Task 1 – Rating Prediction via Prompting

Candidate Name: Misba Yadgiri

Objective
The objective of Task 1 is to evaluate how different prompt engineering strategies affect the ability of a Large Language Model (LLM) to classify Yelp reviews into 1–5 star ratings while producing valid structured JSON output.

Dataset
Source: Yelp Reviews Dataset (Kaggle)
Sample Size: 200 reviews (random sampling)
Fields Used: review_text, actual_stars

Approach
The task was treated as a prompt-based classification problem instead of model training. The LLM was queried using multiple prompt strategies, and predictions were evaluated against ground-truth ratings.

Prompt Designs
Prompt V1 – Naive Prompt: Minimal instruction and weak formatting constraints.
Prompt V2 – Strict JSON Prompt: Explicit JSON-only output with integer rating enforcement.
Prompt V3 – Anchored Prompt: Defined semantic meanings for each star rating with strict output rules.

Evaluation Methodology
Metrics used include Accuracy (comparison of predicted vs actual stars) and JSON Validity Rate (percentage of responses parsable as valid JSON). Exception handling was implemented to ensure robustness.

Results Summary
Prompt V1 showed low JSON validity (~18%) but moderate accuracy (~58%).
Prompt V2 improved JSON validity and accuracy through stricter constraints.
Prompt V3 achieved the highest consistency and overall performance.

Observations
Stricter prompts significantly improve reliability. Anchored definitions improve accuracy and consistency while reducing ambiguity.

Conclusion
Prompt engineering alone can substantially impact LLM performance. Iterative refinement from naive to anchored prompts led to improved accuracy, higher JSON validity, and more reliable outputs.