Name: Jake Brulato
Date: 7 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:

<u>Work Completed:</u>

- Task 3.3. Creation of Prompt Hub (COMPLETE/ADDITIONAL FEATURES ADDED)
    - Summary:
        - Completed the last sprint per requested specifications; uploaded to the internal MRM wiki.
        - Presented to the business Generative AI team on best practices for generation, benchmarking, and training.
        - Recommended upgrading internal LLMs for improved long-context or internal tasks.
    - Lessons Learned:
        - Iterative testing and refinement with Gen AI at Wells Fargo help assess risk before release.
        - Pretrained models may require additional policy and validation checks.
        - Considered complete but serves as a foundation for future model enhancements.
    - Risk & Mitigation:
        - Risk: Public model release may lead to misconfigured parameters or excessive input, causing hallucinations.
        - Mitigation: Human review of LLM outputs is essential. Alternative solutions include using a higher-parameter LLM as a validation checker.
- Task 4.2 Building ReAct and Testing with Public Data (IN PROGRESS/DELAYED)
    - Summary:
        - Built a testing framework using LlamaIndex and Langchain on a personal computer.
        - RAG and API have been tested with functional internal-use code.
        - Package conflicts have delayed internal testing; developers are addressing these while working on task 5.1.
    - Lessons Learned:
        - API specifications vary by package version, requiring adjustments in internal code.
        - RAG improves generation precision but differs from Chain-of-Thought prompting, reducing model evaluation time.

    - Risk & Mitigation:
        - Risk: Unclear queries may lead to hallucinations or irrelevant answers not based on internal RAG or API context.
        - Mitigation: Ensure queries focus on internal data; if no relevant information is found, return “Sorry not found.”
- Task 5.1 Research Chat Template for Visual/Large Language Models (COMPLETE)
    - Summary:
        - Added new requested models for potential prompt hub implementation.
        - Tested Deepseek-R1 distilled Qwen, SmolVLM, and Qwen2.5 (3B & 7B); Qwen is limited to the dev transformer version, unavailable internally.
        - Set up working notebooks for image dimensions, parameters, and multi-image/multimodal inputs.
    - Lessons Learned:
        - Reinforcement learning from deepseeks training methods allows for faster inference and more robust generation compared to normal.
        - SmolVLM’s 256M and 500M struggle with multi-image extraction but match Qwen2-VL-Instruct for single images.
    - Risk & Mitigation:
        - Risk: Deepseek’s capabilities remain untested, and guidelines may differ from Qwen2-VL. SmolVLM’s size may cause hallucinations or errors.
        - Mitigation: Intensive query structuring is needed for Deepseek. For VLMs, test complex images/queries and request dev transformer versions or an approved tokenizer/processor.

<u>Work Scheduled Next Sprint</u>

- Task 4.2 Building ReAct and Testing with Public Data
    - Working environment for testing with code tweaking, all public testing data is loaded and ready for testing. Devbug any errors in generation based on package versions.
- Task 4.3 Add ReAct to Loan Harmonization Models
    - Refine the framework with LangChain/LlamaIndex for project-specific models, compare performance with different versions of Llama
    - Compare chain-of-thought prompting with ReAct reasoning in a clear, efficient format.
- Task 5.2 Testing of Visual Capabilities and Text Generation
    - More complex testing with image, compare and note errors to Qwen2-VL
- Risk & Mitigation:
    - Risk: LangChain/LlamaIndex 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~~ Feb 16th	Backlog
	4.3.	Add ReAct to Loan Harmonization Models	Jan 17th – Mar 28th	In-Progress
	5.	More Generative Model Development	Jan 28th – Mar 28th	In-Progress
✓	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	In-Progress