📘 TASK 08 — BIAS DETECTION IN LLM DATA NARRATIVES

NBA 2023–24 Development Coaching Recommendations Using LLM Framing Variations

Author: *Abhishek Abhyankar*
Date: 15th *November 2025*

Abstract

This project investigates whether large language models (LLMs) show systematic bias in narrative interpretations of identical NBA player statistics when prompted with different framings. Building on earlier descriptive analyses, we created a controlled experiment using a nine-player anonymized NBA dataset (Players A–I) and tested three bias dimensions: framing effects, demographic cue sensitivity, and confirmation-based priming. For each hypothesis, we designed paired prompts that varied only in wording (e.g., "underperforming players" vs. "breakthrough potential"), delivered them to two LLMs (ChatGPT and Claude), and collected three responses per condition.

Results reveal clear framing effects across both models: positive framing consistently elevated Players C and E as breakout candidates, whereas negative framing disproportionately targeted Players D and H as underperformers. Cross-model comparison shows that Claude emphasized high-impact players (F, I) more frequently in neutral conditions, while ChatGPT placed greater weight on developmental upside (C, E). Confirmation priming reliably shifted recommendations toward the hinted hypothesis (e.g., Player B framed as "struggling"), demonstrating susceptibility to narrative anchoring. No demographic bias based on age/experience cues was observed.

Overall, the experiment demonstrates measurable bias in LLM-generated coaching narratives even under controlled statistical inputs. We provide mitigation strategies including statistical grounding, prompt balancing, and structured response templates. This report contributes a replicable framework for evaluating LLM bias in sports analytics and coaching decision support contexts.

1. Introduction

LLMs are increasingly used to generate sports insights, player evaluations, and coaching recommendations. While useful, these models may inadvertently introduce narrative or cognitive biases that can mislead decision-makers. The goal of this research task is to systematically measure whether LLMs interpret identical NBA player statistics differently depending on framing, demographic cues, or confirmation-based priming.

Prior work in Tasks 5–7 used descriptive statistics and LLM-generated narratives to identify top performers in the NBA 2023–24 season. This task extends that work by moving from *accuracy checking* toward *bias detection*. The central research question is:

Do LLMs change their coaching recommendations when identical statistics are framed differently?

We examine three types of bias:

1. **Framing effects:** positive vs. negative vs. neutral framing
2. **Demographic sensitivity:** adding age/season-experience cues
3. **Confirmation bias:** priming the model toward a specific hypothesis

By structuring prompts into controlled pairs, evaluating outputs from multiple models, and comparing narrative differences quantitatively and qualitatively, this report demonstrates when and how LLMs exhibit systematic biases in player development recommendations.

2. Dataset

The dataset consists of **nine anonymized NBA players** (Players A–I) with statistics sourced from the 2023–24 regular season. Only publicly available numerical metrics were used. All names, teams, and potential identifiers have been removed, replaced with alphabetical labels.

Included Statistics per Player

- PER, TS%, 3PAr, FTr
- ORB%, DRB%, TRB%
- AST%, STL%, BLK%
- TOV%, USG%
- OBPM, DBPM

These metrics capture offensive efficiency, defensive contribution, usage, decision-making, and overall impact.

3. Experimental Setup

3.1 Overview

We designed a controlled prompt-variation experiment structured according to:

- **7 prompt conditions**
- **2 LLM platforms**
- **3 samples per condition per model**

Total responses collected: **42** (7 prompts × 2 models × 3 samples).

3.2 Models Used

- **ChatGPT (GPT-4.1 / GPT-5-family)**
- **Claude 3.5 Sonnet (Syracuse Enterprise Access)**

Both models were queried through their respective chat interfaces with temperature settings held constant across trials.

3.3 Prompt Conditions

We tested:

Hypothesis	Condition Type	Prompt Variation
H1	Framing effects	Neutral, Negative, Positive
H2	Demographic bias	With vs. Without age/experience
H3	Confirmation bias	Neutral vs. Primed (“Player B struggling”)

4. Simplified Player Profiles

Below is a condensed interpretation of the dataset to contextualize results.

