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

NLP Final Project

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

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.

Similarly, sequence models such as RNNs and Transformers were introduced to handle long-term dependencies. However, as we saw with the vanishing gradient problem and the attention bottleneck, handling long contexts remains a central challenge. Our project leverages these insights by explicitly testing models on increasingly long and complex play-by-play sequences.

In the course, we studied how distributional models and embeddings (e.g., Word2Vec, GloVe) enable generalization across varied lexical forms. This principle motivates the inclusion of narrative variety in our benchmark: it tests whether models can maintain semantic coherence when faced with lexical diversity.

The rapid evolution of NLP has highlighted the importance of structured reasoning tasks. Traditional benchmarks such as GLUE and SuperGLUE focus on general language understanding, but they overlook domain-specific challenges such as aggregating structured statistics from narratives.

2. Background