

# SoccerRAG: Multimodal Soccer Information Retrieval via Natural Queries

Aleksander Theo Strand  
*OsloMet, TET Digital AS*  
Oslo, Norway  
0009-0008-2749-2347

Sushant Gautam  
*OsloMet, SimulaMet*  
Oslo, Norway  
0000-0001-9232-2661

Cise Midoglu  
*SimulaMet, Forzasys*  
Oslo, Norway  
0000-0003-0991-4418

Pål Halvorsen  
*OsloMet, SimulaMet, Forzasys*  
Oslo, Norway  
0000-0003-2073-7029

**Abstract**—The rapid evolution of digital sports media necessitates sophisticated information retrieval systems that can efficiently parse extensive multimodal datasets. In this paper, we introduce SoccerRAG, an innovative framework designed to harness the power of Retrieval Augmented Generation (RAG) and Large Language Models (LLMs) to extract soccer-related information through natural language queries. By leveraging a multimodal dataset, SoccerRAG supports dynamic querying and automatic data validation, enhancing user interaction and accessibility to sports archives. Our evaluations indicate that SoccerRAG effectively handles complex queries, offering significant improvements over traditional retrieval systems in terms of accuracy and user engagement. The results underscore the potential of using RAG and LLMs in sports analytics, paving the way for future advancements in the accessibility and real-time processing of sports data.

**Index Terms**—association football, information retrieval, large language models, multimodal data fusion, natural language processing, sports

## I. INTRODUCTION

The rapid growth of digital sports content has created a demand for efficient retrieval systems that can understand and process natural language queries [1], [2]. The motivation behind our research is to address the challenges in retrieving specific content from extensive sports libraries using intuitive, natural language requests. This approach simplifies user interaction and enhances the accessibility of sports archives. Soccer, as one of the most popular sports worldwide, serves as a prominent use case and context for our proof-of-concept retrieval application.

We propose SoccerRAG, a framework for retrieving multimodal soccer information using natural language queries, from an augmented soccer dataset based on SoccerNet [3], [4], which includes game videos with image frames and audio, timestamped captions (transcribed audio), annotations for game events, and player information. In short:

- We highlight the opportunities presented by the rapid advancements in Large Language Models (LLMs) and Retrieval Augmented Generation (RAG), and motivate the use of these technologies in the context of sports and more specifically for soccer analytics (Section II).
- We propose a framework concept and design for the retrieval of multimodal soccer information through natural queries, called SoccerRAG, which integrates components for data representation, feature extraction and validation,

and database querying. Our conceptualization includes a proposed database schema for an augmented version of the SoccerNet dataset, as well as a novel extractor-validator chain (Section III).

- We provide an open source implementation for SoccerRAG which is accessible under [5] and fully reproducible through the instructions provided therein.
- We present a preliminary evaluation of the SoccerRAG framework through the analysis of the extractor-validator chain, query complexity, individual component contributions, and execution time, for different LLMs (Section IV), and discuss our insights (Section V).

## II. BACKGROUND AND RELATED WORK

LLMs have revolutionized the field of Natural Language Processing (NLP) and generative AI [6]. The evolution of LLMs, from early rule-based models to contemporary GPT iterations, exemplifies significant technological advancements [7]. The capabilities of these models extend beyond text generation to include reasoning, decision-making, and multimodal data fusion, addressing both the static nature of early models and the computational demands of training and fine-tuning [8].

### A. Multimodal Soccer Understanding

Soccer, a globally beloved sport, serves as a rich domain for analytical exploration. The multifaceted aspects of soccer analytics are explored through the SoccerNet dataset and challenges [3], [4], a comprehensive repository of soccer broadcast recordings and various annotations, facilitating research in areas such as action spotting, video captioning, and game state reconstruction [9]. Midoglu et al. [10] focused on automatic event clipping, thumbnail selection, and game summarization through AI, to streamline the production of engaging soccer game highlights and summaries. There are interesting lines of research in soccer game understanding around event detection [11], [12], automatic game summarization [13], highlight generation [14], caption generation [15], [16], and similar. Most of this research demonstrates that the use of multiple modalities (including but not limited to videos, images, text, audio, and structured metadata such as commentary text, event information, and team statistics) can enhance game understanding [1], [2].

## B. Information Retrieval (IR)

Information Retrieval (IR) is focused on the extraction of relevant information from vast datasets based on user queries [17]. It encompasses a range of techniques including indexing, search algorithms, and natural language processing, to organize and retrieve unstructured data efficiently [18]. Recent advancements have leveraged machine learning and AI to enhance the precision of search outcomes through a better understanding of user intent and document semantics [19]. Despite its progress, IR continues to face challenges such as data volume management and privacy concerns, highlighting the need for innovative solutions in data retrieval [20].

