Objective Assessment of AI Tools for Sports Data Analytics: Maxwell-v2 vs. Generic LLMs

By Palo Galko

August 18, 2024

Introduction

This report presents a comprehensive evaluation of artificial intelligence (AI) systems in the domain of sports data analysis. The study compares the performance of Maxwell v2, a specialized multi-agent AI framework, against traditional Large Language Models (LLMs) such as GPT-4 and Claude 3.5 Sonnet. The evaluation focuses on a diverse set of 43 sports data analysis tasks, ranging from basic statistical analysis to complex predictive modeling and geographic data interpretation.

The primary objectives of this evaluation are to:

- Assess the capabilities of AI systems in handling real-world sports data analysis scenarios
- Compare the performance of a specialized multi-agent system (Maxwell v2) with general-purpose LLMs
- Identify strengths and potential areas for improvement in AI-driven sports analytics

This study is particularly significant as it sheds light on the potential of AI to revolutionize sports data analysis, potentially offering coaches, athletes, and sports analysts powerful tools for deriving insights from complex datasets. The results of this evaluation may have far-reaching implications for the future of AI applications in sports science, training optimization, and performance analysis.

The following sections detail the methodology, present the results of our evaluation, and discuss the implications of our findings. This report aims to provide valuable insights for researchers, sports professionals, and AI practitioners interested in the intersection of artificial intelligence and sports analytics.

Key Findings

- Comparative analysis shows Maxwell v2 outperforming standard LLMs in sports data analysis.
- Maxwell v2's success rates of 87.21% and 84.88% in two test rounds exceeded those of other LLMs tested.
- The unique multi-agent design and tool integration of Maxwell v2 may explain its performance advantage.

Methodology

This evaluation was conducted to compare the performance of an AI agentic framework (Maxwell v2) against plain LLMs on a set of sports data analysis tasks. The methodology was as follows:

1. A set of 43 sports data analysis tasks was developed, requiring code development and execution.
2. Each response was manually evaluated and scored:
    - 1: Accurate and complete result of code execution
    - 0.5: Partially correct result (e.g., correct numerical figures but incorrect plot)
    - 0: Incorrect response
3. The number of automated corrections required to execute the code without errors was recorded (up to 5 iterations).
4. The evaluation process:
    a. LLM or Maxwell v2 develops the code to address the task
    b. Code is executed
    c. If an error occurs, the error is sent back for correction

d. This process continues for up to 5 iterations

About Maxwell v2

Maxwell v2 is an AI agentic framework consisting of 11 individual agents, each handling a specific part of the task and passing the results onto the next agent. These agents collectively work on the solution presented by the user. Key features of Maxwell v2 include:

1. Multi-agent collaboration: 11 specialized agents work together to solve complex tasks.
2. Task-specific handling: Each agent focuses on a particular aspect of the problem.
3. Data processing: Operations are performed on a multi-year nested dataset of training activities.
4. Insights generation: The framework aims to extract insights such as:
   - Performance trends over time
   - Pattern detection
   - Metrics correlation
   - Outliers detection
   - Training load analysis
   - Seasonal performance variations
   - Recovery patterns
5. Tool access: All agents have access to various tools, including:
   - Web search capabilities
   - Code execution
   - Data visualization and plotting
6. Maxwell-v2 is based on an open source AI data analysis library BambooAI developed by Palo Galko.

The goal of Maxwell v2 is to provide comprehensive analysis and insights from sports training datasets, offering a more specialized and potentially more effective approach compared to general-purpose LLMs.

Evaluation Dataset and Task Categorization

The evaluation dataset consists of 43 sports data analysis tasks designed to assess the capabilities of LLMs and the Maxwell-v2 framework. These tasks are divided into 8 distinct categories, each focusing on different aspects of sports performance analysis:

1. Activity Overview and Basic Statistics
2. Performance Trends and Progression
3. Interval and Segment Analysis
4. Environmental and Contextual Impact
5. Physiological Response and Efficiency
6. Workout Structure and Pacing
7. Comparative Analysis
8. Predictive Modeling and Performance Optimization
9. Geographic and Route Analysis

Within each category, the tasks are further classified into three types:

1. **Coach/Athlete Questions**: These tasks simulate questions that a coach or athlete might ask about their training data. They are typically more practical and applied in nature, focusing on specific performance metrics or trends.
2. **Technical Questions**: Each category includes one technical question that a data analyst might pose. These questions often require more advanced statistical analysis or data processing techniques.
3. **Search-Enabled Questions**: Some categories include questions that require external research or information lookup. These tasks test the ability to integrate external knowledge with the given dataset. These tasks were excluded from the evaluation to ensure equal playing field and fairness.

