# From Data to Insights: The Power of LM's in Match Summarization

**Satyavrat Gaur, Pasi Shailendra, Rajdeep Kumar, Rudra Chandra Ghosh, Nitin Sharma**
*Central Research Laboratory, BEL-Ghaziabad, India*
{satyavratgaur, pasishailendra, rajdeepkumar, rudrachandraghosh, nitinsharma}@bel.co.in

## Abstract

The application of Natural Language Processing is progressively extending into many domains as time progresses. We are motivated to evaluate language model's (LMs) capabilities in many real-world domains due to their significant potential. This study examines the use of LMs in sports, explicitly emphasizing their ability to convert data into text and their understanding of cricket. By examining cricket scorecards, a widely played sport on the Indian subcontinent and many other regions, we will evaluate the summaries produced by LMs from several viewpoints. We have collected concise summaries of the scorecards from the ODI World Cup 2023 and assessed them in terms of both factual accuracy and sports-specific significance. We analyze the specific factors that are included in the summaries and those that are excluded. Additionally, it analyzes prevalent mistakes concerning completeness, correctness, and conciseness. We are presenting our findings here and also our dataset and code are available here[1].

## 1 Introduction

Sports contribute over $500 billion annually to the global economy and are crucial for promoting physical health and reducing chronic diseases (Fort and Quirk, 1995; Warburton et al., 2006). They also foster social cohesion by bringing communities together and teaching essential values like teamwork and leadership (Coalter, 2007). Sports enhance personal development through skills such as discipline and resilience, shaping individuals physically and mentally (Holt et al., 2017). Sports play a vital role in improving individual well-being and societal harmony globally. Cricket's popularity surpasses many other sports due to its global reach, boasting 2.5 billion fans worldwide and massive viewership during events like the ICC Cricket World Cup. Its rich history and cultural significance in nations like India, Pakistan, Australia, and England contribute to its enduring appeal. The sport's diverse formats cater to different audiences, from traditional Test matches to fast-paced Twenty-20 games, accommodating varied preferences. Additionally, prestigious leagues such as the IPL and BBL ensure year-round excitement and attract top international players. These factors collectively make cricket a powerhouse in sports, sustaining its widespread popularity and fan engagement globally.

Data-to-text generation involves transforming structured, non-linguistic input like tables, databases, tuples, or graphs into accurate textual descriptions automatically (Reiter and Dale, 1997; Covington, 2001; Gatt and Krahmer, 2018). This process is crucial in various real-world scenarios, such as creating weather forecasts based on meteorological data (Goldberg et al., 1994), summarizing biographical information (Lebret et al., 2016), or generating sports summaries from game statistics (Wiseman et al., 2017). The objective is to convey pertinent details from the input data using natural language, ensuring the generated text faithfully and precisely reflects the original information. Thus, achieving accuracy in representing the source data becomes paramount.

Several methods have been developed to tackle this challenge in Data-to-Text generation. These approaches utilize different strategies such as incorporating the structure of input data (Wiseman et al., 2017; Puduppully et al., 2019; Chen et al., 2020b), employing neural templates (Wiseman et al., 2018), and emphasizing the arrangement of content (Puduppully et al., 2019). The rise of LMs has brought about a profound shift in controllable text generation and data interpretation. Recently, there has been a shift towards utilizing large-scale pre-trained models (Devlin et al., 2018), which have shown notable improvements in fluency and the ability to generalize compared to earlier ap-

[1] https://github.com/satyawork/ODI-WORLDCUP.git

proaches that did not use such models.

In today's world, people regularly handle large amounts of organized data to make decisions and find information. It's crucial now more than ever to present this data in ways that are easy to understand and user-friendly (Zhang et al., 2023; Li, 2023). Conveniently presenting data has sparked interest in techniques that convert intricate data tables into clear, meaningful narratives that meet the specific needs of users (Parikh et al., 2020b; Chen et al., 2020a). These methods can be applied across various fields, such as game strategy planning, financial analysis, and human resources management. Yet, current fine-tuned table-to-text models (Nan et al., 2022a; Liu et al., 2021) are often designed for specific tasks, restricting their flexibility for practical uses in different scenarios.