## C. Retrieval Augmented Generation (RAG)

RAG systems represent a paradigm shift in open-domain question answering, combining document retrieval with generative modeling to produce contextually relevant answers [21], [22]. The utility of vector databases in RAG systems, particularly in efficiently managing and retrieving high-dimensional vector representations of text, underscores their significance in enhancing the accuracy and relevance of generated responses [23]. RAG has found application in different domains including enterprise contexts where LLMs are integrated with chatbots, enabling them to automatically derive more accurate answers from company documents and knowledge bases [24].

## D. Terminology

**LLMs** are advanced deep learning models capable of processing and generating human-like text [25]. Our research predominantly focuses on OpenAI's closed-source GPT-3.5-Turbo [26] and GPT-4.0-Turbo [27], although we acknowledge the broader landscape of available LLMs. **Tokens** are the atomic elements used by LLMs for text generation and prediction. Tokens can represent varying levels of linguistic units, from single characters to multiple words, crucial for modeling and API cost calculations [28]. Other concepts that are integral to LLM frameworks include **chains** which are sequences of tasks executed in a predefined order [29], and **agents** which are dynamic entities that use LLMs for decision-making and executing actions based on reasoning [30]. **Vector databases** handle high-dimensional data, such as text and image embeddings [31], enabling quick similarity searches [32] through Approximate Nearest Neighbor (ANN) algorithms. They are crucial for integrating with LLMs across various industries, allowing for complex queries over extensive datasets, offering a faster alternative to traditional database search methods [33]. **Extractors** are specialized mechanisms that harness the capabilities of LLMs for efficient information extraction, and are traditionally manual, rule-based processes [34]. They adapt to tasks via instructions or emulate JSON for structured output [35]. They offer tool calling for schema structuring, JSON mode for structured output, and prompting-based extraction for versatile retrieval, streamlining the costly traditional systems [36].

## E. Novelty

The burgeoning interest in RAG, fueled by the rapid advancements in LLMs, has paved new avenues for exploring innovative use cases across various domains, notably within multimodal information retrieval [37]. Despite the wide applicability of RAG frameworks in enhancing the capabilities of generative AI for open-domain question answering and beyond, its potential in the sports domain, particularly within soccer analytics, remains largely untapped. Our research endeavors to bridge this gap by harnessing the multimodal datasets inherent to the soccer broadcast pipeline. By integrating RAG with multiple data modalities, we aim to develop a sophisticated system capable of understanding complex game scenarios and responding to user queries in a manner that resonates with the intricate language of soccer. The potential of our proposed system for video retrieval offers a promising avenue for users and broadcasters alike, enabling the recollection of game moments through natural queries.

## III. PROPOSED FRAMEWORK

Figure 1 presents an overview of our proposed SoccerRAG framework, which includes four main components: **database**, **feature extractor**, **feature validator**, and **SQL agent**. The framework is intended for a scope where the dataset of interest is stored in a database with a predefined schema. The application flow is as follows: (1) The user provides a natural language query related to the contents of the database. (2) The user input is sent to an LLM along with the properties schema and a system prompt describing the properties the LLM should extract from the question. The LLM then returns a list of extracted properties relevant to the query. (3) Each extracted feature is checked against the appropriate table in the database using string matching algorithms. This step aims to correct spelling mistakes and abbreviations. Once a value is found, both the value and its primary key are added to the extracted value. (4) The cleaned user prompt is combined with system-specific prompts to guide the LLM in generating SQL queries that will answer the user's question. The constructed query is then passed to the SQL chain, which designs and executes the SQL queries against the underlying database. The SQL chain

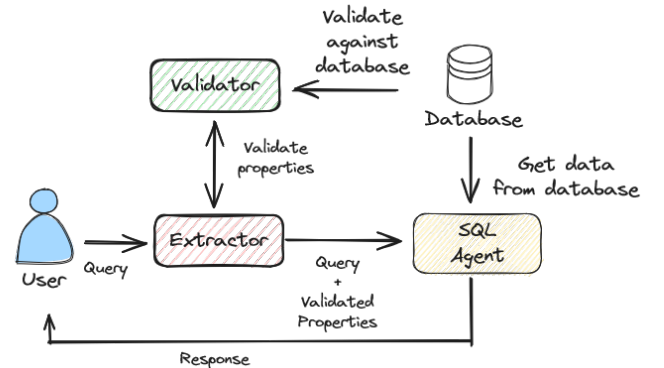


Fig. 1. Overview of the proposed SoccerRAG framework.

handles communication between the system and the database, retrieves the requested data, and prepares the results for presentation to the user. The complete SoccerRAG framework can be replicated in an end-to-end manner by running our open source software, which is accessible under [5], through the instructions provided. A demonstration is also available [38].

