IR-RAG @ SIGIR24: Information Retrieval's Role in RAG Systems

Fabio Petroni
fabio@samaya.ai
Samaya AI
London, United Kingdom

Federico Siciliano
siciliano@diag.uniroma1.it
Sapienza University of Rome
Rome, Italy

Fabrizio Silvestri
fsilvestri@diag.uniroma1.it
Sapienza University of Rome
Rome, Italy

Giovanni Trappolini
trappolini@diag.uniroma1.it
Sapienza University of Rome
Rome, Italy

ABSTRACT

In recent years, Retrieval Augmented Generation (RAG) systems have emerged as a pivotal component in the field of artificial intelligence, gaining significant attention and importance across various domains. These systems, which combine the strengths of information retrieval and generative models, have shown promise in enhancing the capabilities and performance of machine learning applications. However, despite their growing prominence, RAG systems are not without their limitations and continue to be in need of exploration and improvement. This workshop seeks to focus on the critical aspect of information retrieval and its integral role within RAG frameworks. We argue that current efforts have undervalued the role of Information Retrieval (IR) in the RAG and have concentrated their attention on the generative part. As the cornerstone of these systems, IR's effectiveness dramatically influences the overall performance and outcomes of RAG models. We call for papers that will seek to revisit and emphasize the fundamental principles underpinning RAG systems. At the end of the workshop, we aim to have a clearer understanding of how robust information retrieval mechanisms can significantly enhance the capabilities of RAG systems. The workshop will serve as a platform for experts, researchers, and practitioners. We intend to foster discussions, share insights, and encourage research that underscores the vital role of Information Retrieval in the future of generative systems.

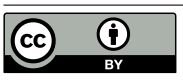
CCS CONCEPTS

• **Information systems** → *Novelty in information retrieval.*

KEYWORDS

Retrieval Augmented Generation, Generative Models, Neural Databases

ACM Reference Format:
Fabio Petroni, Federico Siciliano, Fabrizio Silvestri, and Giovanni Trappolini. 2024. IR-RAG @ SIGIR24: Information Retrieval's Role in RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626772.3657984



SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657984

1 TITLE

IR-RAG @ SIGIR24: Information Retrieval's Role in RAG Systems

2 MOTIVATION

In the ever-evolving landscape of artificial intelligence and machine learning, Retrieval Augmented Generation (RAG) systems have rapidly stood out as one of the main innovations in recent times [5, 11, 13]. These systems, whose goal is to enhance language models [6, 10, 17] by incorporating external knowledge sources, have proven instrumental in a myriad of applications, ranging from natural language processing to recommendation systems. Their ability to leverage vast information repositories and generate coherent, contextually relevant responses significantly advance AI's ability to mimic human thinking through several key aspects:

- **Integration of External Knowledge**: RAG systems mirror human information-seeking behavior by accessing vast databases to enhance understanding of complex queries. This external knowledge integration enables a more comprehensive approach akin to human thought processes.
- **Contextual Understanding**: They excel in grasping the nuances of context, generating accurate and contextually fitting responses similar to human comprehension and judgment.
- **Adaptive Learning**: Reflecting human learning, these systems evolve by assimilating new information, refining their response strategies, and enhancing their decision-making and reasoning capabilities.
- **Decision-Making**: RAG systems mimic human decision-making by evaluating multiple information sources and synthesizing them for well-informed, reasoned responses.
- **Handling Ambiguity and Complexity**: Like humans, RAG systems effectively navigate complex and ambiguous scenarios by integrating diverse information, surpassing the limitations of traditional AI.

Given these capabilities, a workshop focusing on the retrieval component of RAG systems is vital. It will focus on the mechanisms that enable these systems to access and integrate diverse knowledge sources. Understanding and improving this retrieval aspect is crucial for advancing AI towards a level of complexity and adaptability that more closely emulates human cognitive processes, thereby enhancing the overall effectiveness and application scope of RAG systems. Despite the remarkable advancements attributed to RAG systems, they are not without their challenges. RAG systems' pitfalls are varied and significant. Firstly, the quality of the content

