FRI Natural language processing course 2025
Project report

University *of Ljubljana*
Faculty *of Computer and Information Science*

Conversational Agent with Retrieval-Augmented Generation for research assistance

Marko Medved, Sebastijan Trojer, and Matej Vrečar

Abstract

Large language models (LLMs) are increasingly used in conversational agents for academic support, but they often struggle with outdated knowledge and hallucinations in specialized domains. To address this, we developed a conversational assistant using Retrieval-Augmented Generation (RAG), focusing on the field of automatic re-identification. Our system retrieves recent research papers from arXiv and CVF, processes them using domain-adapted query rewriting and relevance-based re-ranking, and integrates this context into language model responses. We implemented and evaluated both a baseline and an attempt of an improved retrieval pipeline, emphasizing relevance over exact matches.

Keywords

Retrieval augmented generation, research assistance, re-identification

Advisor: Aleš Žagar

Introduction

We are developing a conversational agent with Retrieval-Augmented Generation (RAG) to support academic research, focusing on the field of automatic re-identification. The system will leverage a pre-trained chatbot developed in [TODO insert year], enhanced with a RAG mechanism to retrieve and incorporate the latest research papers from arXiv. To keep the model manageable and focused, we will limit the scope of retrieved papers to a specific recent time frame (e.g., one month). The retrieved papers will be processed and integrated into the chatbot's responses, allowing the agent to provide up-to-date and accurate answers about re-identification methods developed after the chatbot's original training. To evaluate our approach an option is to manually check if the answers were relevant for the chosen scientific area and if the papers used for the queries were relevant for the specific query. Another option would be to compare the answers of a chat-bot that was trained with more recent data. Possible pre-trained chatbots that we could use include Mistral 7B or Zephyr.

Related work

Large Language Models (LLMs) have demonstrated strong language understanding and generation capabilities, but they often struggle with hallucinations, outdated knowledge, and limited domain-specific accuracy, especially when handling specialized research queries [1]. Retrieval-Augmented Generation (RAG) addresses these challenges by enhancing LLMs with real-time information retrieval, improving factual accuracy and relevance [2]. Early RAG models followed a basic retrieve-then-generate structure, but more advanced approaches have introduced techniques like query rewriting, reranking, and adaptive retrieval to improve context relevance and coherence [2].

Recent research has shown the effectiveness of RAG-based systems for academic and domain-specific applications. For example, [3] implemented a RAG system using OpenAI Ada embeddings and GPT-3.5 to enhance information retrieval and synthesis for academic documents, demonstrating the importance of precise embedding strategies for research-based chatbots. Afzal et al. [4] focused on domain-specific data by creating the CurriculumQA dataset and optimizing retrieval with tailored evaluation metrics such as relevance, coherence, and faithfulness, highlighting the importance of fine-tuning for specific academic fields. Similarly, Yasin et al. [5] integrated advanced techniques—including fine-tuned embedding models, semantic chunking, and abstract-first retrieval—to improve the accuracy and relevance of scholarly content retrieval.

Evaluating RAG systems remains complex due to the nuanced nature of retrieval and generation quality. Chen et al. [6] introduced the Retrieval-Augmented Generation Benchmark (RGB) to measure RAG performance across four key dimensions: noise robustness, negative rejection, information integration, and counterfactual robustness. Their findings show that while RAG improves LLM accuracy, it still struggles with integrating long-distance information and handling noisy or uncertain evidence. Common evaluation methods for RAG systems include retrieval-focused metrics like MRR (Mean Reciprocal Rank) and nDCG (normalized Discounted Cumulative Gain), as well as generation-focused metrics like BLEU and ROUGE to measure the quality and relevance of generated responses [7].

In our project, we aim to address these issues by developing a RAG-based chatbot for academic research in automatic re-identification—the task of matching individuals across different cameras or settings, a complex problem in computer vision and machine learning. By focusing on recent research papers from arXiv, we seek to enhance the chatbot's ability to retrieve and generate accurate responses about the latest developments in automatic re-identification

Methods

Baseline

Paper Retrieval from arXiv

We begin by retrieving candidate papers using the official arXiv API. For each paper, we extract its title, abstract, and metadata. To enable semantic similarity-based filtering, both the user query and the abstracts are embedded using the *allenai-specter* SentenceTransformer model, which is trained to map scientific documents into a dense vector space. We compute cosine similarity between the query embedding and each abstract embedding, and select the top-k most relevant papers ($k = 30$ by default).

Context Construction

The titles and abstracts of the top-ranked papers are concatenated into a single text block to serve as external context. This context is inserted into the prompt presented to the language model. No chunking, reranking, or content selection is applied at this stage beyond the similarity-based retrieval.

Language Model Inference

We use *Mistral-7B-Instruct-v0.2* as the language model backend. To enable efficient inference on commodity hardware, the model is loaded using 4-bit quantization with *bitsandbytes*, applying NF4 quantization and bfloat16 compute precision. The model is wrapped with Hugging Face's text-generation pipeline and integrated with LangChain to enable prompt-based question answering. Prompt templates include both the constructed context and the current user query. For comparison, we also evaluate answers generated without any retrieved context.

