# 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.

- EVENT\_WEIGHTS: Defines the probability distribution over event types (shots, misses, turnovers, fouls). Basic is biased toward misses and fouls, Medium is balanced, and Hard favors made shots and turnovers. This reshapes the statistical prior of the game log, echoing the role of prior probabilities in probabilistic language models. As in n-gram or log-linear models, skewed distributions lead to systematic biases in prediction, requiring models to adjust expectations dynamically.

- narrative variety: This parameter determines how many phrasing templates are used: one-quarter in Basic, half in Medium, and all templates in Hard. Higher variety increases linguistic diversity and reduces reliance on memorized surface patterns. This reflects the motivation for dense word embeddings: distributional models generalize across varied contexts, but extreme variability stresses a model’s ability to maintain semantic equivalence across forms, a challenge noted in discussions of Zipf’s law and OOV handling.

- VAR events: Retroactive changes (e.g., canceling baskets, downgrading shots) are disabled in Basic, appear at 5% in Medium, and 10% in Hard. These events force the model to revise prior states, akin to structured prediction with retroactive constraints. From the course, this relates to limitations of left-to-right generative models, which cannot easily revise earlier predictions, highlighting the need for architectures with bidirectional context (e.g., BERT) or explicit constraint handling.

- substitution chance: Controls how often substitutions occur: 5% in Basic, 10% in Medium, 15% in Hard. Higher substitution rates expand the set of active participants, requiring long-context tracking. This parallels sequence-labeling tasks in NLP (like NER), where entities may change mid-sequence. RNNs struggle with long dependencies due to vanishing gradients, while Transformers handle longer contexts better but still face quadratic scaling issues.

- adversarial\_assist\_bias: When enabled (Medium, Hard), neutral pass verbs are replaced with ambiguous alternatives such as 'feeds' or 'delivers'. This complicates assist recognition. In NLP terms, this resembles lexical variation in distributional semantics: word embeddings (e.g., Word2Vec, GloVe) place synonyms in nearby regions of the vector space, but rare or misleading lexical choices may increase confusion. Thus, the bias directly stresses a model’s robustness to paraphrasing and lexical variability.

- difficulty\_max\_passes: This parameter sets the maximum number of passes allowed before a shot. Basic games allow up to 5 passes, Medium 3, and Hard only 1. Fewer passes increase ambiguity by forcing quicker offensive plays. Theoretically, this connects to sparse context windows in early n-gram models, where limited history reduces predictive certainty and forces reliance on distributional generalization.

- target\_events: This parameter determines the total number of simulated play-by-play events per game. A lower value (150 in Basic) results in short and simple logs, while higher values (600 in Medium, 900 in Hard) produce longer narratives. From an NLP perspective, longer sequences exacerbate challenges in sequence modeling, as highlighted in the course when discussing vanishing gradients in RNNs and the bottleneck problem in sequence-to-sequence models. Transformers alleviate some of these issues via self-attention, but context length remains a practical limitation.

Detailed Explanation of Parameters:

# 4. Difficulty Parameters

# 5. Results

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

# 6. 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.

# 7. 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.

# 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. |