generated heavily depends on the retrieved documents' relevance, as inaccurate or biased sources can lead to misleading or harmful outputs [9]. Secondly, these systems often struggle with data consistency and coherence. This can result in outputs that are confusing or contradictory [8]. Moreover, the dependency on external databases means that RAG systems can become outdated if the databases are not continuously updated, leading to responses that might be factually incorrect over time [16]. Additionally, the retrieval process adds complexity and computational overhead, potentially leading to slower response times and increased resource consumption. Lastly, there are privacy concerns, as the system's reliance on external data might inadvertently expose sensitive information or violate data use policies [4]. Generally, the pitfalls of Retrieval-Augmented Generation (RAG) systems can be broadly categorized into two main areas: issues with the retrieving process and problems with the generative, Large Language Model (LLM) component. We argue here that in the discourse surrounding Retrieval-Augmented Generation (RAG) systems, the spotlight has predominantly been on the generative component, often overshadowing the equally crucial role of information retrieval. While advancements in Large Language Models (LLMs) captivate public and academic attention with their increasingly sophisticated outputs, the retrieval mechanisms that feed these models have been somewhat undervalued. Recognizing this disparity, our workshop aims to cast a spotlight on information retrieval, an area that, despite being a cornerstone of RAG systems, has been relatively under-explored compared to its generative counterpart. We believe that by dedicating focused attention to enhancing the retrieval mechanisms, we can unlock new levels of efficiency and effectiveness in RAG systems.

3 THEME AND PURPOSE OF THE WORKSHOP

The primary purpose of this workshop is to shift the focus onto the often-overlooked retriever mechanism of Retrieval-Augmented Generation (RAG) systems while pondering the question: "Should research in information retrieval change now that RAG systems exist?". By gathering experts, practitioners, and enthusiasts in a dedicated forum, the workshop seeks to spotlight and deliberate on challenges and innovative ideas associated with the retrieval aspect of RAG systems, aiming to foster a solid community around this critical topic. The intent is to generate a collective effort to better understand and enhance the retrieval mechanisms, ensuring they are given as much importance as the generative components. Through this workshop, we aspire to build a strong foundation and a vibrant community committed to advancing the state of the art in information retrieval for RAG systems. In particular, we are looking for contributions that study:

- **Use Of The Retrieved Context By The LLM:** Recent work [7, 11, 15] has demonstrated that RAG systems are sensible to the order and the nature of the retrieved context. These can be considered preliminary results that pave the way for future research.
- **(Query) Representation Learning:** Improving how queries are represented can significantly enhance the retriever's ability to find relevant documents. This could involve using more advanced natural language processing techniques to understand the context and nuances of the query better.
- **Incorporating Contextual Information:** Including more context in the retrieval process can improve the relevance of the documents retrieved. This could mean taking into account the broader conversation, user preferences, or historical interactions
- **Updating the Document Database:** Keeping the document database up-to-date ensures that the retriever has access to the latest and most relevant information. This is particularly important for topics that are rapidly evolving.
- **Reducing Computational Load:** Optimizing the retriever for speed and efficiency, especially when dealing with large databases, can improve its usability in real-time applications. This might involve techniques for reducing the dimensionality of data or faster search algorithms.
- **Bias Mitigation:** Actively working to identify and mitigate biases in the retrieval process can improve the fairness and reliability of the retrieved content.
- **Cross-Lingual Retrieval Capabilities:** For systems operating in multilingual environments, improving the retriever's ability to handle and retrieve documents in various languages can enhance its effectiveness.
- **Multimodality:** Most of the current research has focused on textual RAG, even though multimodality is highly needed in many applications.
- **Other:** One of the goals of this workshop is to collect new ideas and challenges, so proposals in this sense are very much welcomed.

4 FORMAT

Our workshop will be a full-day workshop. **(a)** Invited speakers (industrial and academic): Invited speakers from both industry and academia. Candidates from various top companies and institutions have already accepted our invitation; **(b)** Selected contributed papers to be presented as oral sessions; **(c)** Accepted papers of slightly lesser significance to be showcased during a poster session; **(d)** An interactive session to share lessons learned; **(e)** Breakout sessions focusing on topics arising from the contributed papers and broader discussions to share insights and lessons learned. Table 1 outlines the schedule for the day's activities. Additionally, crucial dates related to the event are provided in Table 2.

Duration	Activity
15min	Opening
1h	Keynote speaker (1)
45min	Paper presentations (1)
30min	Coffee break
45min	Paper presentations (2)
1h30min	Lunch break
1h	Keynote Speaker (2)
45min	Poster Session
30min	Refreshment break
1h	Breakout session & discussion among participants
30min	Round up and concluding remarks

Table 1: Schedule

Event	Date
Submission deadline	April 25, 2024
Notification	May 23, 2024
Camera-ready versions of accepted papers due	June 22, 2024
IR-RAG Workshop	July 18, 2024

Table 2: Important dates

5 ORGANIZERS ATTENDANCE AND SPECIAL REQUIREMENTS

All the organizers already plan to attend the SIGIR conference onsite, so they will definitely be present if the workshop proposal is accepted. The workshop does not entail any special requirements.

