Basketball Statistics Research Documentation

Objective

The purpose of this assignment is to explore the capability of a Large Language Model (LLM) to answer natural language questions using structured basketball statistics data. The goal is not only to obtain correct answers but also to document the process, identify where the LLM struggles, and capture how prompts were engineered and refined to achieve better accuracy.

I have worked with the 2024–2025 Syracuse University Men's Basketball statistics (overall and conference play). The LLM's mainly ChatGPT and Gemini was prompted with questions across three different ways:

1. **Simple comprehension and direct retrieval questions**
2. **Analytical and judgment-based questions requiring metric definitions**
3. **Running Scripts to ensure the answers provided are correct.**

For each prompt, I documented:

- The natural language question
- The LLM's response
- Issues encountered (misinterpretation, missing metrics, hallucinations)
- Prompt engineering or metric tweaks applied
- Possible reasons for failure or improvement

Simple Questions

Goal: Test if the LLM can correctly extract basic statistics from the table without complex reasoning.

Questions Asked:

1. **How many games did the team play this season?**
   - **Expected Answer:** 33 games.
   - **LLM Response:** Correct (33 games).
   - **Issues:** None.
   - **What Worked:** Simple direct retrieval from the Overall Record line.
2. **What was the team's overall and conference record?**
   - **Expected Answer:** 14–19 overall, 7–13 conference.
   - **LLM Response:** Correct.
   - **Issues:** None.
3. **Who was the team's top scorer and what was his average points per game?**
   - **Expected Answer:** J.J. Starling – 17.8 PPG.

    - **LLM Response:** Correct.
    - **Issues:** None.

4. **Which player had the best 3-point percentage (minimum 30 attempts)?**
    - **Expected Answer:** Jyáre Davis – 43.2% (16/37).
    - **LLM Initial Response:** Listed Eddie Lampkin Jr. (2/5 = 40%).
    - **Issue:** LLM ignored the minimum attempts criterion.
    - **Prompt Adjustment:** Specified "minimum 30 attempts."
    - **Result After Tweak:** Correctly identified Jyáre Davis.
    - **Reason for Failure:** LLM prioritizes highest percentage without filtering for sample size unless explicitly instructed.

Key Takeaways:

- LLM performs well with direct stats.
- Explicit criteria (min attempts, thresholds) are required to avoid hallucination or misleading stats.
- Prompts must guide the LLM to avoid cherry-picking small samples.

Complex Analytical Questions

Goal: Evaluate LLM's ability to synthesize metrics, make judgments, and recommend actions.

Questions Asked:

1. **Who was the most improved player during the season?**
    - **Initial Metric:** Change in scoring average from non-conference to conference play.
    - **LLM Initial Response:** Chose J.J. Starling (most points overall, but no comparison made).
    - **Issue:** LLM did not calculate improvement.
    - **Prompt Adjustment:** Specified: *"Calculate the difference in PPG between overall and conference games and select the player with the largest increase."*
    - **Final Response:** Correctly identified Donnie Freeman (increased from ~13.4 to 14.3 PPG in limited games).
    - **Reason for Failure:** LLM defaults to total performance, not improvement, unless metric is defined.
2. **If the coach wanted to win two more games next season, should they focus on offense or defense?**
    - **Initial LLM Response:** Offense (vague reasoning).
    - **Issue:** Did not analyze scoring margin or turnover stats.
    - **Prompt Adjustment:** Specified:
        - Calculate average scoring margin (–3.1)
        - Compare turnovers per game (–2.7 margin)
        - Compare opponent 3PT% vs. our 3PT%.
    - **Final Response:** Suggested focusing on **defense and turnover reduction**.

    - **Reason for Failure:** LLM often provides surface-level advice without step-by-step guidance; metrics must be spelled out.
3. **Which single player could become a game-changer if improved, and why?**
    - **Initial LLM Response:** Picked J.J. Starling (top scorer).
    - **Issue:** Did not consider efficiency metrics.
    - **Prompt Adjustment:** Specified:
        - Identify player with high usage (FGA) and low efficiency (FG%, TOs).
        - Consider impact on team's margin if efficiency improves.
    - **Final Response:** Correctly identified J.J. Starling as the game-changer due to high volume but 40.7% FG and 26.8% 3PT.
    - **Reason for Failure:** LLM tends to choose highest stats by default unless efficiency context is emphasized.

Key Learnings & Prompt Engineering Insights

1. **Metric Definition is Critical**
    - LLMs cannot infer analytical criteria without explicit instruction.
    - Always define thresholds (e.g., min attempts, improvement metric).
2. **LLM Hallucination Triggers**
    - Prefers extreme values even on small samples.
    - May ignore context like conference vs. overall stats unless prompted.
3. **Effective Prompting Patterns**
    - Use **step-by-step instructions**: "First calculate margin, then compare players."
    - Add **constraints**: "Use players with >30 3PT attempts."
    - Ask **for reasoning or calculations** to validate the answer.
4. **Documentation of Failures is Valuable**
    - Noting *why* the LLM failed (vagueness, missing metric, sample bias) is part of the research outcome.

Conclusion

Over the two reporting periods:

- LLMs excelled in **direct stat retrieval**.
- LLMs struggled with **comparative and judgment tasks** without clear, metric-driven prompts.
- Iterative **prompt refinement and metric definition** led to better accuracy.
- Documenting failure cases provided as much insight as the successful answers.

Task In Progress

Currently, based on the answers, I have seen that Chatgpt has outperformed Gemini in terms of precision but Chatgpt tends to provide vague long answers. I am trying to validate my prompt that would work for both the LLM Models. I tend to complete this task by next week i.e 7th August, 2025.