Input Matters: Evaluating Input Structure's Impact on LLM Summaries of Sports Play-by-Play

Barkavi Sundararajan and Somayajulu Sripada and Ehud Reiter
Department of Computing Science, University of Aberdeen
{b.sundararajan.21, yaji.sripada, e.reiter}@abdn.ac.uk

Abstract

A major concern when deploying LLMs in accuracy-critical domains such as sports reporting is that the generated text may not faithfully reflect the input data. We quantify how input structure affects hallucinations and other factual errors in LLM-generated summaries of NBA play-by-play data, across three formats: row-structured, JSON and unstructured. We manually annotated 3,312 factual errors across 180 game summaries produced by two models, Llama-3.1-70B and Qwen2.5-72B. Input structure has a strong effect: JSON input reduces error rates by 69% for Llama and 65% for Qwen compared to unstructured input, while row-structured input reduces errors by 54% for Llama and 51% for Qwen. A two-way repeated-measures ANOVA shows that input structure accounts for over 80% of the variance in error rates, with Tukey HSD post hoc tests confirming statistically significant differences between all input formats.

1 Introduction

Despite recent advancements, large language models (LLMs) are known to produce factually incorrect output (Jacovi et al., 2025). These errors occur frequently enough to make LLM-based NLG applications unacceptable in accuracy-critical domains, unless deployed in a human-in-the-loop setting where mistakes can be corrected. Sports reporting is one such domain, and in this paper, we investigate the factual accuracy of LLMs for generating NBA game summaries. This is a challenging text generation task because play-by-play data is highly detailed, encoding *entities* (teams, players), *events/actions* (e.g., jump shot, layup, free throw, offensive rebound), *relationships* (e.g., *player – attempts – shot*; *player – commits – foul*), and *attributes* (e.g., scores, time, period, distance).

We hypothesise that structuring the complex play-by-play data is important for producing factually accurate summaries, whereas unstructured input (*garbage in*) yields more factual errors (*garbage out*). To empirically study the impact of input structuring on model output, we evaluate three input formats: unstructured (natural play-by-play descriptions; see Table 7), row-structured tabular format (see Table 1), and hierarchical JSON. Fig. 1 shows a play-by-play event represented as a JSON object.

Recent work by Sui et al. (2024) shows that different tables define structure and formatting in dis-


```
{
  "time": {
    "period": "OT1",
    "clock": "2:13.0"
  },
  "team": "Los Angeles Lakers",
  "play_details": {
    "description": "LeBron James makes
       3-pt jump shot from 25 ft",
    "event": {
      "type": "Shot",
      "action": "Jump Shot",
      "outcome": "Made",
      "distance": "25 ft",
      "points": 3
    },
    "players": {
      "primary_player": "LeBron James"
    }
  },
  "score": {
    "points_scored": "LAL: 3",
    "cumulative": {
      "LAL": 127,
      "ATL": 125
    }
  }
}
```


Figure 1: Hierarchical JSON representation of a single play-by-play event. *(i) Event:* LeBron James makes a three-point jump shot from 25 ft; *(ii) Time:* period (first overtime) and game clock (i.e., time remaining in that period); *(iii) Score:* points scored from the play and the teams' cumulative scores.

Proceedings of the 18th International Natural Language Generation Conference, pages 795–809
October 29–November 2, 2025.

Time	Period	Team	Primary Player	Secondary Player	Play Description
2:56.0	OT1	Los Angeles Lakers	Anthony Davis	N/A	Anthony Davis makes free throw 1 of 2
2:56.0	OT1	Los Angeles Lakers	Anthony Davis	N/A	Anthony Davis makes free throw 2 of 2
2:45.0	OT1	Atlanta Hawks	De'Andre Hunter	LeBron James	De'Andre Hunter misses 2-pt layup from 10 ft (block by LeBron James)
2:44.0	OT1	Atlanta Hawks	N/A	N/A	Offensive rebound by Team
2:33.0	OT1	Atlanta Hawks	Trae Young	N/A	Trae Young makes 2-pt layup from 2 ft
2:13.0	OT1	Los Angeles Lakers	LeBron James	N/A	LeBron James makes 3-pt jump shot from 25 ft

(a) Temporal information, Team and Player details, and Play description in row-structured input.

Event Type	Action Type	Outcome	Distance (ft)	Current Points Scored	LAL Cum. Score	ATL Cum. Score
Shot	Free Throw	Made	N/A	LAL: 1	LAL: 123	ATL: 123
Shot	Free Throw	Made	N/A	LAL: 1	LAL: 124	ATL: 123
Shot	Layup	Missed	10	No points scored	LAL: 124	ATL: 123
Rebound	Offensive	N/A	N/A	No points scored	LAL: 124	ATL: 123
Shot	Layup	Made	2	ATL: 2	LAL: 124	ATL: 125
Shot	Jump Shot	Made	25	LAL: 3	LAL: 127	ATL: 125

(b) Structured event and score details in row-structured input.

Table 1: Excerpt of row-structured play-by-play input representation (partial game). This is a *single table* split into upper and lower panels for readability. 'N/A' indicates that the attribute does not apply. In the released dataset, such fields are stored as `None`.

tinct ways, which can influence how LLMs interpret and process tabular inputs. While their focus was on QA and relatively simple generation tasks (Parikh et al., 2020), we study a more complex scenario: summarising NBA play-by-play data, which is event-driven, temporally ordered, and densely factual with over 450 rows and 13 columns. This makes it a more challenging and underexplored setting for investigating the impact of input structure on LLM-based summarisation.

To investigate this, we evaluate two open-source LLMs, Llama-3.1-70B (Grattafiori et al., 2024) and Qwen2.5-72B (Yang et al., 2024). We use a manual annotation protocol adapted from Thomson et al. (2023) and Sundararajan et al. (2024) to assess how input structure affects factual errors in generated summaries.

The key findings from our study are as follows:

- A two-way repeated-measures ANOVA shows that structured inputs significantly reduce error rates compared with unstructured text, with JSON providing the strongest improvements ($p < 0.05$).
- Error-type analysis shows that Llama-3.1-70B produces more *Number* errors, whereas Qwen2.5-72B produces more *Name* and *Word-subjective* errors. Across both models, *Word-objective* errors dominate; common patterns are event/action substitutions (e.g., mistakenly produces *a pair of free throws* instead of *a layup*).

