# 1. Introduction

Large Language Models (LLMs) have rapidly advanced the boundaries of natural language processing, showcasing remarkable 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.

In this project, we propose a novel evaluation benchmark in the domain of sports analytics. Specifically, we focus on transforming a chronological play-by-play basketball game log into a complete statistical box score formatted as JSON. This task requires temporal reasoning, aggregation of events across a game, and robustness to narrative variability, while demanding strict compliance with a predefined schema. As such, it provides a rigorous testbed for structured reasoning in LLMs.

# 2. Background

The evaluation of LLMs has traditionally relied on broad benchmarks such as GLUE, SuperGLUE, and MMLU, which measure general-purpose reasoning and knowledge retrieval. While these benchmarks have driven impressive progress, they do not fully capture the challenges of domain-specific structured reasoning tasks.

In parallel, 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.

Prior work in information extraction and structured prediction has highlighted the difficulties of enforcing schema consistency. Our project extends this line of inquiry by introducing a dataset and evaluation pipeline specifically tailored to sports analytics, thereby bridging the gap between general benchmarks and real-world structured reasoning tasks.

# 3. Methodology

Our experimental pipeline is composed of three tightly integrated components:

1. \*\*Data Generation\*\* – Using `generate\_data.py`, we simulate basketball games across three difficulty levels (basic, medium, hard). Each simulation yields both a natural language play-by-play log and a ground-truth statistical report. Difficulty levels are defined through parameters such as event complexity, linguistic variety, frequency of substitutions, and the inclusion of retroactive VAR (video assistant referee) corrections.  
2. \*\*Model Evaluation\*\* – With `run\_eval.py`, we query a variety of LLMs using standardized prompts that include rosters and play-by-play logs. The models are tasked with producing the final box score in strict JSON format. Post-processing repairs malformed outputs to ensure schema alignment.  
3. \*\*Scoring and Metrics\*\* – Using `evaluation.py`, model predictions are compared against the ground truth. We apply two complementary scoring modes: field-by-field accuracy (each stat checked independently) and fractional-per-block accuracy (normalizing correctness within team and player blocks).

This methodology ensures that evaluation captures both fine-grained correctness and broader structural alignment. By combining deterministic simulation with systematic evaluation, we provide a reliable benchmark for measuring structured reasoning in LLMs.

# 4. Results

The following tables will present model performance across difficulty levels. For now, placeholders are included pending final experimental results.

# 5. Analysis & Insights

Preliminary findings reveal 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 significantly.

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.

Interestingly, 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 underscore the importance of structured evaluation tasks for exposing specific weaknesses in LLM reasoning abilities.

# 6. Conclusion & Future Work

This project introduced a novel 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 should 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.

By addressing these challenges, we aim to push the boundaries of how LLMs are evaluated and improve their reliability in high-stakes structured reasoning tasks.

# 4. Difficulty Parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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. |

Detailed Explanation of Parameters:

- target\_events: This parameter determines the total number of play-by-play events simulated per game. A low value (e.g., 150 in Basic) results in shorter, simpler logs that are easier for models to parse. A higher value (600 in Medium or 900 in Hard) generates longer games with more opportunities for cumulative errors, forcing models to maintain coherence across extended contexts.

- difficulty\_max\_passes: This parameter sets the maximum number of passes allowed before a shot attempt. With more passes (5 in Basic), the event flow is simpler and predictable. Reducing the number (3 in Medium, 1 in Hard) forces more immediate and ambiguous offensive actions, making it harder for models to detect patterns and correctly assign assists.

- adversarial\_assist\_bias: When enabled, this parameter replaces neutral pass verbs with more ambiguous alternatives. For example, instead of 'passes to', the generator may use 'feeds' or 'delivers'. This adversarial phrasing makes it harder for models to recognize and credit assists. It is disabled in Basic for simplicity, but enabled in Medium and Hard to increase difficulty.

- substitution chance: This parameter controls the probability of substitutions happening during the game. At 5% in Basic, substitutions are rare, so tracking participants is straightforward. At 10% in Medium and 15% in Hard, substitutions become more frequent, introducing more players into the game and requiring the model to handle dynamic rosters.

- VAR events: VAR introduces retroactive changes, such as overturning baskets or converting a 3-point shot into a 2-point shot. In Basic difficulty, VAR is disabled to avoid additional complexity. In Medium, VAR occurs in about 5% of scoring plays, while in Hard it rises to 10%. Handling VAR correctly requires reasoning backwards, which is a significant challenge for LLMs.

- narrative variety: This parameter sets how many linguistic templates are used for describing events. Basic uses only a quarter of the available templates, Medium uses half, and Hard uses the full set. Higher variety increases linguistic diversity, forcing the model to generalize beyond memorized patterns.

- EVENT\_WEIGHTS: These weights determine the relative likelihood of different event types (e.g., turnovers, shots, fouls). In Basic, the distribution favors misses and fouls, producing less structured games. In Medium, the weights are more balanced, creating realistic mixes of events. In Hard, the weights shift toward made shots and turnovers, generating fast-paced, high-stakes situations that make accurate stat tracking more difficult.