Task Completion Summary

I successfully completed the entire workflow for Task 08, including experiment design, LLM response collection, bias analysis, statistical testing, and final reporting. All required deliverables—analysis summaries, chi-square tables, fabrication checks, and ground-truth validations—were finalized and reviewed.

Methodology Overview

The project followed a four-phase methodology: (1) experimental design with predefined hypotheses, (2) structured multi-model LLM data collection, (3) quantitative and qualitative analysis of outputs, and (4) claim validation and bias reporting. Responses were logged with model versions, timestamps, and prompt variations to ensure reproducibility.

Dataset and Experiment Setup

The experiments used anonymized player performance data. Multiple hypotheses tested framing bias, demographic bias, confirmation bias, and selection bias. Each hypothesis included minimally altered prompts, and each prompt was run multiple times. Outputs were analyzed for sentiment, player mentions, recommendation styles, and factual grounding.

Key Findings

Analysis revealed sentiment shifts based on prompt framing (e.g., positive framing produced sentiment scores of 4.0 vs 1.0 under negative framing). Player mentions and recommendation categories varied across prompt types. No hallucinations or contradictions were detected in any hypothesis.

Statistical Analysis Summary

Chi-square tests were performed for key hypothesis pairs. Results showed structural differences in recommendation distributions but with limited statistical power due to small sample sizes. All fabrication rates remained at 0%, reinforcing the consistency of the LLM outputs.

Bias Catalogue

Detected biases included:

- Framing Bias: Sentiment and tone changed depending on positive vs negative framing.

- Demographic Bias: Player references shifted when demographic details were introduced.

- Confirmation Bias: Primed prompts led to responses aligning with implied assumptions.

- Selection Bias: Different statistics were emphasized depending on question framing.

Mitigation Strategies

- Use neutral phrasing and avoid emotionally loaded language in prompts.

- Remove unnecessary demographic cues unless explicitly relevant.

- Provide structured data tables to reduce selective emphasis.

- Use multi-model validation and cross-check claims with ground truth.

Limitations

- Small sample size limits statistical reliability.

- Only a subset of LLMs was tested.

- Real-world demographic data was anonymized, limiting external generalizability.

- Responses may vary with different model versions and temperatures.

Skills Developed

This project strengthened skills in experimental design, Python-based data analysis, regex-based claim validation, LLM behavior analysis, research documentation, statistical reasoning, and ethical AI evaluation. It improved my ability to evaluate model outputs critically and design reproducible studies.

Conclusion

The experiment successfully demonstrated how linguistic framing and contextual cues influence LLM-generated narratives. Although no fabrications were detected, measurable shifts in sentiment, emphasis, and recommendations confirm that LLM outputs are sensitive to prompt variations. The study emphasizes the need for careful prompt engineering, cross-model validation, and awareness of embedded biases in AI systems.