2 Related Work

Prior work on data-to-text generation for basketball (Wiseman et al., 2017; Puduppully et al., 2019; Thomson et al., 2020) has used hierarchical encoder models and related architectures to generate summaries from box scores (aggregated player and team statistics), yet consistent factual errors have been reported (Thomson et al., 2023). The advent of long-context LLMs (up to 128k tokens) (Wu et al., 2024) has enabled studies on richer inputs; for example, Hu et al. (2024) use NBA play-by-play data to compute player and team points. Their work focuses on reasoning and aggregation over play-by-play data, supporting the value of using a single rich dataset for controlled analysis. In contrast, we generate paragraph-style summaries from complete play-by-play data and compare three input structures: unstructured input with 13,000 tokens, row-structured input with 28,000 tokens, and JSON input with 70,000 to 80,000 tokens, producing summaries of 450 to 500 words.

Maynez et al. (2020) demonstrated that automatic metrics are insufficient for measuring hallucinations. Building on Thomson et al. (2023) and Sundararajan et al. (2024), we adapted their manual error annotation protocols to study factual errors at a granular level, making minor task-specific adjustments for the NBA play-by-play data to evaluate the errors in our generated summaries.

3 Experimental Setup

3.1 Game Selection Criteria

Fig. 2 shows our stratified random sampling based on the final point difference (*absolute score difference between two teams*). We collected all NBA games from Dec. 2024–Jan. 2025 to limit potential data contamination (Balloccu et al., 2024), partitioned the games into four strata: ($\leq$ 3, 4–9, 10–19, $\geq$ 20 point differences), and randomly sampled within each stratum to obtain a balanced representation of games, yielding 30 games in total.

3.2 Input Representation and Preprocessing

We processed raw NBA play-by-play logs from Basketball Reference[1] to generate three types of input structures: row-structured, JSON, and unstructured. We performed light cleaning and assigned clear column names, then decomposed each play description into atomic fields (e.g., primary/secondary player, team, action and event type, outcome, distance, time remaining, period, score), applying regular expressions and task-specific rules to normalise values. From these atomic fields, we produced a row-structured table with one row per play (table 1). In parallel, we encoded a hierarchical JSON that nests entities (player, action, team) and their attributes under each play Fig. 1.

For the unstructured variant, we preserved the original natural-language descriptions for both teams and applied only lightweight regex-based cleanup (removing extra spaces around player names and actions) without altering the underlying structure. We release the structured play-by-play datasets (row-structured and JSON)[2].

3.3 Model Details

We selected two open-source LLMs released before the 2024–25 NBA season to mitigate data contamination: Llama-3.1-70B (Jul. 2024; Grattafiori et al. 2024) and Qwen2.5-72B (Sep. 2024; Yang et al. 2024). We accessed Llama-3.1-70B via the Together AI API (Together AI, 2025), and Qwen2.5-72B via Alibaba Cloud's AI platform (Alibaba Cloud, 2025) using a registered account. We used identical decoding settings for both models. The model specifications are shown in Table 8.

[1] https://www.basketball-reference.com/boxscores/pbp/202501010DET.html

[2] https://github.com/BarkaviSJ/nba-input-matters

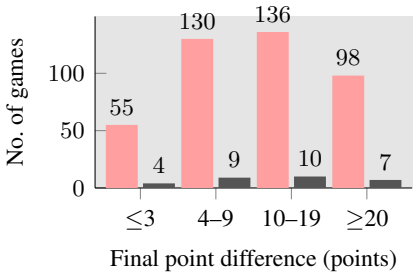

Figure 2: Games by final margin. For each stratum (*absolute score difference:* $\leq$ 3*, 4–9, 10–19,* $\geq$ 20), the left bar shows the total number of NBA games in Dec. 2024 – Jan. 2025, and the right bar shows the stratified sample used in our study. Values above bars are counts.

Prompt. We designed the prompt following best-practice guidance from Google (Google Cloud, 2025) for zero-shot, task-focused generation: a clear role definition, step-by-step instructions, explicit output constraints, and style/tone specifications. We then customised the wording for each input (unstructured, row-structured, JSON) after multiple trials. The final prompt used in our experiments is shown in Fig. 6.

3.4 Annotation Protocol

We evaluate factual and semantic errors using a manual error-annotation protocol rather than automatic metrics such as ROUGE (Lin, 2004) or BLEU (Papineni et al., 2002). We refine the annotation scheme (Thomson et al., 2023) by splitting the original *Word* category into *Word-objective* (verifiable, fact-checkable wording errors) and *Word-subjective* (opinion-based wording not grounded in the play-by-play).

We use seven categories in total: *Number*, *Name*, *Word-objective*, *Word-subjective*, *Context*, *Not Checkable*, and *Other*. These error categories are elaborated in Section 4.2. The manually annotated error counts form the basis for comparing performance across input structures in subsequent statistical analysis.

Notation. For readability in examples and tables, we mark erroneous spans with coloured superscripts indicating the error category.

- **10**$^{\text{U}}$ $\rightarrow$ *6* (U: Number)
- **LeBron James**$^{\text{N}}$ $\rightarrow$ *Anthony Davis* (N: Name)
- **free throw**$^{\text{WO}}$ $\rightarrow$ *layup* (Wo: Word-objective)
- **pivotal moment**$^{\text{WS}}$ (Ws: Word-subjective)

- **Christian Wood**^{C} → *Cam Whitmore* (C: Context)
- **third straight victory**^{X} (X: Not Checkable)
- **other**^{O} (O: Other)

3.5 Hypotheses

We test the following hypotheses about how input structure and model choice influence the number of factual and semantic error rates in LLM-generated summaries:

- **H1 – Input Structure Effect:** Error rates differ significantly across input formats; structured inputs (row-structured and JSON) are expected to produce fewer errors than unstructured input.
- **H2 – Model Effect:** Error rates differ significantly between the two large language models (Llama-3.1-70B and Qwen2.5-72B), with one model expected to perform better overall.
- **H3 – Interaction Effect:** The effect of input structure on error rates depends on the model, suggesting that some models perform better with certain input structures.

To test these hypotheses, we conduct a two-way repeated-measures ANOVA on the total number of annotated errors per 100 words in each summary. Input structure and model are treated as within-subjects factors.

