Literature Survey

Sports History and Records Archives (Group 45)

AUTHORS:
VANSHAJ SHARMA(MT23103)
BHARAT NAGDEV (MT23029)
PULKIT RIHANI (MT23066)
SIDDHANT JHA (MT23097)
LOKESH SAINI (MT23120)
RITESH RAJPUT (MT23075)
Github : https://github.com/siddhantJH/Information-Retrieval-Project

Problem formulation

The identified problem revolves around the lack of a dedicated chatbot focused on retrieving sports history and records, despite the increasing interest and importance of memorable sports events. Our project highlights the need for an efficient and optimized chatbot solution capable of providing users with relevant sports history information through simple interactions. Our proposed solution is utilizing the power of LLMs and RAG model to provide rank based retrieval and relevant results by understanding user's need by feedback mechanism. So far no rank based retrieval model along a chatbot has been developed for the domain and based on our project our system will be leveraging the rank based retrieval of url for query specific tasks along with the url it will extract a short summary for the same along with a special event date and its headline.

Problem Importance

The popularity of sports has surged in recent years, with memorable events becoming cultural touchstones. However, preserving these moments and understanding sports history is crucial. Yet, there's a lack of efficient chatbots dedicated to providing this information, creating a gap in accessibility for enthusiasts.

Our project seeks to address this gap by developing a chatbot focused on sports history retrieval. By leveraging technologies like RAG and LLM, we aim to create an efficient system that delivers insightful information with just a few clicks. This initiative not only enhances accessibility for sports enthusiasts but also contributes to the broader landscape of historical research and sharing knowledge.

Related Work

In the realm of sports history research, there has been a noticeable absence of studies dedicated to capturing the memorable moments in sports history. Instead, the focus has largely been on managing sports information for easy access by end users. Various systems and methodologies have been developed over the years to facilitate this, starting with initiatives like SportsBR and ASMS in the mid-2000s, which focused on video browsing and context-based management, respectively. Later advancements introduced semantic analysis based on ontology and SPARQL, followed by efforts to address data clustering challenges with systems like SSIE. More recent research has delved into video-based retrieval using Big Data analysis and deep learning techniques such as CNNs to enhance retrieval accuracy. These efforts have also seen the development of databases like SportsDB, providing structured sports information to support efficient data management and retrieval.

Despite the advancements in sports data management, there remains a notable gap in research concerning the preservation and retrieval of significant moments in sports history. Previous studies have primarily focused on enhancing accessibility to sports information for users, utilizing various systems and techniques over the years. From early initiatives like SportsBR and ASMS to more recent developments involving semantic analysis and deep learning, the emphasis has been on improving the organization and retrieval of sports data. These efforts have led to the creation of databases like SportsDB, which aim to provide structured sports information to aid in

efficient data management and retrieval. However, there is still room for dedicated research into capturing and preserving the rich history of sports through innovative methodologies and systems. Despite so many attempts till now no model has used RAG and LLM and Machine learning based retrieval systems to provide the rank based retrieval . But our proposed solution solved this problem by using RAG and LLM and reinforcement learning for feedback mechanisms.

References to Related Work

1. SportsDB is an open-source database designed to provide structured sports information for efficient data management and retrieval.
   https://www.thesportsdb.com/

2. This paper presents the design and implementation of the Athletic Sport Management System (ASMS), aiming to develop a computerized system to assist in planning and managing athletic sport activities through context-based information retrieval and enhanced searching capabilities.
   https://www.academia.edu/77319254/Context_Based_Information_Retrieval_of_AthleticSport_Management_System_ASMS_

3. This paper discusses the implementation of semantic retrieval for sports information on the Semantic Web, emphasizing the development of a sports ontology model and a semantic retrieval system using SPARQL query language to enable intelligent retrieval based on semantic relations between sports concepts.
   https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5572265

4. This paper introduces a Summary Sport Information Extraction System (SSIE) aimed at extracting and summarizing statistics from web documents related to athletics.
   https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7363064

