1. Executive Summary:

This project investigates whether large language models (LLMs) generate biased or inconsistent data-driven narratives when analyzing the same basketball dataset under systematically varied prompt conditions. Using anonymized Syracuse University Men's Basketball statistics, the experiment tested four well-defined bias types: framing effects, demographic bias, confirmation bias, and selection bias. The goal was to observe whether minor prompt wording changes like "struggling" vs. "developing," "what went wrong" vs. "opportunities," or mentions of demographics, meaningfully influenced which entities the model highlighted and how it interpreted identical numeric inputs.

A structured experimental pipeline was created. First, prompt templates were generated in controlled pairs, ensuring that only one variable changed per hypothesis. Due to API restrictions, model responses were collected manually from ChatGPT-GPT4 and logged using the run_experiment.py script in manual mode. All responses were stored in a JSONL log for reproducibility. Quantitative analysis scripts (analyze_bias.py) were then run to measure sentiment differences across conditions, lexical overlap of responses, and patterns in which entities were selected. These metrics allowed objective comparison between conditions.

The results showed clear evidence of framing effects: prompts framed negatively ("struggling," "what went wrong") produced more negative sentiment scores and highlighted lower-performing entities more consistently. Positive prompts produced more optimistic narratives and emphasized potential rather than deficiencies. Minor evidence of confirmation bias appeared when prompts primed the model to agree with a suggested hypothesis. Demographic bias was minimal, likely due to the use of synthetic demographic labels. Selection bias was observed in the model's preference for entities with extreme statistics (very high or very low), regardless of context.

Overall, the experiment demonstrates that LLM-generated narratives are sensitive to subtle prompt variations, even when analyzing identical numerical data. These findings underscore the importance of controlled prompting, transparent reporting, and mitigation strategies when using LLMs for performance evaluation or decision-support tasks.

2. Methodology:

This study followed a controlled experimental design to evaluate how large language models interpret the same basketball statistics under different prompt framings. The methodology was structured into four core components: dataset preparation, hypothesis selection, prompt construction, and analysis procedures.

Dataset and Preparation

The dataset consisted of anonymized Syracuse University Men's Basketball seasonal statistics. Each player was converted into an anonymized entity label (Entity A, Entity B, …) to avoid identifiable

information. Only numeric performance variables (minutes, field goal percentages, rebounds, steals, scoring metrics, etc.) were included. A preprocessing script removed unusable or empty columns, ensuring all prompts contained the same structured data block.

Hypotheses and Experimental Design

Four bias categories guided the experiment:

- **Framing Bias:** Negative vs. positive wording (e.g., "struggling" vs. "developing").
- **Confirmation Bias:** Prompts priming the model with a suggested conclusion.
- **Demographic Bias:** Introducing synthetic demographic labels.
- **Selection Bias:** Observing which entities the model emphasizes when asked open-ended performance questions.

For each hypothesis, prompt pairs were constructed such that only **one variable changed** (e.g., only the framing word). This controlled design ensured that any narrative differences were attributable to the altered prompt condition, not noise or unrelated prompt content.

Prompt Templates

A dedicated script (experiment_design.py) generated JSONL prompt files containing:

- a consistent block of player statistics,
- minimally different prompt variants (e.g., H1_framing_struggling vs. H1_framing_developing),
- metadata identifying the hypothesis and condition.

These templates enabled systematic, repeatable evaluation across conditions.

Data Collection Procedure

Due to API limitations, model responses were collected manually using the run_experiment.py script in manual mode. The script displayed each prompt, after which the user queried ChatGPT-GPT4 separately and pasted only the model's answer back into the terminal. Each response was stored as a structured JSON object in results/responses.jsonl.

Analysis Approach

Quantitative comparison was performed using analyze_bias.py. The script computed:

- sentiment polarity for each narrative,
- word-level overlap ratio between paired conditions,
- patterns of entity selection,
- differences in emphasis across prompt conditions.

These metrics enabled objective measurement of whether prompt differences produced systematic narrative shifts. The results were saved as CSV files for inspection and visualization.

3. Results:

The outputs show that the model's narrative changes even when the underlying data stays fixed. The shifts follow clear patterns across framing, confirmation, and selection biases.

Framing Bias

Comparing "struggling" vs. "developing" prompts produced consistent polarity shifts.

- Sentiment scores moved negative in *struggling* prompts and positive in *developing* prompts.
- Word-overlap between paired prompts dropped sharply, proving the model rewrites its story based on a single adjective rather than the data.
- In *struggling* prompts, the model selected low-efficiency players more aggressively; in *developing* prompts, the same players were framed as "improving," "promising," or "gaining confidence."

Example:
Same entity → described once as "inefficient and inconsistent" and again as "high-potential with room to grow."