4 Results

We manually annotated 3,312 errors across 180 play-by-play summaries generated for 30 NBA games (2 models × 3 input structures). Llama-3.1-70B averaged 408 words per summary, whereas Qwen2.5-72B averaged 505 words. To compare factual accuracy across both models and all three input structures, we report normalised error rates, i.e., errors per 100 words (see Table 2). The breakdown by error type is shown in Table 3.

Metric	Llama-3.1-70B	Qwen2.5-72B
Total errors	1,575	1,737
Total words	36,709	45,474
Error rate (per 100 words)	**4.29**	**3.82**

Table 2: Total errors and normalised error rates aggregated across all 30 summaries and three input formats. Normalised error rate = (total errors / total words) ×100.

Error type	Llama-3.1-70B	Qwen2.5-72B
Number^{U}	1.82	0.93
Name^{N}	0.66	0.80
Word-objective^{WO}	1.35	1.25
Word-subjective^{WS}	0.31	0.61
Context^{C}	0.10	0.19
Not checkable^{X}	0.05	0.04
Other^{O}	0.00	0.00

Table 3: Normalised error rates by error type and model, aggregated across all summaries and three input formats. Each rate is computed as (errors of that type / total words) ×100. The sum of all error types equals the overall error rate reported in Table 2, up to rounding.

Source	p (GG)	partial η^2
Input Structure	1.79×10^{-27}	0.813
Model	3.60×10^{-4}	0.056
Input Structure × Model	4.80×10^{-4}	0.050

Table 4: Two-way repeated measures ANOVA. Greenhouse–Geisser corrected p-values (all $p < 0.05$; statistically significant). We also report effect size, partial η^2 for main effects and their interaction. Post hoc Tukey HSD results are in Table 9.

4.1 Analysis by Total Number of Errors

4.1.1 Distribution of Errors

Table 5 shows the distribution of total annotated errors across 30 games, grouped by models (Llama-3.1-70B and Qwen2.5-72B[3]) and input structures (Row, JSON and Unstructured). Both models produced fewer errors in structured inputs: JSON input resulted in the lowest errors (2.17 mean errors for Llama and 2.14 for Qwen), while the row structure had 3.20 for Llama and 2.92 for Qwen. Both models consistently recorded more errors with unstructured inputs (7.05 for Llama and 6.04 for Qwen).

Two-way ANOVA Repeated-Measures. We conducted a two-way repeated-measures ANOVA to examine the effects of input structure and model (two factors as independent variables) on error rates. Each of the 30 games was evaluated under all six conditions (two models × three input structures), resulting in a within-subjects design. The dependent variable was the total number of errors normalised per 100 words.

As shown in Table 4, the input structure had a highly significant effect on error rates (p =

[3]Llama denotes Llama-3.1-70B; Qwen denotes Qwen2.5-72B. These model names refer exclusively to these versions throughout the paper.

Model	Input format	Mean	Median	Mode	Std Dev	Range	Min	Max	Skew	Kurt.	95% CI
Llama	Row	3.20	2.99	2.02	0.98	4.15	1.61	5.76	0.63	0.24	[2.84, 3.57]
Llama	JSON	2.17	2.02	1.32	0.64	2.32	1.32	3.64	0.55	-0.64	[1.93, 2.41]
Llama	Unstructured	7.05	6.83	5.32	1.44	6.56	5.32	11.88	1.69	3.55	[6.51, 7.59]
Qwen	Row	2.92	2.74	2.46	0.58	2.09	2.24	4.33	1.12	0.36	[2.71, 3.14]
Qwen	JSON	2.14	2.00	1.06	0.76	3.03	1.06	4.09	0.84	0.22	[1.85, 2.42]
Qwen	Unstructured	6.04	5.79	5.17	0.87	3.38	4.95	8.33	0.95	0.14	[5.71, 6.37]

Table 5: Descriptive statistics of normalised error rates across two models and three input structures. The table reports measures of central tendency (mean, median, mode), variability (standard deviation, range, minimum, maximum), skewness, kurtosis, and 95% confidence intervals.

1.79×10^{-27}) with a large effect size (partial $\eta^2 = 0.813$), indicating that the way information was structured explains the majority of variance in summary errors. The model main effect was also significant ($p = 3.60 \times 10^{-4}$). The interaction between input structure and model was significant ($p = 4.80 \times 10^{-4}$). These findings highlight the dominant influence of input structure on error rates and motivate post hoc analyses to compare model–input structure combinations.

4.1.2 Post hoc tests

We use Tukey's Honestly Significant Difference (HSD) test to compare the six model-input combinations as distinct conditions, since each produces a unique summary. Based on the results shown in Table 9, we assess each hypothesis defined in section 3.5.

H1: Effect of Input Structure. In Tukey's HSD (Table 9), the input structure significantly affects error rates across all pairwise comparisons (all $p < 0.05$): JSON vs row (mean difference = 0.91), JSON vs unstructured (mean difference = 4.39), and row vs unstructured (mean difference = 3.48). We therefore conclude that input structure significantly influences error rates.

H2: Model Effect. Given the two-way ANOVA showed a significant model and input structure interaction in (Table 4), we assess model differences within each input. Llama vs. Qwen differences are not significant after Tukey correction for JSON ($p = 1.00$) or row ($p = 0.842$), but there is a significant Llama vs. Qwen difference for unstructured (p = 0.0005). Thus, we do not claim a uniform model advantage; model differences appear specifically in the unstructured input. The main effect of model should therefore be interpreted with caution.

H3: Effect of Input Structure and Model Interaction. Consistent with the significant interaction found in the ANOVA, within-model contrasts show a monotonic increase across input structures from JSON to row to unstructured.

- **JSON vs. Unstructured (within model):** As shown in table 9, Llama's errors increase by 4.88 per 100 words when input changes from JSON to unstructured ($p < 0.05$). Qwen shows a 3.90 increase for the same shift ($p < 0.05$).
- **JSON vs. Row and Row vs. Unstructured (within model):** Errors increase when moving from row to unstructured ($p < 0.05$). From JSON to row, Llama shows a significant increase by 1.03 ($p < 0.05$), and Qwen shows a significant increase by 0.78 ($p < 0.05$).
- **Same input across models (between models):** Some cross-model comparisons within the same input are not significant (e.g., Llama's JSON vs. Qwen's JSON: $p = 1.00$; Llama's row vs Qwen's row $p = 0.842$), while unstructured differs $p = 0.0005$.