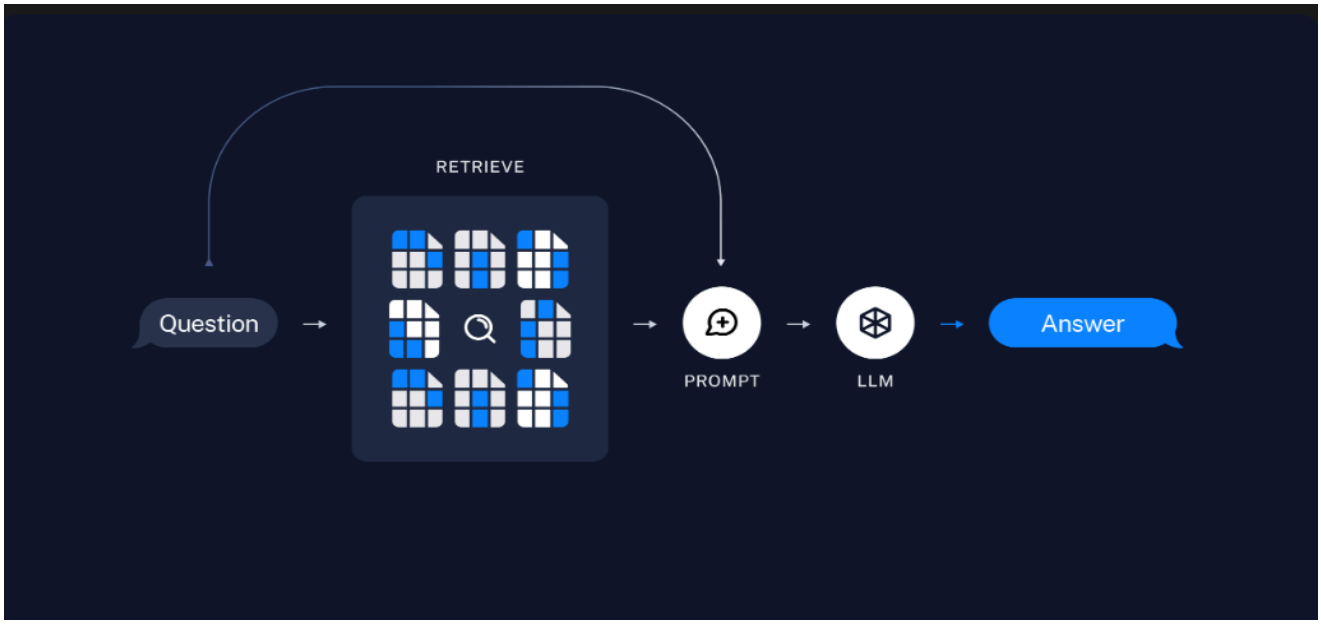
5. This paper presents SportsBR, an advanced sports video browsing and retrieval system leveraging multimodal analysis for event-based browsing and keyword-based retrieval of sports videos.
   https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1527838

