# Initial Planning Report — Task 08: Bias Detection in LLM Data Narratives

Date: October 15, 2025

## 1. Project Overview

This study will investigate whether large language models (LLMs) display systematic bias when generating analytical narratives from identical sports datasets under different prompt framings. Building upon prior research tasks (05–07) that analyzed the 2025 Syracuse Women’s Lacrosse season using Python and LLM-generated narratives, this project transitions to a controlled experiment phase. The focus is to test how question phrasing and contextual cues (e.g., demographic or framing) affect model interpretations of the same statistical facts.

## 2. Dataset Foundation

Data are drawn from prior descriptive and ethical analysis tasks:  
- 2025\_SU\_Lacrosse\_Player\_Stats.csv – player-level performance (goals, assists, turnovers, draws, ground balls)  
- 2025\_SU\_Lacrosse\_Team\_Stats.csv – team-aggregate metrics (total goals, shot %, assists, saves)  
- 2025\_SU\_Lacrosse\_Period\_Stats.csv – per-period scoring trends (used to detect late-game decline)  
- 2025\_SU\_Lacrosse\_Matches\_Stats.csv – match-level outcomes (opponent, home/away, win/loss)  
  
All names are replaced with anonymized identifiers (Player A, Player B, Player C) to ensure privacy and compliance with ethical guidelines.

## 3. Prior Findings Informing Bias Testing

From Tasks 05–07:  
- Team played 19 games.  
- Offense was identified as the primary improvement lever.  
- Player A was cited as top offensive performer.  
- Player B showed highest improvement.  
- Late-game scoring dipped from 74 (1st period) to 47 (4th period).  
These established facts now serve as ground-truth references for detecting potential framing bias in narrative responses.

## 4. Research Hypotheses

H₁ – Framing Bias: LLM outputs differ when describing the same player as “underperforming” vs “developing.”  
H₂ – Demographic Bias: Adding player “experience level” (senior/sophomore) changes recommendations for coaching focus.  
H₃ – Confirmation Bias: Prompts implying a hypothesis (“offense is the issue”) lead LLMs to reinforce that view even when data are neutral.  
H₄ – Selection Bias: LLMs highlight different statistics (goals vs turnovers vs assists) depending on framing.  
H₅ – Cross-Model Variation: GPT-4, Claude, and Gemini produce systematically distinct narrative tones for identical prompts.

## 5. Planned Experimental Design

|  |  |
| --- | --- |
| Phase | Description |
| Prompt Design | Create pairs/sets of prompts varying only in framing (positive vs negative, neutral vs demographic). Example: “Which player needs improvement?” vs “Which player has growth potential?” |
| Model Selection | Use GPT-4 (ChatGPT Plus), Claude 3 (Opus via SU Enterprise), and Gemini Pro. Standardize temperature and system messages for fair comparison. |
| Response Collection | Generate 3–5 outputs per prompt per model. Log to JSON/CSV with model name, temperature, timestamp, and prompt. |
| Quantitative Analysis | Compute sentiment (VADER/TextBlob) and topic frequencies. Run chi-square tests for word distribution differences. |
| Qualitative Analysis | Assess tone, focus areas, and fabrication rate. Flag claims that contradict ground-truth stats. |

## 6. Tools & Environment

Python: pandas, NumPy, scipy.stats, statsmodels, matplotlib  
NLP Libraries: VADER, TextBlob  
Version Control: GitHub repo 'Task\_08\_Bias\_Detection'  
LLM APIs: ChatGPT (GPT-4), Claude.ai (SU license), Gemini Pro

## 7. Ethical Plan

• Sanitize PII from all datasets and outputs.  
• Ensure transparency by labeling all AI-generated text.  
• Avoid reinforcing harmful demographic assumptions.  
• Maintain reproducibility through seeded API calls and documented model versions.

## 8. Summary Statement

This planning phase establishes a reproducible, ethical framework to examine how LLM outputs vary with framing and context on a controlled sports dataset.