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

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

1. Executive Summary

This project evaluates LLMs on converting a chronological play-by-play into a complete **JSON box score**. The task stresses **temporal reasoning**, **event aggregation**, and **strict schema compliance**. We generate synthetic games, prompt the model, repair outputs to the target schema when needed, and score predictions with **two transparent accuracy metrics**.

2. Problem Statement & Motivation

Sports analytics depends on structured data (play-by-play & box scores). Automating the transformation **from narrative play-by-play text to a structured, schema-valid JSON box score** enables faster analysis and downstream tools. The challenge is twofold: the output must (a) **faithfully** reflect the events and totals and (b) **strictly** follow the JSON schema. This project provides a focused **benchmark** for that ability.

3. Task Definition (What the LLM Must Do)

Given a chronological play-by-play of a single basketball game plus team/roster context, the LLM must produce a complete box-score: final score, per-team totals, and per-player stats. We use prompt engineering to ensure that the LLM knows: its task, its output follows a fixed schema, is internally consistent, and needs to reflect what actually happened in the narrative.

4. Data & Simulation

Each game is synthetically simulated and yields two paired artifacts:
(1) a natural-language play-by-play with team metadata (rosters, starting lineup, participants), and (2) a ground-truth box score (team + per-player).
The simulator enforces invariants (e.g., made ≤ attempts for all shot types; team/player totals stay consistent) and models realistic events: substitutions, rebounds, fouls, turnovers/steals, and occasional VAR adjustments (overturn/downgrade recent plays).

Synthetic pairing preserves a clean **source-to-target** mapping for **auditability** and lets us control coverage and difficulty. Substitutions stress entity tracking as participants change on the fly; rebounds and turnovers test possession flow; and VAR introduces limited retroactive edits that must reconcile earlier and later events without breaking invariants.

5. Difficulty Levels

The simulator provides three presets that tune: **target_events** (target game length), **EVENT_WEIGHTS** (weight for prob to event), **narrative variety**, **substitution rate**, **VAR rate**, **max passes** before statistical event, and optional **adversarial assist wording**. Wording changes do **not** alter the underlying stats; they affect how hard the text is to interpret.

- **Basic** - Short logs (150), simple phrasing, few substitutions (5%), no VAR, max passes = 5, no adversarial wording. Why: Establish a baseline: map narrative to schema, aggregate simply, without long-context drift or retroactive edits.
- **Medium** - Mid-length logs (600), moderate substitutions (10%), occasional VAR (5%), richer paraphrasing, max passes = 3, adversarial wording ON. Why: Test robustness to paraphrase, more participants to track, and limited retroactive edits, while keeping complexity manageable.

- **Hard** - Longest logs (900), higher substitution (15%) & VAR (10%), full lexical variety, ambiguous pass verbs and shot descriptions, max passes = 1, and an event mix that increases scoring and opposite order of "passer"-"shooter" in 2pt & 3pt made. Why: Stress long-context memory, consistent entity tracking across many updates, and disambiguation under noisy wording, where structured reasoning typically breaks.
- **Note:** See **Appendices, section 2**, for more details about parameter value selection.

6. Evaluation

We compare each model's reconstructed **JSON box score** to the simulator's **ground truth**, then report **two score methods** from the same list of per-check contributions:

Method 1 — field: Strict per-field counting. **Accuracy** = correct_fields / total_fields × 100. **Checks include:** the **final score**, every **team-stat key**, and every **participant × stat**. **Roster members not in participants** must pass an **all-zeros** check (a zeroed stat map).

Method 2 — fractional_per_block: Block-normalized weighting so each logical block stats contributes ≈ **1.0**: (i) final score, (ii) teamA, (iii) teamB, (iv) playersA, (v) playersB. Within each block, every check has weight 1 / (#checks in block). Finally, do "**weighted fraction × 100**".

Missing-block policy: If a whole team block is missing in the model's JSON, **all** team+player checks for that team count as **incorrect** (denominator preserved) in both methods. Non-participants must be **explicitly all-zero**; missing/partial maps fail that check.

Artifacts & transparency: For every game we save: the **raw model text**, the **repaired JSON** (exact schema), and an optional **details** file with all contributions and block summaries. Per difficulty, "summary.json" aggregates **average/median** accuracy and includes the **formula** and **formula_vars**, so results are **auditable** rather than a black box.

