Research Task 5 - Progress Report #1

Date: July 31, 2025
Student: Austin Anthony Rodrigues
Project: Large Language Models and Sports Statistics Analysis

Executive Summary

I've successfully completed the initial phase of testing Large Language Models' ability to analyze sports statistics using Syracuse Women's Lacrosse 2024 season data. The framework is operational, baseline testing is complete, and the results reveal both promising capabilities and areas requiring deeper investigation.

Project Setup and Data Collection

Dataset Selection

After reviewing the assignment requirements, I chose Syracuse Women's Lacrosse 2024 season statistics for several practical reasons:

- **Complete availability**: Full season player and team statistics accessible
- **Recent performance**: 16-6 record provides both success patterns and improvement opportunities
- **Appropriate scale**: 34 players across 22 games fit well within LLM context limits
- **Personal engagement**: Following a local team made the research more meaningful

Technical Implementation

I developed a Python-based testing framework that:

- Processes official Syracuse statistics into structured format
- Generates progressive difficulty test prompts
- Validates LLM responses against calculated ground truth
- Tracks accuracy patterns and error types
- Exports formatted data for direct LLM testing

Initial Testing Results

Phase 1: Basic Statistical Questions (COMPLETED)

Overall Performance: 100% accuracy (5/5 tests)

Question Type	LLM Response	Expected Answer	Result
Season Record	"16-6"	16-6	✓ Correct
Total Games	"22 games"	22	✓ Correct
Top Scorer	"Meaghan Tyrrell (70 goals)"	Meaghan Tyrrell, 70 goals	✓ Correct
Team Goals	"319 total goals"	319	✓ Correct
Assist Leader	"Emma Ward (37 assists)"	Emma Ward, 37 assists	✓ Correct

Key Observations from Basic Testing

Strengths Identified:

- Perfect accuracy on straightforward data retrieval
- Responses included appropriate context (not just raw numbers)
- Consistent formatting and clear communication
- No evidence of data hallucination or invented statistics

Response Quality Notes:

- LLM naturally provided both player names and statistical values
- The answers were conversational but precise
- Context clues were used appropriately (e.g., "leading scorer" vs "most goals")

Framework Validation

The testing infrastructure proved robust and reliable:

- **Automated validation** successfully caught discrepancies (tested with intentionally wrong responses)
- **Data processing** accurately calculated ground truth statistics
- **Prompt generation** created appropriately formatted questions for LLM testing

- **Results tracking** maintained comprehensive logs for analysis

Methodology Refinements

What's Working Well

1. **Clear data presentation**: Structured statistics in consistent format
2. **Progressive difficulty**: Building from simple to complex questions
3. **Objective validation**: Automated checking removes subjective bias
4. **Comprehensive documentation**: Every test result captured and categorized

Adjustments Made

- Added shooting percentage calculations for more robust player analysis
- Included games played data to enable per-game statistics
- Refined prompt formatting based on initial LLM response patterns
- Enhanced validation logic to handle different response formats

Preliminary Insights

LLM Capabilities Confirmed

- **Data retrieval**: Excellent performance on factual questions
- **Context awareness**: Understood relationship between questions and dataset
- **Communication**: Responses were appropriately detailed and well-formatted

Areas for Further Investigation

- **Mathematical reasoning**: How accurate are multi-step calculations?
- **Comparative analysis**: Can LLMs meaningfully compare player performances?
- **Strategic insights**: Will coaching recommendations be generic or specific?
- **Consistency**: Do responses vary across multiple attempts at same questions?

Challenges Encountered

Technical Issues (Resolved)

- Initial validation logic confusion between game-level and season-total data
- Required refinement of ground truth calculations for player statistics
- Needed to add error handling for different LLM response formats

Data Considerations

- Some players had limited playing time, creating statistical edge cases
- Position classifications required interpretation for meaningful analysis
- Season totals vs. per-game statistics needed careful handling

Research Design Decisions

- Chose to focus on one LLM initially (Claude) for consistency
- Decided to validate each response individually rather than batch processing
- Prioritized accuracy measurement over response time or efficiency

Next Phase Planning (August 15 Target)

Immediate Priorities

1. **Intermediate Testing**: Questions requiring calculations and comparisons
2. **Complex Analysis**: Strategic coaching recommendations and insights
3. **Cross-validation**: Test same prompts with multiple LLMs
4. **Edge Cases**: Challenging scenarios and ambiguous questions

Specific Questions to Test

- "Calculate shooting percentages for players with 20+ shots and rank them"
- "Which players provide the most balanced offensive contributions?"
- "What should Coach Gait focus on to improve from 16-6 to 18-4 next season?"
- "Based on these statistics, what type of player should Syracuse recruit?"

Success Metrics for Phase 2

- **Quantitative**: Accuracy rates on calculation-heavy questions
- **Qualitative**: Reasonableness and specificity of strategic recommendations
- **Consistency**: Variation in responses across multiple attempts
- **Comparative**: Performance differences between LLMs

Personal Reflections

This research has been more engaging than anticipated. Working with real Syracuse data (rather than hypothetical examples) made the testing feel meaningful and connected to actual athletic performance.

The perfect accuracy on basic questions was expected but still satisfying to confirm. What I'm most curious about is whether this performance will be maintained as questions become more complex. Can LLM provide coaching insights that would be useful to the Syracuse coaching staff?

The validation framework development was educational in itself - it forced me to think carefully about what constitutes a "correct" answer, especially for subjective questions about strategy and recommendations.

Technical Deliverables Completed

- **Testing Framework**: Fully operational Python validation system
- **Dataset Processing**: Syracuse 2024 statistics cleaned and formatted
- **Ground Truth Calculations**: All correct answers pre-computed
- **Prompt Library**: Basic questions tested and refined
- **Results Documentation**: Comprehensive logging and analysis
- **GitHub Repository**: Code and documentation ready for submission

Risk Assessment and Mitigation

Potential Challenges Ahead

- **Calculation accuracy**: LLMs may struggle with complex math
- **Domain knowledge**: Sports strategy requires specialized understanding
- **Subjectivity**: How to validate "coaching recommendations"?

Mitigation Strategies

- Develop rubrics for evaluating strategic advice quality
- Test multiple calculation approaches to identify patterns
- Consult Lacrosse Coaching Resources for recommendation validation

Conclusion

Phase 1 has successfully demonstrated that the research framework is sound and that LLMs can handle basic sports statistics analysis with perfect accuracy. The groundwork is laid for more challenging questions that will reveal the true limits and capabilities of these systems in sports analytics.

The next phase will determine whether LLMs can move beyond data retrieval to provide genuine analytical insights that would be valuable to coaches, players, and fans.

Files Submitted:

- Testing_and_Validation.py - Complete testing framework
- syracuse_lacrosse_2024_real.csv - Official team statistics
- basic_testing_results.json - Raw test results and validation
- README.md - Project overview and methodology

Next Report Due: August 15, 2025