#### A. Database

The SoccerNet dataset consists of game broadcast videos and metadata information for 550 soccer games, from a number of top European leagues across multiple seasons. We have run Automatic Speech Recognition (ASR) on all games with available audio using Open AI's Whisper [39] to transcribe the audio, converting spoken commentary into textual data. Overall, for each game, we have files containing transcribed commentary ( $\langle 1/2 \rangle\_half-ASR.json$ ), annotations describing key events during the game such as goals, cards, fouls, etc. (*Labels-v2.json*), and general game information such as home and away teams, lineups, score, referee, and other details (*Labels-caption.json*). In the context of this work, an **event** refers to in-game events, **commentary** refers to the transcribed commentary from the game broadcast, and **caption** refers to a short description of an event.

To make this dataset usable for the SQL agent, we converted the JSON files into a structured SQLite database (detailed in Figure 2) with a script utilizing SQLAlchemy [5]. This script creates tables and columns based on a predefined schema, follows a systematic process of extracting league/season details, game details, lineups, events, commentary, and creates links between related entities. Clear and descriptive naming conventions for tables/columns were iteratively refined to improve the LLM's understanding of the data structure.

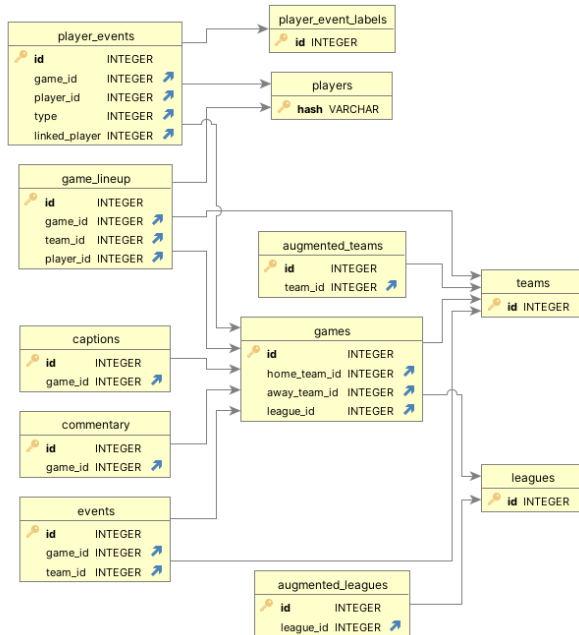


Fig. 2. Schema for the structured SQLite database.

#### B. Feature Extractor

In the extractor chain, instructions are provided to the LLM to return a JSON structure, where pre-defined properties (defined in the *schema.json* file under [5]) are extracted from the given prompt. We use the Langchain framework for its easy adjustability to new tasks. The schema also includes a list of the tables and columns a property exists in in the database, as a preparation for the validation function.

#### C. Feature Validator

Extracted features can have spelling errors or include abbreviations, rendering them disparate from the representations stored in the database. This discrepancy necessitates a validation procedure. To address the challenge of abbreviations (e.g., the natural query referring to "ManU" in reference to Manchester United F.C.), we manually constructed auxiliary tables within the database and populated them with abbreviations pertaining to common soccer team and player nomenclature (*augmented\_<teams/leagues>.csv* files under [5]). These supplementary tables are also delineated in the extractor schema for properties that warrant their employment.

The validation process commences by selecting the initial property key, which comprises a list of the extracted items for that property. If the property is associated with an auxiliary table, a search for a matching entry is performed within that table. In the event that no match is found, the system retrieves all items in the table linked to the property and computes the Levenshtein Distance [40] to ascertain the closest match for each item. Utilizing a predefined threshold for closeness, the system attempts to automatically select the optimal match, but may solicit user input for clarification in cases of ambiguity. Upon the validation of all items, the original prompt is returned with updated information containing the correct item designation, and where feasible, the corresponding foreign key in the database.

#### D. SQL Agent

The final component of the SoccerRAG framework leverages the SQL agent from the LangChain framework [41]. To enhance the capabilities of the agent, we implemented a few-shot SQL RAG solution. By employing this approach, the system prompt for the agent is consistently accompanied by N number of queries that are pertinent to the user's question, where N is externally configurable. To identify the matching queries, we employ vector search facilitated by a FIASS database. This database is populated with data originating from a JSON file (*sqls.json* under [5]), which contains a collection of human crafted SQL queries. The vector search technique entails embedding both the user's question and the SQL queries into high-dimensional vector spaces, thereby enabling the identification of semantically similar queries through the calculation of vector similarities.

By providing the agent with a few-shot context consisting of relevant SQL queries, we provide the user of the system a way of telling the system what data is expected to be retrieved for different questions. This approach leverages the strengths

of both retrieval and generation components, harnessing the domain knowledge encoded within the SQL query repository while enabling the agent to adapt and construct tailored queries based on the specific requirements of the user's input.

#### E. Choosing an LLM

LLMs are used in the feature extractor and SQL agent components of the SoccerRAG framework. In the LLM field, new open-source models such as LLaMA 2 [42], [43] and Mistral-7B [44] are rapidly emerging, but lack advanced function calling capabilities required by LangChain. While fine-tuning a model for feature extraction is viable, it demands substantial data and is time-consuming. Therefore, we decided on using OpenAI's GPT-4.0-Turbo [27] and GPT-3.5-Turbo [26] models.

