GenAI for Software Development: Assignment 3

Benjamin Tremblay
bptremblay@wm.edu

Rowan Miller
rvmiller@wm.edu

1 Introduction

Throughout the semester, we have built up to prompt engineering, and the different strategies that can be used to prompt Large Language Models. In this assignment, we compare multiple AI models using several prompt engineering strategies to evaluate the effectiveness of said strategies.

2 Dataset Preparation

We chose (number) different AI models to use in our experiment: gpt-4o-mini, Codestral-2501, Llama-4-Scout-17B-16E-Instruct, and Ministral-3B. This allows for a wider look at the AI landscape, and chance to analyze modern models.

If not otherwise stated, the temperature used in prompting is .7. This provides a standard basis to test lower and higher temperatures, and lets us test the AI's on an even playing field. The ipynb files in the GitHub repository were used to produce the prompt responses and gather our information. We constructed a CSV file called Problems_Explored where we gathered and organized our data in a viewer friendly way, which we will refer to during the analysis.

3 Analysis

Using the Large Language Models mentioned above, we tried a variety of prompt engineering strategies to answer various questions that one might ask to get an idea of the effectiveness of each prompt. Considering that there is no standardized answer to compare the LLM's answers with, we will provide an analysis for each problem, rating prompting strategies and models to each other.