The results demonstrate that input structure significantly affects performance. Errors for both models increase from JSON to row to unstructured, but by different magnitudes; therefore, any model advantage is conditional on input (negligible for JSON and row, present for unstructured).

4.2 Analysis by Error Category

As mentioned in Table 3, we compute the error rate by normalising errors per hundred words for each category. Fig. 3 presents a comparison of error rates for each category across three inputs. The bar chart also provides an overview of errors made by two models. Out of the seven error categories, we can observe that number, name, and word objective errors are predominant in all combinations. While

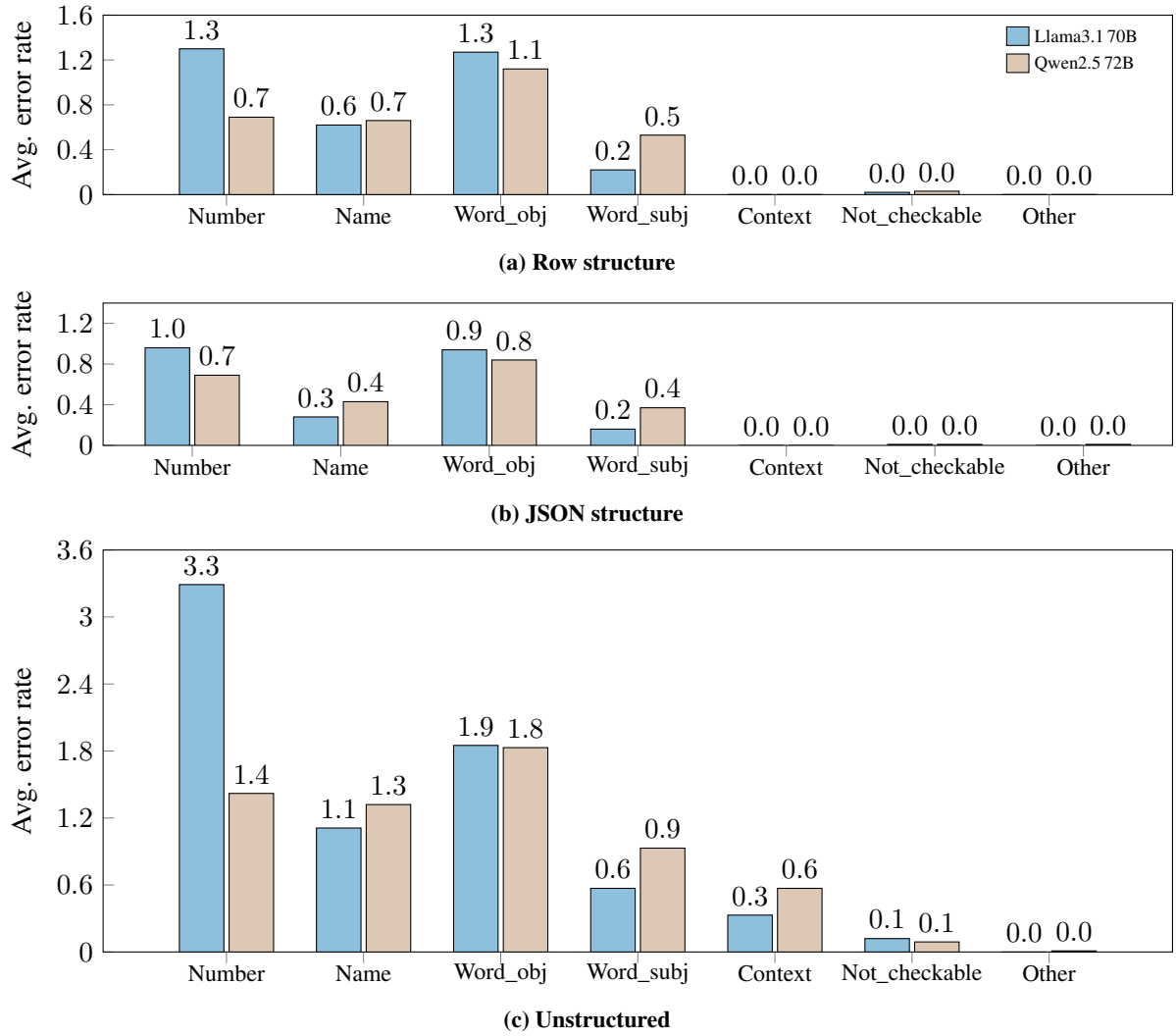

Figure 3: Average error rates (errors per 100 words) by error type and input structure. Each panel reports *average error rates across the 30 summaries* for that input format (Row, JSON, Unstructured), comparing Llama-3.1-70B and Qwen2.5-72B across seven error categories. Structured inputs (Row and JSON) yield significantly fewer total errors than unstructured inputs, with JSON providing the largest reduction.

word subjective and context are specific to certain models and input structures.

In addition to the main hypotheses we tested in section 4.1, we conduct a two-way ANOVA with repeated measures for the five main error categories to understand the significance between these specific error types, models, and input structures.

Number errors. Numbers encode key information in basketball play-by-play data: it has point increments, cumulative scores, shot counts (two-pointer, three-pointer), shot distances (22ft, 2 ft), and period details in ordinals (first, second, third, and fourth). As shown in Fig. 3, both models exhibit their highest number-error rates on unstructured input, and lower rates on row and JSON. On the structured inputs, Llama commits higher number errors than Qwen. As reported in Table 10, a two-way repeated-measures ANOVA on number-error rates also shows significant model, input structure and model×input interaction (all $p < 0.05$)

Name errors. Name errors occur when a model misnames players or teams in the summary. For example, swapping the winning team's name or attributing an action to the wrong player. Qwen 2.5 72B makes more of these mistakes than Llama 3.1 70B even on JSON input (see Fig. 3b). In Table 11, a two-way repeated-measures ANOVA shows significant main effects of model and input (both $p < 0.05$), with no significant model×input interaction.

Word-objective errors. Word-objective errors are predominant in both models across all input

structures. These errors occur when the model generates an incorrect description of actions in the game, e.g., '*calling a turnover a rebound*' or claim a player '*orchestrated a fast-break opportunity that led to a layup*' even when the team never lost possession. They also commit verb errors, such as '*outscore*', '*tied the game*' or '*cut the deficit*' when the plays show no such changes. These mistakes highlight how even small wording errors can misrepresent the game moments. A two-way repeated-measures in Table 12 shows a significant input effect ($p < 0.05$), with non-significant model and model×input interaction.