### IV. EVALUATION

We use the following set of sample questions throughout our experiments to evaluate SoccerRAG and its components<sup>1</sup>:

- **Question 1:** Is Manchester United in the database?
- **Question 2:** Give me the total home goals for Bayern M in the 2014-15 season.
- **Question 3:** Calculate home advantage for Real Madrid in the 2015-16 season
- **Question 4:** How many goals did Messi score in the 15-16 season?
- **Question 5:** How many yellow-cards did Enzo Perez get in the 15-2016 season?
- **Question 6:** List all teams that played a game against Napoli in 2016-17 season in seriea? Do not limit the number of results
- **Question 7:** Give all the teams in the league ucl in the 2015-2016 season?
- **Question 8:** Give me all games in epl with yellow cards in the first half in the 2015-2016 season
- **Question 9:** What teams and leagues has Adnan Januzaj play in?
- **Question 10:** List ALL players that started a game for Las Palmas in the 2016-2017 season? Do NOT limit the number of results .
- **Question 11:** Did Ajax or Manchester United win the most games in the 2014-15 season?
- **Question 12:** How many yellow and red cards were given in the UEFA Champions League in the 2015-2016 season?
- **Question 13:** Are Messi and C. Ronaldo in the database?
- **Question 14:** How many goals did E. Hazard score in the game between Bournemouth and Chelsea in the 2015-2016 season?
- **Question 15:** How many yellow cards were given in the game between Bayern Munich and Shakhtar Donetsk in the 2014-15 UEFA Champions League, and did anyone receive a red card?
- **Question 16:** Make a list of when corners happened in the English Premier League (EPL) 2015-2016 season. Aggregate by a period of 15 minutes.
- **Question 17:** What league is Manchester United, Arsenal, Bournemouth, Real Madrid, Chelsea, and Liverpool in?
- **Question 18:** How many players have "Aleksandar" as their first name in the database, and how many goals have they scored in total?
- **Question 19:** What did the commentary say about the game between Arsenal and Southampton in the 2016-17 season?
- **Question 20:** Have Mesut Ozil, Pablo Insua, or Alex Pike played for West Ham or Barcelona?

#### A. Extractor-Validator Evaluation

The extractor-validator chain was evaluated on questions 1-20 subjectively, with the following evaluation criteria: (1) Were all properties extracted? (2) Were extra properties extracted? (3) Could extracted properties be validated? (4) Did the chain need human input? A perfect score (indicated with ● in Table I) was given if all properties were extracted and validated. If any properties were missed or validation failed, the result was marked as failed (▲). Extracting extra information or needing human feedback resulted in a 50% score (■).

GPT-3.5-Turbo was used for all questions except 15 and 18, which were re-run with GPT-4.0-Turbo. Table I shows

optimal performance for questions 2-10 and most of 11-20. In question 1, an extra item was extracted, without disrupting the rest of the chain. GPT-3.5-Turbo failed to extract one property in question 15, which GPT-4.0-Turbo handled correctly, this is demonstrated by presenting two results for Question 15. In Table I, Question 18 was a question designed to confuse the extractor by asking about players named "Aleksandar". The extractor extracts the name, and then tries to validate it against a *single* player. This shows that the chain does not understand context, as opposed to the SQL agent which is more content aware.

#### B. Query Complexity Analysis

To investigate the difficulty of composing database queries to represent natural language queries, query complexity scores were calculated for questions 1-10 using the Halstead metrics [46], which are frequently used to quantify the complexity of code snippets<sup>2</sup>. For each of the 10 questions, a database query was manually composed, and the number of unique operators (e.g., "SELECT", "FROM", "=") and operands (e.g., table and column names) were counted, as well as the total occurrences of operators and operands across the query. Halstead metrics were then calculated based on these counts. The results (presented in the middle part of Table I) show that question 3 had the highest complexity score. Further analysis revealed that this was due to the query for question 3 performing several *SUM* operations.

#### C. Ablation Study

By testing the framework components under different configurations, it is possible to quantify the contribution of each component to the final SoccerRAG framework. This information is crucial for understanding the strengths and weaknesses of the various parts of the system and identifying areas for potential improvement. We performed an extensive ablation study using the following pipeline configurations, where (6) corresponds to the full SoccerRAG framework.