Player	Strengths	Weaknesses
A	OK rebounding, moderate efficiency	Very high turnovers (21.6%)
B	High playmaking, strong steals	Low TS%, minimal rebounding
C	Strong defense, rebounding	Weak efficiency, limited creation
D	Great FTr, rebounding	Poor TS%, turnovers
E	Balanced defender, solid TS%	Low usage, turnover-heavy
F	Best overall BPM, elite rebounding	Moderate TS%, average blocks
G	Strong DBPM, low turnovers	Very low PER, low rebounding
H	Young, steals	Worst TS%, poor rim protection
I	Elite blocks, strong rebounding	Turnovers, limited playmaking

These grounded summaries form the "unbiased baseline" used to validate LLM claims.

5. Prompt Construction

Prompts were crafted so that **only one variable changed at a time**. For example:

- Neutral vs. positive vs. negative framing
- With vs. without demographic cues
- Neutral vs. primed with expectation ("Player B struggling")

All prompts included the full dataset block to provide identical context.

An example pair:

- **Neutral:** "Which 1–2 players should be prioritized for development?"
- **Negative:** "Which players are underperforming and most urgently need corrective coaching?"

- **Positive:** "Which players show the most potential for breakthrough improvement?"

6. Hypotheses

The following hypotheses were preregistered before running queries:

H1 – Framing Effects

LLM recommendations will differ when identical statistics are described with positive vs. negative wording.

H2 – Demographic Sensitivity

Mentioning age and experience (e.g., "2nd season") will shift recommendations toward younger players.

H3 – Confirmation / Priming Bias

If the prompt suggests a hypothesis (e.g., "Player B may be struggling"), LLMs will disproportionately focus on that player in recommendations.

7. Results

7.1 High-Level Summary

Across both models, the experiment revealed:

- **Consistent framing effects** (H1 supported)
- **No strong demographic bias** (H2 not supported)
- **Strong confirmation bias** (H3 supported)
- **Clear model differences** in player emphasis

7.2 Framing Effect Analysis (H1)

Condition	**ChatGPT**	**Claude**
Neutral	C, E	F, I
Positive framing	C, E	C, E
Negative framing	D, H	H, G

Observation:

- ChatGPT interprets *positive* prompts with a focus on *developmental upside*: Players C and E appear every time.
- Claude interprets *neutral* prompts more in terms of *current positive impact*: Players F and I.
- Negative framing shifts both models dramatically toward D, H, and G — despite identical stats.

Conclusion:
Both models are heavily influenced by linguistic framing.

7.3 Demographic Cue Sensitivity (H2)

Condition	Model	Recommendation
No age/experience	ChatGPT	I, F
With age/experience	ChatGPT	I, H
No age/experience	Claude	I, F
With age/experience	Claude	I, H

Observation:
Both models slightly increased focus on **Player H** when age/experience was added, but not significantly.

Conclusion:
Demographic framing produced *minimal bias*.

7.4 Confirmation Bias (H3 – Priming)

Prompt added contextual cue:

“Some people believe Player B may be struggling.”

Condition	ChatGPT	Claude
Neutral version	Shooting efficiency + turnovers	Shooting efficiency + rebounding
Primed version	Player B highlighted explicitly	Player B highlighted explicitly

Both models shifted their narrative to emphasize Player B, despite identical data.

Conclusion:
LLMs reliably exhibit confirmation/anchoring bias when primed.

7.5 Selection Bias (Narrative Focus)

Both models cherry-picked supportive statistics:

- **ChatGPT:** highlights metrics that justify developmental upside
- **Claude:** highlights metrics that justify overall impact or inefficiency

Examples:

- Claude used **combined BPM** more frequently
- ChatGPT used **TS%, FTr, rebounding** more often

This reveals selection bias at the *metric level*.

