Assignment: Advanced AI Systems for Yelp Reviews

Overview

This assignment explores how AI systems can go beyond simple text classification to demonstrate reasoning, robustness, and actionable insights in the context of **Yelp reviews**.

You will begin with baseline sentiment classification and progressively tackle harder challenges: structured prompting, reasoning vs direct answers, multi-task outputs, and robustness to domain shift. Some tasks may not have perfect solutions — part of the evaluation is how you **design, experiment, and justify** your approaches.

Dataset

- **Primary**: Yelp Review Full (650k train, 50k test)
- **Optional domain-shift test sets**: Amazon Reviews, IMDB Reviews (available on HuggingFace)
- Labels: {1, 2, 3, 4, 5} stars

Tasks

1. Zero-Shot & Few-Shot Prompting Under Constraints

- Use an instruction-tuned LLM (e.g., GPT-4, Claude, Llama-3).
- Prompt the model to classify reviews into 1–5 stars.
- Output must be in strict JSON format:
  `{ "stars": 3, "explanation": "The review was neutral, with mild praise but no enthusiasm." }`
- Do it for both zero-shot and few-shot prompting and compare.
- Evaluate:
    - Accuracy / Macro-F1 on a sample subset
    - Format compliance rate (valid JSON %)

---

2. Chain-of-Thought vs Direct Answers

- Compare two prompting strategies:
    - **Direct**: “This review is X stars.”
    - **Reasoning**: “Explain reasoning, then output stars.”
- Analyze: Does chain-of-thought improve or hurt classification?
- Report both **accuracy** and **error types** (e.g., reasoning mismatch).

---

3. Multi-Objective AI Assistant

Extend the task to make the system useful for businesses:

- For each review, generate:
    - Star rating (1–5)
    - Extracted **key complaint/compliment**
    - A short, polite **business response**
- Since ground truth doesn’t exist, design an **evaluation protocol**:
    - You may use human annotation, or
    - LLM-as-judge evaluation (e.g., GPT evaluating outputs for faithfulness/actionability).
- Report findings and **examples of success/failure**.

---

4. Domain Shift & Robustness

- Train/fine-tune on Yelp → test on Yelp AND **Amazon/IMDB** reviews.
- Report:
    - Drop in performance (Yelp → other domains)
    - How adversarial inputs affect predictions
- Suggest at least **one mitigation**

Deliverables

1. **Code / Notebook(s)**
    - All experiments, prompts, and evaluation scripts
    - Robustness/adversarial dataset creation
2. **Report (4–5 pages)**
    - Prompt iterations & insights
    - Results tables (accuracy, JSON compliance, robustness, etc.)

    - Comparative analysis: direct vs CoT, classification vs assistant outputs
    - Evaluation of domain shift
    - Discussion of trade-offs and limitations

Notes

- Some tasks are intentionally **open-ended** and may not have a perfect solution.
- The goal is not just accuracy, but **designing, analyzing, and justifying approaches**.
- Use **smaller subsets** for fast iteration if needed.