Results of "Gemini-2.5-pro" on 150 examples (50 on each difficulty level, that in "data"**):**
(See the folder named "data-details" in folder "Appendices" for more details)
Methods: 1 = "field", 2 = "fractional_per_block**" |** Metrics: A = Average, M = Median

basic				medium				hard				Total			
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
95.5	95.7	94.4	94.9	80.3	79.8	51	44.3	56.1	55.8	27	25.9	77.3	79.8	57.5	45.5

8. Analysis & Insights

LLMs do well on short, simple games, but as logs grow longer and more complex, we see misattributed team/player stats, such that the hardest cases expose long-context limits. Details, per game, are saved for exploration and analysis of misattributes.

9. Limitations

The setup focuses on single-game narratives and a fixed schema; it does not evaluate multi-game aggregation, injury/time-on-court modeling, or retrieval from external databases. These choices keep the task well-controlled but limit generality to broader analytics workflows.

10. Conclusion & Future Work

We built a controlled dataset and evaluation framework for converting play-by-play text into schema-valid JSON with transparent scoring. LLMs handle basic games easily, but struggle on long, statistically loaded, with high lexical variety and retroactive edits (because of VAR). **Future work:** (1) Fine-tune on structured basketball data to better reflect realistic play and statistics. (2) Check for results of the LLM while integrating retrieval or tool-based reasoning.

Appendices

1. Performance Comparison

We compare multiple LLMs across providers on the same 15-example set per difficulty (Basic/Medium/Hard).
Methods: 1 = "field" (strict per-field match), 2 = "fractional_per_block" (block-normalized: final score, team A/B, players A/B).
Metrics: A = Average, M = Median.

1.1. Table for Performance Comparison

- Rows are models within a provider;
  moving down generally means a larger/stronger model.
- Columns move from Basic → Medium → Hard (left to right),
  and from Method 1 → Method 2 (within each difficulty).
- **OOC (Out of credits):** We exhausted Anthropic API credits mid-run.
  Having already purchased $5 + 10$ top-ups for earlier experiments, we chose not to buy additional credits; therefore, Anthropic results in this comparison are incomplete.
  This is a budget constraint, not a model issue.
  In separate preliminary tests, all Anthropic models listed in this table gave valid responses, but those results are not included here due to the credit limit.

	basic				medium				hard				Total			
	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 (more details in "Gemini - 15 examples" folder, in " Performance Comparison" folder, in "Appendices" folder)																
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 (more details in "OpenAI - 15 examples" folder, in " Performance Comparison" folder, in "Appendices"																
gpt-4o-mini	55.5	53.3	31.9	31.7	29.9	29.9	15.3	15.5	19.3	19.6	9.6	10	34.9	29.9	18.9	15.5
gpt-4o	71.8	72.2	55	59.1	41.9	43.5	20.5	22.3	26.5	26.6	13.2	13.4	46.7	43.5	29.6	22.3
Anthropic (more details in "Anthropic - 15 examples" folder, in " Performance Comparison" folder, in "Appendices"																
claude-sonnet-4-20250514	80.8	80	62.2	62.4	55.7	54.6	28.3	28.7	37.7	37.8	16.3	16.4	58	54.6	35.6	28.7
claude-opus-4-20250514	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
claude-opus-4-1-20250805	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

1.2. Key takeaways

1. **Difficulty gradient:**
   performance drops consistently from **Basic** → **Medium** → **Hard**. The Hard set stresses long context, more players, more retroactive edits (VAR), and more adversarial wording (e.g., ambiguous pass verbs), which increase block-level penalties.

2. **Metric effect:**
   Method 2 is systematically lower than Method 1 because it averages **five blocks**:
   (i) final score, (ii) team A, (iii) team B, (iv) players A, (v) players B,
   So **each block carries ~20%** of the grade.
   Because of that, a single error in the **final score** drags down ~20% of Method-2 accuracy, whereas in Method 1, it will only have a small weight, which is equal to other smaller mistakes.
   The same applies similarly to the **team blocks**: each mistake results in 1/15%,
   And if, for example, three stats are wrong, you lose another 1/3% at once.

   Also, mistakes in these blocks are more likely to occur because:

   - **Aggregation pressure:** Final score and team totals depend on aggregating many events; one misread (e.g., 3→2 after VAR) propagates to multiple fields and the final.

   - **Retroactive edits (VAR):** Late corrections force the model to revise earlier assumptions; if it doesn't, final and team totals drift.

