NLP Final Project

Submitted by:  
Lior Ben Sidi – 207490913  
Ido Avital – 207850280

# 1. Introduction

LLMs have advanced the boundaries of natural language processing, showcasing capabilities in text understanding, reasoning, and generation. Despite these achievements, evaluating their performance on tasks that require structured reasoning and strict adherence to formal schemas remains a key challenge.

We propose an evaluation benchmark in the domain of sports analytics.

We focus on transforming a chronological play-by-play basketball game log into a complete statistical box score formatted as JSON.

The task requires temporal reasoning, aggregation of events across a game, and robustness to narrative variability, and requiring strict compliance with a predefined schema.

It provides a testbed for structured reasoning in LLMs.

# 2. Background

Sports analytics has emerged as a field where structured data plays a central role.

Box scores, play-by-play logs, and advanced statistics are indispensable for performance analysis, prediction, and decision-making.

Automating the transformation of unstructured narratives into structured statistical representations offers both academic and practical value, yet poses difficulties for LLMs due to the need for consistency, aggregation, and error-free schema adherence.

Our project extends this line of inquiry by introducing a dataset and evaluation pipeline specifically tailored to sports analytics, bridging the gap between general benchmarks and real-world structured reasoning tasks.

# 3. Methodology

Our pipeline has three components:

• Data generation (`generate\_data.py`): simulate games at three difficulty levels; produce the play‑by‑play log and ground‑truth box score (JSON).

• Model evaluation (`run\_eval.py`): prompt models with rosters + log to output a final box score as strict JSON; minimally repair formatting if needed.

• Scoring (`evaluation.py`): compare predictions to ground truth with two metrics: field‑level accuracy and block‑normalized accuracy.

# 4. Difficulty Parameters

# 5. Results

Performance tables will appear in Appendix A after experiments.

# 6. Analysis & Insights

We observe consistent trends across models and difficulty levels. On basic examples, most models are able to correctly aggregate statistics and adhere to JSON formatting. However, as difficulty increases, performance deteriorates.

Common errors include malformed JSON outputs, misaligned team or player statistics, and degenerate all-zero reports when the model fails to parse the log. Substitution events and VAR corrections introduce additional challenges, often leading to mismatched participants or incorrect score adjustments.

models with native support for JSON output (e.g., GPT-4o, Gemini Pro) demonstrated stronger robustness in schema adherence, though they still struggled with reasoning over long narratives. These insights highlight structured evaluation tasks for exposing specific weaknesses in LLM reasoning abilities.

# 7. Conclusion & Future Work

We introduce a benchmark for evaluating structured reasoning in LLMs, centered around the task of converting basketball play-by-play logs into box scores. The dataset and evaluation framework reveal that while current models perform well on simple cases, they struggle with complex, long-context reasoning.

Future work: explore fine-tuning models on structured sports data, incorporating retrieval or symbolic reasoning tools, and enforcing stricter schema validation mechanisms. Beyond sports, the framework can be generalized to other domains where narrative-to-structure transformation is critical, such as legal case summaries or clinical notes.

Addressing these challenges can push the boundaries of how LLMs are evaluated and improve their reliability in high-stakes structured reasoning tasks.

# Appendices

## Appendix A: Difficulty Parameters Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Basic | Medium | Hard | Explanation |
| target\_events | 150 | 600 | 900 | Controls how many events are generated. Low = short/simple logs, High = long/complex games. |
| difficulty\_max\_passes | 5 | 3 | 1 | Maximum passes before a shot. More passes create simpler logs, fewer passes make parsing harder. |
| adversarial\_assist\_bias | False | True | True | If True, uses ambiguous verbs for passes, making assist detection harder. |
| substitution chance | 5% | 10% | 15% | Probability of substitutions. More subs → more players appear, harder tracking of participants. |
| VAR events | Disabled | 5% chance | 10% chance | VAR cancels/changes plays. Adds complexity and requires the model to undo/reason backwards. |
| narrative variety | ¼ of phrases | ½ of phrases | All phrases | How many wording templates are sampled. Higher = more linguistic diversity, harder for LLMs. |
| EVENT\_WEIGHTS | Bias to misses & fouls | Balanced | Bias to made shots & turnovers | Weighted distribution of event types. Shapes the overall game difficulty and ambiguity. |