Word-subjective errors. Word-subjective errors occur when a model adds opinion-based wording to an objective summary. Examples include calling a contest a '*classic match*,' praising a team for showing 'flashes of brilliance' or 'relentless efforts', describing a play as a 'pivotal moment' or mentioning it a 'memorable game in this season.' These phrases reflect personal judgement and depend on the annotator's interpretation rather than verifiable facts. They fall outside the factual error category, but we captured this error type to ensure that any subjective or opinion-based wording is flagged. Qwen generated more word subjective errors than Llama across all inputs (see Fig. 3). Table 13 shows significant main effects of input and model ($p < 0.05$), with no significant model×input interaction.

Context errors. Context errors occur when a summary leads readers to misinterpret the game, such as describing actions by players who did not participate in that period or game. We observed this error only in Qwen on unstructured inputs (Fig. 3 and Fig. 5), where the model hallucinated player names when the input lacked complete and structured data (see Table 14).

Not checkable and Other errors. Not checkable errors occur when a summary states facts that cannot be verified from the game's per-period, box-score, or play-by-play data. Even if the information lies outside these sources (for example, 'third straight victory'), we still flag it as *Not checkable*. 'Other' error type cover any mistakes that do not fit in the previous categories. Both of these errors occurred infrequently and showed no significant variation across input structures or models, so we did not perform further in-depth analysis on them.

5 Discussion

Fig. 5 presents partial-game summaries with error annotations to demonstrate a few examples of factual errors produced by our models for three different inputs (taken from the first overtime of the game shown in Table 1 and Table 7).

5.1 Input structure

We observe multiple factual errors in the partial game summary based on unstructured input (see Section 4.2). In this example, the Llama model struggles to produce a factual summary: it describes incorrect actions such as 'a pair of three-pointers' and 'a dunk and a layup,' uses wrong player names, and reports incorrect scores. Number errors are predominant; for instance, when the model generates 'tie the game at 130–129,' it fails to reason correctly, as the actual score was 123.

Fig. 5 shows that Qwen on *unstructured input* struggles to generate the correct player names for the teams in the game. For example, it outputs Frank Ntilikina and Christian Wood, who were not on the Cavaliers or the Rockets, which constitutes a context error. One cause of these name and context errors is that the original play-by-play data (Section 3.2) represents players with initials and surnames (e.g.,'L. James' for LeBron James), which can lead the model to produce incorrect player names. More importantly, the unstructured data lacks atomic values: specific events and actions are not separated into meaningful columns or key-value pairs (see Table 7).

The summaries from *row-structured inputs* capture play-by-play data and events by clearly segregating entities (player names, team names), and action types (jump shots, layups, rebounds, distance shots, incremental points, cumulative scores) into meaningful column names with atomic values. This structure we hypothesise helps to preserve the information and reduce the factual errors to some extent. For example, the row-structured partial summary in Section 4.2 correctly identifies several events (free throws, jump shot, three-pointer, tying the game) but still makes two word-objective errors, describing Young's jump shot as a layup. Due to the challenging nature of over 450 plays in a game, even the hierarchical JSON struggles to render the right order of actions in some examples.

The summaries generated from *hierarchical JSON* data (see Fig. 1) proved more factually accurate than the row-structure input. By encoding

Partial-game summary from Llama-Unstructured input

The Lakers fought back, however, with James and Davis leading the charge. James hit a pair of three-pointers[WO], and Davis added a dunk and a layup[WO] to tie the game at 130[U]-129[U].
Throughout the game, James and Davis were the Lakers' top performers, combining for 63[U] points. Young and Bogdanovic[N] led the Hawks, combining for 61[U]points.
List of errors.

- a pair of three-pointers[WO]: incorrect word objective phrases from events, correct event/phrase is 'a three-pointer'.
- a dunk and a layup[WO]: incorrect events (word objective), the player made a 'free-throw'.
- 130[U]: incorrect score (number), the correct number/score is 123.
- 129[U]: incorrect score (number), the correct number/score is 123.
- 63[U]: incorrect score (number), the correct number/score is 77.
- Bogdanovic[N]: incorrect player name, the correct player name is Hunter.
- 61[U]: incorrect score (number), the correct number/score is 757.

Partial-game summary from Qwen-Row input

The game saw several lead changes, with Anthony Davis making a pair of free throws to tie the game at 123. LeBron James then hit a three-pointer to give the Lakers a 127-125 lead, but Trae Young answered with a layup[WO] to tie it again.
The Lakers regained the lead with a three-pointer from Gabe Vincent, but the Hawks stayed close with a jump shot from Dyson Daniels. LeBron James finished with a double[WO]-double, contributing significantly on both ends of the court.
List of errors.

- with a layup[WO]: incorrect events (word objective), the player made 'a jump shot.'
- double[WO]: incorrect word (objective), LeBron James finished with a 'triple-double.'

Partial-game summary from Qwen-JSON input

The Hawks took an early lead with a layup from Onyeka Okongwu, but the Lakers quickly tied it[WO] with a three-pointer from Max Christie.
Throughout the game, LeBron James and Anthony Davis were the standout players for the Lakers, contributing significantly with scoring, rebounds, and assists.
List of errors.

- tied it[WO]: word objective error, the Lakers 'responded with' a three-pointer, they did not tie the game.

Partial-game summary from Llama-JSON input

In overtime, the Hawks took the lead early, with Onyeka Okongwu making a layup to put them up 121-119. The Lakers fought back, with James making a 3-pointer to put them up 127-125. The Hawks regained the lead, with Trae Young making a 3-pointer to put them up 134-132 with 8 seconds left. The Lakers were unable to score on their final possession, and the Hawks emerged victorious.
List of errors.

- No errors in this partial summary.

