**Narrative-to-Box-Score:  
Evaluating LLMs on Structured Reasoning in Sports Analytics**

*Course: Natural Language Processing, Technion – Israel Institute of Technology*Submitted by: Lior Ben Sidi – 207490913, Ido Avital – 207850280  
This report summarizes the project at a non-technical level, focusing on problem framing, approach, and evaluation design.

**For more technical information, we recommend reading the "README.md" file.**

**Executive Summary**

This project evaluates how well a Large Language Model (LLM) can convert a natural-language, chronological play-by-play description of a basketball game into a complete, structured box-score in JSON format. The task stresses temporal reasoning (events over time), aggregation (rolling up plays to final statistics), and strict adherence to a predefined schema. The pipeline generates synthetic games, queries an LLM, repairs the model’s output to match the schema when needed, and scores the result with clear, interpretable metrics.

**Problem Statement & Motivation**

Sports analytics relies on structured data such as box scores and play-by-play logs. Automating the transformation from narrative descriptions to structured statistics is valuable for performance analysis and decision-making. However, it is can be challenging for language models because the output must be both faithful to the narrative, such that it will represent the true stats, and also consistent with a strict schema of JSON.  
This project provides a focused testbed for that challenge.

**Task Definition (What the Model Must Do)**

Given a chronological play-by-play of a single basketball game plus team/roster context, the model must produce a complete box-score: final score, per-team totals, and per-player statistics. The output must follow a fixed schema, be internally consistent, and reflect what actually happened in the narrative.

**Data & Simulation**

Games are synthetically simulated to produce two paired artifacts: (1) a natural-language play-by-play narrative with team metadata (rosters, starting lineup, participants), and (2) a ground-truth statistical report (team and per-player stats). The simulation enforces core safety invariants such as “attempts ≥ made” across all shot types and handles realistic basketball phenomena like substitutions, rebounds, fouls, and occasional video-review (VAR) adjustments that may overturn or modify recent plays while keeping the data consistent.

Rationale. Synthetic pairing lets us control difficulty and coverage while keeping a clean “source → target” mapping for auditing. Substitutions dynamically update the set of participants, testing entity tracking; rebounds and turnovers test possession flow; and VAR introduces limited retroactive edits to probe whether models can reconcile earlier and later statements without breaking invariants. (Optionally using a fixed random seed supports reproducibility of specific game scenarios.)

**Difficulty Levels**

To stress different aspects of reasoning, the simulator offers three presets. Internally, these presets adjust event mix, narrative variety, substitution/VAR rates, maximum passes before a shot, optional assist-wording ambiguity, and the target number of events (target game length). The wording does not change the underlying statistics, but it **does** change how hard the text is to interpret.

* **Basic** — Shorter games, simpler phrasing, few substitutions, no VAR.  
  **Why:** Isolates baseline skills: mapping narrative to schema, simple aggregation, and exact JSON formatting—without long-context drift or retroactive edits.
* **Medium** — Longer narratives, moderate substitutions, few VAR, richer wording.  
  **Why:** Adds paraphrase diversity and modest roster churn (more **participants**), plus occasional retroactive corrections, to test robustness beyond the baseline while keeping complexity manageable.
* **Hard** — Many events, higher substitution/VAR rates, broader lexicon, and more adversarial wording (e.g., ambiguous pass verbs); often tighter possession windows (fewer allowed regular passes) to force quicker plays that effects the stats.  
  **Why:** Stresses long-context memory, entity consistency across many updates, and disambiguation under noisy wording—exactly where structured reasoning and strict schema adherence tend to break.

**Evaluation**

The evaluation compares the model’s structured box-score to the known ground truth, using two complementary views:

• Per-field accuracy: counts how many individual fields match exactly.

• Block-normalized accuracy: evaluates correctness within logical blocks (final score, team stats, player stats) so that each block contributes proportionally.

Both views are computed from the same underlying comparison pass. Results are recorded per game and summarized per difficulty, with transparent formulas so the accuracy can be audited rather than treated as a black box.

**Main results on the most powerful LLM from Google (without "tools"):**

**…**

**Limitations & Future Directions**

The setup focuses on single-game narratives and a fixed schema; it does not evaluate broader tasks such as multi-game aggregation, injury/time-on-court modeling, or retrieval from external databases. Future work may include expanding the schema, tightening validation rules, and exploring methods (e.g., tool use or retrieval) that may help models maintain consistency over longer narratives and more complex scenarios, and check the detailed results.

**Conclusion**

By pairing synthetic, auditably generated game narratives with strict, transparent evaluation criteria, this project offers a clean way to study structured reasoning in LLMs within a realistic sports analytics setting.

**5. Analysis & Insights**

Initial observations indicate that LLMs handle simple, short examples relatively well, often producing correct aggregation. However, on longer and more complex examples, models frequently struggle with 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.

**Appendices**

For other models we have 15 examples to compare in the following table:  
methods: 1 = field, method 2 = fractional\_per\_block  
metrices: Average = A, Median = M

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **basic** | | | | **medium** | | | | **hard** | | | | **General** | | | |
| **Method 1** | | **Method 2** | | **Method 1** | | **Method 2** | | **Method 1** | | **Method 2** | | **Method 1** | | **Method 2** | |
| **A** | **M** | **A** | **M** | **A** | **M** | **A** | **M** | **A** | **M** | **A** | **M** | **A** | **M** | **A** | **M** |
| **Gemini** | | | | | | | | | | | | | | | | |
| **gemini/gemini-1.5-flash** | **64.6** | **64.1** | **40.3** | **39.5** | **38.5** | **38.3** | **18.5** | **18** | **23.6** | **23** | **11** | **10.4** | **42.2** | **38.3** | **23.3** | **18** |
| **gemini/gemini-1.5-pro** | **71.1** | **68.9** | **43.8** | **44.4** | **48.6** | **46.5** | **23.3** | **22.9** | **32.0** | **32.4** | **14.6** | **14.8** | **50.6** | **46.5** | **27.2** | **22.9** |
| **gemini/gemini-2.5-flash** | **91.8** | **93.3** | **84** | **90.4** | **65.4** | **64.8** | **47.2** | **51.9** | **41.1** | **41.1** | **20.3** | **20.3** | **72.3** | **72.9** | **58** | **58.2** |
| **gemini/gemini-2.5-pro** | **94.4** | **96.9** | **87.6** | **95.3** | **80.5** | **81.9** | **48.8** | **41.9** | **59.3** | **59** | **27** | **26.8** | **78.1** | **81.9** | **54.4** | **41.9** |
| **OpenAI** | | | | | | | | | | | | | | | | |
| **gpt-4o-mini** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **gpt-4o** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **gpt-5-mini** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **gpt-5** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Anthropic** | | | | | | | | | | | | | | | | |
| **claude-sonnet -4-20250514** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **claude-opus -4-20250514** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **claude-opus -4-1-20250805** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Grok** | | | | | | | | | | | | | | | | |
| **xai/grok-3-mini** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **xai/grok-3** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **xai/grok-4-0709** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **xai/grok-code-fast-1** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

בתיקיית נספחים תוכלו למצוא גם ...