   - **Roster churn (participants/substitutions):** Who's on court changes; if players aren't tracked correctly, team blocks and player blocks desynchronize (missing players, all-zeros, or double-counting).

   - **Invariants and dependencies:** Constraints like *attempts* ≥ *made* and "team totals = sum of players" are easy to violate under long context and paraphrased wording.

   - **Partial/invalid outputs:** Truncated JSON or omitted sections count as a near-zero for that entire block in Method 2, but only penalize a subset of fields in Method 1.

   - Using ambiguous verbs and a variety of templates can lead to incorrect attribution, which directly affects team and player blocks.

3. **Scaling helps:**
   within each provider, upgrading to a more capable model almost always improves scores.

4. **Medians vs Averages:**
   If **M >> A**, the model is competent but **unstable**;
   if **A ≈ M**, performance is **steady** across the 15 games.

1.3. Bottom line

Upgrading to stronger models improves both **accuracy** and **stability**, but **Hard** remains challenging due to long context, paraphrase variety, substitutions, and VAR

Method 2 (block-normalized) highlights these weaknesses more clearly.
It emphasizes complete, end-to-end accuracy rather than just field-level matches.

2. More Detailed Parameter Value Selection

2.1. Parameter Value Selection Table

Parameter	Basic	Medium	Hard	Short 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 (don't effect the stats), 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 chance	Disabled	5%	10%	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 & opposite order of "passer", "shooter" in 2pt, 3pt made shots	Weighted distribution of event types shapes the overall game difficulty and ambiguity. More "important" statistical events and the use of opposite order of players in the event of made shots may cause the LLM to get confused between the roles of the player who made the assist and the player who scores.

Event Type	**basic - weights**	**Medium - weights**	**Hard - weights**
turnover_by_bad_pass	4	3	2
steal	5	5	5
timeout	4	3	2
assist_and_score_2pt	3	7	9
assist_and_score_2pt_opposite	0	5	7
assist_and_score_3pt	3	7	9
assist_and_score_3pt_opposite	0	5	7
miss_2pt_from_pass	10	8	6
block_on_2pt_shot	8	6	4
shooting_foul_2pt	8	5	3
miss_3pt_from_pass	9	7	5
block_on_3pt_shot	8	6	4
shooting_foul_3pt	7	5	3

2.2. Detailed Explanation of Parameters

target_events:
This parameter determines the total number of simulated play-by-play events per game. A lower value (150 in Basic) yields short, simple logs, while higher values (600 in Medium, 900 in Hard) produce longer narratives. From the course, longer sequences intensify long-range dependency issues in recurrent models (vanishing gradients). Transformers mitigate some of this via self-attention.

difficulty_max_passes:
This parameter sets the maximum number of passes allowed before a shot. Basic games allow up to 5 passes, Medium 3, and Hard only 1. Fewer passes increase ambiguity by forcing quicker offensive plays. Theoretically, this connects to sparse context windows in early n-gram models, where limited history reduces predictive certainty and forces reliance on distributional generalization.

adversarial_assist_bias:
When enabled (Medium, Hard), neutral pass verbs are replaced with ambiguous alternatives such as 'feeds' or 'delivers'. This complicates assist recognition. In NLP terms, this resembles lexical variation in distributional semantics: word embeddings (e.g., Word2Vec, GloVe) place synonyms in nearby regions of the vector space, but rare or misleading lexical choices may increase confusion. Thus, the bias directly stresses a model's robustness to paraphrasing and lexical variability.

substitution chance:
Determines how often player substitutions happen in the simulation (5% in Basic, 10% in Medium, 15% in Hard). Each substitution changes who is on the court, so the model must keep an up-to-date list of active players and attribute every event to someone who is actually playing. Higher substitution rates mean more lineup switches, more bookkeeping, and more chances to mix players up or credit stats to someone who already checked out. From the course perspective, this stresses long-context tracking: maintaining consistent assignments over many steps is harder as the sequence grows.

VAR events:
Retroactive changes (e.g., canceling baskets, downgrading shots) are disabled in Basic, appear at 5% in Medium, and 10% in Hard. These events force the model to revise prior states, akin to structured prediction with retroactive constraints. From the course, this relates to limitations of left-to-right generative models, which cannot easily revise earlier predictions, highlighting the need for architectures with bidirectional context (e.g., BERT) or explicit constraint handling.

narrative variety:
We control how many phrasing templates the generator uses: one-quarter in Basic, half in Medium, and all templates in Hard. Higher variety increases linguistic diversity and tests generalization beyond surface patterns. The motivation follows the lectures on distributional representations and dense word embeddings (e.g., Word2Vec/GloVe), which capture similarity from context, and on subword tokenization, which reduces OOV sensitivity.

