

Information Retrieval and Text Mining

Project 2

Alexandru-Radu Handac
alexandru-radu.handac@s.unibuc.ro

Abstract

This project explores the use of Retrieval-Augmented Generation (RAG) to combine information retrieval (IR) and large language models (LLMs) for Romanian football-related queries. By evaluating IR-only, LLM-only, and IR + LLM approaches, the results show that RAG ensures more accurate and contextually relevant responses by leveraging the strengths of both components. Additionally, experiments on prompt design and temperature settings demonstrate their impact on response quality and variation.

1 Introduction

Retrieval-Augmented Generation (RAG) is an effective method for solving tasks by integrating information retrieval (IR) with generative language models (LLMs). This project investigates the differences between IR-only, LLM-only, and IR + LLM approaches in terms of accuracy and reliability. IR-only systems excel at retrieving relevant context but cannot provide detailed or complete answers. On the other hand, LLM-only models can generate fluent responses but often struggle with factual accuracy or hallucination. By combining the strengths of both, the RAG approach produces responses that are more accurate, contextually relevant, and grounded in reliable information.

2 Related Work

There has been growing interest in language models for underrepresented languages like Romanian. OpenLLM-Ro (Masala et al., 2024) introduces RoLlama2, a model designed for Romanian using the Llama2 architecture. It addresses challenges like limited training data and offers foundational and instruction-tuned versions, improving Romanian-specific tasks.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) combines retrieval and language generation to handle knowledge-intensive tasks. RAG

retrieves relevant documents and integrates them into the generation process, improving accuracy and providing factual and explainable results for tasks like question answering.

3 Method

Our method leverages a retrieval-augmented generation (RAG) approach, combining a sentence transformer model for document retrieval and OpenLLM-Ro (Masala et al., 2024) for response generation. This design aligns with the structure of existing RAG frameworks, such as those proposed by Lewis et al. (2020) (Lewis et al., 2020), which demonstrated the advantages of combining non-parametric retrieval with parametric language models.

3.1 Dataset

The dataset consists of 7 text files, primarily focused on Romanian football. These files provide information about notable football players, teams, rumors about player transfers, and descriptions of key personalities such as Gheorghe Hagi and Sir Alex Ferguson. The resources were sourced from the Instagram page "gensportmedia," which specializes in sports news, and Wikipedia for details about the personalities.

The total dataset size is approximately 15,233 characters, 2963 words, and 116 sentences.

Expectations: Considering the small size and specific focus of the dataset, the model is expected to perform well for well-defined football-related queries. However, it may struggle with ambiguous or broader questions due to the limited coverage of the dataset. The retrieval-augmented approach (IR + LLM) is likely to improve factual accuracy by dynamically retrieving relevant information from the dataset before generating responses.

3.2 Preprocessing

Minimal preprocessing was applied to the dataset. Text files were split into paragraphs using the new-line delimiter (`\n\n`) to divide the content into smaller, contextually meaningful chunks. No additional steps, such as lowercasing, stopword removal, or stemming, were performed.

3.3 Method and Results

Step 1: Dataset Preparation. We processed Romanian football-related text by splitting documents into paragraphs for granular retrieval, ensuring optimized semantic similarity and more precise retrieval.

Step 2: Information Retrieval. A sentence transformer model was used to generate embeddings for documents. Top- n documents were retrieved using cosine similarity, allowing the system to focus on the most relevant context for each query.

Step 3: LLM Selection. The generative component utilized a Romanian-specific large language model, optimized for the domain and loaded in a quantized 4-bit precision format for faster inference without sacrificing response quality.

Step 4: Ensemble Design. Retrieved documents were passed to the LLM using tailored prompts that explicitly instructed the model to minimize hallucination and focus on fact-based answers. This integration ensured accurate, contextually grounded responses.

