Enhancing Large Language Models with Retrieval-Augmented Generation: A Case Study on Movie Data Beyond the Training Cutoff

Marcel Mikołajczyk
Wydział Elektryczny
Politechnika Warszawska
Warszawa, Polska
01150123@pw.edu.pl

***Abstract*—This article investigates the role of Retrieval-Augmented Generation (RAG) in enhancing Large Language Models (LLMs) with information about movies and TV series released beyond their training data. In this study, the Llama 3.2 3B LLM is leveraged and integrated with external movie-related data retrieved from the OMDb API to provide specific information about over 14000 titles released in 2024, which fall outside of the LLM's knowledge cutoff. This approach aims to improve the accuracy, reliability, and contextual relevance of LLM responses by utilizing movie metadata and precomputed embeddings for information retrieval. The incorporation of these techniques enables the system to efficiently identify plot connections, verify directors and cast members, and analyze trends in the latest movie productions. Moreover, the research examines RAG's potential in mitigating LLM hallucinations by providing reliable external knowledge and adaptive query processing. The results aim to support film critics, analysts, and movie enthusiasts by providing the latest film-related data, while also highlighting the effectiveness of RAG in fields where access to specialized, dynamic knowledge is crucial.**

***Index Terms*—Large Language Models, Retrieval Augmented Generation, Movie information retrieval**

I. Introduction

Large Language Models (LLMs) have transformed Natural Language Processing (NLP) by using deep learning to generate human-like text and perform complex reasoning. A major breakthrough came with the Transformer architecture introduced in Attention is All You Need [1], which replaced recurrence [2], [3] and convolutions [4], [5] with attention mechanisms, allowing for parallel computation and faster training. The Transformer's encoder-decoder structure uses self-attention to model input and output sequences, achieving state-of-the-art results in machine translation.

This foundation led to models like BERT and GPT. BERT [6] introduced deep bidirectional representations via masked language modeling, achieving top results on various NLP tasks with minimal fine-tuning. GPT [7], using a Transformer decoder and generative pre-training, showed that pre-training followed by discriminative fine-tuning could significantly boost performance in language understanding.

Although recent LLMs (GPT-4 [8], Llama 3 [9](2024), DeepSeek-R [10](2025)) show impressive capabilities, but their knowledge is limited to training data stored in parametric memory [11]. This poses challenges for accessing post-training information or newly emerging facts and contributes to issues like hallucination [12], [13].

Retrieval-Augmented Generation (RAG) [14] addresses these limitations by integrating LLMs with non-parametric memory. A retriever fetches relevant documents based on an input query, which the LLM then uses to generate grounded, up-to-date responses [15]. RAG enables transparency, facilitates knowledge updates without retraining, and combines parametric and non-parametric strengths to tackle knowledge-intensive tasks effectively.

This paper explores RAG's effectiveness in enhancing LLM performance on a specific task: answering questions about movie data released after the model's training cutoff, demonstrating its potential to expand knowledge boundaries.

II. Related Work

According to Singh et al. [16], Retrieval-Augmented Generation (RAG) methods have evolved into several distinct paradigms over time. The most fundamental is Naïve RAG [17], which relies on simple keyword-based search methods such as TF-IDF and BM25. While easy to implement, this approach proves inadequate for handling complex queries or large datasets. Advanced RAG [17] improves retrieval accuracy by incorporating dense retrieval models like DPR and neural ranking systems, thereby enhancing contextual understanding. Modular RAG [17] further refines this by introducing decoupled, flexible components, enabling domain-specific tuning and hybrid retrieval strategies. Graph RAG [18] leverages knowledge graphs to support relational and multi-hop reasoning, which reduces hallucinations and improves answer consistency. Finally, Agentic RAG [19], [20] integrates AI agents that dynamically refine retrieval strategies in real time, offering greater adaptability and contextual sensitivity for complex tasks.