EVENT_WEIGHTS:
Defines the probability distribution over event types.
Basic favors misses/fouls → lower scoring and fewer state changes.
Medium is balanced. Hard increases made shots and opposite order of "passer"-"shooter" in 2pt & 3pt made shots, leading to more scoring swings and more frequent state updates.
Tuning these base rates controls difficulty: for example, more scoring events → more updates to track about score/assists/made shots → higher risk of cumulative attribution errors.

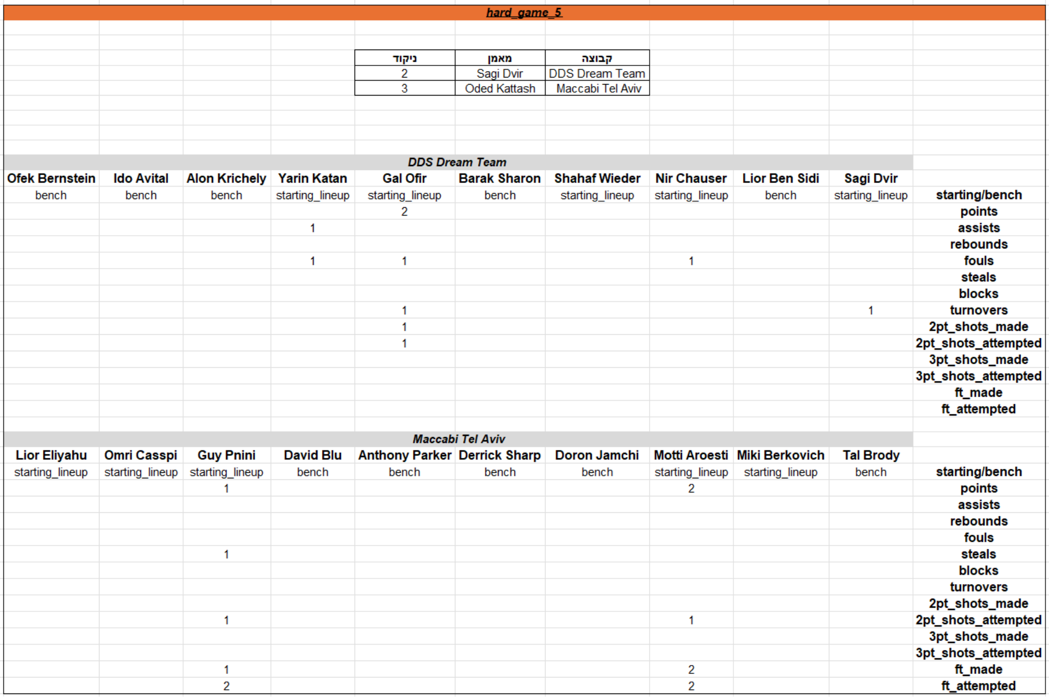
3. Human baseline (3 games)

Before evaluating LLMs, we tested whether a human can consistently reconstruct a box score from a play-by-play log.
We asked a friend to analyze three short games using our schema, updating only:

1. The final team score.
2. Per-player stats.

We then compared their reports to the simulator's ground truth.