ROUGE Score Evaluation. The IR + LLM responses were evaluated using the ROUGE-L metric. For example:

- **Question:** *A marcat vreodata Banel Nicolita un autogol?*
- **Ground Truth:** *Da, Banel Nicolita a marcat un autogol impotriva echipei Real Madrid.*
- **IR + LLM Response:** *Da, Banel Nicolita a marcat un autogol in timpul unui meci din Liga Campionilor 2006-2007 dintre Real Madrid si FCSB.*
- **ROUGE-L Score:** 0.5625

Comparison of Approaches. To demonstrate the differences between the IR-only, LLM-only, and IR + LLM approaches, we used the following question:

Question: *Cine sunt inculpatii in dosarul Valiza?*

- **Information Retrieval (IR) Response:**

Totuși, DNA a declanșat urmărirea penală față de Becali, Teia Sponte și alții, iar dosarul a ajuns în instanță la sfârșitul anului 2008. Primul termen a avut loc la 21 ianuarie 2009 la Curtea de Apel Cluj. În continuare, procesul a fost mutat la Înalta Curte de Casație și Justiție, ca urmare a cererii avocaților inculpaților.

- **LLM-Only Response:** *Ion Mihalache, Ion Popescu-Muscel, Gheorghe Tatarascu, Alexandru Marghiloman, Constantin Argețoiu, Alexandru Vaida-Voevod, Ion Bratianu, Ion I.C. Bratianu, Gheorghe Tatarascu, Alexandru Marghiloman, Constantin Argețoiu, Alexandru Vaida-Voevod, Ion Bratianu, Ion I.C. Bratianu.*

- **IR + LLM Response:** *George Becali, Teia Sponte si Victor Piturca.*

These results highlight the advantages of the IR + LLM approach. While IR alone retrieves relevant but incomplete context, and LLM alone may generate incoherent or inaccurate outputs, the combined approach leverages IR to provide grounded context and LLM to generate coherent, accurate answers.

Temperature Tuning. To evaluate the impact of different temperature values on the quality and variation of the generated responses, we tested the system with a range of temperatures. The results for the question *Din ce cauza a demisionat Dorinel Munteanu?* are as follows:

- **Response with temperature 0.1:** *Dorinel Munteanu si-a dat demisia pentru ca era nemulțumit de problemele financiare ale echipei.*
- **Response with temperature 0.2:** *Dorinel Munteanu si-a dat demisia pentru ca era nemulțumit de problemele financiare ale echipei.*
- **Response with temperature 0.5:** *Dorinel Munteanu și-a dat demisia pentru ca era nemulțumit de problemele financiare ale echipei.*
- **Response with temperature 0.7:** *Dorinel Munteanu și-a dat demisia pentru că era nemulțumit de problemele financiare ale echipei.*
- **Response with temperature 1.0:** *Dorinel Munteanu si-a dat demisia din functia de*

antrenor al echipei Otelul Galati pentru ca era nemultumit de situatia financiara a echipei.

Analysis: Lower temperature values (e.g., 0.1 and 0.2) produced more deterministic responses, with minimal variation across outputs. As the temperature increased, the system generated more diverse and detailed responses. For instance, at temperature 1.0, the model provided a more elaborate answer, including specific details about the context.

4 Future Work

Future developments of this project could include deploying the system as a chatbot for Romanian football enthusiasts, capable of answering user queries in real time. In addition, fine-tuning the model on a specific football domain dataset could improve accuracy and reduce hallucinations.

5 Conclusion

This project demonstrates the effectiveness of Retrieval-Augmented Generation (RAG) for addressing knowledge-intensive tasks. The IR + LLM approach outperforms IR-only and LLM-only setups by combining accurate retrieval with coherent generation. While the system performs well for Romanian football-related queries, its scalability and adaptability to other domains or languages require further investigation. Future work could focus on fine-tuning for specific datasets and deploying the system as a chatbot for real-time applications.

Limitations

This project was limited by the inability to utilize multiple GPUs in the Kaggle environment and the 16GB VRAM constraint, which impacted inference speed and scalability. While the system can easily accommodate more text files, faster responses and larger datasets would require better GPU resources. Additionally, extending the system to other languages would need fine-tuning, as the current model is optimized for Romanian text.