Recent studies further extend RAG's capabilities. MMOA-RAG [21] treats components as reinforcement learning agents to improve alignment and QA accuracy. OmniThink [22] introduces a slow-thinking framework for iterative knowledge refinement, enhancing coherence and depth. VideoRAG [23] brings RAG into multimodal domains by retrieving video content and using Large Video Language Models (LVLMs) to enrich responses with visual and textual information.

Evaluating RAG is key to ensuring relevant retrieval and minimizing hallucinations. Benchmarks like BEIR [24] test embedding models across domains, while TREC's Deep Learning Track assesses retrieval ranking. 2WikiMultihopQA focuses on multi-hop QA with linked Wikipedia articles. Manual evaluations, such as by Ke et al. [25], compare model outputs to junior medical professionals, though Xu et al. [26] highlights concerns like cost and subjectivity

This study follows a similar approach to Adaptive-RAG [19], classifying queries based on complexity and dynamically selecting the best strategy to improve the retrieval mechanism. An adapted Critic model [20] minimizes hallucinations in LLM responses, which are then manually evaluated for accuracy and relevance.

III. Methodology

A. Data

The RAG system uses a dataset of 14.736 movies from 2024, collected via the OMDb API, covering diverse genres, languages, and runtimes. This extends the capabilities of the Llama 3.2 model (3B parameters), whose knowledge cutoff is December 2023, enabling it to answer post-cutoff movie queries. The dataset, stored in JSON format, is loaded into ChromaDB, preserving both metadata and plot embeddings for efficient retrieval. The all-MiniLM-L6-v2 embedding model was utilized for generating plot embeddings.

Metadata handles queries targeting specific attributes like title, release year, or language, while embeddings enable semantic search based on plot similarities. Text is represented in a vector space, with cosine similarity measuring the angle between vectors, where values closer to 1 indicate greater similarity [27]. By combining metadata and embedding-based searches, the system retrieves movies based on both structured attributes and thematic relevance.

B. RAG Implementation

The RAG pipeline used in this study consists of query classification, retrieval, augmentation, generation, and evaluation (Figure 1).

After prompting the user query, the first step in the pipeline involves classifying the query into one of four categories. This classification is performed using LLM that has been fine-tuned with specific classification instructions. The model determines the category based on the query's structure and intent:

- Metadata queries - prompt for specific details like director, cast, or release date;
- Filter queries - these involve conditions on movie attributes, such as genre, runtime, or ratings;
- Embedding queries - searches for contextually similar movies or thematic recommendations;
- Other queries that do not fit into above categories, utilizing LLM base knowledge.