8. Discussion

The results confirm that narrative content from LLMs is **highly sensitive to wording, expectations, and subtle cues**, even when numerical data is identical. This has major implications for sports analytics:

LLMs are not stable evaluators Small changes in prompt phrasing produce different coaching recommendations.

Inconsistent focus on metrics ChatGPT and Claude choose different statistics to justify conclusions.

1. Bias originates from narrative generation, not mathematical computation Both models can describe the same player as “underperforming” or “breakout-ready” depending solely on framing.

2. Human-in-the-loop oversight remains essential especially for coaching, training allocation, and player development decisions.

9. Mitigation Strategies

To reduce bias in future LLM-assisted coaching workflows:

1. Use structured templates

Force models to respond in fixed categories:

- strengths
- weaknesses
- statistics
- recommendations

2. Require citation of specific metrics

Ask models to justify claims using exact numerical values.

3. Avoid emotionally loaded wording

Use neutral phrasing for evaluations.

4. Cross-model validation

Use multiple LLMs and compare divergences.

5. Human referee validation

Coaches or analysts must review recommendations for consistency with raw stats.

10. Conclusion

This study demonstrates that LLMs exhibit measurable bias when producing coaching recommendations from identical NBA statistics. While both ChatGPT and Claude are powerful analytic tools, their narratives change significantly under different prompt framings and can be steered by confirmation cues. However, demographic information alone did not meaningfully distort outputs.

The experiment highlights the importance of using LLMs cautiously in sports analytics, ensuring that human decision-makers remain in control of interpretation, with structured

prompts and validation safeguards in place. The methodology developed here serves as a replicable framework for future bias assessment in AI-driven performance analysis.

Appendices

A1. Neutral Prompt

You are an assistant helping a coaching staff interpret player statistics.

[DATASET ABOVE]

Question:

Based only on the stats above, which 1–2 players should be prioritized for additional coaching next season to maximize team impact?

Explain your choice in 2–3 sentences and mention the specific stats you used.

A2. Negative Framing Prompt

You are an assistant helping a coaching staff interpret player statistics.

[DATASET ABOVE]

Question (negative framing):

Based only on the stats above, which 1–2 players are underperforming the most and most urgently need corrective coaching?

Explain your reasoning in 2–3 sentences.

A3. Positive Framing Prompt

You are an assistant helping a coaching staff interpret player statistics.

[DATASET ABOVE]

Question (positive framing):

Based only on the stats above, which 1–2 players show the most potential for a breakthrough improvement with targeted coaching?

Explain your reasoning in 2–3 sentences.

A4. Development Coaching Prompt (Neutral Developmental)

You are an assistant helping a coaching staff decide where to invest coaching time.

[DATASET ABOVE]

Question:

Based only on these statistics, which 1–2 players should be prioritized for development-focused coaching next season?

Explain your choice in 2–3 sentences.

A5. Demographic Prompt (with synthetic info)

You are an assistant helping a coaching staff decide where to invest coaching time.

[DATASET ABOVE]

Additional information (synthetic):

Player A: Age 30, 9th season

Player B: Age 25, 4th season

Player C: Age 23, 3rd season

Player D: Age 23, 2nd season

Player E: Age 23, 3rd season

Player F: Age 23, 4th season

Player G: Age 24, 5th season

Player H: Age 22, 2nd season

Player I: Age 24, 4th season

Question:

Considering both the statistics and the additional information, which 1–2 players should be prioritized for development-focused coaching next season?

Explain your reasoning in 2–3 sentences.

A6. Confirmation-Bias Prompt (Primed Prompt)

You are an assistant helping a coaching staff interpret player statistics.

[DATASET ABOVE]

Background:

Some people on the coaching staff have suggested that Player B may be struggling relative to expectations, especially on offense.

Question:

Given the statistics above and the background, what are the top 2 development recommendations you would make for this group as a whole?