Figure 4: Examples of partial game summaries with error annotations, illustrating errors from the same game and period under the three input types, Row, JSON, and Unstructured (see Table 1). Corrections are included for clarity. *Notation:* superscripts N, U, WO, WS, C, O, and X indicate the error categories: Name, Number, Word-objective, Word-subjective, Context, Other, and Not-checkable; this is used here for readability; see the Annotation Protocol Section 3.4 for definitions.

each event as a nested key-value pair, where time, team, play details, and score are separate fields, this clear separation preserves the exact relationship between the actions taken. Most names, numbers, and actions are correct (see Section 4.2 and Fig. 5), though James's 'jump shot' is misrendered as a

Category	Total errors	Both agree	A1 only	A2 only	Agreement (%)	Precision (%)	Recall (%)	Unbiased F_1 (%)
Number	112	108	0	4	96.4	100.0	98.2	98.2
Name	21	15	1	5	71.4	93.8	84.4	84.4
Word_objective	23	13	4	6	56.5	76.5	72.4	72.4
Word_subjective	13	4	2	7	30.7	66.7	51.5	51.5
Context	9	6	1	2	66.6	85.7	80.4	80.4
Not_checkable	8	5	0	3	62.5	100.0	81.2	81.2
Other	2	0	2	0	0.0	0.0	N/A	N/A
Overall	188	151	10	27	**80.3**	**93.8**	**89.3**	**89.3**

Table 6: Inter-Annotator Agreement by Error Category. Agreement is highest for *Number* errors and lowest for *Word-subjective* errors.

layup.

5.2 Models

In terms of the summary style, Llama focuses on the first few plays and last few plays in each quarter. They also produce around 8 to 12 numerical figures to summarize one quarter, compared to Qwen's 5–6 for the same period. For example,

> The Rockets responded with a **10**^{U}-**0**^{U} run, capped off by a dunk from Cam Whitmore, to take a 45-**35**^{U} lead.
> **10**^{U}-**0**^{U}: number errors, the correct runs during that time is 6-7 run. Llama makes these computation quite often and results in more number errors.
> **35**^{U}: number error, the opponent team score is 32.

Qwen makes slightly more name mistakes than Llama (see Fig. 3 and Fig. 5). We observed a few instances where both models showed bias towards star players in their summaries. For example,

> The Lakers extended their lead with a series of baskets from **LeBron James**^{N} and Anthony Davis.

In fact, the baskets came from Knecht and Davis, but both models incorrectly substituted LeBron James, reflecting a tendency, though not frequent, to favour well-known players in their summaries.

Importance of prompt. The instructions given in the prompt also play an important role in reducing a few errors in the summaries. After several trials, we applied the prompt mentioned in (Fig. 6) for all our generations. Qwen's summaries adhered to the instructions (narrative style, structure, constraints), whereas Llama followed the instructions for narrative style and structure but struggled with constraints such as calculating individual player points (even when instructed not to) and maintaining the mandatory word limit of 450 words. The content focus differs for both models.

5.3 Inter-annotator agreement

One of the authors manually annotated errors in all 180 generated summaries, and a second annotator independently annotated 9 summaries (5% of the dataset) to assess inter-annotator agreement. We provided detailed guidelines describing error categories, examples, and instructions for marking errors. Annotators worked independently and marked errors at the token/phrase level.

Agreement was measured as exact-match overlap: for each category, we computed precision, recall, and unbiased F1 by alternately treating each annotator as the reference and averaging. The overall agreement was 80.3% with unbiased F1 of 0.89, indicating strong agreement despite some variation across categories.

6 Conclusion

Factual inaccuracies remain a major challenge for deploying LLMs in real-world applications. Our study shows that structured inputs, especially hierarchical JSON, substantially reduce errors in NBA play-by-play summaries: error rates dropped by up to 69% for Llama and 65% for Qwen compared to unstructured text. A two-way repeated-measures ANOVA confirmed that input structure explained over 80% of the variance in error rates, with significant differences across all formats. As future work, we will develop an LLM-as-Judge framework to complement costly manual evaluation by automatically assessing atomic factual claims against play-by-play input logs. We will also extend this work to ice hockey play-by-play data to identify general data-structuring features that minimise factual inaccuracies across multiple sports datasets.

Limitations

We acknowledge some limitations in this work. First, we only looked at NBA basketball game summaries. Second, we required annotators to have qualitative NBA knowledge to understand the play-by-play data. This requirement improved annotation accuracy but significantly reduced the pool of eligible annotators and limited scalability. Each summary took approximately 30 to 60 minutes to annotate, depending on its complexity, and one eligible annotator withdrew because of this time constraint. Third, we evaluated only two models (Llama and Qwen), which limits the generalizability of our findings to other large language models.

Ethics Statement

We received approval from the University of Aberdeen Ethics Review Board to perform the inter-annotation experiment. This study is based entirely on publicly available NBA play-by-play data, which we preserved in its original structure without introducing additional bias. For error annotation, one author performed a full manual review of all 180 summaries, and a second annotator, who volunteered and provided informed consent, reviewed 9 summaries to help validate consistency. Annotators received detailed guidelines and example annotations, were free to withdraw at any time without penalty, and were compensated for their time as previously agreed upon before the annotation process.

Acknowledgements

We thank Javier González Corbelle for his hard work in helping with the annotations in this paper. We thank the anonymous reviewers for their detailed feedback and suggestions, which have significantly improved this work. We also thank the NLG (CLAN) reading group at the University of Aberdeen for their invaluable feedback.

References

Alibaba Cloud. 2025. Alibaba cloud ai platform. https://www.alibabacloud.com/.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Google Cloud. 2025. Overview of prompting strategies | generative ai on vertex ai. https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/prompt-design-strategies.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Wenlin Yao, Hassan Foroosh, Dong Yu, and Fei Liu. 2024. When reasoning meets information aggregation: A case study with sports narratives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4293–4308, Miami, Florida, USA. Association for Computational Linguistics.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, and 7 others. 2025. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *Preprint*, arXiv:2501.03200.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6908–6915.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 645–654, New York, NY, USA. Association for Computing Machinery.

Barkavi Sundararajan, Yaji Sripada, and Ehud Reiter. 2024. Improving factual accuracy of neural table-to-text output by addressing input problems in ToTTo. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7350–7376, Mexico City, Mexico. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Somayajulu Sripada. 2020. SportSett:basketball - a robust and maintainable data-set for natural language generation. In *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 32–40, Santiago de Compostela, Spain. Association for Computational Lingustics.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Comput. Speech Lang.*, 80(C).