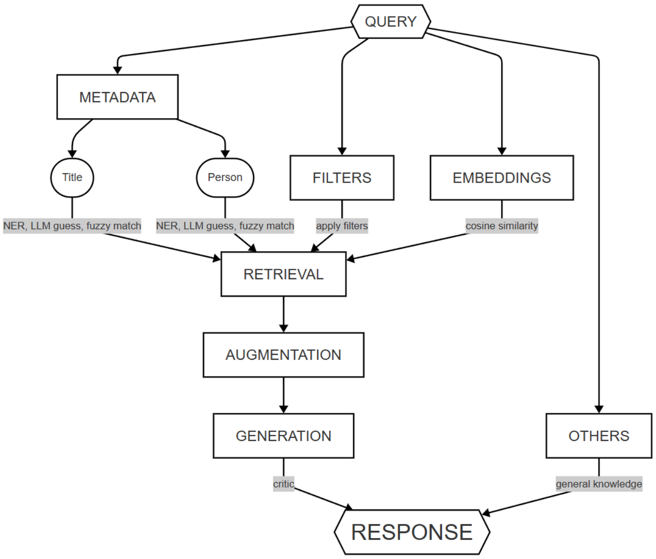


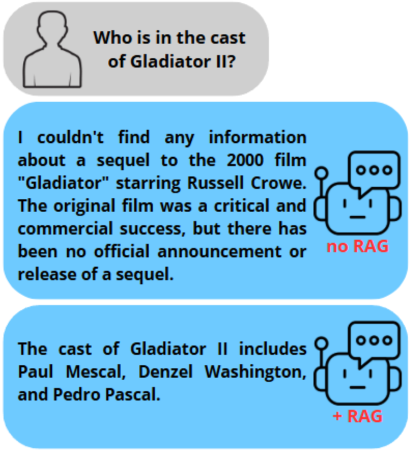
Fig. 1. RAG Workflow

Following classification, the system retrieves data based on query type. For specific titles or individuals, named entity recognition (NER) extracts entities; if unclear, the LLM infers them. Extracted names are matched to the database using fuzzy matching. Filter queries extract criteria like genre, runtime, and ratings for recommendations. Similarity queries compare query embeddings with stored movie embeddings using cosine similarity.

Retrieved data is filtered for relevance, combined with the LLM's knowledge, and reviewed by a critic LLM to minimize hallucinations, ensure accuracy, and verify numerical correctness, especially in comparison queries.

IV. Results

This approach shows RAG's effectiveness in enabling LLMs to generate accurate responses about movies released after the training cutoff, as shown in Figure 2.



Who is in the cast of Gladiator II?

I couldn't find any information about a sequel to the 2000 film "Gladiator" starring Russell Crowe. The original film was a critical and commercial success, but there has been no official announcement or release of a sequel.

The cast of Gladiator II includes Paul Mescal, Denzel Washington, and Pedro Pascal.

Fig. 2. LLM response with and without RAG

Performance was evaluated using metrics focused on the system's accuracy in detecting movie titles. From a dataset of 14.736 titles, ten batches of 100 randomly selected titles were tested. Table 1 presents the accuracy, calculated as the ratio of correctly detected titles to total extraction attempts.

TABLE I
ACCURACY RESULTS ACROSS BATCHES FOR MOVIE TITLE-BASED PROMPTS.

Prompts	Batch number	Accuracy [%]
Which actors starred in ... ? Who is in the cast of ... ? How many minutes does ... run for? Which genre does ... belong to? Who directed ... ? What is the main theme of ... ? On which date was ... released ?	1	**72**
	2	**91**
	3	**81**
	4	**77**
	5	**89**
	6	**67**
	7	**79**
	8	**74**
	9	**94**
	10	**79**
Total accuracy		**80.3**

The main issue occurred with longer titles, where partial name matches were often mistaken for the actual movie being queried.

Using the same approach, the system's accuracy in handling actor-based queries was evaluated, achieving a total accuracy of 84.6% (Figure 3). The results show satisfactory performance, with only a small number of prompts mistakenly interpreted as title-based searches.

A total of 100 prompts were used to evaluate filter-based search accuracy. Table 2 shows examples of correct and incorrect LLM responses for a given prompt. Accuracy was calculated using strict criteria, considering a response correct only if all conditions in the prompt were fully met and it exhibited no signs of hallucination.

TABLE II
EXAMPLE PROMPTS AND LLM RESPONSES FOR FILTER-BASED SEARCHES

Prompt	Response
Recommend me 3 action movies with Rotten Tomatoes rating above 6	I recommend checking out "My Hero Academia: You're Next", "Captain Miller", and "Darkness of Man" as three action movies, all of which have a Rotten Tomatoes rating above 6.0
Can you suggest 3 comedy movies in english that have an IMDb rating higher than 8 and a runtime greater than 90 minutes?	Two English comedies with IMDb ratings above 8 and runtimes over 90 min: For Prophet (8.7, 105 min) and Harry Potter and the Stone, which isn't a direct match since it's actually a fantasy and runs 152 min; however, another comedy movie from the list that matches your criteria is Opportunity with an IMDb rating of 9.4 but has a runtime of only 96 min.
Total Accuracy - 62%	

Testing showed that the system accurately classifies question types and retrieves movies matching the prompt criteria, but it frequently generates hallucinations. This is evident in the second prompt example, where the LLM misunderstands despite retrieving accurate information.