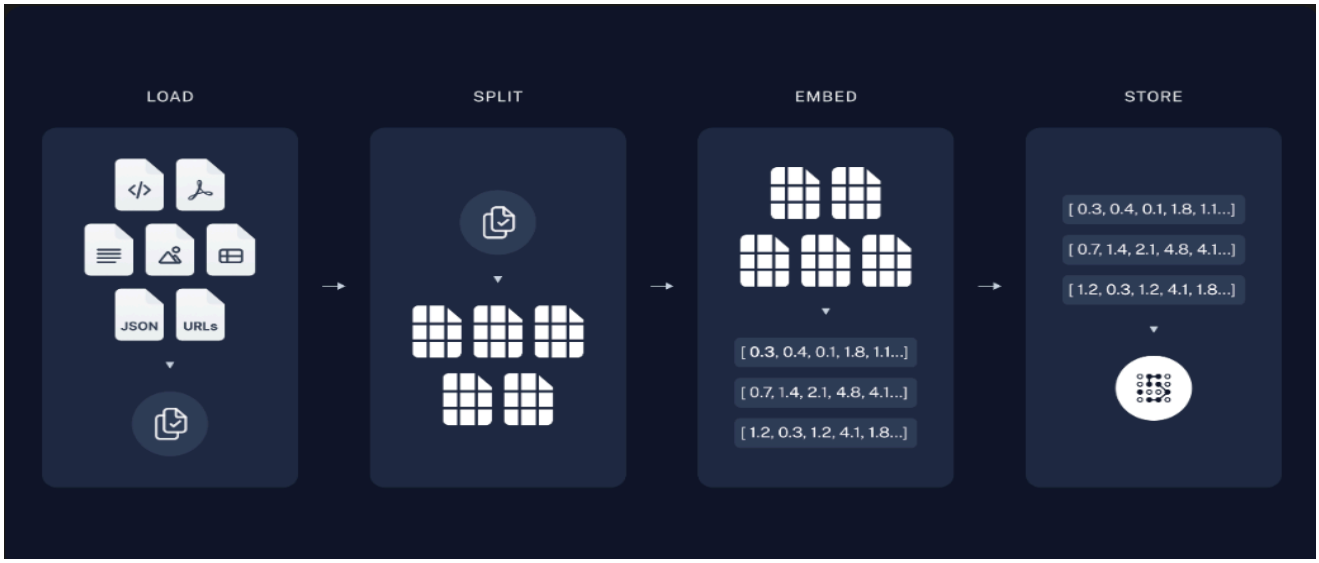
6. This paper introduces an integrated infrastructure for team sports analysis, seamlessly combining real-time data stream analysis and sketch-based video retrieval.
   https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8622592

Methodology

The project's framework has been carefully designed to achieve its intended goals. Our main objective is to seamlessly incorporate a Retrieve and Generate (RAG) based Language Model (LLM) to extract concise summaries related to significant sports moments. In developing the retrieval system, we are prioritizing the implementation of Gemini Pro LLM.

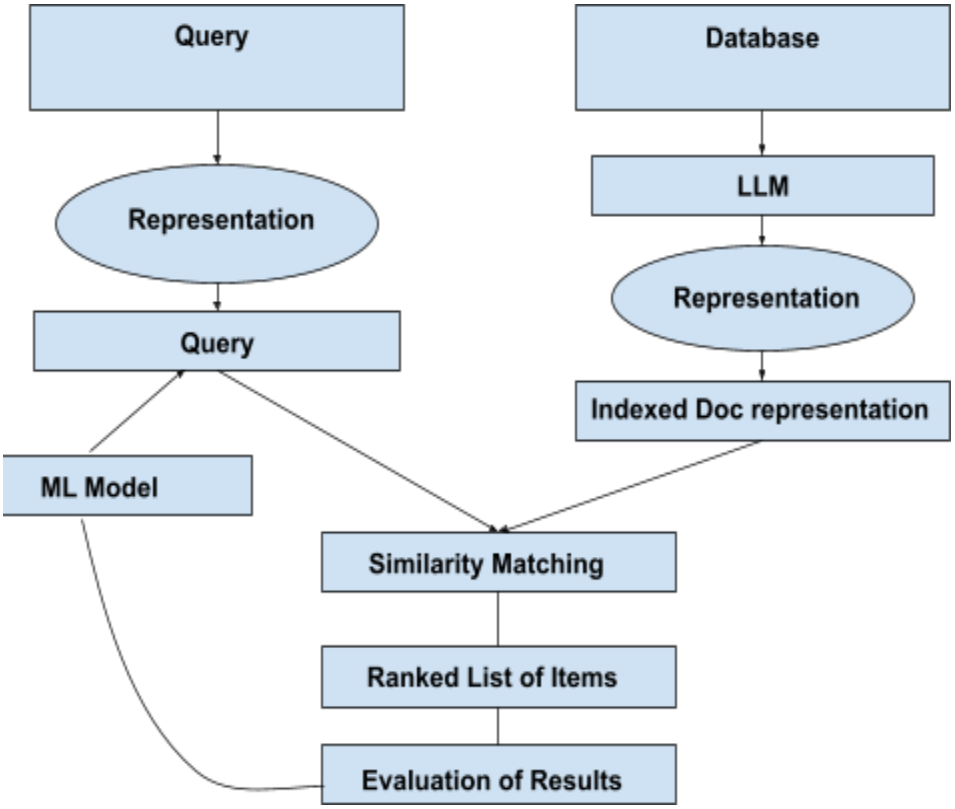
- **Integration of RAG-based LLM Model:** The project focuses on integrating a Retrieve and Generate (RAG) based Language Model (LLM) for the retrieval of summaries related to sports memorable events. The RAG model combines the capabilities of retrieval-based models and generative models to provide more accurate and contextually relevant summaries.
- **Formulation of the Retrieval System:** The project formulates a retrieval system named Gemini Pro LLM, leveraging advanced natural language processing (NLP) techniques. This system aims to enhance the retrieval process by incorporating features such as semantic understanding, context-awareness, and summarization capabilities.
- **Implementation of Langchain RAG Model:** The Langchain RAG model is implemented as part of the project to facilitate the retrieval and generation of summaries from a given input query. This model utilizes a hierarchical structure to capture both local and global context information, resulting in more coherent and informative summaries.
- **Google Colab RAG Model Implementation:** The project utilizes Google Colab as the platform for implementing the RAG model.
- **Data Sources:** The project leverages data from reputable sources such as Wikipedia ("Sport in India") and Indianetzone ("Memorable Events in Indian Cricket") to train and evaluate the RAG-based LLM model.
  https://en.wikipedia.org/wiki/Sport_in_India
  https://www.indianetzone.com/28/memorable_events_indian_cricket.htm
- **Gradio Frontend:** The project incorporates a Gradio frontend interface to facilitate user interaction with the retrieval system.
- **Frontend Development:** The project involves the development of a frontend interface using HTML, CSS, and JavaScript. This frontend interface is designed to provide a visually appealing and intuitive user experience, allowing users to interact seamlessly with the retrieval system.
- **Backend Development:** The project implements a backend server using Node.js and Express.js to handle requests from the frontend interface. The backend server manages the retrieval process, communicates with the RAG model, and delivers the generated summaries to the frontend for display.