3.1. Example — hard_game_5: 100% accuracy (no discrepancies):



hard_game_5

ניקוד	מאמן	קבוצה
2	Sagi Dvir	DDS Dream Team
3	Oded Kattash	Maccabi Tel Aviv

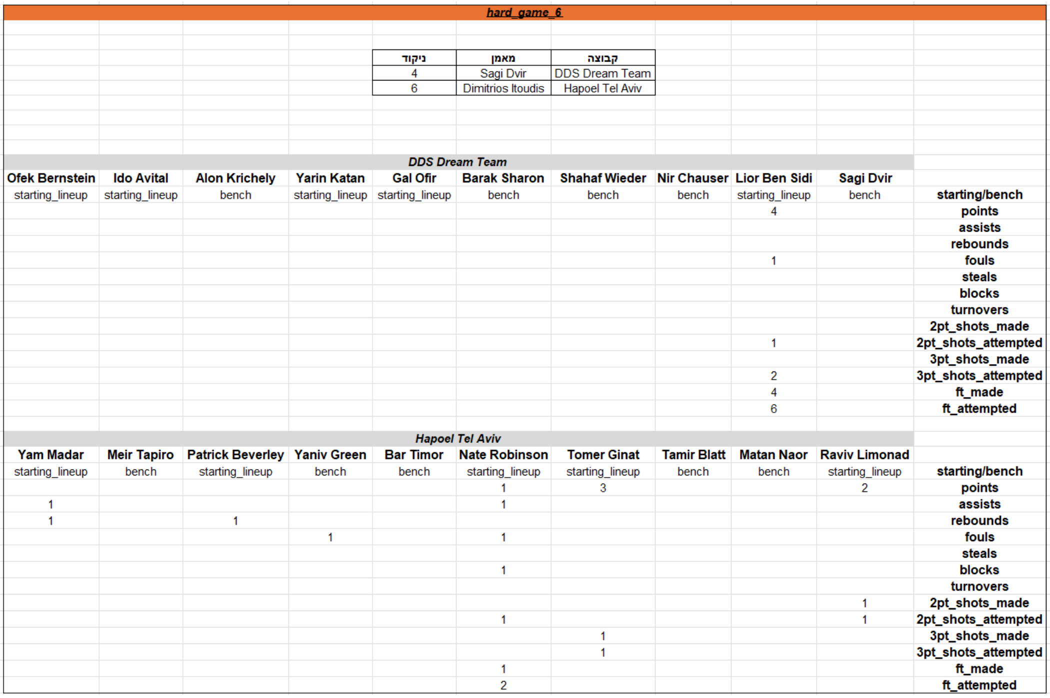
DDS Dream Team

Ofek Bernstein	Ido Avital	Alon Krichely	Yarin Katan	Gal Ofir	Barak Sharon	Shahaf Wieder	Nir Chauser	Lior Ben Sidi	Sagi Dvir	
bench	bench	bench	starting_lineup	starting_lineup	bench	starting_lineup	starting_lineup	bench	starting_lineup	starting/bench
				2						points
			1							assists
										rebounds
			1	1			1			fouls
										steals
										blocks
				1					1	turnovers
				1						2pt_shots_made
				1						2pt_shots_attempted
										3pt_shots_made
										3pt_shots_attempted
										ft_made
										ft_attempted

Maccabi Tel Aviv

Lior Eliyahu	Omri Casspi	Guy Pnini	David Blu	Anthony Parker	Derrick Sharp	Doron Jamchi	Motti Aroesti	Miki Berkovich	Tal Brody	
starting_lineup	starting_lineup	starting_lineup	bench	bench	bench	bench	starting_lineup	starting_lineup	bench	starting/bench
		1					2			points
										assists
										rebounds
										fouls
		1								steals
										blocks
										turnovers
										2pt_shots_made
		1					1			2pt_shots_attempted
										3pt_shots_made
										3pt_shots_attempted
		1					2			ft_made
		2					2			ft_attempted

