Name: Jake Brulato

Date: 12/3/2024

Project Title: Prompt Engineering With Re+Act Framework and Layer Skipping Large Language Models

Mentor and Company: Anwesha Bhattacharyya (anwesha.bhattacharyya@wellsfargo.com)

Company: Wells Fargo

Dates of Internship: 10/28/2024 – 03/28/2025

Project Objective: The deliverables have largely remained consistent with the initial proposal. The primary goals include creating a collection of prompts to effectively evaluate the capabilities of LLMs on fundamental tasks and researching and consolidating the best prompting guidelines into a single resource to enhance prompting practices. Additionally, the project incorporates the ReAct framework to address issues such as hallucination and error propagation in chain-of-thought reasoning, improving interpretability and performance on interactive decision-making benchmarks like ALFWorld and WebShop.

Methodology: **Primarily, we utilize public data to fine-tune multiple LLMs before implementation. This process involves researching various engineering hubs and rewriting their code implementations to create a fully localized system. The system follows guidelines to either generate answers from preloaded knowledge or leverage a Retrieval Augmented Generation (RAG) architecture. While initial testing and adjustments are conducted using public data, we transition to internal data after tuning to ensure security and compliance. Final testing and refinements are carried out in the company's proprietary engine, where extensive tweaking and guard-railing occur.**

Major Tasks: (in major chunks that may (or may not) be ~ 2 weeks of work and takes you to end of internship)

Task 1.1 Project Introduction and Initial Prompts for Prompt Hub(November 19th – November 29th, 2024)

- Look at verified Wells Fargo LLMs for deployment and create a structure proposal for prompt hub.
- Access LLMs in online Python compute engine and start working to create prompt hub
- Read more about the goal of why multiple LLMS are needed and the research is important to Wells Fargo.
- Creating initial prompts based on these LLMS and record how they all act differently when assigned the same prompt.
- Playing with the existing LLM's in Wells Fargo internal network, seeing how they particuallry coded their LLM's and tweaking my current code to theirs.
- Goal is to make a one stop solution for developers, validators, and researchers so they can quickly identify how to structure their prompts among the different models.
- At this current point it is initially completed but is more of a process that goes along the whole project.

Task 1.2 More Instruction Tuning, Engineering, and integrating ReAct (Dec 2nd ,2024– Mar 28th , 2025)

- Researching Chain-of-thought prompting and action plan generation.
- Introduce and read more about ReAct (Reason + Action = React) and its two interactive benchmarks.
- Going back and tweaking prompts when needed for specific roles assignments. Should be a back and forth with the initial prompt hub.
- Goal of the ReAct framework is to mimic a similar case for Wells Fargo, tweaking the reasoning for the business need by taking both private and public information to output an answer with 1-2 context examples.
- Depending on timeframe of ReAct implementation, may introduce layer skipping to the mix for specific Llama models by Meta.

Task 1.3 Completing ReAct and Final Presentations (Jan 20th – Mar 28th, 2025)

- Complete ReAct for internal models and start consolidating final research presentation.
- Tweak any findings for presentation to the Natural Language Processing team either the third or fourth week of March. Advanced Technologies for Modeling members will also be present as they are being worked with in conjunction to my the Natural Language Processing Team.

Outcomes Expected: Leveraging the research on LLMs tailored to specific business needs can significantly enhance the likelihood of developing finely tuned models for multiple divisions. By employing specialized LLMs with advanced reasoning capabilities, the Model Risk Framework at Wells Fargo can be further optimized. One specific business need is enabling validators to review documents more efficiently, reducing processing time and improving accuracy. Additionally, these LLMs can be utilized to rigorously test model boundaries, ensuring robust performance and compliance with organizational standards.

Potential Risks and Strategies to Overcome :

Wells Fargo operates under strict security regulations, which can present several potential obstacles that may delay project progress. One significant challenge has been the lack of immediate access to research materials or environment variables within the internal Python cluster. Each LLM model requires a specific environment configuration, and this limitation often causes delays. Additionally, websites such as Hugging Face and Mistral, which host crucial model card documentation, are blocked due to the presence of chatbots, requiring me to request access for each blocked resource. The time zone difference with the Advanced Technologies for Modeling team in India has also led to occasional delays in receiving progress updates. Furthermore, while my primary focus is this project, there are instances when my mentor or functional manager requests immediate assistance with other tasks, which slightly elongates the timeline.

To overcome these obstacles, I have implemented strategies such as proactively requesting necessary permissions for blocked resources, documenting detailed access requirements in advance, and scheduling regular progress updates that accommodate time zone differences. Additionally, I prioritize

effective communication with my mentor and manager to balance urgent requests with ongoing project responsibilities.