** York College – CUNY**

**YorkGPT Project Proposal**

|  |  |
| --- | --- |
| **Project Title** | YorkGPT Chat LLM |
| **Project Sponsor** | Internship Stipends supported by the S-STEM program |
| **Project Manager** | Rafael Nunez; Director Web and Business Intelligence Information Technology |
| **Mentor** | Dr. Abu Kamruzzaman; Faculty, Business & Economics |
| **Developer** | Darryl Nurse & Shashank Sigdel; S-STEM scholars |
| **Proposed Start Date** | June 1st 2024 |
| **Project Duration** | **Summer 2024** |

Table 1: Project Proposal Information

**The Project Proposal has five sections:**

I. Background of the project

II. Objectives of the project

III. Expected results of the project

IV. Project implementation and management

**I. BACKGROUND**

Analyze the ability of Large Language Models (LLMs) to maintain data accuracy and integrity during Retrieval Augmented Generation. By investigating their capabilities to generate consistent information and assessing the success rate of factual data returned to the user, this project seeks to provide insight into the reliability of LLMs in text-based and chat-based contexts. When seeking to implement LLM solutions in systems with dynamic and pertinent data, maintaining data accuracy and integrity is a priority, to ensure users can rely on the service.

**II. OBJECTIVES**

1. To produce quantifiable insights into the dependability, constraints, and range of Retrieval Augmented Generation in the York GPT LLM.
2. To achieve a response accuracy rate of approximately 90% in test queries related to York College information.
3. To integrate a seamless, secure interface for users to query the LLM system.
4. To attain a high positive feedback score from users on the usefulness of the product.

**III. PROJECT IMPLEMENTATION AND MANAGEMENT PLAN**

**A. Expected project results**

This project is expected to yield an integrated application which allows members of the York College Body to efficiently ask questions to a York-specialized LLM-powered chatbot and receive a response related to their query, regarding information about York College, such as course schedules, upcoming events, and student and faculty resources. This service, once implemented, will improve the ability of students and faculty to quickly and efficiently access new and relevant information relating to their courses, campus events, opportunities, and resources. The system will also serve to reduce the workload of in-person informational faculty, as students will be able to efficiently find information directly from the website, through the chatbot. There will also be a feedback method in place, where users can grade the usefulness of the generated response. This information will serve as a guideline to improve the training and retrieval data of the LLM.

**B. Project activities and work plan**

YorkGPT will be developed and implemented by Darryl Nurse and Shashank Sigdel, under the guidance of Dr. Abu Kamruzzaman, and management of Rafael Nunez.

**Proposed Project Timeline:**

**June**

* Determining Pre-Trained Model.
* Determining Training Platform.
* Collecting Training Dataset in Document Store.
* Training LLM on Dataset.

**July**

* Continued Training LLM on Dataset.
* Packaging Model with Ollama for Langchain Interface Integration.
* Setting up York Website Data Scraper and Cleaner.
* Automation of Data Pipeline.
* Developing LLM Feedback Loop for Model Finetuning.
* Initializing Vector Store for Retrieval Augmented Generation.
* Assembling Application for Backend Testing.

**August**

* Structuring Security Details.
* Developing User Interface for YorkGPT on York Website.
* Hosting Backend on York System.
* Frontend and Backend Connection Testing.
* Final Adjustments and Deployment, feedback testing and improvement.

Project Tech Stack

* Langchain
* Ollama
* York Website API

**Project Tasks**

|  |  |
| --- | --- |
| Darryl Nurse – LLM Developer | Shashank Sigdel – Data Engineer |
| LLM Training and Finetuning | York Website API Development |
| LLM Testing and Optimization | Data Collection |
| Langchain Application Development | Data Handler Development |
| RAG Testing and Optimization | Database Setup for LLM |
| Feedback Loop Integration | Data Pipeline Automation |
| Scalability and Performance Assessment | |
| Security Testing and Development | |
| Client Interface Development | |
| York Backend System Integration | |
| Application Integration and Testing | |

Table 2: Overview of The Division of Tasks Between Developers

**C. Project Beneficiaries**

The people expected to benefit from this project are students, both incoming and continuing, teachers and other faculty, and those related to the mentioned groups.

**D. Implementing agent management of project**

During the development stage the project will be managed by the mentor and project manager as needed. Production implementation will be managed by the project manager.

**Initial implementation requirement stated below:**

**From:** Thitima Srivatanakul <tsrivatanakul@york.cuny.edu>   
**Sent:** Thursday, May 16, 2024 4:34 PM  
**To:** darryl.nurse@yorkmail.cuny.edu; shashank.sigdel@yorkmail.cuny.edu  
**Cc:** Abu Kamruzzaman <akamruzzaman@york.cuny.edu>; Radoslaw Wojciechowski <rwojciechowski@york.cuny.edu>  
**Subject:** York GPT Project - Stipend

Dear Darryl, Dear Shashank,

I understand that you are working on the York GPT Project under the mentorship of Dr. Abu and seeking an internship to work on this project with our IT department.

I'm happy to inform you that the S-STEM program will be able to provide you with a stipend for this internship with IT.