Recent studies show that LMs can achieve performance comparable to state-of-the-art fine-tuned models in tasks like answering table questions and fact-checking. Yet, there is still much uncharted territory regarding LMs' ability to effectively generate text from tabular data to meet users' information needs.

We have created a dataset named ICC CRICKWORLD CUP sourced from reputable sports analytics repositories, encompassing 12 distinct categories concerning the ICC Cricket World Cup 2023. We tested several small-scale language models (SLMs) on this dataset to generate five types of summaries. These models were adopted for the resource constraint scenario. The resulting summaries underwent evaluation using various metrics supplemented by human assessment for enhanced accuracy. Additionally, we have documented our findings on the performance of these summaries, highlighting instances of both success and areas needing improvement. This initiative introduces a cricket dataset structured with tables, challenging LMs to derive meaningful insights from these data points. Future advancements could leverage larger parameter LMs to refine this approach, potentially extending its application to diverse domains beyond cricket.

Previously, significant advancements have been made in converting tabular data to text and summarizing it. Two primary types of summaries can be generated from tabular data:

- Query-Based Summary: These type of method involves querying specific information from the table, selecting particular cells, and generating textual data based on the chosen fields. This type of summary is tailored to particular interests or queries within the data.
- Data-Based Summary: This approach provides the entire structured dataset to a language model (LM), which then generates textual data based on the comprehensive dataset. This method allows for a more holistic summary of the entire dataset, capturing broader insights and trends.

Both methods have their distinct advantages and applications, enhancing the ability to derive meaningful narratives from structured data.

In our work, we collected data points from reputable and reliable cricket sources to create a Cricket Summarization Dataset. The dataset includes five types of summaries: an overall match summary, a bowling perspective summary, a batting perspective summary, and summaries from the viewpoints of Team 1 supporters and Team 2 supporters. These summaries were prepared by a few cricket experts. Additionally, multiple LLM models were used to generate summaries, which were then evaluated using various metrics. We also documented the mistakes made by the LLMs in different scenarios.

## 2 Related Work

Natural Language Processing (NLP) for summarization focuses on automatically condensing large volumes of text into shorter, coherent summaries while retaining key information. This involves techniques like extractive and abstractive summarization, where models either select important sentences or generate new, concise summaries. NLP-based summarization is widely used in areas like news aggregation, research papers, and legal documents to provide quick and efficient insights.

Horasan and Bilen (2020) researched the summarization of news articles about sports, demonstrating the impact of NLP in the sports domain. The study highlighted the increasing use of machine learning and deep learning techniques in sports analysis. Mahajan et al. (2024) showcased the application of machine learning for shot detection in cricket, further illustrating the integration of advanced computational methods in sports. Additionally, other research efforts, such as those by Hussain et al. (2024) have focused on analyzing videos and audios of cricket matches to derive in-

| Reference | JSON | Summary | Bowler Summary | Batting Summary | Team 1 Supporter | Team 2 Supporter | Match | Team 1 | Team 2 |
|---|---|---|---|---|---|---|---|---|---|
| https://www.icc-cricket.com/tournaments/cricketworldcup/matches/228846/india-vs-australia | [ { "match": "India vs Australia, Final - Live Cricket Score, Commentary", "series": "ICC Cricket World Cup 2023", ... ..India Innings 240-10 (50 Ov) ", { "batter": " Rohit (c) ", "wicket": " c Head b Maxwell ... | Australia secured a convincing victory over India by 6 wickets in a cricket match... ..contributions from Kohli (54 runs) and Rahul (66 runs),... ...Travis Head's spectacular century (137 runs) led | ...the bowlers played a significant role in determining the outcome. Australia's bowling attack, led by Mitchell Starc ...with Bumrah and Siraj both scalping 2 wicket each... ultimately overshadowing India's bowling efforts. | ...batters from both teams showcased notable performances ...but India's batting falter against Australia's disciplined bowling attack ...overshad-owed their efforts, leading Australia to a comfortable win. | ...team posted a competitive total of 240 runs, with notable contributions from Rohit Sharma ... Although India's bowlers, led by Jasprit ... managed to pick up crucial wickets ... Despite their efforts, India couldn't | ... Australia demonstrated dominance both with bat and ball ... Head's stellar performance, supported by ... excellent skills, restricting India's batting lineup ... compreh-ensive team effort from Australia, ... | India vs Australia | India | Australia |