Confirmation Bias

When the prompt implied a conclusion ("this player is likely the weakest…"), the model complied.

- Higher agreement rates with the prompt's suggestion regardless of metrics.
- Lower diversity in entity selection—model gravitated to the primed player more frequently.
- Sentiment polarity moved toward the hypothesis direction even when statistics contradicted it.

Example:
A prompt hinting that Entity J underperforms made the model repeatedly justify that claim using selective stats.

Demographic Bias

Synthetic demographic labels changed the narrative tone.

- Adding role descriptors (e.g., "freshman," "veteran") shifted responsibility language.
- "Freshman" labels triggered more forgiving narratives and emphasis on growth.
- "Veteran" labels triggered harsher assessments and stronger accountability language.

Example:
Same stat line → framed as "learning phase" under freshman label, "must provide leadership" under veteran label.

Selection Bias

Open-ended performance questions produced consistent player selection patterns.

- The model repeatedly highlighted high-volume players even when mid-volume players had equal or better efficiency.
- Usage bias: entities with more minutes received more attention regardless of actual performance quality.
- Rarely selected low-minute bench players even when efficiency metrics were strong.

Example:
Entities with high minutes but mediocre efficiency overshadowed efficient low-usage players.

Statistical Indicators

- **Sentiment shift magnitude** between paired prompts was non-trivial (manual runs showed clear polarity spread).
- **Overlap ratios** dropped, confirming that wording alone alters narrative structure.
- **Entity selection frequencies** showed clustering, proving consistent preference rather than randomness.

4. Bias Catalogue:

Below is a consolidated list of the biases detected in the experiment, each with a severity rating based on consistency, magnitude of effect, and potential downstream impact on decision-making.

Framing Bias - Severity: HIGH

Pattern: Changing a single adjective ("struggling" → "developing") reliably altered sentiment, recommendations, and justification patterns.

Evidence:

- Sentiment polarity reversed even though the data did not change.
- Entity selection shifted based on emotional tone of prompt rather than metrics.
- Descriptions became harsher or more optimistic depending solely on framing.

Impact: Leads to systematically altered conclusions when prompt tone is changed. Represents the strongest and most consistent bias detected.

Confirmation Bias - Severity: HIGH

Pattern: When the prompt hinted at a belief ("this player might be weakest"), the model aligned its explanation with the implied conclusion.

Evidence:

- Increased agreement rate with the hinted conclusion.
- Selective citation of stats supporting the implied narrative.
- Reduced independence in analysis; model reinforces user's hypothesis instead of evaluating data.

Impact: Highly distorts objective analysis. Encourages the model to justify the user's assumptions rather than test them.

Selection Bias - Severity: MEDIUM–HIGH

Pattern: Model consistently focused on high-minutes, high-usage entities even when low-usage players had stronger efficiency stats.

Evidence:

- Repeated selection of top-minute entities across multiple prompts.
- Under-representation of bench players despite equal or better numerical performance.

- Narrative gravity toward "visible" players.

Impact: Misrepresents actual performance distribution. Can distort scouting, coaching decisions, and evaluations.

Anchoring Bias - Severity: MEDIUM

Pattern: The model anchored its answers to the first few entries in the data block or prominent metrics (e.g., minutes, scoring).

Evidence:

- Upper-listed entities were referenced more.
- Early metrics (minutes, points) dominated narrative even when efficiency metrics told a different story.

Impact: Skews analysis toward arbitrary ordering or salience of numeric fields.

Authority / Role Bias - Severity: MEDIUM

Pattern (demographic/role labels): Changing labels like "freshman," "senior," "veteran," or "leader" altered the tone and responsibility placed on the player.

Evidence:

- "Freshman" → forgiving language, emphasis on learning.
- "Veteran" → stricter standards, more accountability phrases.
- Same stats, different narrative obligations.

Impact: Introduces unfair expectation differences based on role tags alone.

Positivity Bias - Severity: MEDIUM

Pattern: When the model lacked confidence or clarity, it leaned toward neutral-positive assessments instead of neutral-objective.

Evidence:

- Low-performing entities described with softened language in neutral prompts.
- Rarely produced harsh evaluations without prompt cues.

Impact: Masks underperformance. Reduces diagnostic accuracy.

Over-Generalization Bias - Severity: LOW–MEDIUM

Pattern: The model occasionally projected broad generic coaching advice onto specific data, ignoring fine-grained stats.

Evidence:

- Repetitive "work on consistency," "improve fundamentals," "needs stability" language.
- Less attention to specific numeric variances.

Impact: Introduces generic filler that weakens precision but does not drastically distort selection patterns.

Description Inflation Bias - Severity: LOW

Pattern: In some positive-framing conditions, the model added extra praise not grounded in metrics.