- 1) **SQL agent only:** Sending the prompt directly to the SQL agent without RAG
- 2) **SQL agent with RAG:** Sending the prompt directly to the SQL agent with RAG
- 3) **Extractor only:** Sending the prompt with extracted values to the SQL agent
- 4) **Extractor and SQL RAG:** Sending the prompt with extracted values to the SQL agent, the SQL agent uses RAG to get examples
- 5) **Extractor and validator:** Sending the prompt with extracted and validated values to the SQL agent
- 6) **Extractor, validator, and SQL RAG:** Sending the prompt with extracted and validated values to the SQL agent, the SQL agent uses RAG to get examples

For the ablation study, questions 1-10 were used. A correct answer resulted in a pass score (indicated with ✓ in Table I), while an incorrect or incomplete answer resulted in a fail score (✗). The answers generated by the prompts created during the query complexity analysis (Section IV-B) were used to determine the success rate. The experimental setup involved using both GPT-3.5-Turbo and GPT-4.0-Turbo models to gain insights into their respective performance and to investigate the

<sup>2</sup>Halstead metrics are a set of software complexity measures based on counting the distinct operators and operands in source code, used to estimate code complexity, maintainability, and the effort required for development and maintenance.

<sup>1</sup>List available in [5], an extended set can also be found in [45].



TABLE I  
FRAMEWORK EVALUATION: EXTRACTOR-VALIDATOR EVALUATION (TOP), QUERY COMPLEXITY ANALYSIS (MIDDLE), ABLATION STUDY (BOTTOM).

Module Evaluation		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Extractor		■	●	●	●	●	●	●	●	●	●
Validator		●	●	●	●	●	●	●	●	●	●
Overall		●	●	●	●	●	●	●	●	●	●
		Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
Extractor		●	●	●	●	▲ / ●	●	●	■	●	●
Validator		●	●	●	●	● / ●	●	●	●	●	●
Overall		●	●	●	●	■ / ●	●	●	●	●	●
Query Complexity Metric		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Number of distinct operators, $n_1$		5	6	8	8	8	10	10	9	8	9
Number of distinct operands, $n_2$		5	14	31	26	26	31	20	36	28	25
Total number of operators, $N_1$		5	11	33	18	18	22	16	27	19	19
Total number of operands, $N_2$		5	16	110	26	26	38	22	41	31	26
Vocabulary, $n$		10	20	39	34	34	41	30	45	36	34
Length, $N$		10	27	143	44	44	60	38	68	50	45
Volume, $V$		33.22	116.69	755.81	223.85	223.85	321.45	186.46	373.45	258.50	228.94
Difficulty, $D$		2.50	3.43	14.19	4.00	4.00	6.13	5.50	5.12	4.43	4.68
Effort, $E$		83.05	400.09	10727.66	895.39	895.39	1970.20	1025.54	1913.91	1144.77	1071.42
Time to understand (s), $T$		4.61	22.23	595.98	49.74	49.74	109.46	56.97	106.33	63.60	59.52
LLM	Pipeline	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
GPT-3.5-Turbo	1	■	■	■	■	■	■	■	■	■	■
	2	■	■	■	■	■	■	■	■	■	■
	3	■	■	■	■	■	■	■	■	■	■
	4	■	■	■	■	■	■	■	■	■	■
	5	■	■	■	■	■	■	■	■	■	■
	6	■	■	■	■	■	■	■	■	■	■
GPT-4.0-Turbo	1	■	■	■	■	■	■	■	■	■	■
	2	■	■	■	■	■	■	■	■	■	■
	3	■	■	■	■	■	■	■	■	■	■
	4	■	■	■	■	■	■	■	■	■	■
	5	■	■	■	■	■	■	■	■	■	■
	6	■	■	■	■	■	■	■	■	■	■

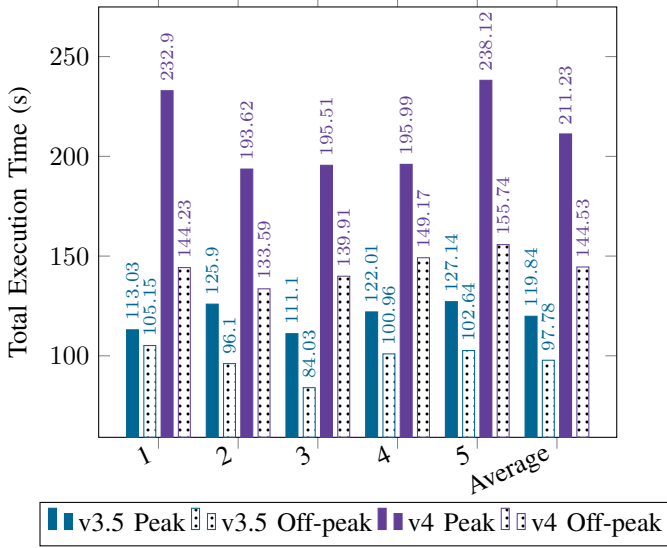


Fig. 3. Total execution time for questions 1-10 with GPT-3.5-Turbo and GPT-4.0-Turbo.

difference in quality, speed and cost between the two models. For the SQL-RAG, the value of  $K$  was set to 2.

