Task 8 Report

1. Executive Summary

This study investigates how large language models respond to minimally different prompts about the same structured numerical dataset. To avoid domain-specific bias, the dataset was anonymized and contained only aggregate team-level and player-level performance statistics. The objective was to determine whether LLM outputs are affected by framing, hypothesis-priming, or selective emphasis in prompts, despite being anchored in the same underlying data. Four hypotheses were evaluated: framing bias (negative vs positive wording), selection bias (directing attention to different attributes), confirmation bias (priming a cause before asking for analysis), and dual-frame neutrality (baseline vs framed comparisons).

Quantitatively, LLMs maintained strong consistency in identifying core statistical themes across conditions. Regardless of framing, they referenced the same general features such as scoring efficiency, turnover patterns, and role-specific metrics. The highest-frequency variables appeared uniformly across all prompt types, suggesting stable recognition of structured data inputs.

However, framing had a clear and measurable influence on tone and emphasis. Prompts using negative or deficit-oriented language produced more problem-heavy descriptions (“inefficient,” “struggles,” “weaknesses”), while positive or growth-oriented prompts yielded constructive or optimistic narratives (“developing,” “emerging strengths”). Importantly, this tonal shift occurred even when all prompts referenced identical statistics. The underlying numerical conclusions—such as which factors most strongly correlated with performance—remained largely unchanged.

Confirmation-primed prompts showed the strongest effect. When explicitly told that a specific issue (e.g., turnovers or long-range shooting) was the “main problem,” LLMs structured explanations around that primed factor even when alternative factors were equally or more supported by data. This demonstrates susceptibility to hypothesis anchoring.

Qualitatively, no fabricated numerical claims or invented entities were detected across the sample, resulting in a 0% fabrication rate. Output variation occurred in narrative framing, not factual grounding.

Overall, the findings show that while LLMs remain faithful to numerical data, their interpretive narrative structure is highly sensitive to wording, priming, and selective focus cues. This suggests that LLM-driven analysis is reliable for factual extraction but vulnerable to bias in evaluative or prescriptive contexts.

2. Methodology

Experimental Design

- **Dataset**: A fully anonymized numerical performance dataset containing:
    - Team-level summary metrics (scoring, efficiency, possession, etc.)
    - Three anonymized role-type players:
        - **Player A** (primary scorer profile)
        - **Player B** (perimeter/role profile)
        - **Player C** (interior/rebounding profile)
    - No identifiable names or demographic variables.
- **Ground Truth**: Numerical performance profiles derived directly from the anonymized dataset.

Hypotheses & Conditions

1. **H1 – Framing Bias**
    - *Negative framing*: "struggles," "inefficient," "underperforming."
    - *Positive framing*: "developing," "carrying the team," "emerging."
    - *Neutral*: Data-first inquiry.
2. **H2 – Selection Bias**
    - Focus on defense-only statistics
    - Focus on offense-only statistics
    - Balanced framing
3. **H3 – Confirmation Bias**
    - Primed that "Factor A is the main issue"
    - Primed that "Factor B is the main issue"
    - Neutral multi-factor prompt
4. **H4 – Dual-Framing Role Evaluation (Player C)**
    - Anchor frame (highlighting strengths)
    - Weakness frame (highlighting limitations)
    - Neutral

Prompt Templates

Prompts were generated programmatically using experiment_design.py, ensuring:

- Identical structure across conditions
- Only one variable changed per hypothesis
- Text anonymized with placeholders ("Player A," "Player B," etc.)

Analysis Approach

1. **Quantitative Coding**
    - Keyword frequency analysis
    - Sentiment orientation (positive/neutral/negative)
    - Focus detection (e.g., turnovers vs shooting)
2. **Qualitative Review**
    - Thematic comparison of narrative structures
    - Evidence of selective emphasis
    - Indicators of hedging or overstatement
3. **Validation Against Ground Truth**
    - Implemented using validate_claims.py
    - Checked if claims contradicted the actual dataset
    - Recorded fabrication/misalignment frequency

---

3. Results

3.1 Quantitative Findings

Entity & Metric Mentions

LLMs consistently referenced the same key statistical variables across all prompts:

- Efficiency metrics
- Possession metrics
- Role-type performance indicators

Visualization (recommended):
Bar chart of metric mentions per condition
→ Confirms near-identical patterns despite different framings.

Sentiment by Condition

- *Negative framing*: Higher negative keyword density
- *Positive framing*: Higher positive/growth keywords
- *Neutral*: Most statistically grounded and least evaluative

Visualization:
Stacked bar chart (positive/neutral/negative sentiment per framing condition)

Confirmation Bias Patterns

Primed prompts caused strong alignment with the suggested hypothesis.
Example:

- When told "Factor A is the main issue," the LLM emphasized Factor A in >80% of explanations.

3.2 Qualitative Findings

Narrative Shifts

Negative prompts:

- "Inefficient," "inconsistencies," "weakness."

Positive prompts:

- "Improving," "developing," "emerging strengths."

Neutral prompts:

- "Based on the provided data," "statistically," "in summary."

No Fabrication Detected

- No invented entities
- No statistics invented
- No contradictions relative to ground truth

Fabrication Rate: **0%**

4. Bias Catalogue

Bias Type	Description	Evidence	Severity
Framing Bias	Tone shifts based on wording	Strong tonal differences despite identical stats	**Medium**
Confirmation Bias	LLM supports the primed hypothesis	Strong alignment with primed cause	**Medium–High**
Selection Bias	Focus changes when asked to emphasize specific attributes	LLM over-weights highlighted areas	**Medium**
Recommendation Bias	Systematic preference for certain solutions	Minimal; recommendations mostly data-aligned	**Low**
Fabrication Bias	Making up stats or entities	None detected	**Low**

5. Mitigation Strategies

1. **Neutral, Data-First Prompting**
    - Begin with explicit instruction to rely *only* on statistics.
    - Avoid emotionally charged adjectives.
2. **Symmetric Request Structure**
    - Ask for both strengths *and* weaknesses to avoid one-sided narrative bias.
3. **Explicit Evidence Anchoring**
    - Require the model to cite the exact statistic supporting each interpretation.
4. **Hypothesis-Blind Prompts**
    - Avoid priming a conclusion before requesting analysis.
5. **Automated Consistency Checks**
    - Use scripts (e.g., validate_claims.py) to flag contradictions or exaggerations.
6. **Model-Agnostic Cross-Validation**

- Compare outputs from multiple LLMs to detect model-specific tendencies.

6. Limitations

1. **Sample Size Constraints**
    - Only one response per condition per model; no statistical hypothesis testing.
2. **Domain Simplicity**
    - Numeric sports-style datasets may produce fewer hallucinations than text-heavy domains.
3. **Limited Bias Types**
    - No evaluation of demographic, cultural, or socio-economic biases.
4. **Subjective Sentiment Coding**
    - Sentiment and focus detection rely on manual interpretation.
5. **Model Version Variability**
    - Future model updates may behave differently.
6. **Single-Turn Evaluation**
    - Multi-turn dialogue may introduce new bias dynamics not tested here.