Figure 1: Snapshot of the dataset with all 10 fields with 5 different summaries.

sights. Despite these advancements, there remains a scarcity of NLP-focused work specifically targeting cricket, indicating a potential area for further exploration and development.

On the other hand, multiple researchers have extensively studied text summarization, as highlighted in a survey Hussain et al. (2024), which documented progress in this task and identified common pitfalls that LM;s encounter when summarizing long texts. Base models like BERT (Lewis et al., 2019), T5 (Raffel et al., 2023), RoBERTa (Liu et al., 2019), and BART (Yu et al., 2021) have shown promising results in summarization tasks . Further, fine-tuned versions of these models have been applied to tabular data to generate various summaries. For instance, Andrejczuk et al. (2022) explored table-to-text generation with TabT5 model based on T5 pre-trained model, while Liu et al. (2022) and Zhao et al. (2023) focused on pre-training techniques for table summarization and query-focused summarization of tabular data. Large language models also show this type of capability.

Tabular data presents additional complexities compared to regular text summarization, and several studies have addressed table-to-text summarization. A systematic review by Osuji et al. (2024) covers the datasets used for this task, such as ToTTo (Parikh et al., 2020b) by (Parikh et al., 2020a; Nan et al., 2022b) and others containing paired data of tables and their respective summaries. These datasets primarily consist of paired data, with some additional information based on specific cases. Similarly, we generated a dataset of summaries along with a comparative analysis to explore this domain further.

## 3 Dataset

We curated a comprehensive ICC Cricket World Cup 2023 dataset from reputable sports analytics repositories, encompassing detailed records of each team's batting and bowling performances and powerplay logs. Additionally, the dataset includes summarizations of match logs from various viewpoints: a general match summary, summaries from the perspectives of bowling and batting performances, and summaries specific to the team batting first and last. Figure 1 shows an overview of dataset fields. These summaries were written by a writer with good knowledge of the 2023 ODI World Cup. Also, all summaries are cross-verified for any errors, and it is confirmed that all noteworthy performances are included in the summary. Each match entry in the dataset is accompanied by essential details such as match ID, date-time, the team batting first, and the team batting last. The dataset comprises 48 matches, each with summaries and additional pertinent information.

### 3.1 Summaries

Our analysis, informed by cricket experts and various sports articles, identified five potential summarization perspectives:

- **Normal Summary:** This summary condenses the entire match, providing a neutral overview of the match.

- **Bowling Summary:** This summary focuses on the bowling performances of both teams, highlighting noteworthy bowling spells and statistics.

- **Batting Summary:** This summary emphasizes the batting performances of both teams, summarizing crucial innings and statistics.
- **Team 1 Supporter Summary:** This summary presents the perspective of that team supporter who batted first, focusing on their team's performance and positive aspects of the game.
- **Team 2 Supporter Summary:** This summary presents the perspective of that team supporter who batted second, focusing on their team's performance and positive aspects of the game.

The types of summary presented above avoid subjective evaluations of performance, as these depend heavily on the context of the match. Factors like pitch and ground conditions, which are not included in the input data, significantly influence performance. Therefore, determining whether a performance is 'good' or 'bad' relies on analyzing the scorecard data and understanding the person generating the summary, whether a large language model (LM) or a human.