Together AI. 2025. Together ai api documentation. `https://api.together.xyz/`.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Yingsheng Wu, Yuxuan Gu, Xiaocheng Feng, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024. Extending context window of large language models from a distributional perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7288–7297. Association for Computational Linguistics.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Appendices

This appendix provides supplementary figures, tables, and prompts referenced in the main paper. We include the unstructured play-by-play input (Table 7) for reference. The model hyperparameters and decoding settings used to generate the 180 game summaries are shown in Table 8. We also report post-hoc Tukey HSD tests and a full set of ANOVA tables for each error type (number, name, word-objective, word-subjective, and context). These results complement the summary statistics in the main text by presenting full p-values and effect sizes. To further illustrate the annotation protocol, we provide extended annotated examples of game summaries with error labels (Fig. 5), as well as the full evaluation prompt template used for generation (Fig. 6).

Time	LA Lakers	N/A	Score	N/A	Atlanta
5:00.0	Start of 1st overtime	N/A	N/A	N/A	N/A
5:00.0	Jump ball: A. Davis vs. O. Okongwu (J. Johnson gains possession)	N/A	N/A	N/A	N/A
3:05.0	N/A	N/A	122–123	N/A	Turnover by J. Johnson (bad pass; steal by G. Vincent)
2:56.0	Shooting foul by O. Okongwu (drawn by A. Davis)	N/A	122–123	N/A	N/A
2:56.0	N/A	N/A	122–123	N/A	C. Capela enters the game for O. Okongwu
2:56.0	A. Davis makes free throw 1 of 2	1.0	123–123	N/A	N/A
2:56.0	A. Davis makes free throw 2 of 2	1.0	124–123	N/A	N/A
2:45.0	N/A	N/A	124–123	N/A	D. Hunter misses 2-pt layup from 10 ft (block by L. James)
2:44.0	N/A	N/A	124–123	N/A	Offensive rebound by Team
2:33.0	N/A	N/A	124–125	2.0	T. Young makes 2-pt layup from 2 ft
2:13.0	L. James makes 3-pt jump shot from 25 ft	3.0	127–125	N/A	N/A

Table 7: Excerpt of unstructured play-by-play input (partial game). Raw log text from Basketball-Reference (Overtime). No typed schema in the header; event semantics are embedded in free-text team columns, and cells are *non-atomic*; one span may combine event type, participants, outcome, distance, and points. Fields that do not apply to a given event are shown as N/A for reference. Columns show team descriptions, per-event points, and the cumulative score.

Model	Parameters	Release	Prompt	Temperature	Top-p	Top-k	API Platform
Llama-3.1-70B	70B	Jul. 2024	Zero-shot	0	0.95	50	Together AI API
Qwen2.5-72B	72B	Sep. 2024	Zero-shot	0	0.95	50	Qwen (Alibaba)

Table 8: Model specification and decoding settings used in all runs.

Group 1	Group 2	Mean Diff	p (Tukey-adjusted)	Lower	Upper	Significant
(a) Input Structure						
JSON	Row	0.9102	0.000	0.4938	1.3265	**Yes**
JSON	Unstructured	4.3913	0.000	3.9750	4.8077	**Yes**
Row	Unstructured	3.4812	0.000	3.0648	3.8975	**Yes**
(b) Model						
Llama	Qwen	-0.4412	0.165	-1.0651	0.1827	No
(c) Interaction: Input Structure × Model						
Llama_JSON	Llama_Row	1.0360	0.0003	0.3492	1.7228	**Yes**
Llama_JSON	Llama_Unstructured	4.8803	0.0000	4.1935	5.5671	**Yes**
Llama_JSON	Qwen_JSON	-0.0313	1.0000	-0.7181	0.6555	No
Llama_JSON	Qwen_Row	0.7530	0.0226	0.0662	1.4398	**Yes**
Llama_JSON	Qwen_Unstructured	3.8710	0.0000	3.1842	4.5578	**Yes**
Llama_Row	Llama_Unstructured	3.8443	0.0000	3.1575	4.5311	**Yes**
Llama_Row	Qwen_JSON	-1.0673	0.0002	-1.7541	-0.3805	**Yes**
Llama_Row	Qwen_Row	-0.2830	0.8424	-0.9698	0.4038	No
Llama_Row	Qwen_Unstructured	2.8350	0.0000	2.1482	3.5218	**Yes**
Llama_Unstructured	Qwen_JSON	-4.9117	0.0000	-5.5985	-4.2249	**Yes**
Llama_Unstructured	Qwen_Row	-4.1273	0.0000	-4.8141	-3.4405	**Yes**
Llama_Unstructured	Qwen_Unstructured	-1.0093	0.0005	-1.6961	-0.3225	**Yes**
Qwen_JSON	Qwen_Row	0.7843	0.0151	0.0975	1.4711	**Yes**
Qwen_JSON	Qwen_Unstructured	3.9023	0.0000	3.2155	4.5891	**Yes**
Qwen_Row	Qwen_Unstructured	3.1180	0.0000	2.4312	3.8048	**Yes**

Table 9: Tukey HSD post hoc test results. The table reports pairwise differences in mean error rates between groups, with Tukey-adjusted p-values (p-adj), 95% confidence intervals (Lower, Upper), and a significance flag (Significant). Mean Diff is computed as Group 1 minus Group 2, so positive values indicate Group 1 has a higher mean than Group 2. **Section (a)** compares input structure conditions (Row, JSON, and Unstructured) and shows significant pairwise differences. **Section (b)** compares Models (Llama vs. Qwen). This contrast is not significant, indicating no reliable difference in overall error rates. **Section (c)** shows Interaction contrasts between specific Model and Input Structure combinations (e.g., Llama_JSON vs. Llama_Row). Several significant interactions exist between Model and Input Structure combinations.

Source	SS	DF	MS	F	p-unc	η_p^2
Input Structure	84.24	2	42.12	57.55	6.54×10^{-20}	0.398
Model	37.90	1	37.90	51.80	1.77×10^{-11}	0.229
Model × Input Structure	21.32	2	10.66	14.57	1.41×10^{-6}	0.143

Table 10: Two-way repeated-measures ANOVA on error rates, reporting SS: Sum of Squares, DF: Degrees of Freedom, MS: Mean Square; F: F-statistic, uncorrected p-values (p-unc), and partial effect sizes (η_p^2). Input structure shows the largest effect, followed by Model and their interaction; all are statistically significant.

