Research Task 5 – Progress Report #2

Date: August 15, 2025
Student: Austin Anthony Rodrigues
Project: Large Language Models and Sports Statistics Analysis

1. Introduction

This second phase of the research builds on the July 31, 2025 findings from Part 1, where Large Language Models (LLMs) demonstrated perfect accuracy in basic factual retrieval tasks for Syracuse University Women's Lacrosse 2024 season statistics.

In Part 2, the focus shifted toward:

- **Intermediate analytical calculations**
- **Multi-step reasoning**
- **Strategic insight generation**

The aim was to evaluate whether LLMs could maintain accuracy and deliver meaningful recommendations when confronted with more complex, real-world sports analytics scenarios.

2. Objectives

The primary objectives for this phase were:

1. Test LLM performance on **threshold-based list generation** and derived statistical calculations.
2. Assess strategic recommendations using a **custom evaluation rubric** (Specificity, Actionability, Plausibility).
3. Identify strengths, weaknesses, and consistent patterns in LLM analytical behavior.

3. Methodology

3.1 Dataset

Same dataset as Part 1:

- Syracuse Women's Lacrosse 2024 official statistics
- 34 players, full-season stats (goals, assists, shots, shooting %, games played)
- Team record: 16–6 (Home: 9–2 | Away: 5–2 | Conference: 9–1)

3.2 Testing Framework

The Python-based validation system from Part 1 was extended to handle:

- Calculated metric validation (e.g., shooting %)
- Threshold-based filtering accuracy checks
- Manual rubric scoring for open-ended strategic responses

3.3 LLM Tested

- **Claude Sonnet 4** remained the primary LLM
- Prompts were generated from intermediate_prompts.txt and strategic_prompts.txt

4. Results

4.1 Intermediate Analytical Questions

Question	Expected Output	LLM Output	Accuracy
Calculate shooting % for top 3 scorers	Within ±0.5% of actual	All 3 correct	Pass
List players with ≥10 goals	9 players	8 players	Fail

Observation:
While numeric calculation accuracy was high, threshold-based player list generation occasionally excluded valid entries. This suggests that sorting and filtering logic in LLM reasoning may need explicit instructions.

4.2 Strategic Insight Questions

Evaluated using a 5-point rubric for Specificity, Actionability, and Plausibility.

Prompt	Specificity	Actionability	Plausibility	Verdict
"How can Syracuse improve from 16–6 to 18–4?"	2	5	2	Fail – Lacked balance between offense/defense; overly general offensive focus
"Which non-goal improvements would have the biggest impact?"	4	4	4	Pass – Suggested draw control %, turnover reduction, and defensive clears

Observation:
Strategic prompts with **explicit statistical framing** received higher rubric scores. Generic prompts without constraints tended to elicit vague or unbalanced recommendations.

4.3 Summary of Findings

Strengths:

- Accurate when given explicit formulas or direct calculation instructions.
- Clear formatting and explanation in outputs.
- Capable of producing actionable, contextually relevant recommendations for targeted prompts.

Weaknesses:

- Occasional omission of qualifying players in list-based outputs.
- Inconsistent specificity in broader strategic advice.
- Tendency toward offense-heavy recommendations without defensive context unless prompted.

Patterns:

- Formula-driven prompts → High accuracy & reliability.
- Open-ended prompts → Variable quality, heavily dependent on prompt specificity.

5. Comparative Insights (Part 1 vs Part 2)

Capability	Part 1	Part 2
Basic Data Retrieval	100% accuracy	N/A
Calculations	N/A	High accuracy
Threshold Lists	N/A	Moderate accuracy (1/2 correct)
Strategic Analysis	N/A	Mixed results (1 pass, 1 fail)

6. Recommendations

1. **Prompt Engineering:** Include explicit metric definitions and thresholds in the question to reduce list errors.
2. **Balanced Context:** Incorporate both offensive and defensive statistics in strategic prompts.
3. **Multi-LLM Testing:** Cross-validate results with ChatGPT and GitHub Copilot for consistency.
4. **Domain-Specific Fine-Tuning:** Explore training/fine-tuning with lacrosse-specific strategic case studies.

7. Next Steps

- Run identical Part 2 tests on ChatGPT and Copilot.
- Introduce **defensive and possession metrics** into prompts.
- Automate rubric scoring using the validation framework.
- Prepare a **combined final report** consolidating both parts for publication.

8. Conclusion

The second phase of testing demonstrates that while LLMs can handle intermediate sports analytics tasks effectively, strategic reasoning quality varies depending on prompt clarity and contextual framing. The system excels in deterministic calculations but remains susceptible to incomplete outputs and generic advice when handling open-ended queries.

The findings from Part 2 provide a clear roadmap for refining prompt design and validation processes, laying the foundation for the final stage of this research.

All data sources are publicly available team statistics.

Contact: arodr173@syr.edu

Course: Research Methods

Instructor: Dr. J. R. Strome

Timeline: July 2025 – August 2025