To give input to LM, we provide a list of instructions in Table 1. These instructions are developed by following best practice as suggested by Amatriain (2024) and are improved iteratively so that some issues noticed during experiments can be handled using a prompt, like if "...in 1 paragraph" is not appended at the end of the prompt. LM gives point or tabular data, which is not required in the summary.

The match log follows this sequence: first, the batting log of the team that batted first, followed by the bowling log of the opposing team. Next, it includes the fall of wickets for the first batting team and the powerplays of the first batting team. Then, the log continues with the batting log of the team that batted last, followed by the bowling log of the opposing team. Finally, it records the fall of wickets for the last batting team and the powerplays of the last batting team.

# 4 Experiment

## 4.1 Experimental Setup

Each LM was evaluated with a consistent temperature value of 0.1 to encourage factual accuracy in the generated summaries, and a maximum token length of 4096 to avoid imposing constraints on the summary length. The dataset consisted of JSON-formatted text containing scorecards for all 48 matches from the ODI World Cup 2023. Five different types of summaries were generated for each match, with specific instructions provided via prompts. LangChain and Hugging Face pipelines facilitated the text summarization process, with LangChain management workflow execution and coordination, and Hugging Face providing access to the various language models. The experiments were conducted on an NVIDIA RTX 5000 GPU with 16 GB of memory and 9728 CUDA cores.

## 4.2 Models

We have conducted experiments using several large language models, implementing resource-efficient techniques like quantization. Due to resource constraints, we utilized a maximum of 13 billion models, all quantized. The models we have explored for the experimentations are: **LLaMA 2** (Touvron et al., 2023) 7B chat model with 4-bit and 8-bit quantization and a full-fledged model. **LLaMA 3** 8B instruct model (AI@Meta, 2024) is used with 4-bit and 8-bit quantization, and a Non-quantized model is also used. **Mistral 7B** (Jiang et al., 2023) instruction-tuned model with quantization is used. **Vicuna-7B** (Zheng et al., 2023), **Phi-3-Mini** (Abdin et al., 2024).

## 4.3 Automated Evaluation

To evaluate the performance of our text generation tasks, we leverage several established summarization metrics commonly used across various NLP applications, including paraphrasing, automatic summarization, and machine translation. Those are **BLEU** (Papineni et al., 2002), **ROUGE-L** (Lin, 2004), **BERTScore** (Zhang et al., 2019). By employing these complementary metrics, we comprehensively understand how well our text generation models perform in terms of factual accuracy, content coverage (recall), fluency, and semantic similarity to the reference text.

## 4.4 Human Evaluation

Automatic scoring methods mentioned above are great for checking factual overlap in summaries, but they can't tell the whole picture. Human judgment is essential for an excellent summary. Humans can see if the summary captures the key ideas and meaning, works for a specific audience (considering their knowledge and goals based on the cricketing context and summary author), and even catch factual errors that automatic metrics might miss.

| Task | Instruction |
|---|---|
| Summary | As a sports journalist, give textual summary of above match from data provided above in 1 paragraph |
| Bowling summary | As a sports journalist, give summary of bowler's performance of both teams in 1 paragraph |
| Batting summary | As a sports journalist, give a summary of batter's performance of both teams in 1 paragraph |
| Team1 supporter | As team1 supporter, give summary of the above match data in one paragraph |
| Team2 supporter | As team2 supporter, give summary of the above match data in one paragraph |

Table 1: List of prompt that were given to generate five types of summaries, where Team1 refers to the team that batted first, and Team2 refers to the team that batted second.

In this evaluation, we ask individuals with good knowledge of the sport and who have closely observed the matches of the 2023 ODI Cricket World Cup held in India. They read the summaries generated by LMs, noting any inconsistencies observed. Inconsistencies could be of any type, but we primarily focus on the summary's completeness, correctness, and conciseness. These three aspects are addressed as follows:

- **Completeness**: The summary should capture all important performances from the scorecard. All notable performances should be present, covering players from both teams, including bowlers and batters.
- **Correctness**: The summary should have minimal false information. We also try to understand why false information arises—whether due to wrongly related information within the provided context or because the LM generated irrelevant information.
- **Conciseness**: The summary shouldn't include information not important enough to be in a cricket summary. For instance, if a player's performance didn't significantly impact the match, it shouldn't be mentioned in the summary.