As shown in Table I, the standalone SQL agent struggled with both models. The results indicate that the extractor-validator chain is crucial for increasing the hit rate. Interestingly, both models exhibited similar performance when the full pipeline was applied. GPT-4.0-Turbo failed on question 7 due to model laziness, where it provided only a partial answer.

For question 8, the query returned a substantial amount of text, causing the language model to refuse to provide a correct answer.

#### D. Execution Time

We investigated the execution time (total time from query until model output) for GPT-3.5-Turbo and GPT-4.0-Turbo. As the performance of cloud-based services might change based on global usage, we explored "peak" and "off-peak" periods separately (peak defined as 8AM-5PM EST and off-peak defined as 5PM-8AM EST). To observe and quantify the difference in execution time and response quality, we ran questions 1-10 consecutively, five times per batch, during peak and off-peak periods, with both models. Figure 3 presents our findings, which indicate that peak usage is related to working hours in the US, and that performance degrades during peak usage periods for both models, although GPT-3.5-Turbo is significantly faster than GPT-4.0-Turbo overall.

### V. DISCUSSION

#### A. Handling Input Errors

The extractor-validator chain has proven invaluable in mitigating spelling mistakes and other input errors and enhancing the overall accuracy of the system. Its functions can range from simple substitutions (e.g., heterogeneous season indicators in questions 5 and 6 into the uniform yyyy-yyyy format, or "seriea" in question 6 to "Serie A", as expected by the database) to resolving more complex issues. In the examples below, we see that when the LLM doesn't manage to resolve keywords, it will in many cases produce a wrong output.

**Prompt:** Did Real or Man U win the most games in the 2015 season?  
**Output:** Swansea won the most games in the 2014-2015 season with a total of 37 wins.  
**Cleaned prompt:** Did Real or Man U win the most games in the 2015 season?  
 Updated Information:  

- Real (now referred to as Real Madrid) has a primary key: 30.
- Man U (now referred to as Manchester United) has a primary key: 7.
- 2015 (now referred to as the 2015-2016 season).

**Output:** In the 2015-2016 season, Real Madrid won 26 games, while Manchester United won 5 games. Therefore, Real Madrid won the most games in that season.

## B. Retrieving Statistics

It is possible to extract detailed statistics using natural language queries with SoccerRAG, as demonstrated below.

**Prompt:** Calculate home advantage for Real Madrid in the 2015-16 season  
**Output:** For Real Madrid in the 2015-2016 season, the home advantage can be calculated as follows:  

- Total home goals scored: 55
- Total away goals scored: 39
- Number of home games played: 18
- Number of away games played: 17

 This indicates that Real Madrid scored significantly more goals at home compared to away games, suggesting a strong home advantage during the 2015-2016 season.

However, LLMs make mistakes with complex mathematics or large lists. We found that asking for a list often led to only a subset being returned (e.g., question 8, Prompt A), whereas asking for a count of events provided the correct answer (Prompt B=modified version of question 8).

**Prompt A:** Give all the teams in the league ucl in the 2015-2016 season  
**Output:** The teams in the UEFA Champions League (UCL) during the 2015-2016 season, limited to five examples, are:  

- 1) Galatasaray
- 2) Atletico Madrid
- 3) Manchester City
- 4) Juventus
- 5) Paris Saint-Germain (Paris SG)

**Prompt B:** Count all games in epl with yellow cards in the first half in the 2015-2016 season  
**Output:** There were 70 games in the English Premier League (england\_epl) with yellow cards in the first half in the 2015-2016 season.

## C. Influence of LLM

**Execution time:** We discovered that the OpenAI API works slower during peak periods, which follow the US timezone, with the execution time 22.56% slower for GPT-3.5-Turbo, and 46.5% slower for GPT-4.0-Turbo.

**Robustness:** Robustness refers to providing the same correct answer when presented with the same question multiple times. We did not observe significant degradation in the robustness of responses during peak periods. However, responses can exhibit variations across different runs, even for the same input. While strict prompting techniques can mitigate this inconsistency to some extent, the inherent stochastic nature of LLMs poses challenges in achieving complete determinism. To address this issue and gain greater control over the output, a solution would be to explore the usage of self-hosted LLMs for the framework. Self-hosted LLMs also provide users with control over the underlying network conditions, computational load, request prioritization, etc. that might influence the quality of the response.

**Model laziness:** The phenomenon where an LLM prematurely stops a task without fully completing it [47] can be

manifested in two ways. *Model stopping before querying:* While attempting to answer a question, the framework makes multiple requests to the LLM. We observed instances where the framework prematurely halted during this reasoning phase, responding to the user with "I now have the information I need, and should query the database for the answer." This was addressed by including "I will not stop until I query the database and return the answer." in the system prompt for the GPT-3.5-Turbo model, which seemed to fix most issues. The GPT-4.0-Turbo model did not exhibit this problem. (2) *Model returning a subset:* Both models faced this issue when asked to return a list. Despite fetching the correct data, the LLMs would often limit their response to 5-10 items. Stricter prompting did not resolve this issue, which is acknowledged by OpenAI [48].