Ethical Statement

This project could potentially be misused to spread misinformation if biased or incomplete datasets are provided, as the model generates responses based solely on the input data. Bias may also arise from the dataset, which focuses on Romanian

football content sourced from "gensportmedia" and Wikipedia.

References

- Patrick Lewis, Ethan Perez, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.
- Mihai Masala, Denis C. Ilie-Ablachim, Dragos Corlatescu, et al. 2024. Openllm-ro - technical report on open-source romanian llms. *arXiv preprint arXiv:2405.07703*.

A Example Appendix

List of Files:

- banel.txt
- dorinel_pleaca.txt
- dosarul_valiza.txt
- dragusin_psg.txt
- hagi.txt
- sir_alex.txt
- steaua_qarabag.txt

List of Queries:

- A marcat vreodata Banel Nicolita un autogol?
- Din ce cauza a demisionat Dorinel Munteanu?
- Care este suma de transfer zvonita pentru Radu Dragusin?
- Cate titluri de campioana a Angliei a castigat Sir Alex Ferguson cu Manchester United?
- Cine sunt inculpatii in dosarul valiza?
- Care este porecla lui Gica Hagi data de catre romani?
- Este adevarat ca a castigat echipa FCSB impotriva celor de la Qarabag?

List of Prompts:

Folosind doar informatiile oferite mai jos, raspunde clar si exact la intrebare.

Nu ghici si nu adauga informatii din afara sursei. Foloseste textul exact acolo unde este relevant.

Exemple de raspunsuri corecte:

- Daca intrebarea este 'Cine este...', raspunsul tau ar putea fi 'Persoana care a facut ceva' sau 'Num...
- Daca intrebarea este 'Care este...', raspunsul tau ar putea fi 'Este...'.
- Daca intrebarea este 'Unde este...', raspunsul tau ar putea fi 'Locatia este...'.

Informatii disponibile:

{docs}

Intrebare: {query}

Raspuns:

Folosind doar informatiile oferite mai jos, raspunde clar si exact la intrebare.\n

f"{context}"

Intrebare: {query}\n

Raspuns:

Results:

- **Question:** A marcat vreodata Banel Nicolita un autogol?

Ground Truth: Da, Banel Nicolita a marcat un autogol impotriva echipei Real Madrid.

IR + LLM Response: Da, Banel Nicolita a marcat un autogol in timpul unui meci din Liga Campionilor 2006-2007 dintre Real Madrid si FCSB.

ROUGE-L: 0.5625

- **Question:** Din ce cauza a demisionat Dorinel Munteanu?

Ground Truth: Dorinel Munteanu a demisionat din cauza unor probleme financiare ale clubului.

IR + LLM Response: Dorinel Munteanu a demisionat din cauza problemelor financiare ale echipei.

ROUGE-L: 0.7619

- **Question:** Care este suma de transfer zvonita pentru Radu Dragusin?

Ground Truth: PSG ar putea plati 50 de milioane de euro pentru transferul lui Radu Dragusin.

IR + LLM Response: 50 de milioane de euro.

ROUGE-L: 0.5263

- **Question:** Cate titluri de campioana a Angliei a castigat Sir Alex Ferguson cu Manchester United?

Ground Truth: Noua titluri de campioana a Angliei.

IR + LLM Response: Noua titluri de campioana a Angliei a castigat Sir Alex Ferguson cu Manchester United.

ROUGE-L: 0.6000

- **Question:** Cine sunt inculpatii in dosarul Valiza?

Ground Truth: Inculpatii din dosarul Valiza sunt George Becali, Teia Ponte si Victor Piturca.

IR + LLM Response: George Becali, Teia Sponte si Victor Piturca.

ROUGE-L: 0.6316

- **Question:** Care este porecla lui Gica Hagi data de catre romani?

Ground Truth: Porecla acestuia este regele.

IR + LLM Response: Regele

ROUGE-L: 0.1091

- **Question:** Este adevarat ca a castigat echipa FCSB impotriva celor de la Qarabag?

Ground Truth: Da, FCSB a castigat impotriva celor de la Qarabag.

IR + LLM Response: Da, FCSB a castigat impotriva celor de la Qarabag.

ROUGE-L: 1.0000