- Use the loading function like webloader provided by langchain
- Use the text splitter mechanism like Recursive provided by langchain
- Converting splitted data to vector embedding by gemini embedding
- used Chroma db vector store to store the vector embeddings and convert into vector index
- This vector index and gemini pro LLM will be used in RAG chain pipeline provided by langchain to return the output of the query given by the user

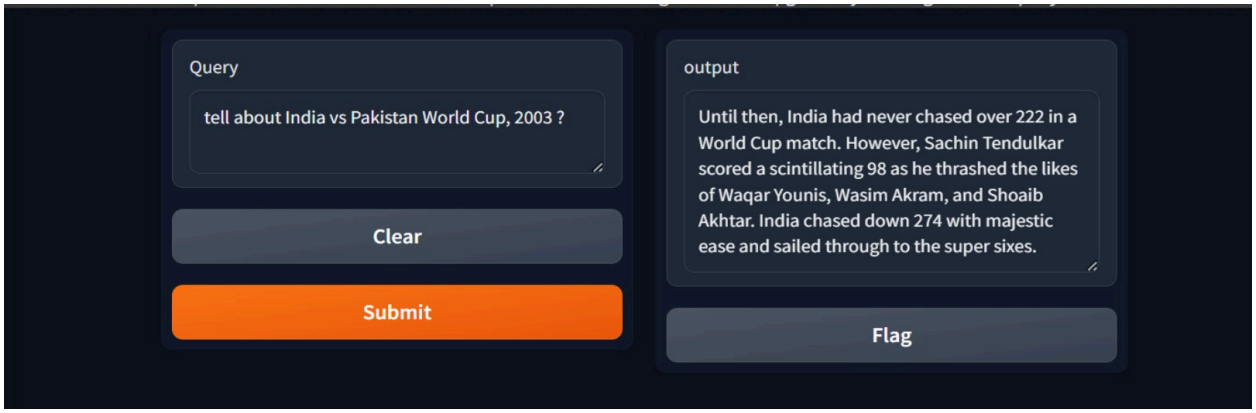
Project Pipeline

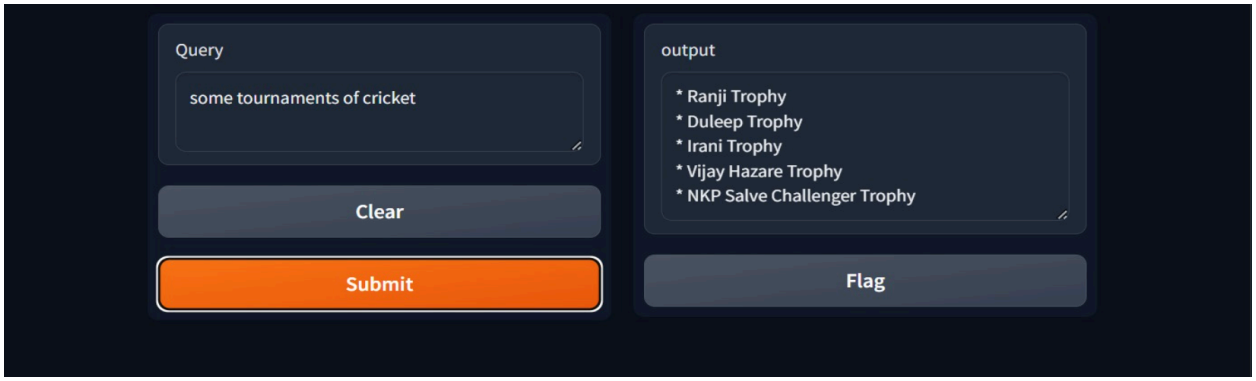


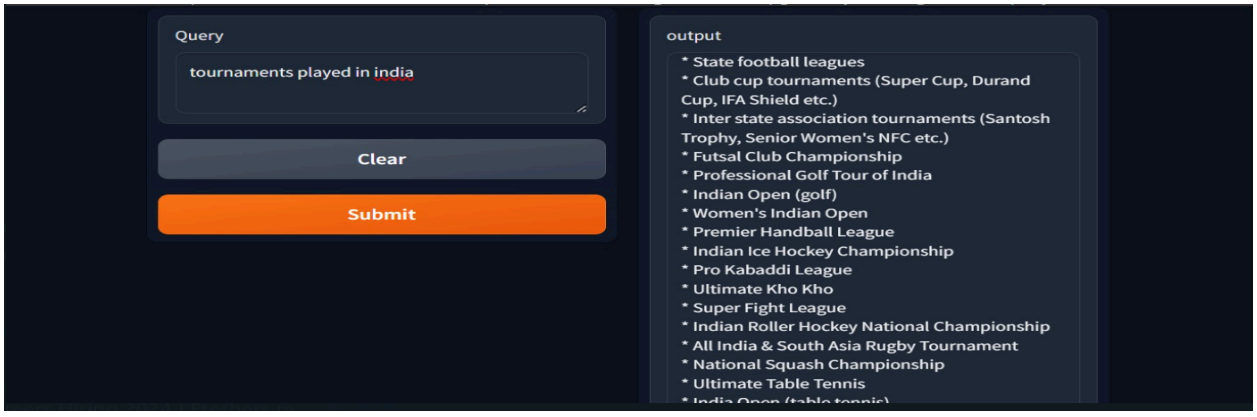
The pipeline of our project shown above is explained below:

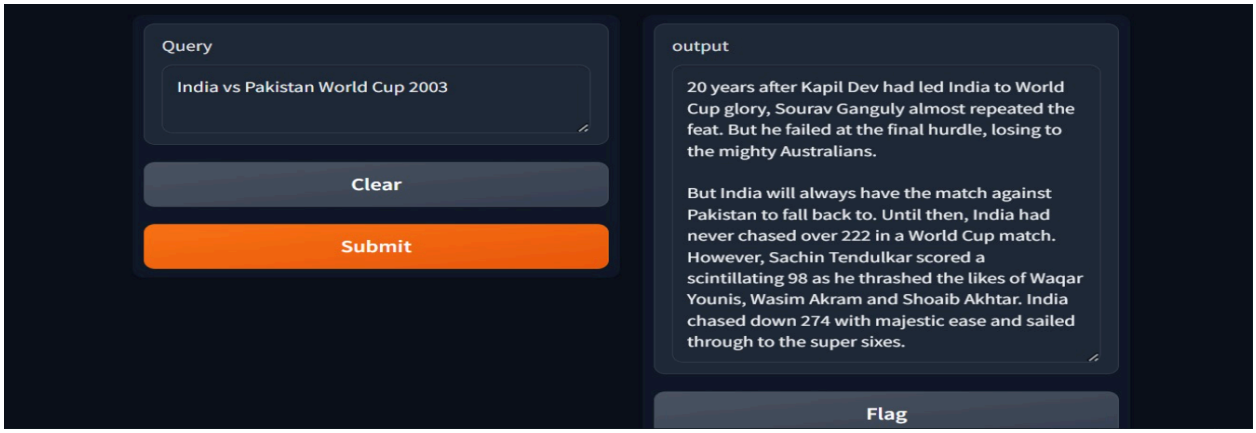
- Text Query from user
- Changing representation as per internal representation
- Giving Query to RAG model
- RAG model the returns the most relevant documents from vector database prepared from a large database built using web scraping on data available on the web related to sports events and send it to LLM
- Giving Query and Documents returned from the RAG model to fine tuned LLM
- Converting results from LLM to internal representation
- Similarity matching
- Ranking results based on knowledge graph , user cache and user search history
- Based on users direct and indirect feedback evaluation will be done
- Using Machine learning model we will be updating query
- Again fetching results from our information retrieval model based on updated query
- Showing results to user again

Results


Query
tell about India vs Pakistan World Cup, 2003 ?
Clear
Submit
output
Until then, India had never chased over 222 in a World Cup match. However, Sachin Tendulkar scored a scintillating 98 as he thrashed the likes of Waqar Younis, Wasim Akram, and Shoaib Akhtar. India chased down 274 with majestic ease and sailed through to the super sixes.
Flag


Query
some tournaments of cricket
Clear
Submit
output
* Ranji Trophy
* Duleep Trophy
* Irani Trophy
* Vijay Hazare Trophy
* NKP Salve Challenger Trophy
Flag


Query
tournaments played in india
Clear
Submit
output
* State football leagues
* Club cup tournaments (Super Cup, Durand Cup, IFA Shield etc.)
* Inter state association tournaments (Santosh Trophy, Senior Women's NFC etc.)
* Futsal Club Championship
* Professional Golf Tour of India
* Indian Open (golf)
* Women's Indian Open
* Premier Handball League
* Indian Ice Hockey Championship
* Pro Kabaddi League
* Ultimate Kho Kho
* Super Fight League
* Indian Roller Hockey National Championship
* All India & South Asia Rugby Tournament
* National Squash Championship
* Ultimate Table Tennis


Query
India vs Pakistan World Cup 2003
Clear
Submit
output
20 years after Kapil Dev had led India to World Cup glory, Sourav Ganguly almost repeated the feat. But he failed at the final hurdle, losing to the mighty Australians.
But India will always have the match against Pakistan to fall back to. Until then, India had never chased over 222 in a World Cup match. However, Sachin Tendulkar scored a scintillating 98 as he thrashed the likes of Waqar Younis, Wasim Akram and Shoaib Akhtar. India chased down 274 with majestic ease and sailed through to the super sixes.
Flag

Evaluation Parameters : Usefulness of Method

Citation :- https://analyticsindiamag.com/metrics-for-reinforcement-learning/ , https://aisera.com/blog/llm-evaluation/

1. Query Optimization using Machine Learning
    - Dispersion across Time (DT)
    - Short-term Risk across Time (SRT)
    - Long-term Risk across Time (LRT)
    - Dispersion across Runs (DR)
    - Risk across Runs (RR)
    - Dispersion across Fixed-Policy Rollouts (DF)
    - Risk across Fixed-Policy Rollouts (RF)

2. Data Retrieval Using Large Language Model
    - Reliability
    - Efficiency
    - Bias Detection
    - User Trust
    - Fine-Tuning

3. Rank Based Retrieval
    - F1 score
    - Precision
    - Recall
    - NDCG