PHASE - 3

Introduction

This phase evaluates the performance of three LLMs, GPT-5, Gemini Pro, and Gemma3:4b across three tasks in the Technology domain: Code Summarization, API Question Answering, and API Documentation Generation.
The goal is to identify performance patterns using a rubric-based approach.

Methodology

Three tasks were selected within the technology domain to evaluate how Large Language Models (LLMs) can improve developer workflows through automation and enhancement. Each task was approached using three distinct prompting strategies, CLEAR, Few-shot, and Chain-of-Thought (CoT) to test the effect of prompt engineering on output quality.
Each prompt was then run across three different LLMs: GPT-5, Gemini Pro, and Gemma3:4b, resulting in nine outputs per task (3 prompt styles × 3 models). This produced a total of 27 outputs across all tasks.
To ensure objectivity, outputs were randomized and evaluated blindly using a 4-point rubric across four key metrics:

- Accuracy - factual correctness of the output
- Completeness - coverage of all required elements in the task
- Coherence - logical flow and readability
- Domain Appropriateness & Language - technical precision and context relevance

In addition, a flag variable was used to track the presence of bias or hallucinations, especially for factual or API-related tasks.
The evaluation followed a standardized scoring scale:
4 = Exceptional (exceeds professional standards)
3 = Proficient (meets professional standards)
2 = Developing (noticeable errors, needs revision)
1 = Inadequate (major issues or missing elements)
This structured methodology ensured consistency and reliability in assessing the relative performance of different LLMs and prompt engineering techniques.

Results

Key findings from the excel sheet:

- GPT-5 and Gemini Pro performed almost equally well, both achieving an average score of 3.95, excelling in reasoning-intensive tasks such as Code Summarization and API Documentation.
- Gemma3:4b, while significantly smaller, achieved a respectable average score of 3.64, demonstrating its potential for cost-effective deployment.
- Across prompt styles, Chain-of-Thought (CoT) consistently produced the most accurate and complete outputs for reasoning tasks, while CLEAR was optimal for concise, direct question answering.
- No hallucinations or bias were observed in this evaluation, likely due to the small and controlled dataset. However, minor issues were noted with Gemma occasionally omitting key details in Task 2 of question answering.