## 5 Results

Our curated dataset of cricket World Cup match summaries served as the gold standard because of multiple verifications and validation for evaluating summaries generated by different LMs. These LM summaries were assessed using the previously-mentioned automated evaluation metrics, and the tabular result is present in Table 2.

The results indicate a correlation between high metric scores and similarity to the human-crafted summaries. LMs such as Llama2, Llama3, and Mistral achieved promising results in terms of similarity based on these metrics. Phi3 mini and Small performed averagely, while Vicuna exhibited the lowest similarity scores.

We conducted a human evaluation with a cricket expert to validate these findings further. We evaluate 1440 summaries generated by LM's. This evaluation aimed to assess the actual correctness and completeness of the summaries beyond the limitations of automatic metrics. The results of this human evaluation and the identified causes of errors in some summaries are discussed in detail below.

The following are some common pitfalls and errors where the model struggles and its outputs are compromised:

- In many instances model incorrectly states a team's victory margin as "won by 1 run" instead of using the correct phrasing based on the scenario. In cricket, the margin of victory depends on whether the team batted first or second, such as "won by 8 runs" or "won by 5 wickets." it is observed that LMs often struggle with choosing the appropriate phrasing, especially when the chasing team wins, frequently outputting "win by 1 runs" instead of the correct "win by 5 wickets."
- When a bowler from the winning team performs extraordinarily, the model mistakenly attributes that performance to a bowler from the opposing team, particularly when no notable performances are observed from the los-

| Model | | Rouge-l | | Bleu | | BERT | |
|---|---|---|---|---|---|---|---|
| | P | R | F1 | | P | R | F1 |
| Llama-2-7b-chat-hf | 0.2438 | 0.2031 | 0.2183 | 0.4692 | 0.6502 | 0.6646 | 0.6569 |
| Llama-3-7b-chat-hf | 0.2406 | 0.2162 | 0.2256 | 0.4864 | 0.6534 | 0.6738 | 0.6631 |
| Mistral-7B-Instruct-v0.1 | 0.1882 | 0.1285 | 0.1492 | 0.3618 | 0.5890 | 0.6187 | 0.6031 |
| Mistral-7B-Instruct-v0.2 | 0.2671 | 0.1991 | 0.2239 | 0.4475 | 0.6383 | 0.6697 | 0.6530 |
| Phi-3-mini-128k-instruct | 0.2580 | 0.0792 | 0.1054 | 0.1989 | 0.5273 | 0.5182 | 0.5223 |
| Phi-3-small-128k-instruct | 0.2580 | 0.0792 | 0.1054 | 0.1989 | 0.5273 | 0.5182 | 0.5223 |
| vicuna-7b-v1.5 | 0.0771 | 0.1268 | 0.0792 | 0.1268 | 0.4335 | 0.3777 | 0.3996 |

Table 2: Performance metrics of various language models (LMs) evaluated on the ICC CRICK-WORLD CUP dataset. The table has four columns, 1st one is the name of the column and rest three are primary evaluation metrics: 1) **Rouge-l**, with three sub-columns Precision (P), Recall (R), and F1-score (F1); 2) **Bleu Score**; and 3) **BERT Score**,with three sub columns Precision, Recall, and F1-score presented as P, R and F1. The results highlight the comparative performance of model's accurate and contextually relevant cricket match summaries.

ing team's bowlers.

- A common and frequent error observed by an article reader is that the model confuses the terms "each" and "both" in its output. This leads to incorrect summaries where two bowlers listed consecutively in the scorecard are inaccurately stated to have taken the same number of wickets. This issue is mostly observed in the bowlers' specific summaries and never in the batting summaries.

