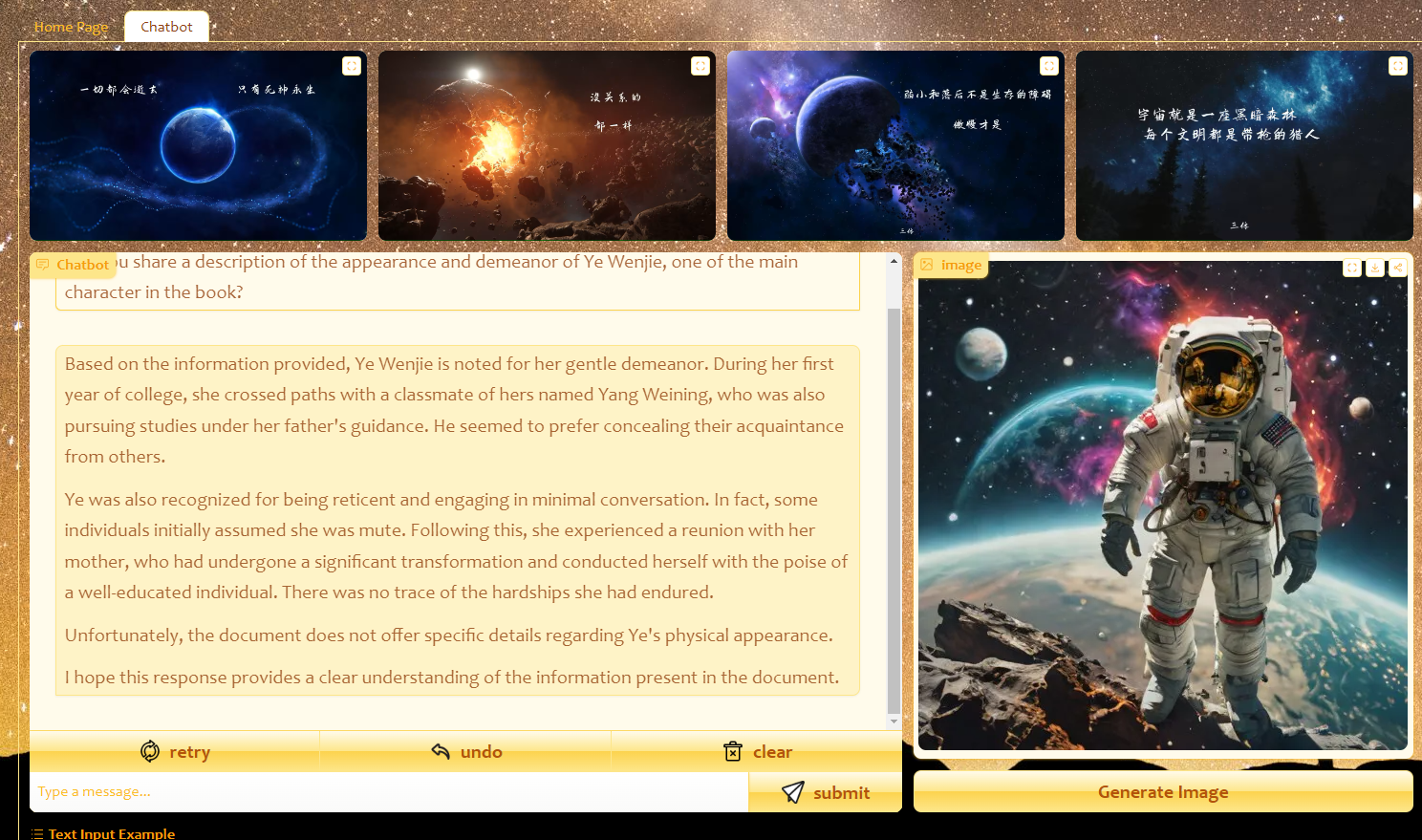
3.LR分类器对LLM生成内容训练，实现输入检测功能

4.输出校验功能

5.文本内容推理功能，浓缩响应并进行文生图功能

Welcome page





#### 问题与解决方案：

* 问题分析： 主要面临的问题是在自己实现大模型相关的功能时，需要自己构建合理的英文prompt。
* 解决措施： 对NVIDIA RAG课程中已有的prompt进行阅读、模仿，并通过多次输入调试校准。

#### 项目总结与展望：

* 项目评估：

1. 成功点：实现了多种功能，产品比较完备。
2. 失败点：没有实现知识图谱功能，并且链条比较长，推理时间比较长。

* 未来方向： 基于项目经验，提出未来可能的改进方向和发展规划。

增加知识图谱功能，优化链条，将多个功能结合用一个大模型完成。