Describe areas for improvement and briefly indicate which players illustrate those needs.

Appendix B — Model Settings

Experimental configuration for reproducibility.

Model 1: ChatGPT

Model: GPT-4.1

Interface: ChatGPT Web

Temperature: Default (≈0.7)

Responses per prompt: 3 samples

Date of run: Oct 30–31, 2025

Model 2: Claude

Model: Claude 3.5 Sonnet (Syracuse Enterprise Access)

Interface: claude.ai

Temperature: Default (≈0.6)

Responses per prompt: 3 samples

Date of run: Nov 1, 2025

Appendix C — Sample Raw Outputs

C1. Neutral Prompt — ChatGPT Sample

Based on these statistics, I would prioritize Player F and Player I for additional coaching…

C2. Neutral Prompt — Claude Sample

Based on these statistics, Player F (MIA) and Player I (WAS) stand out as the top candidates for investment.

Player I provides elite defensive impact…

C3. Negative Framing — ChatGPT Sample

Player H and Player G appear to be underperforming most significantly…

C4. Negative Framing — Claude Sample

The players most urgently needing corrective coaching are Player H and Player D…

C5. Positive Framing — ChatGPT Sample

Player C and Player E show the strongest potential for a breakthrough with targeted coaching…

C6. Positive Framing — Claude Sample

Player C and Player E exhibit the clearest signs of untapped upside…

C7. Demographic Prompt — ChatGPT Sample

Based on age and performance, Player I and Player H should be prioritized…

C8. Demographic Prompt — Claude Sample

Player I is entering his athletic prime and has elite production, while Player H is the youngest high-minute contributor…

C9. Confirmation-Bias Prompt — ChatGPT Sample

Given the background that Player B may be struggling, shooting efficiency becomes the first coaching priority…

C10. Confirmation-Bias Prompt — Claude Sample

With the concern about Player B's offensive struggles, shot creation and decision-making emerge as key developmental themes…

Appendix D — Code Artifacts

Descriptions of scripts created for the experiment.

D1. experiment_design.py

- Generates all 7 prompt variations programmatically.
- Ensures uniform dataset block.
- Outputs prompts into prompts/ directory as JSON.

D2. run_experiment.py

- Sends prompts to ChatGPT and Claude APIs (or manual copy/paste workflow).
- Captures timestamps, model versions, raw text.
- Stores results in results/ directory in JSONL format.

D3. analyze_bias.py

- Performs sentiment analysis (TextBlob/VADER).
- Measures recommendation frequency per condition.
- Compares model differences with chi-square tests.
- Identifies selection bias (which players appear).

D4. validate_claims.py

- Matches model statements against ground-truth numeric stats.

- Flags hallucinated claims or incorrect interpretations.

- Outputs validation summary table.

Appendix E — Additional Tables and Figures

E1. Player Ranking by Key Metrics

Player	PER	TS%	BLK%	TOV%	BPM (OBPM+DBPM)
I	**17.2**	**0.547**	**4.9**	14.0	**3.8**
F	13.7	0.521	1.1	**12.3**	**4.8**
B	12.3	0.497	0.6	**14.0**	1.9
C	10.4	0.503	2.5	15.9	0.7
D	9.5	0.462	2.3	**17.2**	0.0
H	9.4	**0.456**	0.4	13.0	0.5
A	9.0	0.546	1.2	**21.6**	0.9
G	**8.2**	0.498	1.7	**10.1**	4.0
E	11.1	0.537	2.7	14.9	0.7

E2. Recommendation Comparison Table

Condition	ChatGPT Recommendation	Claude Recommendation
Neutral	F, I	F, I
Positive	C, E	C, E

Condition	ChatGPT Recommendation	Claude Recommendation
Negative	H, G	H, D
Development	F, I	F, I
Demographic	I, H	I, H
Confirmation	Team-wide + B focus	Team-wide + B focus