Fynd Assessment Task-1: Rating vs Prompt Evaluation

I had taken the different prompting strategies that affect the performance of a Large Language Model (LLM) on Yelp review star-rating prediction.
We compare 3 prompt versions:

1. **Direct Prompting**
2. **Chain-of-Thought Prompting**
3. **Few-Shot Prompting**

Each version is tested on a sample of 200 reviews, and evaluated across:

- Accuracy (Predicted vs Actual rating)
- JSON Validity Rate (the model return valid JSON?)
- Reliability / Consistency (the model return same rating across repeated runs?)

Approach:

The task was completed using a structured workflow:

1. **Dataset Sampling**: 200 reviews were randomly selected from the Yelp dataset.
2. **Prompt Iteration**: Three different prompt designs were created.
3. **Model Execution**: Each review was evaluated using each prompt version.
4. **JSON Extraction & Parsing**: Regex-based fallback logic was implemented to handle malformed outputs.
5. **Evaluation Metrics**: Accuracy, JSON validity rate, and response consistency were calculated.
6. **Comparison**: Performance of all 3 prompt strategies was compared side-by-side.

Prompt Design :

Prompt Version 1 — Direct Prompting (Baseline)

```
Rate this Yelp review from 1–5.

Return ONLY valid JSON:

{

  "predicted_stars": <1-5>,

  "explanation": "<brief reason>"

}

Review: "<review>"
```

Why This Design?

- Serves as a baseline for evaluating improvements.

- Direct and simple, minimal model instruction.
- Evaluates basic LLM JSON-following ability.

System Behaviour

- Fastest response time
- Moderate accuracy
- JSON validity is low because the LLM sometimes outputs extra text
- Useful to understand the raw LLM capability without guidance

Prompt Version 2 — Chain-of-Thought Prompting

1. Identify sentiment.

2. Identify positive/negative keywords.

3. Decide a star rating (1-5).

4. Return ONLY valid JSON.

JSON format:

```
{
  "predicted_stars": <1-5>,
  "explanation": "<why>"
}
```

Review: "<review>"

Why This Design?

- Encourages structured analysis → improves accuracy
- Breaks down reasoning into steps
- Useful for sentiment-heavy tasks like Yelp rating prediction

System Behavior

- Highest reasoning quality
- Improved prediction accuracy
- JSON validity decreases because the LLM sometimes prints reasoning outside JSON
- Requires robust JSON extraction logic

Prompt Version 3 — Few-Shot Prompting

You are an expert sentiment classifier.

Example 1:

Review: "Terrible service and cold food."

Output: {"predicted_stars": 1, "explanation": "Very bad experience"}

Example 2:

Review: "Amazing food! Loved the ambience."

Output: {"predicted_stars": 5, "explanation": "Highly positive sentiment"}

Return valid JSON only.

Review: "<review>"

Why this Design?

- Provides demonstrations for the model to replicate
- Strongly improves JSON formatting
- Leads to highest consistency in responses
- Reduces hallucination and unnecessary explanations

System Behaviour

- Best JSON validity rate
- Very high consistency
- Accuracy close to or better than CoT prompting
- Best overall prompt for structured tasks

Evaluation Methodology

Each prompt version was tested on ~200 reviews.
For each review:

- The LLM was called 3 times → consistency measurement
- JSON extraction was performed via:
    - fenced-code block detection
    - substring detection { ... }
    - cleanup for quotes/trailing commas

Comparison Table:

Prompt Version	**Accuracy**	**JSON Validity**	**Consistency**
Direct	0.63	**0.71**	**0.54**
Chain-of-Thought	**0.71**	**0.66**	**0.49**
Few-Shot	**0.69**	**0.93**	**0.81**

Short Discussion on Approaches:

Direct Prompt

- Works as a baseline
- Lacks structure → JSON validity is low
- Accuracy is moderate
- Good for quick prototyping, not production

Chain-of-Thought Prompt

- Improved accuracy due to deeper reasoning
- More verbose → more JSON breaks
- Good for high accuracy needs when formatting is not strict
- Requires JSON parser fallback mechanisms

Few-Shot Prompt

- Best JSON validity & consistency
- High accuracy because examples guide the model
- Most stable across all 200 samples
- Recommended for real-world deployments.