NLP Final Project

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

# 1. Introduction

In this project, we designed a novel evaluation challenge for Large Language Models (LLMs). The task requires models to convert a chronological play-by-play basketball game log into a full statistical box score, formatted strictly as JSON. This is a challenging problem since it involves temporal reasoning, aggregation of events, robustness to long contexts, and adherence to strict output formatting.

# 2. Background

Recent advances in Large Language Models have shown remarkable abilities in text understanding, generation, and reasoning. However, evaluating their capacity for structured reasoning tasks such as producing valid statistical reports from narrative descriptions remains an open problem. Prior work has mostly focused on general NLP benchmarks, but domain-specific structured evaluations such as sports analytics have received little attention.

# 3. Methodology

We implemented a full pipeline consisting of three main components:

1. Data Generation: Using `generate\_data.py`, we simulate basketball games across three difficulty levels (basic, medium, hard). Each game produces a narrative play-by-play log and a ground-truth box score with team and player statistics.  
2. Model Evaluation: Using `run\_eval.py`, we prompt different LLMs to reconstruct the final box score from the play-by-play narrative. Outputs are repaired and normalized to the exact schema.  
3. Scoring: Using `evaluation.py`, we compare the reconstructed outputs against the ground truth. Two scoring modes are applied: field-by-field accuracy and fractional-per-block accuracy.

# 4. Results

The following table is reserved for accuracy results of different models across difficulty levels. It will be filled after experiments are completed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# 5. Analysis & Insights

Initial observations indicate that LLMs handle simple, short examples relatively well, often producing correct aggregations and adhering to the JSON format. However, on longer and more complex examples, models frequently struggle with malformed JSON, misaligned statistics, or degenerate all-zero reports. The hardest cases expose significant limitations in long-context reasoning.

# 6. Conclusion & Future Work

We created a challenging dataset and evaluation framework for assessing structured reasoning of LLMs in the context of sports analytics. Our results highlight that while modern LLMs can handle basic aggregation tasks, they still struggle with long-context structured reasoning. Future work includes fine-tuning on structured sports data, integrating retrieval or tool-based reasoning, and enforcing stricter schema validation.