- The LM often confuses cricket terminology. For example, a top-order batsman is incorrectly identified as the top-scorer batsman, a three-wicket haul is mistaken for a hat-trick, and if a fielder catches a ball, the summary incorrectly states that the fielder took the wicket.

- In the summary from the perspective of a team's supporter, even if the match is one-sided and the team is clearly losing, the model inaccurately uses adjectives such as "thrilling" and "exciting" to describe the match. This misrepresentation occurs despite the obvious lack of competitiveness, leading to an unrealistic portrayal of the match's nature.

- From the perspective of a team's supporter, if the supporting team sets an easy score to chase, the model erroneously claims that the team has given a tough score to chase.

- From the perspective of a team's supporter, the model often erroneously claims that an easy score set by the team is a tough score to chase. This phenomenon also extends to individual performances, where any bowler's performance is misrepresented as an economical spell, or phrases like "gave a good fight" are inaccurately attributed to the whole team or individual players, despite their actual performance.

Table 3 shows examples of the above observations. Complete documentation of the above work can be seen in our repository, where we annotate all the summaries generated by LM.

## 6 Limitation

Although the summary generated could be flawed, cricket summarizing could have multiple limitations. This limitation starts from the input data set.

Several conditions significantly impact a cricket match, such as pitch conditions, weather, and pressure situations, yet the scorecard doesn't reflect these. The pitch can influence the match outcome substantially, but details about its condition are absent. Similarly, weather conditions like rain or humidity, which can alter the course of a game, are not included. The pressure of the match situation, whether a tense chase or a dominant performance, isn't conveyed through numbers alone.

Player contributions in a cricket match extend beyond just batting or bowling performances. Outstanding fielding efforts, 'like saving crucial runs' or 'taking spectacular catches,' can significantly impact the game's outcome. Additionally, tactical decisions by the captain and support from the coaching staff, such as field placements, DRS (Decision Review System) calls, or strategic bowling changes, are crucial but not documented in the scorecard. The scorecard also fails to capture team

| Match | Generated Summary | Errors Explanation |
|---|---|---|
| Ind vs Aus, Final Nov 19, 02:00 PM | ...What a thrilling match! India's Rohit Sharma top-scored with 47 runs, but it was Rahul's 66 runs that kept us in the game ... | Rohit Sharma was not top-scorer, KL Rahul was the top scorer with 66 runs.<br>Reason: Rohit sharma was a top-order player, so possibly LM got confused with top-order and top-scored |
| Ind vs NZ, 1st Semi-Final Nov 15, 02:00 PM | ... Kane Williamson gave them a glimmer of hope, but Shami's triple-wicket haul and Bumrah's wicket-taking spell ensured India's dominance. .. | Shami took 7 wickets, not a triple-wicket haul, and not only Bumrah but also Siraj and Kuldeep took 1 wicket and Kuldeep's economy is also low.<br>Reason: The use of the triple wicket hall is unclear, Bumrah's name is mentioned above in the list be this could be the reason his performance is mentioned in the summary. |
| Ind vs RSA, 37th Match Nov 05, 02:00 PM | ...Jadeja picked up five wickets, including the crucial ones of Temba Bavuma, Rassie van der Dussen, and Kagiso Rabada, to finish with impressive figures ... | Jadeja didn't pick the wicket of Rassie van der Dussen.<br>Reason: The majority of wickets are taken by Jadeja, so the summary favors Jadeja. also, adjacent wickets were taken by Jadeja in the batting order. |
| Ind vs SL, 33rd Match Nov 05, 02:00 PM | a Shubman Gill and Kohli shared a 193-run partnership, while Shreyas Iyer's 82 and Ravindra Jadeja's 35 helped India post a massive total of 357/8. ... | Shubman Gill and Kohli didn't have a partnership of 193 runs, they had 189 run partnerships.<br>Reason: Shubman Gill's wicket fell when India's score was 193, maybe they correlated it with this data. |
| Ind vs Eng, 29th Match Oct 29, 2:00 PM | ...22 runs in his seven overs. Kuldeep Yadav and Ravindra Jadeja also chipped in with two wickets each, while Mohammed Siraj bowled economically, conceding just 33 runs in his six overs ... | Ravindra Jadeja didn't take 2 wicket in this match and comparatively Kuldeep(eco. 3) and Jadeja(eco. 2.3) were more economical.<br>Reason: Whenever two bowlers who are adjacent to each other in order were sometimes clubbed together, the words "each" or "both" were used, often leading to incorrect answers. |
| Ind vs Aus, 5th Match Oct 08, 02:00 PM | ... Hazlewood's 1 not out was a rare bright spot. Our bowlers, particularly Bumrah and Jadeja, did their best to restrict India, but Kohli's 85 and KL Rahul's 97 proved too much for us.... | A lot of error , Hazlewood performance in not noteworthy. Bumrah and Jadeja are from team India but are mentioned as performers of New Zealand.<br>Reason: Hazlewood was the last batter from Australia maybe this is the reason for putting him in the summary. Bumra was the best performer from the bowling side, maybe this is the reason for this error. |
| Ind vs Ban, 17th Match Oct 19, 02:00 PM | ...but ultimately couldn't prevent India from winning the match by 5 runs. It was a dominant performance by India, and they will ... | Information is correct throughout but representation is wrong. Gernally's team wins while chasing, his winning is told by a number of wickets. But here India's winning is told by the number of winnings. |
| Ind vs Ban, 17th Match Oct 19, 02:00 PM | ...scoring 51 off 43 balls and Das contributing 66 off 82 balls. Mehindy Hasan Miraz and Towhid Hridoy also chipped in with useful runs, while Mushfiqur Rahim ... | Mehindy Hasan Miraz didn't contribute noteworthy runs(3)<br>Reason: Mehindy Hasan Miraz gives the best bowling from Bangladesh, so his bowling performance dominated in summary so his batting performance got mentioned in summary. |

