Final Project Report – NLP Course

# 1. Introduction / Motivation

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 difficult because it requires temporal reasoning, aggregation of events, and robustness to long contexts, as well as adherence to strict output formatting.

# 2. Dataset Description

The dataset consists of synthetic basketball games generated with Python scripts. Each example includes:  
- Team rosters (head coach, starting lineup, bench)  
- A narrative play-by-play log of events  
- The ground truth final JSON report with full team and player statistics  
  
The dataset is split into three difficulty levels:  
- Basic: short logs, simple scoring  
- Medium: longer logs with more complex events  
- Hard: long, ambiguous logs requiring advanced reasoning  
  
We generated at least 150 examples, each tagged with its difficulty level.

# 3. Evaluation Method

We implemented `evaluation.py` to measure the accuracy of LLM outputs compared to ground truth reports. The evaluation supports two modes:  
- Field-by-field comparison: Each stat is checked individually.  
- Fractional per block: Each team/player contributes proportionally, rewarding partial correctness.  
  
Special care is taken to penalize missing players, mismatched scores, or incorrect stats, while rewarding correct all-zero stats for non-participants.

# 4. Models Tested

We evaluated multiple LLMs using LiteLLM as the unified API. The following models were tested (results will be filled later):  
  
- GPT-4o: Average accuracy (field) = \_\_\_% ; (fractional) = \_\_\_%  
- o3-mini: Average accuracy (field) = \_\_\_% ; (fractional) = \_\_\_%  
- Gemini 1.5-pro: Average accuracy (field) = \_\_\_% ; (fractional) = \_\_\_%  
- Gemini 2.5-flash: Average accuracy (field) = \_\_\_% ; (fractional) = \_\_\_%  
- Claude Opus 4: Average accuracy (field) = \_\_\_% ; (fractional) = \_\_\_%  
- Grok-3: Average accuracy (field) = \_\_\_% ; (fractional) = \_\_\_%

# 5. Analysis & Insights

Typical strengths observed:  
- Accurate stat aggregation on short/basic examples  
- Correct handling of JSON format when supported natively  
  
Typical weaknesses observed:  
- Models often produced malformed JSON  
- All-zero reports when failing to parse game logs  
- Nested or misplaced team objects  
  
Harder examples exposed limitations in long-context reasoning and structured output control.

# 6. Conclusion & Future Work

We created a challenging dataset and evaluation framework for measuring structured reasoning of LLMs in sports analytics. Our results highlight that while modern LLMs can handle simple aggregation, they still struggle with long-context structured reasoning. Future directions include fine-tuning models on structured sports data, integrating retrieval or tool-based reasoning, and enforcing JSON schemas more robustly.