TASK 1 – Rating Prediction via Prompting

This task focuses on evaluating how different prompt designs influence the performance of a large language model when classifying Yelp reviews into star ratings. The objective was to predict ratings from one to five stars using prompting alone and to analyze the impact of prompt structure on accuracy, output consistency, and JSON validity.

Prompt Iterations and Design Decisions

Prompt Version 1 was designed as a basic classification prompt. It directly instructed the model to read a Yelp review and predict a star rating while returning the output in JSON format. This prompt served as a baseline to understand the model's default behavior without additional guidance. However, the results showed very low accuracy and frequent JSON formatting issues. The lack of explicit sentiment-to-rating mapping caused the model to behave inconsistently, particularly for neutral or mixed reviews.

Prompt Version 2 was introduced to address the limitations of the initial prompt. This version included explicit rating criteria that clearly defined what constitutes each star level based on sentiment intensity and customer satisfaction. The purpose of this change was to reduce ambiguity and guide the model toward a more consistent interpretation of review sentiment. As a result, Prompt Version 2 achieved higher accuracy and a better JSON validity rate compared to the baseline.

Prompt Version 3 further refined the approach by enforcing strict JSON-only output and encouraging internal reasoning before producing the final response. The motivation behind this improvement was to increase output reliability and minimize parsing failures, which is critical for real-world applications. While the accuracy did not surpass Prompt Version 2, this version showed improved output stability and better adherence to the required response format.

Evaluation Results and Comparison

A quantitative comparison of the three prompt versions was conducted using accuracy and JSON validity rate as evaluation metrics. Prompt Version 1 achieved an accuracy of 0.02 with a JSON validity rate of 0.02. Prompt Version 2 achieved the highest accuracy of 0.115 and a JSON validity rate of 0.18. Prompt Version 3 achieved an accuracy of 0.05 with a JSON validity rate of 0.08.

Discussion and Observations

Overall, Prompt Version 2 performed best in terms of accuracy, indicating that explicit rating guidelines significantly enhance model understanding. Prompt Version 3 demonstrated better reliability in output structure, making it more suitable for production scenarios where strict formatting is required. The results clearly indicate that prompt engineering has a strong impact on model behavior, even without fine-tuning.

Evaluation Constraints

Due to large language model inference latency and API constraints, the evaluation was conducted on a representative subset of the Yelp reviews dataset. This approach ensured efficient experimentation while still allowing meaningful comparison across different prompt strategies.