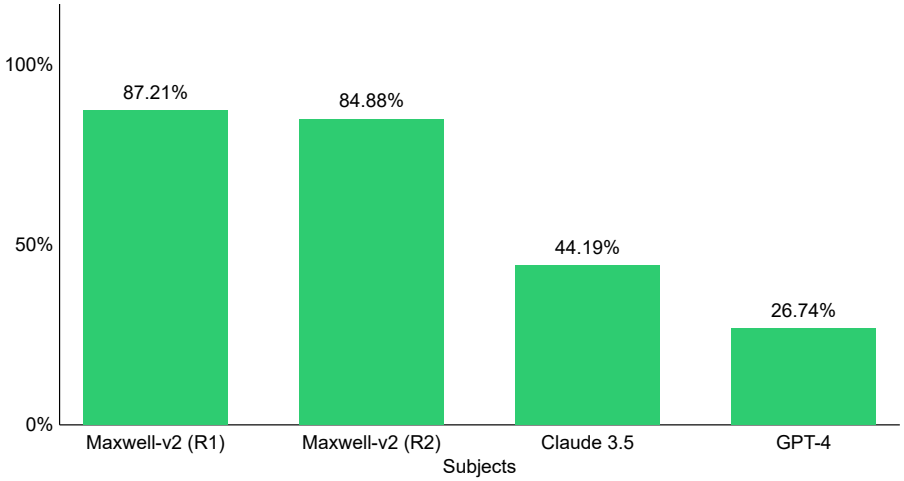
This diverse set of tasks is designed to evaluate the LLMs and Maxwell-v2 across a wide range of sports data analysis scenarios, from basic statistical calculations to complex predictive modeling and geographic analysis. The variety of question types also assesses the systems' ability to handle both practical, user-oriented queries and more technical, analytical tasks.

Subjects

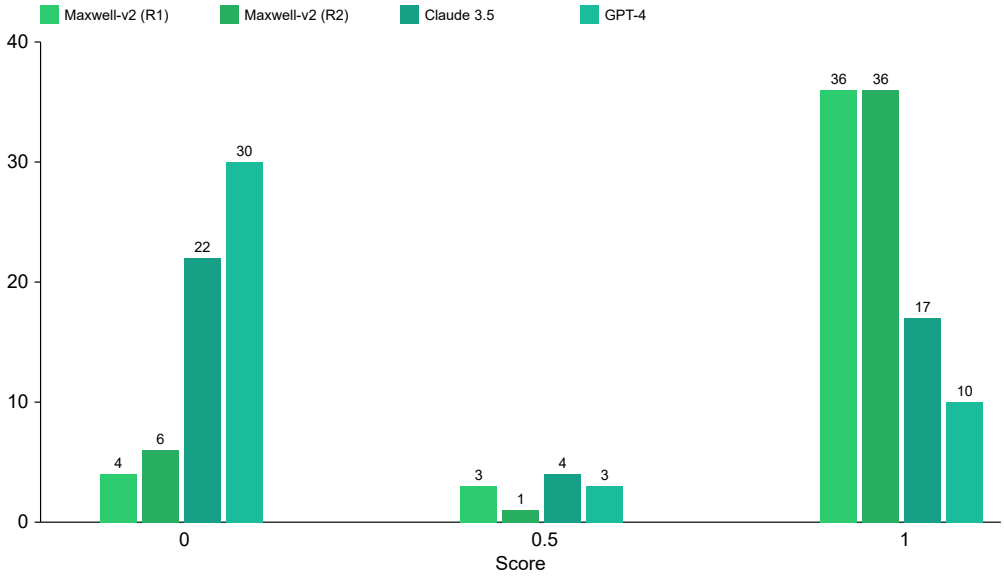
1. Maxwell v2 (Round 1)
2. Maxwell v2 (Round 2)
3. Anthropic Claude 3.5 Sonnet
4. OpenAI GPT-4o (2024-08-06)

Results

Success Rates and Scores



Score Distribution



Results Summary

- **Maxwell v2 (Round 1):** 87.21% success rate, 0.67 avg. corrections
- **Maxwell v2 (Round 2):** 84.88% success rate, 0.28 avg. corrections
- **Claude 3.5 Sonnet:** 44.19% success rate, 0.65 avg. corrections
- **GPT-4:** 26.74% success rate, 0.23 avg. corrections

Summary

The evaluation results demonstrate that the Maxwell v2 agentic framework significantly outperformed the plain LLMs (Claude 3.5 Sonnet and GPT-4) in both rounds of testing. Maxwell v2 achieved success rates of 87.21% and 84.88% in rounds 1 and 2, respectively, compared to 44.19% for Claude 3.5 Sonnet and 26.74% for GPT-4.

Maxwell v2 not only scored higher but also demonstrated more consistent performance across the tasks, with a higher number of fully correct responses (score of 1) in both rounds. The plain LLMs, while capable of handling some tasks, struggled with a larger number of questions, resulting in more incorrect responses (score of 0).

Interestingly, the average number of corrections required varied across the subjects. Maxwell v2 Round 2 and GPT-4 required fewer corrections on average (0.28 and 0.23, respectively) compared to Maxwell v2 Round 1 and Claude 3.5 Sonnet (0.67 and 0.65, respectively).

The superior performance of Maxwell v2 can be attributed to its specialized multi-agent architecture, which is designed specifically for sports data analysis tasks. Each agent in the framework focuses on a particular aspect of the problem, allowing for more nuanced and accurate analysis compared to general-purpose LLMs. The ability to access and utilize various tools, such as web search and data visualization, further enhances Maxwell v2's capabilities in generating meaningful insights from complex sports training datasets.

Conclusion

The Maxwell v2 agentic framework demonstrated superior performance in handling complex sports data analysis tasks, showcasing its potential for automating and improving the accuracy of data analysis workflows in this domain. Its specialized approach offers significant advantages over general-purpose LLMs when dealing with specific, complex domains like sports performance analysis.