1. Objective

State that this week's goal is to design and pre-register hypotheses for detecting bias in LLM-generated narratives using the verified Syracuse Women's Lacrosse 2025 dataset.
Mention that this expands on prior work from Tasks 05–07 where you validated descriptive statistics via Python and compared LLM responses.

2. Dataset and Ground Truth

- Dataset: Syracuse Women's Lacrosse 2025 (season summary + player stats).

- Ground Truth verified through:

    - `pure_python_stats.py`, `pandas_stats.py`, `polars_stats.py`

    - Descriptive Statistics and LLM Validation Report (Task 05 & 06)
      README
      Research Task 06

- Key validated numbers (e.g., 19 games, 234 goals, 112 assists, 346 points).

3. Hypotheses

ID	Hypothesis	Bias Type	Expected Observation
H1	"Struggling player" vs "Developing player" framing changes recommendations.	Framing Bias	Tone and target player shift.
H2	Adding demographics (senior/junior) alters who is labeled as "high potential."	Demographic Bias	References to seniority increase.
H3	Priming LLM with "I think Player B is MVP" induces agreement.	Confirmation Bias	More frequent affirmations.
H4	"Who performed worst?" vs "Who needs development?" affects sentiment.	Framing/Sentiment Bias	Shift in polarity scores.

4. Prompt Design

Show four paired conditions with anonymized data:

```
Base Data:
Player A — 34 Goals
Player B — 46 Assists, 76 Points
Player C — 4 GWG
```

Then list Prompt 1–4 pairs exactly as in the Task 08 example

Research_Task_08

.
 Use neutral, positive, and negative framings plus a demographic variant.
 Ensure all player names are "Player A/B/C."

5. Data Collection Plan

- Models: GPT-4, Claude 3, Gemini 1.5.

- Samples: 5 runs per prompt (temperature 0.7).

- Log fields: `model`, `prompt_id`, `response_text`, `timestamp`, `sentiment_score`.

- Save in `results/Week1_logs/`.

6. Planned Analysis Methods

- Quantitative → Sentiment (VADER/TextBlob), entity frequency, chi-square tests.

- Qualitative → Tone comparison, bias phrases, hallucinations.

- Validation → Cross-check against Python outputs for factual accuracy.