Assignment 6 - Prompt Engineering Benchmark Framework

Group Code: LLM_Agents_Tom_Igor_Roie

Submitters:

Igor Nazarenko, ID 322029158
Roie Gilad, ID 312169543

GitHub Repository:

https://github.com/roiegilad8/LLM_Agent_Orchestration_HW6

Self-Assigned Grade: 95/100

December 16, 2025

1 Project Overview

This project implements a comprehensive Prompt Engineering Benchmark Framework designed to scientifically evaluate the effectiveness of different prompting strategies across multiple Large Language Models (LLMs).

The system orchestrates a comparative analysis of:

- **3 Models:** OpenAI GPT-4o, xAI Grok-2, and Perplexity Sonar.
- **4 Techniques:** Baseline, Few-Shot, Chain-of-Thought (CoT), and ReAct.

The framework processes a dataset of 100 questions per combination (1,200 total inference calls), utilizes an automated fuzzy-matching grading engine to score accuracy against ground truth, and generates detailed visualizations and cost analysis reports. The entire system is supported by a production-grade CI/CD pipeline, ensuring code quality and reproducibility.

2 Self-Assessment Justification

2.1 Grade: 95/100

2.1.1 Strengths - What We Did Exceptionally Well

Our project demonstrates several areas of excellence that justify this high grade:

Robust Architecture: Built a modular, extensible framework that separates data preparation, grading, and analysis. The C4 diagrams in ARCHITECTURE.md clearly document the system design.

Comprehensive Analysis: Implemented a full CI/CD pipeline with GitHub Actions, strict linting (Ruff/Black), and pre-commit hooks to ensure code quality and consistency.

Production-Grade Quality: We produced complete documentation including a detailed PRD with KPIs and acceptance criteria, three comprehensive ADRs documenting key architectural decisions, a professional README with installation instructions (including Ollama setup for Linux/macOS/Windows), cost breakdown table, and an analysis notebook with statistical insights and visualizations.

Visual Documentation: Generated professional visualizations (heatmaps, bar charts) and included screenshots in the README to provide immediate visual proof of results.

2.1.2 Weaknesses - Areas for Improvement

We acknowledge several areas that prevented a perfect score:

API Cost Management: While we analyzed costs, the current implementation runs sequentially. Future versions could optimize token usage or use batch APIs to reduce expense.

Grading Nuance: The fuzzy matching algorithm is effective but could be improved with an LLM-as-a-Judge approach for more semantic understanding of ”correctness” beyond string similarity.

2.1.3 Time Investment & Effort

This project required approximately **40-50 hours**.
Focus: Significant effort was placed on the data pipeline robustness (data_prep_FINAL.py) and the automated grading logic (compare_results_FIXED.py) to ensure fair comparisons.

2.1.4 Innovation & Unique Aspects

Our project demonstrates several innovative elements:

Multi-Model Comparison: Successfully orchestrated and normalized outputs from three distinct API providers (OpenAI, xAI, Perplexity) into a single unified benchmark.

2.1.5 Learning Outcomes

This project provided valuable insights:

Deepened understanding of how different prompt techniques (Few-Shot vs. CoT) impact model performance differently across various LLM architectures.

Mastered the integration of Python automation with CI/CD pipelines for reproducible research.

Academic Integrity Declaration

We, the undersigned, declare that this submission is our own original work. We have:

- Completed this assignment independently as a group
- Properly cited all external sources, libraries, and tools used
- Not copied code from other students or unauthorized sources
- Not shared our code with other students
- Followed all academic integrity guidelines of the course

We understand that violations of academic integrity may result in penalties including failure of the assignment or course.

Signatures:

Igor Nazarenko, ID 322029158 Date: December 16, 2025

Roie Gilad, ID 312169543 Date: December 16, 2025