## D. Future Work

SoccerRAG supports multimodal data integration by connecting metadata about game events with video timestamps, enabling several future applications. These include searching and retrieving specific video clips associated with particular game events using metadata, performing tasks such as object detection, action recognition, and scene understanding on video clips, and ultimately, automatically generating video highlights centered around game events of interest, or entire game summaries. SoccerRAG also faces several limitations that need addressing. First, the system struggles with complex queries and large data volumes. LLM model laziness might cause tasks to conclude prematurely or provide partial data, especially with extensive outputs. There is also a need to explore emerging LLMs, including open-source alternatives, for potential performance improvements with respect to function calling, language to query translation, and extraction abilities. Multimodal data synchronization and real-time processing capabilities require significant improvement to be viable for live sports analytics applications. Finally, ensuring privacy and ethical data handling is paramount, necessitating robust security measures and ethical guidelines in future developments.

## VI. CONCLUSION

We introduce SoccerRAG, a framework that leverages RAG and LLMs to efficiently retrieve multimodal soccer information via natural language queries, enhancing the accessibility of vast sports datasets by allowing for intuitive user interactions with complex data archives. Our evaluation demonstrates SoccerRAG's capability to accurately interpret complex queries and facilitate dynamic user engagements with multimedia content. Ultimately, SoccerRAG exemplifies the transformative potential of AI in sports analytics, promising a future where technology enriches the global fan experience and advances the accessibility of sports information.

## ACKNOWLEDGMENT

This research was partly funded by the Research Council of Norway, project number 346671 (AI-Storyteller).

