Rachel Huang
GenAI Project 3 Model Comparison and Prompting Techniques Report

Optimal Model for Response Quality
OpenAI's GPT-4o seemed to have the most accurate and detailed responses, especially for the later tasks that were more complex. This is because GPT-4o has over 200 billion parameters, while in comparison, the next largest model, Gemini-2.0-Flash, has approximately 40 billion parameters. An increase in the number of model parameters allows the model to learn more patterns and capture more details from its data.

Ranked second for response quality, Gemini-2.0-Flash also seemed to be able to provide more information in its responses because of its longer responses. This was especially apparent for more complex prompting strategies such as chain of thought and prompt chaining. For example, in row 63 of Results_Report.xslx where chain of thought is used, the Gemini response is much more documented with comments and step by step explanations as compared to the corresponding Mistral response. In addition, the Gemini response provides more error handling and print statements along the way to make its response more clear to the user.

Optimal Prompting Technique for Efficiency
Zero shot prompting seemed to be able to solve the earlier tasks that involved short functions in a succinct manner, without elongated responses from the other prompting strategies that would still result in essentially the same solution. For example, for the first task where the LLM's are prompted to summarize the short input Java function, the results of few shot prompting with 3 examples did not seem to be significantly different/more accurate than the zero shot prompting response for this task.

Optimal Model for Response Time
Ollama gives significantly shorter and more concise responses compared to GPT-4o, which is due to the fact that Ollama is a smaller model that runs locally and depends on local resources. On the other hand, GPT-4o is a significantly larger model that uses cloud-based resources. Ollama's responses are still accurate, but seem to have trouble adhering to prompting strategies. Thus, Ollama fails, relative to the other models, to generate more complex and detailed responses that would be more efficient and explain its thought process to the user.

Optimal Prompting Technique for Response Quality
The most effective prompting technique seemed to be chain of thought. Making the model detail its reasoning process step-by-step led to more accurate and comprehensive answers. This was especially apparent in the later tasks that were more complex.

For example, this is clear in row 23 of Results_Report.xslx. When GPT-4o is prompted to explain its step-by-step process, the user gets a detailed description of each part of the solution to the task. The results of other prompting strategies do not usually give the user such a clear and concise explanation of its solution and each part of the solution. In addition, chain of

thought is more likely to be able to correct itself, if necessary, during its step-by-step process, as it is explaining how it arrived at its solution.

Comparing and Contrasting Models (Vector Embeddings)
When other models were compared against Gemini-2.0-Flash, their vector embedding scores were usually higher and above 80%. This seems to suggest that most of the model responses had the most in common with Gemini, relative to the other models.