Name: Jake Brulato
Date: 17 Jan, 2025
Project Title: Prompt Engineering and Framework Conversion for MRM
Mentor and Company: Anwesha Bhattacharyya, Wells Fargo
Period Covered: 01 Jan – 17 January 2025

Project Objective: Developing a centralized repository of tailored prompts for Large Language Models, specifically optimized for the DSAI team within Wells Fargo's MRM sector, incorporating a ReAct (Reason + Action) framework for open-source LLMs after prompt hub creation.

Progress Summary:

Work Completed:

- Task 3.3. Creation of Prompt Hub (IN PROGRESS/DELAYED)
    - Summary: The Prompt Hub is a centralized resource being developed to optimize prompt usage across models. Its purpose is to streamline model-specific performance for varied generation types and user requirements. Progress highlights include:
        - Centralizing details such as model size, creation dates, training data, and token limits.
        - Iterative testing to develop structured, efficient prompts with examples and variations for usability.
        - Adjustments are ongoing to ensure the tool is accessible to users with minimal experience. Frequent feedback has delayed timelines due to evolving business needs.
    - Lessons Learned:
        - Different models excel in specific contexts:
            - Faster inference for real-time applications.
            - Superior generation for domain-specific outputs.
            - Enhanced ability to interpret text and visual tasks.
        - As part of Wells Fargo's internal LLM initiative, significant effort is being invested in minimizing barriers for end-users to enhance productivity.
    - Risk & Mitigation:
        - Risk: Improper prompts can lead to model hallucinations, resulting in fabricated or inaccurate outputs that could disrupt decision-making.
        - Mitigation: Introduce visual indicators to distinguish accurate outputs from hallucinations. Document common pitfalls and provide structured guidance for optimal prompt usage.
- Task 4.1 Meet with Team to Discuss ReAct for Loan Harmonization (COMPLETE)
        - Summary: The team met to discuss the integration of the ReAct framework for a shelved loan harmonization project. Key outcomes included:

        - Reassessment of structure with Newer LLMs were identified as a better fit compared to previously used models.
        - Previous errors were reviewed with the goal of transitioning to ReAct prompting instead of zero- or few-shot approaches.
        - Initial setup was completed using LangChain to enable external and retrieval-based tasks.
    - Lessons Learned:
        - Other business units reported limited success with similar models, leading to delays or project cancellations.
        - Many validators lack familiarity with model functionality. Providing structured guidance and implementing ReAct could significantly improve project outcomes.
    - Risk & Mitigation:
        - Risk: External information retrieval via LangChain may be restricted, limiting functionality. Additionally, document structures may hinder performance with longer contexts.
        - Mitigation: Develop document chunking strategies to optimize retrieval-augmented generation (RAG) and ensure compatibility with internal systems.

<u>Work Scheduled Next Sprint</u>

- Task 3.3. Creation of Prompt Hub
    - Update presentations based on feedback, adding detailed explanations for users with little experience.
    - Create both simplified and detailed versions of presentations while integrating ReAct framework principles.
- Task 4.2 Building ReAct and Testing with Public Data
    - Fine-tune testing data from external sources for compatibility with the online compute platform.
    - Test on loan harmonization documents and develop an initial framework.
- Task 4.3 Add ReAct to Loan Harmonization Models
    - Refine the framework with LangChain for project-specific models.
    - Compare chain-of-thought prompting with ReAct reasoning in a clear, efficient format.
- Risk & Mitigation:
    - Risk: LangChain may face internal restrictions on external retrieval, despite indications from team members that it is approved.
    - Mitigation: Advocate for enabling LangChain's retrieval feature or develop a RAG pipeline to enhance model reasoning capabilities.

Work Package Updates:

	WP	Work Package Title	Scheduled Dates	Status
~~✓~~	~~1.~~	~~Onboarding~~	~~28th Oct – Nov 22nd~~	~~Complete~~
~~✓~~	~~2.~~	~~Testing performance of Large Language Models and Configuration~~	~~Nov 25th – Dec 13th~~	~~Complete~~
	3.	Creation of Prompt Hub for DSAI and MRM	Dec 16th – Jan 30th	In-Progress
~~✓~~	~~3.1.~~	~~Research Model Specifications and Prompting Techniques~~	~~Dec 16th – Dec 18th~~	~~Complete~~
~~✓~~	~~3.2.~~	~~Testing Prompts around research~~	~~Dec 16th – Dec 20th~~	~~Complete~~
	3.3.	Creation of Prompt Hub	Dec 18th ~~– Dec 30th~~ Jan 30th	Backlog
	4.	Implement ReAct Framework	Jan 1st – Mar 28th	In-Progress
✓	~~4.1.~~	~~Meet with Team to Discuss ReAct for Loan Harmonization~~	~~Jan 13th – Jan 17th~~	~~Complete~~
	4.2.	Building ReAct and Testing with Public Data	Jan 6th – Jan 17th	In-Progress
	4.3.	Add ReAct to Loan Harmonization Models	Jan 17th – Mar 28th	In-Progress