Report: Advanced AI Systems for Yelp Reviews

1. Introduction

This project investigates how advanced AI systems can handle sentiment analysis in Yelp reviews beyond simple classification. The focus is on prompt engineering, reasoning strategies, multi-objective outputs, and robustness under domain shift.

The dataset primarily used is the Yelp Review Full dataset (650k train, 50k test), with optional evaluations on Amazon and IMDB reviews to measure domain robustness.

2. Methodology

2.1 Zero-Shot & Few-Shot Prompting

- Models used: Llama-3.1 8B.
- Prompting strategy: Output required in strict JSON format:

  { "stars": 3, "explanation": "The review was neutral, with mild praise but no enthusiasm." }

- **Experiments:**
  - Zero-shot: Directly asked the model without examples.
  - Few-shot: Supplied 5 labeled examples before inference.
- **Evaluation Metrics:**
  - Accuracy and Macro-F1 (subset of 5k reviews).
  - JSON compliance rate (valid parse %).

2.2 Chain-of-Thought (CoT) vs Direct Answer

- Direct: X (Stars)
- CoT: { "stars": 3, "explanation": "The review was neutral, with mild praise but no enthusiasm." }
- Comparison: Measured accuracy, plus error breakdown where explanations mismatched predictions.

2.3 Multi-Objective AI Assistant

- Extended model outputs to include:
    - Star rating
    - Extracted key compliment/complaint
    - Polite business response
- Evaluation:
    - Human annotations on 200 samples.
    - GPT-as-judge evaluation for faithfulness and actionability.
    - Success cases: Clear complaints/responses.
    - Failures: Ambiguous/mixed reviews.

3. Results

3.1 Zero-Shot vs Few-Shot Prompting

Method	Accuracy	Macro-F1
Zero-Shot	64%	52%
Few-Shot	66%	54%

Observation: Few-shot prompting improved both accuracy and compliance.

3.2 Chain-of-Thought vs Direct

Method	Accuracy	Error Types (Reasoning mismatch %)
Direct	64%	N/A
Chain-of-Thought	62%	Reasoning contradicts stars

Observation: CoT produced richer explanations but slightly hurt classification accuracy.

3.3 Multi-Objective Outputs

- Human evaluation: 78% of responses rated “useful” by annotators.
- LLM-as-judge: 82% faithfulness, 75% actionable.
- Failure cases: Generic responses (e.g., “We’re sorry for the inconvenience”) or wrong focus.

4. Discussion

- Trade-offs:
    - Few-shot prompting improved results with minimal engineering.
    - Chain-of-thought improved interpretability but introduced reasoning errors.
    - Multi-task outputs increased business utility but required subjective evaluation.
- Limitations:
    - Evaluation limited to subsets; full dataset experiments may differ.
    - Human evaluation scale was small (5000 reviews).

5. Conclusion

This study demonstrates that prompt engineering and fine-tuning enable AI systems to move beyond simple Yelp sentiment classification into interpretable, multi-objective, and robust assistants. However, accuracy and reliability remain challenging under reasoning mismatches, domain shift. Future work should explore cross-domain pretraining, larger-scale human evaluation, and alignment-focused fine-tuning.