Source	SS	DF	MS	F	p-unc	η_p^2
Input Structure	23.15	2	11.57	73.48	7.35×10^{-24}	0.458
Model	0.78	1	0.78	4.98	0.027	0.028
Model × Input Structure	0.22	2	0.11	0.71	0.493	0.008

Table 11: Two-way repeated-measures ANOVA on name error rates, reporting SS: Sum of Squares, DF: Degrees of Freedom, MS: Mean Square; F: F-statistic, uncorrected p-values (p-unc) and partial eta-squared (η_p^2). Input structure shows a large, significant effect; model is small but significant; the interaction is not significant.

Source	SS	DF	MS	F	p-unc	η_p^2
Input Structure	28.35	2	14.17	69.11	8.14×10^{-23}	0.443
Model	0.37	1	0.37	1.79	0.183	0.010
Model × Input Structure	0.13	2	0.06	0.31	0.737	0.004

Table 12: Two-way repeated-measures ANOVA on word objective error rates, reporting SS: Sum of Squares, DF: Degrees of Freedom, MS: Mean Square; F: F-statistic, uncorrected p-values (p-unc) and partial eta-squared (η_p^2). Input structure shows a large, significant effect; model and the interaction are not significant with negligible effect sizes.

Source	SS	DF	MS	F	p-unc	η_p^2
Input Structure	7.88	2	3.94	44.61	2.29×10^{-16}	0.339
Model	3.91	1	3.91	44.34	3.48×10^{-10}	0.203
Model × Input Structure	0.17	2	0.09	0.99	0.375	0.011

Table 13: Two-way repeated-measures ANOVA on word-subjective error rates, reporting SS: Sum of Squares, DF: Degrees of Freedom, MS: Mean Square; F: F-statistic, uncorrected p-values (p-unc) and partial eta-squared (η_p^2). Input structure and model show significant effects with moderate effect sizes; the interaction is not significant with a negligible effect size.

Source	SS	DF	MS	F	p-unc	η_p^2
Input Structure	8.02	2	4.01	54.20	5.03×10^{-19}	0.384
Model	0.30	1	0.30	4.01	0.047	0.023
Model × Input Structure	0.59	2	0.30	4.01	0.020	0.044

Table 14: Two-way repeated-measures ANOVA on context error rates, reporting SS: Sum of Squares, DF: Degrees of Freedom, MS: Mean Square; F: F-statistic, uncorrected p-values (p-unc) and partial eta-squared (η_p^2). Input structure shows a large, significant effect; model is small but significant; the interaction is also significant with a small effect size.

Partial-game summary from Qwen-Unstructured input

Jalen Green's[N] two free throws[WO] brought the Rockets within 12[U]-8[U], and a subsequent 3-pointer by Frank Ntilikina[C] tied the game at[WO] 14[U]-14[U]. The quarter ended with Cleveland leading[WO] 32-36[U], thanks to a flurry of 3-pointers from both sides[WO], including a buzzer-beater by Christian Wood[C], who hit a 3-pointer to cap off a strong first quarter for Houston.

List of errors:

- Jalen Green's[N]: incorrect name, should be Cam Whitmore.
- free throws[WO]: incorrect action (word objective), the player made two layups.
- 12[U]-8[U]: incorrect scores (number), the correct numbers/scores are 25-22.
- leading[WO]: incorrect word, Cleveland were trailing.
- 36[U]: incorrect score (number), it should be 39
- flurry of 3-pointers from both sides[WO]: word objective error, there were only a couple of three-pointers by Rockets.
- Christian Wood[C]: incorrect context, he did not play for Rockets or NBA in the 2025 season. Cam Whitmore is the correct player

Partial-game summary from Qwen-Row input

The Lakers' bench provided a spark[WS], with Max Christie[N] and Shake Milton combining for a series of clutch plays. A critical turnover by Deandre Ayton[N] and a subsequent three-pointer from Austin Reaves extended the Lakers' lead to 88-75 as the quarter ended.

List of errors:

- provided a spark[WS]: subjective phrase, they combined for 6 points at the end of Q3. Some may agree scoring over 10 points to be spark.
- Max Christie[N]: he was not a bench player and he did not contribute during the clutch play of that quarter. the right bench player is Christian Koloko.
- Deandre Ayton[N]: incorrect name, Turnover was made by Deni Avdija.

Partial-game summary from Llama-JSON input

The fourth quarter saw the Trail Blazers make a late push, with Anfernee Simons making a 3-pointer and Scoot Henderson adding a jump shot to cut the deficit to 94-86. However, the Lakers were able to hold them off, with LeBron James making a layup[WO] and Max Christie adding a dunk to give them a 103-97 lead.

List of errors:

- layup[WO]: word objective error, he made a jump shot.

Figure 5: More examples of partial game summaries with error annotations to demonstrate errors faced in different input structures (Row, JSON and Unstructured). Corrections are included here for clarity. *Notation:* superscripts N, U, WO, WS, C, O, and X indicate the error categories: Name, Number, Word-objective, Word-subjective, Context, Other, and Not-checkable; this is used here for readability; see the Annotation Protocol Section 3.4 for definitions.

NBA Play-by-Play Game Summary Prompt

You are a professional basketball reporter hired by my company to cover NBA games. You are provided with a hierarchical JSON play-by-play dataset, which includes structured team information and a chronological list of plays. Each play entry contains key details such as time remaining, the team involved, play description, event type, player names, and score updates.

Your task is to analyse the data and identify the four most significant moments in each quarter, key scoring plays, momentum shifts, lead changes, game ties, and clutch performances. Prioritise impactful events such as three-pointers, layups, dunks, free throws, and possessions immediately following turnovers or rebounds. Ignore minor plays that do not significantly influence the game's flow.

Write a fluent, engaging, and objective game summary of exactly 450 words, structured around these defining moments. The narrative should clearly convey why these moments were significant, seamlessly integrating them into the game's overall storyline. It should flow naturally from start to finish, capturing the game's pace and intensity.

Mention the final score and quarter-ending scores for context, but do not rigidly structure the summary by time. Discuss key players in relation to the game-defining plays, without calculating total points. Maintain a smooth, journalistic writing style. Do not write in bullet points, lists, or unnecessary formatting. The summary must be factually accurate, strictly based on the provided data, and completely free from external assumptions, while maintaining a compelling and professional tone.

Quarter-End Context: Mention the total score contextually within the narrative as each quarter concludes. DO NOT list the quarter scores separately at the very end of the summary.

Figure 6: Prompt used for generating NBA game summaries from hierarchical JSON play-by-play data. The first paragraph specifies the input format, while the remaining instructions are consistent across all runs.