System Behavior

The system retains a basic conversational memory by concatenating the most recent user query with the current one. This helps preserve short-term context without implementing a full dialogue memory or rewriting mechanism. No fine-tuning is performed, and no re-ranking, filtering, or external knowledge base is used. This baseline is intended to evaluate performance with minimal architectural complexity.

Attempt at an Improved pipeline for retrieval

A modified RAG system was implemented to explore potential improvements in relevance and accuracy for person re-identification from video. The updated pipeline introduces several architectural changes aimed at more targeted retrieval and domain-aligned generation:

- *Two-stage query rewriting*: Queries are rewritten separately for retrieval (search) and semantic matching, allowing distinct optimization for recall and ranking.
- *Domain-specific augmentation*: Rewritten queries are enriched with terms such as *person re-identification* and restricted to the *cs.CV* category to better focus the search on relevant literature.
- *Local database integration*: A local CVF paper database supplements arXiv results, expanding the overall coverage of relevant academic sources.
- *Unified re-ranking*: Retrieved documents are encoded using *allenai-specter* and re-ranked via cosine similarity against the match-stage query.
- *Task-specific prompting*: Prompts used during response generation are tailored for the Re-ID task to better align outputs with domain-specific needs.

Figure 1 presents an overview of the enhanced pipeline, including the two-stage query rewriting, integration of a local CVF database, and re-ranking strategy. These modifications were evaluated against the baseline architecture, with comparative analysis focusing on retrieval relevance.

Preliminary results

To evaluate the effectiveness of our RAG-enhanced approach, we tested it on scientific questions related to papers published after December 2023. The goal was to determine whether the system could incorporate recent research not seen during the original model's training.

Across a range of prompts, results showed that while the system often failed to retrieve the exact target paper with top-$k = 5$, it still produced more informative and current responses by leveraging similar recent articles. Increasing k to 10 improved retrieval in several cases—for example, retrieving a nearly identical sentiment analysis paper by the same authors.

Refined queries also led to improved results. For a fast and robust covariance estimator, adding terms like "low contamination regime" and "near-linear sample complexity" helped surface the correct article. In contrast, generic queries often led the model to irrelevant domains, such as quantum mechanics.

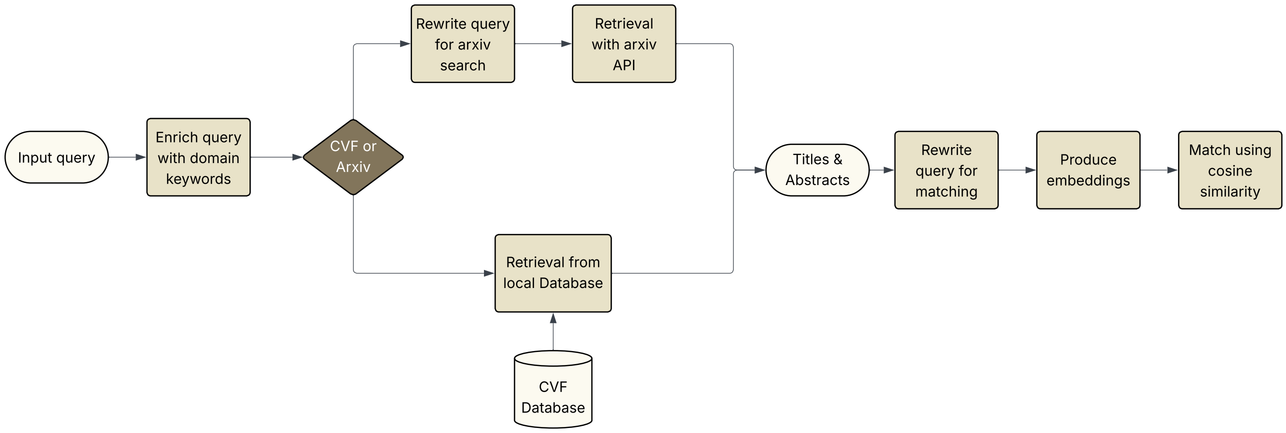

Figure 1. Overview of the enhanced retrieval pipeline.

For topics like NP-complete problems and object detection, specificity again proved critical. While exact matches were rare, the RAG system offered responses more reflective of recent research trends than the baseline. However, it occasionally exhibited erratic behavior, such as generating its own sub-questions, which significantly increased inference time.

Overall, these preliminary findings, detailed further in *results/testing_the_model.docx*, suggest that RAG enhances response relevance and currency, particularly with higher k and well-formed queries. However, retrieval control and relevance filtering remain areas for future improvement. Note that these results were collected using the baseline method.

Results

Evaluation description

As part of the evaluation protocol, we simulate the retrieval task by selecting a specific target paper and generating two types of queries per paper:

- *Specific* query: designed to closely reflect the paper's title or main content.
- *Broad* query: incorporates more general or related terms to capture a wider set of relevant documents.

This setup allows us to assess the retrieval system's ability to rank truly relevant papers near the top of the search results, which simulates real-world scenarios where researchers look for both precise and comprehensive literature coverage.