3.2. Example — hard_game_6: 100% accuracy (no discrepancies):



hard_game_6

ניקוד	מאמן	קבוצה
4	Sagi Dvir	DDS Dream Team
6	Dimitrios Itoudis	Hapoel Tel Aviv

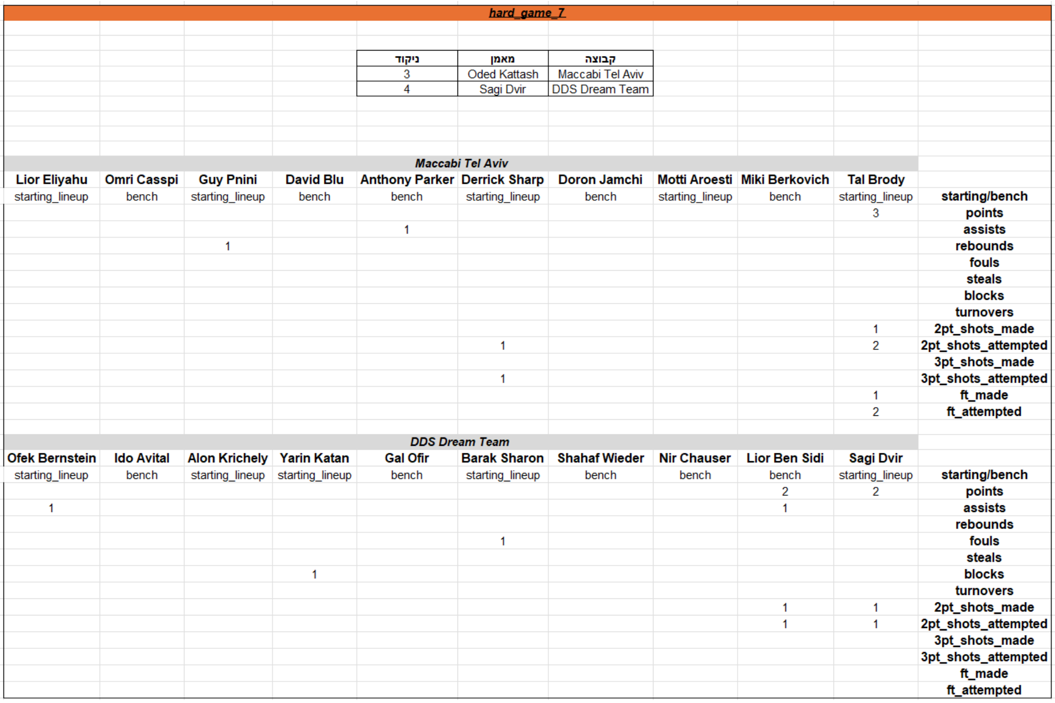
DDS Dream Team

Ofek Bernstein	Ido Avital	Alon Krichely	Yarin Katan	Gal Ofir	Barak Sharon	Shahaf Wieder	Nir Chauser	Lior Ben Sidi	Sagi Dvir	
starting_lineup	starting_lineup	bench	starting_lineup	starting_lineup	bench	bench	bench	starting_lineup	bench	starting/bench
								4		points
										assists
										rebounds
								1		fouls
										steals
										blocks
										turnovers
										2pt_shots_made
								1		2pt_shots_attempted
										3pt_shots_made
								2		3pt_shots_attempted
								4		ft_made
								6		ft_attempted

Hapoel Tel Aviv

Yam Madar	Meir Tapiro	Patrick Beverley	Yaniv Green	Bar Timor	Nate Robinson	Tomer Ginat	Tamir Blatt	Matan Naor	Raviv Limonad	
starting_lineup	bench	starting_lineup	bench	bench	starting_lineup	starting_lineup	bench	bench	starting_lineup	starting/bench
					1	3			2	points
1					1					assists
1		1								rebounds
			1		1					fouls
										steals
					1					blocks
										turnovers
									1	2pt_shots_made
					1				1	2pt_shots_attempted
						1				3pt_shots_made
						1				3pt_shots_attempted
					1					ft_made
					2					ft_attempted

3.3. Example — hard_game_7: 100% accuracy (no discrepancies):



hard_game_7

ניקוד	מאמן	קבוצה
3	Oded Kattash	Maccabi Tel Aviv
4	Sagi Dvir	DDS Dream Team

Maccabi Tel Aviv

Lior Eliyahu	Omri Casspi	Guy Pnini	David Blu	Anthony Parker	Derrick Sharp	Doron Jamchi	Motti Aroesti	Miki Berkovich	Tal Brody	
starting_lineup	bench	starting_lineup	bench	bench	starting_lineup	bench	starting_lineup	bench	starting_lineup	**starting/bench**
									3	**points**
				1						**assists**
		1								**rebounds**
										fouls
										steals
										blocks
										turnovers
									1	**2pt_shots_made**
					1				2	**2pt_shots_attempted**
										3pt_shots_made
					1					**3pt_shots_attempted**
									1	**ft_made**
									2	**ft_attempted**

DDS Dream Team

Ofek Bernstein	Ido Avital	Alon Krichely	Yarin Katan	Gal Ofir	Barak Sharon	Shahaf Wieder	Nir Chauser	Lior Ben Sidi	Sagi Dvir	
starting_lineup	bench	starting_lineup	starting_lineup	bench	starting_lineup	bench	bench	bench	starting_lineup	**starting/bench**
								2	2	**points**
1								1		**assists**
										rebounds
					1					**fouls**
										steals
			1							**blocks**
										turnovers
								1	1	**2pt_shots_made**
								1	1	**2pt_shots_attempted**
										3pt_shots_made
										3pt_shots_attempted
										ft_made
										ft_attempted

3.4. Conclusion:

Across all three games, the human annotator achieved **100% accuracy**, confirming the task is **human-solvable and well-defined** under our schema and providing a clear **upper-bound reference** for model evaluation.