Research Task 08 - Initial Planning Report

1. Objective

To design and run a controlled experiment testing whether large language models (LLMs) such as GPT-4o, Claude 3.5, and Gemini 1.5 show bias when interpreting the same sports performance data under different prompt framings. The goal is to measure how wording or context affects which players are emphasized and how performance is described.

---

2. Dataset

The dataset will use anonymized statistics from the Seattle Sounders' recent season, focusing on three players labeled **Player A**, **Player B**, and **Player C**. Variables include goals, assists, shots, and defensive recoveries. No real player names or personal identifiers will be used, and all data will be stored locally (not committed to GitHub).

---

3. Hypotheses

Framing Bias: Asking "Which player underperformed this season?" vs. "Which player shows the most improvement potential?" will lead to different recommendations.
Confirmation Bias: If a prompt begins "Defense was the main problem," the model will favor defensive issues even if statistics show otherwise.
Demographic Bias: Adding "senior" or "rookie" labels may change how LLMs frame player performance.
Selection Bias: Models may focus more on offensive metrics (goals, assists) than defensive ones when identifying "top performers."

---

4. Planned Method

- Develop five paired prompt versions (neutral, positive, negative, demographic, confirmation).
- Query **GPT-4o**, **Claude 3.5 Sonnet**, and **Gemini 1.5 Pro** with identical data under each condition.
- Collect 3 – 5 responses per prompt to control randomness.
- Log each run with model name, timestamp, and prompt version.
- Use Python (Pandas/Polars) to analyze:
    - Which players are mentioned most.
    - Sentiment/tone differences.
    - Changes in recommendations under different framings.