Bias Detection in LLM Data Narratives – Initial Planning Report

1. Introduction

This report outlines the initial plan for Research Task 8, which investigates potential biases in data narratives produced by large language models (LLMs). The goal is to determine whether models like GPT-4 and Claude generate different interpretations of the same dataset depending on how prompts are framed or whether demographic cues are included.

2. Initial Planning and Dataset

The dataset selected for this experiment is the official 2024 Syracuse Women's Lacrosse team statistics, including goals, assists, shot accuracy, and defensive metrics. Earlier tasks provided validated descriptive analyses and LLM-generated narratives that form the baseline for this study. Initial planning involved reviewing prior work, identifying key variables (player performance, position, team outcomes), and testing small prompt variations ('struggling' vs 'developing') to confirm that framing affects tone and recommendations.

3. Objectives and Hypotheses

The project aims to measure how framing, demographic emphasis, and hypothesis priming affect LLM-generated narratives.

Key hypotheses:

- H1: Positive vs. negative framing changes recommendations.

- H2: Mentioning demographics shifts which players are emphasized.

- H3: Hypothesis-primed prompts lead to higher agreement (confirmation bias).

- H4: Selection bias occurs in which data points are highlighted.

4. Experimental Approach

Paired prompts differing only by framing will be created for each hypothesis. GPT-4 and Claude will each be queried 3–5 times per prompt. All outputs, prompts, and metadata (model version, temperature, timestamp) will be logged in structured CSV/JSON files. Responses will be analyzed using sentiment scoring, word-frequency analysis, and statistical testing to identify significant narrative shifts.

5. Tools and Analysis Methods

- LLMs: GPT-4, Claude

- Python: Pandas, NumPy, Matplotlib for data and visual analysis

- Validation: Compare outputs to factual dataset values to detect inaccuracies

6. Ethics and Deliverables

Data are public and anonymized, eliminating privacy concerns. All LLM-generated text will be labeled, and randomness controlled. Deliverables include a bias-mapping report, visual comparisons, and prompt-engineering recommendations for reducing bias.

7. Next Steps and Starting Plan

The first stage will focus on re-structuring the Syracuse Women's Lacrosse dataset into clear tables for prompt input and validation. This will involve verifying numerical accuracy, ensuring all key player metrics are consistent, and preparing data snippets for prompt embedding. Next, prompt templates will be drafted for each hypothesis, emphasizing minimal language variation between test conditions. Initial test runs will begin with GPT-4 using neutral and positively framed prompts to establish baseline sentiment scores. After successful validation, the same prompts will be applied to Claude for cross-model comparison. By the end of Week 1, all prompt pairs, logging structures, and evaluation scripts will be finalized for systematic data collection.