Name: Jake Brulato
Date: 28 Feb, 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 4.2 Building ReAct and Testing with Public Data (Complete)
    - Summary:
        - Internal package conflicts are still occurring, and came to conclusion that the tools for the ReAct agent needed to be built in-house.
        - Created a RAG framework using Llama 2 on local machine with defined functions to process multiple document types using external data to test and chunk it.
        - Implemented a multi-step "thought", "action, and "reflection" based on the user query.
    - Lessons Learned:
        - Llama 2 may not be strong enough to handle multi-document processing and will need to be changed to Llama 3.1 or at least Llama 3.
        - Different embedding models change the speed of processing chunked data, using a better model is recommended.
    - Risk & Mitigation:
        - Risk: File structure internally may differ from normal a normal pdf or csv in the future, possibly causing issues in the document processing/chunking of data.
        - Mitigation: Added to documentation example file usage and how it would work if used with internal documents.
- Task 5.2 Testing of Visual Capabilities and Text Generation (Complete)
    - Summary:
        - Completed testing of Deepseek-R1 distilled Qwen, SmolVLM, and Qwen2.5 (3B & 7B) with completed examples for use by the ATOM GenAI team.
        - Consolidated to a report showing the text capabilities, nuances, and image/video performance across all models with potential recommendations.
    - Lessons Learned:
        - Model performance isn't entirely limited to the parameter amount used to train the model.

        - Smaller models can have the potential to do great in one task compared to another that can do multiple tasks well.
    - Risk & Mitigation:
        - Risk: Qwen 2.5 and Distilled Deepseek are still very experimental and found to be prone to some hallucinations when prompted in long context.
        - Mitigation: Wrote out specifics that a person should avoid when model is in use, provided proper files for image resizing and processing as well.

Work Scheduled Next Sprint

- Task 4.3 Add ReAct to Loan Harmonization Models
    - Refine the user query to completely follow the Loan Harmonization project guidelines based on example work from another division.
    - Use RAG for the text extraction of fed files through previous outputs. Make it think critically and adapt with each iteration.
- Task 6. Present Findings on Generative Models in AIA Seminar
    - Present in the weekly seminar that MRM hosts the proper guidelines and functionality of core models used in the prompt hub.
- Risk & Mitigation:
    - Risk: Metadata and guidelines may be too complex for the current model to process in a long document string. May cause hallucinations if not properly checked.
    - Mitigation: Pull similarity score from chunks that were used and train a model if need be. If scores are too low, request more resources or switch to another model with a better training method.

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 Feb 16th Feb 22th~~	~~Complete~~

	4.3.	Add ReAct to Loan Harmonization Models	Jan 17th – Mar 28th	In-Progress
~~✓~~	~~5.~~	~~More Generative Model Development~~	~~Jan 28th – Mar 28th~~	~~Complete~~
~~✓~~	~~5.1.~~	~~Research Chat Template for Visual/Large Language Models~~	~~Jan 31st – Feb 5th~~	~~Complete~~
~~✓~~	~~5.2.~~	~~Testing of Visual Capabilities and Text Generation~~	~~Feb 5th – Feb 14th~~	~~Complete~~
	6.	Present Findings on Generative Models in AIA Seminar	Feb 28th – Mar 13th	In-Progress