Table 3: This table summarizes common errors observed when requesting text summaries from a large language model. The first column identifies the match (i.e., match in ODI Worldcup 2023 men). The second column presents a snippet from the summary with the error highlighted in red. Finally, the last column provides the correct statement and proposes possible reasons for the error.

morale or the influence of specific players on team dynamics, which can be pivotal in determining the match's result.

Generating a summary using LM can misinterpret cricketing terms and vocabulary. In our experiments, cricket has a unique set of terms and jargon that the model may not accurately understand. For instance, terms like "top-order," "five-wicket hall," or "runout" have specific meanings in cricket, and a model might misinterpret these if they lack proper context. Additionally, subtle nuances and the significance of particular statistics or events in a match might not be fully captured, leading to summaries that miss critical details or convey incorrect information about the game's flow and key moments. Also, bias toward the batting team is noted at many points.

We can increase the number of trials within those chosen values to ensure generalizability. This allows for robust comparisons and reduces the risk of overfitting the model to specific conditions. Similarly, when working with text summarization tasks, we can leverage various prompting techniques to address common issues like the ones described in the above section. However, a key challenge lies in data coverage. The 2023 ODI World Cup scorecard did not encompass every cricketing situation an LM might encounter. This includes scenarios like the Duckworth-Lewis method (DLS) for rain-affected matches, super overs for tied scores, the "impact player" rule used in the IPL, and even the diverse formats themselves (Test matches, T20, the ultra-fast T10 format). Each scenario involves distinct cricket rules and regulations, and comprehensive data encompassing all these variations is crucial for training an LM to generate accurate summaries across the cricketing spectrum.

## 7 Conclusion and Future Scope

In this paper, we present an exploratory experiment designed to test the capability of language models to convert cricket scorecards into summaries. Our study underscores the potential errors in the generated summaries, focusing on issues related to completeness, correctness, and conciseness. We include qualitative examples of typical errors and explore potential reasons for their occurrence. After extensive analysis, we conclude that LM summaries can contain errors, necessitating cross-verification. Specifically, similar types of data can confuse models, and selecting appropriate adjectives can be challenging. Further observations are detailed in the observations section. Testing this hypothesis motivates future studies in table-to-text summarization within the sports domain.

