Task 08 – Initial Planning Report

Bias Detection in LLM Data Narratives

Name: Neha Chandrakant Sharma
Date: October 15, 2025
Dataset: SU Women’s Lacrosse 2024 Season Statistics

Objective

The purpose of this task is to design and plan an experiment that detects potential **biases in large language models (LLMs)** when analyzing identical datasets under different prompt framings. This includes identifying how **framing, demographic, and confirmation biases** might influence the narratives generated by models like ChatGPT (GPT-4), Claude 3.5, and Gemini 1.5.

Planned Dataset and Ground Truth

I will reuse the **SU Women's Lacrosse 2024 season dataset** from Task 05, which contains quantitative player performance metrics such as goals, assists, turnovers, minutes played, and class (year). Ground-truth statistics—means, standard deviations, and player aggregates—were previously computed in Task 04 using Python, Pandas, and Polars. These will serve as the factual baseline for validating LLM statements.

Research Hypotheses

I plan to test three primary hypotheses:

ID	Hypothesis	Bias Type
H1	Positive vs. negative framing will alter the tone and recommendations in LLM outputs.	Framing Bias
H2	Mentioning player demographics (e.g., class year) will change which players are emphasized.	Demographic Bias
H3	Priming the model with a conclusion (e.g., "prove defense was weak") will lead to selective evidence use.	Confirmation Bias

Experimental Design Outline

- **Models:** GPT-4 (ChatGPT), Claude 3.5, Gemini 1.5
- **Prompt Structure:** Pairs of minimally different prompts (e.g., "Who underperformed?" vs "Who has potential to improve?")
- **Runs per Model:** 3 – 5 replicates to control for randomness
- **Metrics:**
  - Sentiment analysis (TextBlob/VADER)
  - Frequency of player mentions
  - Fact validation against descriptive statistics

All prompts and results will be logged in JSON format with timestamps and model metadata for reproducibility.

Expected Deliverables

- experiment_design.py – to auto-generate all prompt variations
- run_experiment.py – to query LLMs and log results
- analyze_bias.py – to calculate sentiment and entity frequency differences
- validate_claims.py – to cross-check model statements with statistical truth
- REPORT.md – final bias-detection report summarizing findings

Next Steps

1. Finalize prompt templates for each hypothesis by October 18.
2. Begin collecting model responses (Week 2).
3. Run quantitative analysis and sentiment scoring (Week 3).
4. Compile findings and visualizations for final report (Week 4).

Reflection

At this stage, my focus is on ensuring **experimental control and reproducibility**. I anticipate challenges in quantifying subtle language shifts and plan to address them by combining automated sentiment scoring with manual thematic review.
This initial planning establishes the foundation for a systematic and ethical study of LLM bias using my validated sports dataset.