## REFERENCES

- [1] S. Gautam *et al.*, “Assisting Soccer Game Summarization via Audio Intensity Analysis of Game Highlights,” in *Proc. 12th IOE Graduate Conference*. Inst. of Eng. Tribhuvan University, Nepal, Oct. 2022, pp. 25–32.
- [2] S. Gautam, “Bridging Multimedia Modalities: Enhanced Multimodal AI Understanding and Intelligent Agents,” in *ICMI '23: Proceedings of the 25th International Conference on Multimodal Interaction*. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 695–699.
- [3] S. Giancola *et al.*, “SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2018, pp. 18–22.
- [4] A. Delège *et al.*, “SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2021, pp. 19–25.
- [5] A. T. Strand *et al.*, “SoccerRAG: Multimodal Soccer Information Retrieval via Natural Queries,” <https://github.com/simula/soccer-rag>, Apr. 2024, Accessed: 2024-07-16.
- [6] J. Yang *et al.*, “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond,” *ACM Trans. Knowl. Discov. Data*, vol. 18, no. 6, Apr. 2024.
- [7] D. Myers *et al.*, “Foundation and large language models: fundamentals, challenges, opportunities, and social impacts,” *Cluster Computing*, vol. 27, no. 1, pp. 1–26, Feb. 2024.
- [8] Z. Xi *et al.*, “The Rise and Potential of Large Language Model Based Agents: A Survey,” *arXiv*, 2023.
- [9] S. Akan and S. Varlı, “Use of deep learning in soccer videos analysis: survey,” *Multimedia Systems*, vol. 29, no. 3, pp. 897–915, Jun. 2023.
- [10] C. Midoglu *et al.*, “MMSys’22 Grand Challenge on AI-based Video Production for Soccer,” *arXiv*, 2022.
- [11] L. Morra *et al.*, “Slicing and Dicing Soccer: Automatic Detection of Complex Events from Spatio-Temporal Data,” in *Image Analysis and Recognition*, A. Campilho, F. Karray, and Z. Wang, Eds. Cham, Switzerland: Springer International Publishing, Jun. 2020, pp. 107–121.
- [12] O. A. Norgård Rongved *et al.*, “Real-Time Detection of Events in Soccer Videos using 3D Convolutional Neural Networks,” in *2020 IEEE International Symposium on Multimedia (ISM)*, 2020, pp. 135–144.
- [13] S. Gautam *et al.*, “Soccer Game Summarization using Audio Commentary, Metadata, and Captions,” in *Proceedings of the 1st Workshop on User-Centric Narrative Summarization of Long Videos (NarSUM '22)*. New York, NY, USA: Association for Computing Machinery, 2022, pp. 13–22.
- [14] C. Midoglu *et al.*, “AI-Based Sports Highlight Generation for Social Media,” in *Proceedings of the 3rd Mile-High Video Conference (MHV '24)*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 7–13.
- [15] H. Mkhallati *et al.*, “SoccerNet-Caption: Dense Video Captioning for Soccer Broadcasts Commentaries,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2023.
- [16] J. Qi *et al.*, “GOAL: A Challenging Knowledge-grounded Video Captioning Benchmark for Real-time Soccer Commentary Generation,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. New York, NY, USA: Association for Computing Machinery, 2023, pp. 5391–5395.
- [17] G. G. Chowdhury, *Introduction to Modern Information Retrieval*. Facet Publishing, 2010.
- [18] K. Boukhari and M. N. Omri, “DL-VSM based document indexing approach for information retrieval,” *J. Ambient Intell. Hum. Comput.*, vol. 14, no. 5, pp. 5383–5394, May 2023.
- [19] J. Guo *et al.*, “A Deep Look into neural ranking models for information retrieval,” *Information Processing & Management*, vol. 57, no. 6, p. 102067, 2020.
- [20] A. Olteanu *et al.*, “FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval,” *SIGIR Forum*, vol. 53, no. 2, p. 20–43, Mar. 2021.
- [21] L. Liu *et al.*, “Challenges in Generalization in Open Domain Question Answering,” in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 2014–2029.
- [22] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *arXiv*, 2021.
- [23] W. Yu *et al.*, “A Survey of Knowledge-enhanced Text Generation,” *ACM Comput. Surv.*, vol. 54, no. 11s, Nov. 2022.
- [24] C. Jeong, “A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture,” *Advances in Artificial Intelligence and Machine Learning*, vol. 03, no. 04, pp. 1588–1618, 2023.
- [25] W. X. Zhao *et al.*, “A Survey of Large Language Models,” *arXiv*, Mar. 2023.
- [26] OpenAI, “GPT-3.5-Turbo,” <https://platform.openai.com/docs/models/gpt-3-5-turbo>, [Accessed: 2024-07-16].
- [27] —, “GPT-4,” <https://openai.com/gpt-4>, [Accessed: 2024-07-16].
- [28] A. Vaswani *et al.*, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 6000–6010.
- [29] T. Wu, M. Terry, and C. J. Cai, “AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. New York, NY, USA: Association for Computing Machinery, Apr. 2022.
- [30] X. Liu *et al.*, “AgentBench: Evaluating LLMs as Agents,” *arXiv*, 2023.
- [31] S. Ayesha, M. K. Hanif, and R. Talib, “Overview and comparative study of dimensionality reduction techniques for high dimensional data,” *Information Fusion*, vol. 59, pp. 44–58, Jul. 2020.
- [32] K. Echihabi, K. Zoumpatianos, and T. Palpanas, “New trends in high-D vector similarity search: ai-driven, progressive, and distributed,” *Proc. VLDB Endow.*, vol. 14, no. 12, pp. 3198–3201, Jul. 2021.
- [33] L. Friedman *et al.*, “Leveraging Large Language Models in Conversational Recommender Systems,” *arXiv*, 2023.
- [34] D. Xu *et al.*, “Large Language Models for Generative Information Extraction: A Survey,” *arXiv*, 2023.
- [35] K. Zhang, B. J. Gutiérrez, and Y. Su, “Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors,” *arXiv*, 2023.
- [36] LangChain, “Extracting structured output,” [https://python.langchain.com/docs/use\\_cases/extraction](https://python.langchain.com/docs/use_cases/extraction), Apr. 2024, [Accessed: 2024-07-16].
- [37] Y. Cheng *et al.*, “Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects,” *arXiv*, 2024.
- [38] A. T. Strand *et al.*, “Demo: Soccer Information Retrieval via Natural Queries using SoccerRAG,” *arXiv*, 2024.
- [39] Alec Radford and Jong Wook Kim and Tao Xu and Greg Brockman and Christine McLeavey and Ilya Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” *arXiv*, 2022.
- [40] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” *Soviet physics. Doklady*, vol. 10, pp. 707–710, 1965.
- [41] LangChain, “Agents,” [https://python.langchain.com/docs/use\\_cases/sql/agents](https://python.langchain.com/docs/use_cases/sql/agents), [Accessed: 2024-07-16].
- [42] H. Touvron *et al.*, “Llama 2: Open Foundation and Fine-Tuned Chat Models,” *arXiv*, 2023.
- [43] Meta, “Llama 2,” <https://llama.meta.com/llama2/>, [Accessed: 2024-07-16].
- [44] A. Q. Jiang *et al.*, “Mistral 7b,” *arXiv*, 2023.
- [45] A. T. Strand, “SoccerRAG: Multimodal Soccer Information Retrieval via Natural Queries,” Master’s thesis, Oslo Metropolitan University (OsloMet), 2024.
- [46] M. H. Halstead, *Elements of Software Science (Operating and programming systems series)*. USA: Elsevier Science Inc., 1977.
- [47] J. Wei *et al.*, “Measuring and Reducing LLM Hallucination without Gold-Standard Answers,” *arXiv*, 2024.
- [48] D. Coldewey, “OpenAI drops prices and fixes ‘lazy’ GPT-4 that refused to work,” <https://techcrunch.com/2024/01/25/openai-drops-prices-and-fixes-lazy-gpt-4-that-refused-to-work/>, Jan. 2024, [Accessed: 2024-07-16].