This work opens several promising areas for future research in domain-specific table summarization. Analyzing the performance of high-parameter LMs on such tasks could lead to improvements in robustness and accuracy. As datasets grow and incorporate summaries with varying word counts and output lengths, more nuanced human evaluation metrics may be developed, offering deeper insights from human assessments. With the increasing size and capabilities of models, the ongoing advancement of LMs provides a valuable opportunity to enhance text summarization, translation, and content generation tasks. Fine-tuning LMs on domain-specific datasets, including sports, could unlock the potential for high-performing models that are less prone to errors than generic ones due to their deeper understanding of the specific domain.

## References


Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI@Meta. 2024. Llama 3 model card.

Xavier Amatriain. 2024. Prompt design and engineering: Introduction and advanced methods.

Ewa Andrejczuk, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, and Yasemin Altun. 2022. Table-to-text generation and pre-training with tabt5.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. *arXiv preprint arXiv:2004.10404*.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020b. Kgpt: Knowledge-grounded pre-training for data-to-text generation. *arXiv preprint arXiv:2010.02307*.

Fred Coalter. 2007. *A wider social role for sport: Who's keeping the score?* Routledge.

Michael A Covington. 2001. Building natural language generation systems. *Language*, 77(3):611–612.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rodney Fort and James Quirk. 1995. Cross-subsidization, incentives, and outcomes in professional team sports leagues. *Journal of Economic literature*, 33(3):1265–1299.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Eli Goldberg, Norbert Driedger, and Richard I Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

Nicholas L Holt, Kacey C Neely, Linda G Slater, Martin Camiré, Jean Côté, Jessica Fraser-Thomas, Dany MacDonald, Leisha Strachan, and Katherine A Tamminen. 2017. A grounded theory of positive youth development through sport based on results from a qualitative meta-study. *International review of sport and exercise psychology*, 10(1):1–49.

Fahrettin Horasan and Burhan Bilen. 2020. Extractive text summarization system for news texts. *International Journal of Applied Mathematics Electronics and Computers*, 8(4):179–184.

Altaf Hussain, Noman Khan, Muhammad Munsif, Min Kim, and Sung Baik. 2024. Medium scale benchmark for cricket excited actions understanding.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Hongxin Li. 2023. Jingran su, yuntao chen, qing li, and zhaoxiang zhang. sheetcopilot: Bringing software productivity to the next level through large language models. *arXiv preprint arXiv:2305.19308*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. Tapex: Table pre-training via learning a neural sql executor.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Milind Mahajan, Shreekar Kulkarni, Mitesh Kulkarni, Abhishek Sabale, and Aakanksha Thakar. 2024. Deep learning in cricket: A comprehensive survey of shot detection and performance analysis. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, 2(06):1748–1754.

Linyong Nan, Lorenzo Jaime Yu Flores, Yilun Zhao, Yixin Liu, Luke Benson, Weijin Zou, and Dragomir Radev. 2022a. R2d2: Robust data-to-text with replacement detection. *arXiv preprint arXiv:2205.12467*.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. 2022b. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Chinonso Cynthia Osuji, Thiago Castro Ferreira, and Brian Davis. 2024. A systematic review of data-to-text nlg.

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.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020a. 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.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020b. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6908–6915.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Darren ER Warburton, Crystal Whitney Nicol, and Shannon SD Bredin. 2006. Health benefits of physical activity: the evidence. *Cmaj*, 174(6):801–809.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2018. Learning neural templates for text generation. *arXiv preprint arXiv:1808.10122*.

Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. 2021. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. 2023. Data-copilot: Bridging billions of data and humans with autonomous workflow. *arXiv preprint arXiv:2306.07209*.

Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Ruizhe Chen, Xiangru Tang, Yumo Xu, Dragomir Radev, and Arman Cohan. 2023. Qtsumm: Query-focused summarization over tabular data.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.