A stipend of $2,000 will be awarded upon the completion of the 12-week internship between June - August 2024, contingent on a satisfactory evaluation by your mentor (Dr. Abu) and the IT department.

*This offer is pending approval of your acceptance into the IT internship.*

**You will need to maintain a weekly logbook, attend scheduled trainings and meetings as needed, and work 20 hours per week.**

If you have any questions, feel free to let us know.

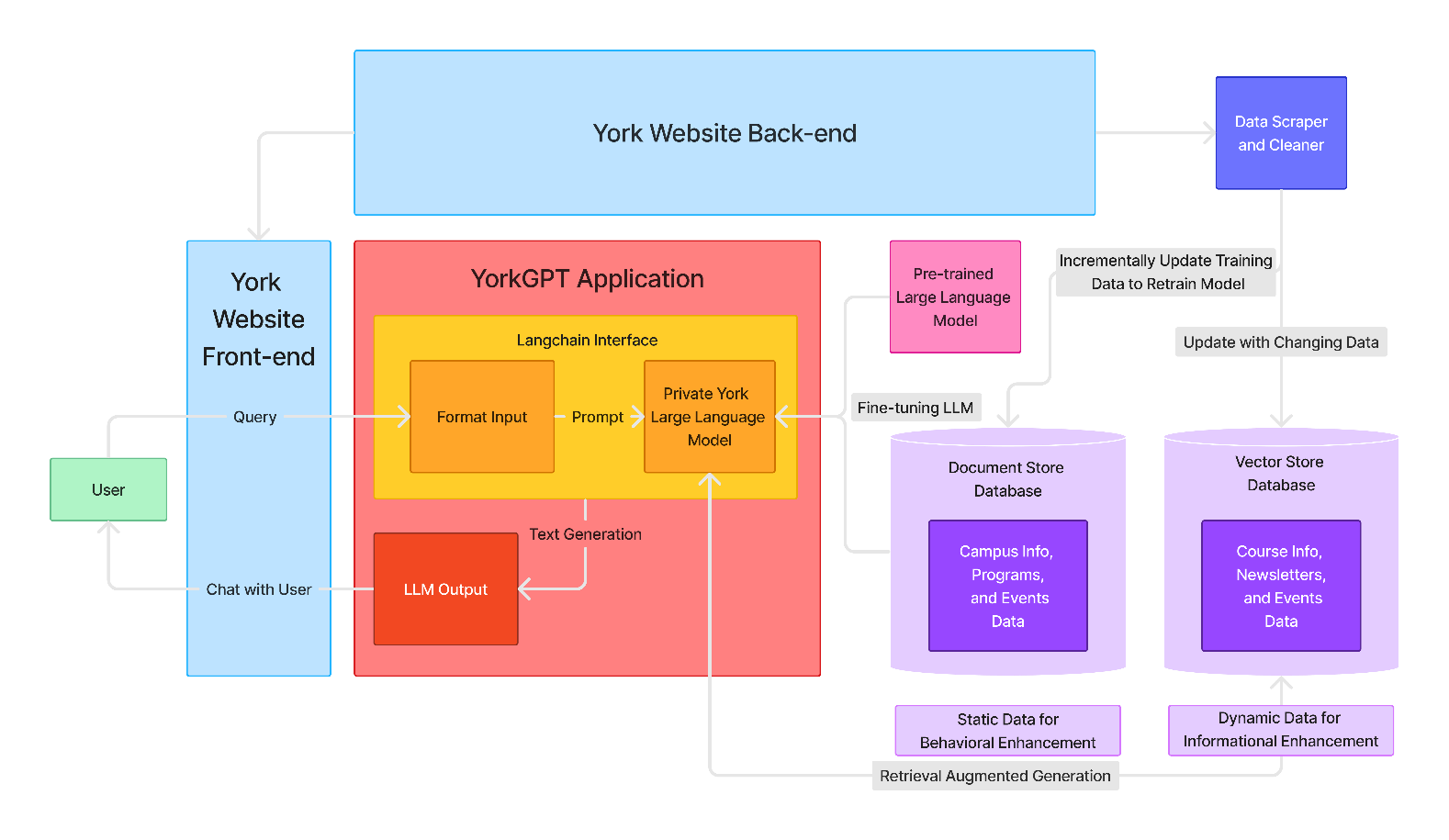
Best,

Thitima

**IV. PROJECT MONITORING AND EVALUATION**

The project will be initially monitored by the project mentor and project manager. Project sponsor will be notified with the updates as time progresses.

Once project is handed over to IT the IT team will be responsible to monitor and update as needed.

APPENDIX: Check attached “YorkGPT System Design” for Full Image

Proposed YorkGPT System Design

System Design Overview

The YorkGPT System uses Langchain to create an interface to provide responses and retrieval methods for the Private Large Language Model (LLM) trained on York College data. This data will be stored in the Document Store Database, which will contain static information about York College. This information will serve to directly bolster the LLMs efficiently in York College context. Using Langchain, the LLM will also be able to perform Retrieval Augmented Generation (RAG) through the Vector Store Database, which will contain dynamic information about York, such as new events and opportunities. These databases will be operated upon by the Data Scraper and Cleaner Mechanism, which will obtain the data from York Website and siphon it into the correct database.