To ensure a fair comparison with the baseline system, we did not include our local CVF paper database during evaluation, as the baseline does not have access to it.

Metrics for retrieval evaluation

Since our primary focus was on improving retrieval performance, we evaluated the system using metrics inspired by common recognition and retrieval tasks.

The Cumulative Matching Characteristic (CMC)-like metric measures the probability that the correct match appears within the top-k retrieved results.

It is computed by checking, for each query, whether the correct item is found at rank 1, rank 2, and so on up to rank k, then averaging these results across all queries.

This metric is particularly useful here because it directly reflects the likelihood that a user or system will find a correct identity within a small set of top candidates, which is crucial for efficient person re-identification in large video databases.

The mean Average Precision (mAP)-like metric evaluates the retrieval quality across all relevant items for each query by calculating the average precision at the ranks where relevant results appear, then averaging this precision across all queries.

This captures both precision and recall over the entire ranked list, offering a more comprehensive measure of retrieval effectiveness.

Together, these metrics give complementary views of system performance:

- **CMC** focuses on the position of the first correct match, indicating practical usability by measuring how quickly a correct result appears in the ranked list.
- **mAP** reflects overall retrieval accuracy and ranking quality by considering all relevant results throughout the list, capturing both precision and recall.

This combination makes them well-suited to assess person re-identification and related retrieval tasks.

Evaluation Results

To estimate uncertainty in our evaluation, we used non-parametric bootstrapping over the query set with 1000 resamples. For each resample, we computed both CMC curves and mean Average Precision (mAP). The reported values in Table 1 correspond to the mean and standard deviation across these samples.

For the CMC plots (Figure 2), we show the mean curve along with confidence intervals. These intervals represent the $[15^{\text{th}}, 85^{\text{th}}]$ percentiles (i.e., a 70% confidence interval) computed at each rank position based on bootstrapped CMC values.

As seen in Table 1 and Figure 2, our current method underperforms the baseline across all metrics, with the largest gap observed in the specific query setting. This indicates that the current approach may not be capturing the essential characteristics needed.

Improving performance may require exploring different models or techniques that better represent the relationship between queries and target papers, particularly in more precise retrieval scenarios.

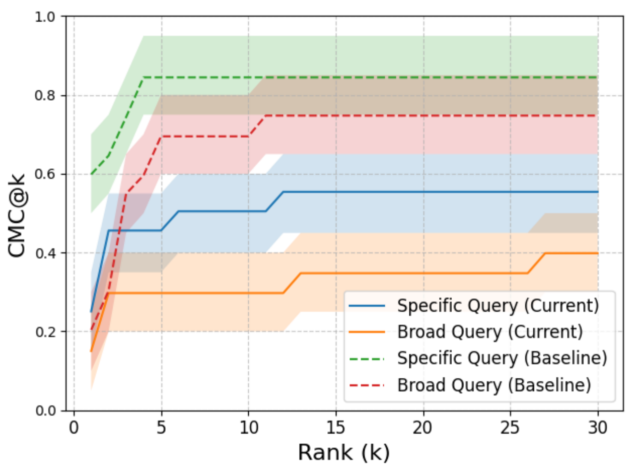

Figure 2

Table 1. Retrieval performance for specific and broad queries. Best results in bold.

Method	Rank-1	Rank-5	mAP
Specific Queries			
Current	0.254 ± 0.097	0.450 ± 0.112	0.365 ± 0.093
Baseline	**0.600 ± 0.110**	**0.850 ± 0.081**	**0.683 ± 0.091**
Broad Queries			
Current	0.150 ± 0.082	0.301 ± 0.103	0.231 ± 0.083
Baseline	**0.199 ± 0.090**	**0.698 ± 0.103**	**0.369 ± 0.078**

Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

References

[1] Nimit Kandpal, Myle Ott, Xuezhi Li, Mike Lewis, Luke Zettlemoyer, and Dani Yogatama. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, 2023.

[2] Yuan Gao, Zihan Zhang, Xiaodong Liu, Furu Wei, and Jianfeng Gao. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024.

[3] Maryamah Maryamah, Muhammad Maula Irfani, Edric Boby Tri Raharjo, Netri Alia Rahmi, Mohammad Ghani, and Indra Kharisma Raharjana. Chatbots in academia: A retrieval-augmented generation approach for improved efficient information access. In *2024 16th International Conference on Knowledge and Smart Technology (KST)*, pages 259–264, 2024.

[4] Anum Afzal, Juraj Vladika, Gentrit Fazlija, Andrei Staradubets, and Florian Matthes. Towards optimizing a retrieval augmented generation using large language model on academic data. *arXiv preprint arXiv:2411.08438*, 2024.

[5] Ahmet Yasin Aytar, Kemal Kilic, and Kamer Kaya. A retrieval-augmented generation framework for academic literature navigation in data science. *arXiv e-prints*, pages arXiv–2412, 2024.

[6] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, Beijing, China, 2024. {jiawei2020, hongyu, xianpei, sunle}@iscas.ac.cn.

[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.