NLP Final Project

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

The goal is to measure structured reasoning under realistic, noisy narratives.

We build a benchmark that asks LLMs to turn a basketball play‑by‑play into a strict JSON box score — a task that requires temporal reasoning, aggregation, and schema compliance.

# 1. Introduction

Sports analytics depends on reliable box scores; automating narrative→structure is both useful and difficult.

General benchmarks (e.g., GLUE/SuperGLUE/MMLU) miss domain‑specific structured aggregation tasks.

# 2. Background

# 3. Methodology

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

• Model evaluation (`run\_eval.py`): prompt LLMs to output a final box score as JSON; minimally repair format.

• Data generation (`generate\_data.py`): simulate games at three difficulty levels; produce log + ground truth.

Pipeline:

Our experimental pipeline is composed of three tightly integrated components:

# 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.

- EVENT\_WEIGHTS: different priors over event types; shape ambiguity and difficulty.

Results table (5×5) is placed in Appendix A and will be filled after experiments.

- narrative variety: ¼/½/All templates; higher lexical diversity reduces template‑matching.

Models with native JSON support fail less on format, but still struggle with long‑context reasoning.

Performance drops with difficulty: long logs, substitutions, VAR, and varied phrasing cause JSON errors and stat misalignments.

# 7. Conclusion & Future Work

# Appendices

Next: richer validation, retrieval/tools, and domain‑tuned training.

Our benchmark exposes limits of LLM structured reasoning in sports.

## 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. |