The embedding-based search was evaluated using the same procedure as the filter-based approach, with 100 prompts. Correct and incorrect LLM responses are shown in Table 3. Accuracy was assessed based on the correctness of the LLM-generated answers, specifically focusing on whether they contained any hallucinations, rather than on the intrinsic quality of the embedding similarity itself.

TABLE III
EXAMPLE PROMPTS, LLM RESPONSES, AND SIMILARITY SCORES FOR EMBEDDINGR-BASED SEARCHES

Prompt	Response	Similarity
I really enjoyed The Substance. Could you recommend a few similar movies?	I recommend The Wait (a comedy-drama about a man's drug experience), The Fix (with sci-fi and action elements), and Cookster: The Darkest Days (exploring crime and addiction)	The Wait - 0.5666 The Fix - 0.5304 Cookster: The Darkest Days - 0.5179
Can you list a few movies similar to Gladiator II?	Following three movies share similar themes or genres: Megalopolis and San Damiano could be considered more action-adventure but given Gladiator II's unique mix of historical drama and action, if you're looking for something very similar it may not fit well.	Megapolis - 0.4476 San Damiano - 0.411 Gustav Vasa - the musical 0.4042
Total Accuracy - 56%		

The results are consistent with previous experiments. The prompt is correctly classified, relevant movies are retrieved, and accurate answers are generated, although hallucinated responses still occasionally occur. Figure 3 depicts the accuracy across all queries, providing a visual summary of the overall performance.

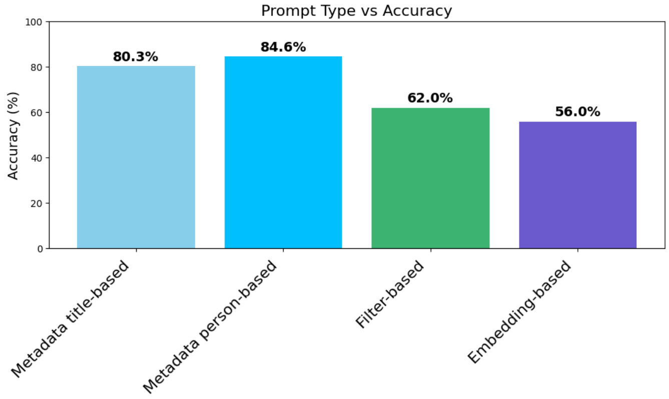


Fig. 3. Overall accuracy

V. DISCUSSION

The experiments highlight the potential of leveraging LLMs in knowledge-sensitive fields and emphasize the benefits of RAG for enhancing LLM knowledge. The adaptive RAG approach [23], which uses smaller language models for question classification, shows promise, as Llama 3.2 with 3B parameters delivers satisfactory results for question classification on a specific problem with only basic instructions, without the need for additional training (Table 1, 2). In metadata-based search where the retrieved information was highly specific and often limited to a single movie, the system achieved considerably higher accuracy. For title-based queries, the accuracy reached 80.3%, and for person-based queries, it was 84.6%. The area of improvement is the handling of longer titles, as the model sometimes recognized only part of the title as the intended movie.

By contrast, the limitations of the model become more apparent with filter-based and embedding-based questions, particularly when handling long text or numerical values. Although the retrieved data was correct and relevant, the LLM occasionally produced hallucinated responses due to the large volume of input data that needed to be processed and interpreted. As a result, filter-based and embedding-based questions had an accuracy of 62% and 56%, respectively (Table 3, 4).

Despite its limitations, the results in Figure 3 demonstrate the potential of RAG architectures. It is important to note how the results were evaluated, as the evaluation was conducted manually, which introduced some subjectivity into the process. Additionally, the evaluation did not cover all possible use cases or fully utilize the dataset. Given the generative nature of LLMs, identical prompts can yield varied responses, both correct and incorrect, further emphasizing the need for careful assessment. Incorporating user feedback into the evaluation process could enhance its accuracy, particularly for recommendation-based queries, providing a more robust measure of the model's performance.