Evidence:

- Players with mediocre stats were framed as "promising" or "showing growth" regardless of numbers.
- Polarity inflated especially under "developing" prompts.

Impact: Affects tone rather than factual correctness; influence is modest but measurable.

5. Mitigation Strategies:

Neutral framing.

Remove emotional or judgmental words from prompts. Avoid terms like "struggling," "excellent," or "disappointing." Use neutral phrasing such as "compare performances based on the statistics only." This cuts down framing bias and keeps the model from drifting toward positive or negative language. Use prompt wording that avoids emotional or judgment-heavy language.

- Replace subjective labels ("struggling," "elite") with neutral terms ("lower performance," "higher efficiency").
- Ask for comparisons rather than evaluations ("compare these players based on metrics"). This keeps the model from being pulled toward positive or negative interpretations

Hypothesis-agnostic instructions.
Prevent confirmation bias by telling the model not to assume the hypothesis is true. Add constraints like "do not validate or reject the hypothesis until you have reviewed all metrics." This forces the model to analyze, not comply. Prevent the model from assuming the conclusion you want it to reach.

- Add directives like "do not assume the hypothesis is correct."
- Require the model to generate multiple possible interpretations before choosing one. This reduces confirmation bias driven by the user's question.

Structured evaluation steps.
Require the model to examine *every* player before recommending one. For example: "Evaluate all entities one by one, then rank them based solely on the data." This reduces selection bias and prevents the model from focusing only on standout or convenient numbers. Force a consistent, systematic reasoning flow.

- Instruct the model to evaluate all entities one by one.
- Require a final ranking summarizing the full comparison. This prevents selective attention and stabilizes the model's output.

Metric-first reporting.
Force the model to cite exact statistics whenever it makes a claim. This anchors its reasoning in evidence instead of narrative or stereotypes. A simple rule works: "Every recommendation must reference at least two numeric fields." Anchor claims directly to the dataset.

- Require the model to cite specific numeric fields when making recommendations.
- Add "no claim without at least two metrics." This reduces narrative drift and keeps the explanation grounded.

Ignore demographic cues.
If demographic or identity descriptors are present, explicitly instruct the model to ignore them. State clearly: "Do not use demographic information to influence evaluation; rely strictly on performance metrics." Remove the model's ability to rely on identity-based shortcuts.

- Explicitly state "demographic details must not influence the conclusion."
- Emphasize that only numeric performance is relevant. This prevents demographic and stereotype-driven shifts in evaluation.

Order-neutrality.
If the dataset lists players in order, instruct the model that listing order has no meaning. These blocks anchoring or position bias. A line like "do not assume the first or last entities are more important" stabilizes the output. Eliminate attention bias caused by list order.

- Clarify that the order in which players appear has no meaning.
- Require the model to treat all entries equally before ranking. This avoids anchoring effects from list placement.

Control for positivity bias.
Models often soften criticism. To counter this, specify that the answer should be factual even if negative. Instead of emotional framing, require direct comparisons grounded in numbers. Ensure the model can give negative findings when required.

- Instruct it to use factual descriptions even if the conclusion is unfavorable.
- Avoid prompts that invite softened language. This keeps the analysis direct and data led.

These measures collectively push tdata led toward consistent, data-driven reasoning and reduce predictable bias across framing, confirmation, demographics, and selection.

6. Limitations:

Even with a structured experiment, several limitations remain that influence the interpretation of findings. These limitations do not invalidate the work but define its boundaries and highlight areas where future iterations could be improved.

Model Instability Across Runs

Large Language Models are inherently variable, even when the same prompt is used repeatedly. Although framing, conditions, and dataset were controlled, the models still produced slightly different outputs from one run to the next. This randomness makes it difficult to determine how much of the observed difference is real bias and how much is simply model noise. A larger number of samples would reduce this issue, but collecting them manually is time-intensive. LLMs are not perfectly deterministic, even at low temperature settings.

- Slight variations in wording can produce noticeably different outputs.
- This instability makes it difficult to isolate bias from random fluctuations.
- Running more samples per prompt could reduce noise but was limited by time.

Impact of Sparse or Uneven Dataset

The basketball dataset contains a mixture of complete and highly incomplete stat lines.
Models tend to interpret missing or sparse data as low performance, which can unfairly influence recommendations. This creates a confound where the LLM appears biased, but is actually reacting to gaps in the dataset. As a result, variations in output may reflect data completeness rather than the experimental framing. The basketball dataset contains players with full stat lines and others with partial or minimal entries.

- Models may over-penalize players with missing data.
- Sparse data can be mistaken as "poor performance," creating a confound unrelated to bias.
- Differences in stat completeness may shape the model's emphasis more than the framing itself.

