

Research Task 08 — Mid-Project Progress Report

Date: November 1, 2025

Prepared by: Kruti Kotadia

Objective

To detect potential biases in large-language-model (LLM) narratives by comparing how models describe the same dataset under different prompt framings (positive, negative, demographic, confirmation, and neutral).

Progress Summary

Over the past two weeks, I finalized my experimental setup and began data collection.

- Created five prompt templates representing distinct framing conditions.
 - Tested prompts on GPT-4o and Claude 3.5 Sonnet, with plans to include Gemini 1.5 Pro next week.
 - Each prompt was run three times per model to capture variability.
 - Responses were logged with metadata (model, version, condition, timestamp) and stored locally.
 - A `.gitignore` file ensures that raw data and JSON logs remain untracked.
-

Preliminary Observations

- Early runs show tone shifts: positive frames emphasize growth, while negative frames highlight weaknesses even with identical data.
 - Some demographic prompts show subtle preference toward “senior” players.
 - No fabricated statistics observed so far, but sentiment intensity differs across models.
-

Next Steps (Before Nov 15)

- Complete all runs across three models (15–20 responses total).
 - Quantify sentiment and player-mention frequency with Python (Pandas + Polars).
 - Create summary tables and simple bar-chart visuals.
 - Draft the final bias-detection report with mitigation recommendations.
-

Tools & Skills Used

Python, Pandas, OpenAI API, Claude API, Jupyter Notebook, GitHub for version control, and Excel for early data tabulation.