ChatRPI RCOS Project Proposal

Vision:

The vision for the ChatRPI is to create an open-source search engine tailored specifically to Rensselaer Polytechnic Institute, serving as a powerful information retrieval tool for students, faculty, and staff. The project will consist of several key phases, starting with the collection and preprocessing of data from university websites (data scraping), followed by the development of a robust search engine core capable of efficient and relevant retrieval. Machine learning will play a pivotal role, enhancing query understanding, ranking, and personalization, while also addressing quality assurance concerns like anomaly detection and spam filtering. We will build off previous Language Endowed Intelligent Agents and directly improve on those models to be beneficial for ChatRPI specifically. The user interface will feature a user-friendly web extension application with a search bar and user feedback mechanisms. Furthermore, we aim to create user profiles and ensure compliance with legal and ethical considerations such as data privacy and copyright. The project's long-term vision encompasses continuous improvement, updates with the data scraping as more RPI webpages are created, user adoption, and potential integration with other university systems, ultimately establishing the search engine as a vital resource for the university community. Flexibility and adaptability will be key as the project evolves and responds to user needs and emerging technologies.

In the initial phase of the project, data collection for three web domains is the focal point. For RPI websites, a text scraping approach is employed across all affiliated sites, and a custom API is developed for easy access to the extracted data. Collected data undergoes preprocessing to address issues such as irrelevant information and standardize formats. We will test and integrate a variety of LLM models such as GPT 4.0, WebGPT, Llama2, and Falcon 180B to see which one will be best suited for ChatRPI. The integration of these models will require both custom APIs for RPI websites and academic papers. Testing involves deploying a customized version of the "retrieval" plugin, connecting to external APIs, and selecting a vector database for efficient data retrieval. In the final phase, rigorous testing using ChatBot and Dimon evaluates chatbot performance, conversational flow, and user experience, with tailored test scenarios ensuring robustness. The chatbots are then integrated into a user-friendly Google Chrome extension, and iterative feedback guides optimizations for enhanced performance and user satisfaction.

Stack:

- Python
- Pytorch to train
- SQL to store user data
- React for front end website
- Django to serve the web application

Goals:

- Finish making a scraping program that will take all texts from every RPI affiliated website. The stored data should be structured so that all the texts from a single website also contain the original URL that it came from.
- Using all the scraped data we will feed OpenAI language model using its API.
- Create a google chrome extension that uses OpenAI's API that we gave the scraped data to create a smaller scale implementation of our project.

Milestones:

Milestones for January

- Research how the web retrievals work for different LLMs
  - Be sure to document this process
- Get everyone set up with Jira

Milestones for Feburary

- Start and finish scrapping data
- Add all the data from the scraping program to OpenAI's API.
- Start implementing the google chrome extension in junction multiple LLMs
  - Document everything!!!

Milestones for March

- Continue working with web retrievals for the extension
  - Start thinking about which LLMs works best with ChatRPI specifically and document the reasons why
- Finalize the UI for the extension

Milestones for April

- Pick the best LLM for ChatRPI and state why
- Make RCOS presentation and have multiple demonstrations ready

Current Team Members (name, email, discord, github, credits):

- Solomon Starkes, starks4@rpi.edu, snackingchubba, solostarkes, 2 credits (project lead)
- Munzir Abdelgadir, abdelm7@rpi.edu, wunzir, Wunzir, 1 credit
- Sidy Thiam, thiams@rpi.edu, mangobi2, sthiam915, 1 credit
- Elnuman Logman, logmae@rpi.edu, e_man0, Elog0, 4 credits
- Brandon Boston, bostob@rpi.edu, Book of M, BrandonBoston-ASC5,4 credits