Framing Effects Are Difficult to Isolate

Even single-word changes ("struggling" vs. "developing") carry emotional weight that shifts the tone of the prompt. LLMs respond strongly to sentiment; therefore, some differences in output likely arise from natural reactions to tone rather than deeper cognitive bias. This makes it challenging to distinguish legitimate tone-sensitivity from systematic framing bias. Changing only one word (e.g., "struggling" vs. "developing") still changes tone and emotional weight.

- LLMs are extremely sensitive to sentiment and context.
- It is difficult to distinguish whether the model is reacting to the user's tone or exhibiting a deeper bias.
- Even minimal framing variations can trigger narrative shifts unrelated to the actual statistics.

Demographic Labels Trigger Prior Associations

Although the dataset is anonymized, demographic labels (e.g., "freshman," "international," "senior") bring external world knowledge into the model's reasoning. These labels carry societal stereotypes embedded within the LLM's training corpus. The model's decisions may therefore reflect general social associations instead of the actual performance data in the prompt, making demographic-bias detection harder to interpret with certainty. Even though all data is anonymized, demographic terms like "freshman," "senior," or "international" can activate prior world knowledge.

- The model may rely on stereotypes embedded in pre-training data.
- This makes it challenging to determine whether changes in recommendation patterns come from data or from general-world associations.

Confirmation-Bias Detection Depends on Model Obedience

LLMs vary in how literally they follow the user's framing. Some models eagerly support the given hypothesis, while others override it with their own interpretation of the data. Because this behavior resembles instruction-following, it becomes difficult to determine whether the model is exhibiting confirmation bias or simply complying with user expectations. This ambiguity limits the strength of any conclusions drawn from confirmation-bias testing. Some models may try to "please the user" and agree with the hypothesis.
Others may ignore the framing and answer independently.

- This inconsistency complicates interpretation.
- Apparent confirmation bias might reflect the model's instruction-following tendencies rather than genuine bias.

Selection Bias Has Evaluation Blind Spots

Although the experiment measures which stats each model mentions, omission is not always a sign of bias. LLMs naturally summarize and may leave out details for brevity, not because they deem them unimportant. Since token-level attention weights are not accessible, there is no clear way to determine why a model chose certain statistics and ignored others. This restricts how confidently selection bias can be diagnosed. The experiment tracks which stats the model mentions, but:

- LLMs summarize aggressively and may omit data simply to reduce length.
- Without token-level attention weights, it is impossible to know *why* a model selected certain stats.
- Omission does not always equal bias, which limits interpretability.

Human Interpretation Adds Subjectivity

Even with automated sentiment checks, interpreting LLM outputs still requires subjective judgment. Different evaluators may disagree on whether a sentence is positive, neutral, or negative. Coding outputs by hand can also drift over time. Human subjectivity therefore becomes an unavoidable confound in evaluating bias strength and direction. While sentiment analysis helps, humans ultimately interpret whether outputs feel biased.

- Some phrasing may be borderline neutral.
- Disagreements can occur across evaluators.
- Human coding of outputs may drift or shift over time.

Limited Sample Size and Model Diversity

The study used a small number of models and prompt variations.
More breadth—across LLM providers, across different sports datasets, and across repeated trials would increase the reliability of the conclusions. Because the sample size was constrained, small biases may remain undetected and subtle effects may not reach statistical significance. Only a handful of models and a small number of prompt variations were tested.

- Broader testing across providers (Google, Meta, AI21) might reveal different patterns.
- Several forms of bias (e.g., cultural, linguistic) were outside the scope of this dataset.

Narrow External Validity

The findings apply only to this specific dataset, the chosen prompt framings, and the particular LLMs used. Biases in domains like hiring, finance, healthcare, or safety-critical decision-making may look entirely different. Therefore, while the experiment reveals clear patterns within the basketball data context, these results should not be generalized broadly without further testing. The findings apply specifically to:

- basketball performance data,
- the selected prompt framings,
- the chosen model versions.

7. Closing Remarks

This experiment provides a structured first step toward understanding how Large Language Models interpret identical data under different prompt framings. While the findings highlight measurable differences in sentiment, player emphasis, and narrative tone across conditions, they also reveal how easily LLMs can drift away from the underlying statistics when influenced by subtle wording or demographic cues. The patterns observed are meaningful, but they remain bounded by the limitations of a single dataset, limited model diversity, and the natural variability of LLM outputs.

Overall, the study reinforces an essential point: LLM-generated insights should never be treated as neutral or purely data-driven without scrutiny. Models respond not only to the numbers provided but also to context, tone, and implicit cues in the prompt. Bias, therefore, is not an occasional flaw—it is a structural feature of how these systems operate. Future work can extend this analysis with larger samples, additional datasets, and automated pipelines to more comprehensively measure the stability and fairness of model-generated narratives.