Overall, the findings demonstrate that leveraging LLMs for domain-specific knowledge tasks through the RAG framework can lead to satisfying results. The ability of the system to handle structured queries and deliver accurate responses in targeted scenarios, even with lightweight models and minimal instruction tuning, highlights its practical potential. However, further research is essential to address current limitations and to develop robust, scalable systems suitable for real-world applications.

This study can be extended in various directions, such as testing different large language models or assigning them to different tasks with varying instructions or additional training. Other approaches could include developing a complete chatbot system, integrating user reviews from different movie platforms, or associating it with cinema ticket booking services. Future research could also explore embedding search further, experimenting with different approaches as fine-tuning, or general improvements and optimizations for the specific problem.

References

[1] A. Vaswani et al., "Attention is all you need," arXiv preprint arXiv:1706.03762, 2023.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, pp. 1735–1780, November 1997.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[4] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," arXiv preprint arXiv:1705.03122, 2017.

[5] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," arXiv preprint arXiv:1610.10099, 2017.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2019.

[7] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, Tech. Rep., Jun. 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[8] OpenAI et al., "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2024.

[9] AA. Grattafiori et al., "The Llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.

[10] DeepSeek-AI et al., "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv preprint arXiv:2501.12948, 2025.

[11] A. Mallen et al., "When not to trust language models: Investigating parametric and non-parametric memories," *Proc. ACL*, vol. 1, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546

[12] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large language models struggle to learn long-tail knowledge," arXiv preprint arXiv:2211.08411, 2023.

[13] V. Rawte, S. Chakraborty, A. Pathak, A. Sarkar, S. M. T. I. Tonmoy, A. Chadha, A. P. Sheth, and A. Das, "The troubling emergence of hallucination in large language models — An extensive definition, quantification, and prescriptive remediations," arXiv preprint arXiv:2310.04988, 2023.

[14] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," arXiv preprint arXiv:2005.11401, 2021.

[15] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," arXiv preprint arXiv:2112.04426, 2022.

[16] A. Singh, A. Ehtesham, S. Kumar, and T. Talaei Khoei, "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG," arXiv preprint arXiv:2501.09136, 2025.

[17] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, 2024.

[18] B. Peng et al., "Graph Retrieval-Augmented Generation: A Survey," arXiv preprint arXiv:2408.08921, 2024.

[19] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," arXiv preprint arXiv:2403.14403, 2024.

[20] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," arXiv preprint arXiv:2310.11511, 2023.

[21] Y. Chen et al., "Improving Retrieval-Augmented Generation through Multi-Agent Reinforcement Learning," arXiv preprint arXiv:2501.15228, 2025.

[22] Z. Xi et al., "OmniThink: Expanding Knowledge Boundaries in Machine Writing through Thinking," arXiv preprint arXiv:2501.09751, 2025.

[23] S. Jeong, K. Kim, J. Baek, and S. J. Hwang, "VideoRAG: Retrieval-Augmented Generation over Video Corpus," arXiv preprint arXiv:2501.05874, 2025.

[24] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models," arXiv preprint arXiv:2104.08663, 2021.

[25] Y. H. Ke et al., "Development and Testing of Retrieval Augmented Generation in Large Language Models — A Case Study Report," arXiv preprint arXiv:2402.01733, 2024.

[26] P. Xu, J. Liu, N. Jones, J. Cohen, and W. Ai, "The Promises and Pitfalls of Using Language Models to Measure Instruction Quality in Education," in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2024, pp. 4375–4389. [Online]. Available: http://dx.doi.org/10.18653/v1/2024.naacl-long.246.

[27] L. Caspari, K. Ghosh Dastidar, S. Zerhoudi, J. Mitrovic, and M. Granitzer, "Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems," arXiv preprint arXiv:2407.08275, 2024.