6 ORGANIZERS

Fabio Petroni is co-founder at Samaya AI, a company building an AI-powered knowledge-discovery platform. Before that, he was Researcher at the FAIR team of Meta AI and the R&D department at Thomson Reuters, where he focused on Knowledge-Intensive Natural Language Processing. He was the Sponsor Chair for EuroSys 2023, Local Chair for AKBC 2022, and co-organized the Workshop on Representation Learning for NLP hosted at ACL 2020. He is one of the authors of the original paper on RAG [11], of the Lost in the middle paper [12] and Autoregressive Search Engines [3].

Federico Siciliano is a PostDoc in Data Science at Sapienza University of Rome. He is part of the RSTLess research group at the Sapienza University of Rome, which focuses on Robust, Safe, and Transparent Deep Learning. Siciliano has publications in the domains of Information Retrieval and Explainable Artificial Intelligence, featuring conferences such as ECIR, RecSys, and IJCNN. He has also been a Program Committee member for the 5th Knowledge-aware and Conversational Recommender Systems Workshop (KaRS 2023) at RecSys 2023, the 38th AAAI Conference on Artificial Intelligence (AAAI-24) and the 46th European Conference on Information Retrieval (ECIR 2024). Siciliano is a Ph.D. Candidate in Data Science from the University of Sapienza, with a thesis on "Architectural Components of Trustworthy Artificial Intelligence". He is one of the authors of Reinforced Retrieval Augmented Machine Learning (RRAML) [1].

Fabrizio Silvestri. *(h-index: 45; citations: 6,655; i10-index: 104)* is a Full Professor at the Department of Computer, Control and Management Engineering at Sapienza University of Rome. His research interests focus on Artificial Intelligence, particularly machine learning applied to web search problems and natural language processing. He has authored more than 150 papers in international journals and conference proceedings and holds nine industrial patents. Silvestri has been recognized with a "test-of-time" award at the ECIR 2018 conference for an article published in 2007. He also received three best paper awards and other international recognitions. Silvestri spent eight years in industrial research laboratories, including Yahoo! and Facebook. At Facebook AI, he directed research groups to develop artificial intelligence techniques to combat malicious actors who use the Facebook platform for malicious purposes, such as hate speech, misinformation, and terrorism. Silvestri has experience in organizing numerous workshops and conferences, and he will be one of the General Chairs of ECIR 2025 in Lucca and one of the Program Committee Chairs of CIKM 2026 in Rome. Silvestri holds a Ph.D. in computer science from the University of Pisa, with a thesis on "High-Performance Issues in Web Search Engines: Algorithms and Techniques". He is one of the authors of Multimodal Neural Databases [18] and RRAML [1].

Giovanni Trappolini is an Assistant Professor at the Department of Computer, Control and Management Engineering at Sapienza University of Rome. He received his Ph.D. in Machine Learning in 2022 under the supervision of Professor Emanuele Rodolà and was awarded the title of Sapienza's honor student. His research interests lie at the intersection of geometric and multimodal deep learning, with a particular focus on information retrieval. He has a solid track record of publications in the most important machine learning conferences and has the privilege of partnering with numerous world-renowned institutions, including Stanford, Technion, Meta, Amazon, and UniPi, among others. He is the creator of Fauno [2], the largest (and best-performing) Italian large language model (LLM) to date, and one of the authors of Multimodal Neural Databases [18] and The Power of Noise [7]. Last year, he also organized the FLIRT [14] workshop at SIGIR.

7 PC MEMBERS

Potential PC members for reviewing paper submissions:

- Aditya Sanghi, University Of Toronto | Autodesk
- Aleksandra Piktus, Sapienza University of Rome | Cohere
- Alon Halevy, Meta
- Andrea Bacciu, Sapienza University of Rome | Amazon
- Bora Edizel, Warner Bros. Discovery
- Charles L. A. Clarke, University of Waterloo
- Franco Maria Nardini, ISTI-CNR
- Iftah Gamzu, Amazon
- James Thorne, KAIST
- Konstantina Palla, Spotify
- Nicola Tonellotto, University of Pisa
- Oleksandr Pryymak
- Prabhat Agarwal, Pinterest
- Raffaele Perego, ISTI-CNR
- Ricardo Baeza-Yates, Pompeu Fabra University
- Shoval Lagziel, Amazon
- Wang-Chiew Tan, Meta

8 SELECTION PROCESS

The workshop invites submissions of papers ranging from two to six pages via an open call for contributions. We welcome diverse submissions, including reports detailing original research, preliminary research findings, proposals for innovative work, and position papers. All submitted papers will undergo a rigorous peer review process led by the Program Committee.

9 TARGET

The workshop is tailored to engage a diverse audience comprising researchers, practitioners, and enthusiasts invested in the domain of RAG systems but also more broadly in the realms of Information Retrieval (IR) and Generative Machine Learning (GML). Academics,

industry professionals, graduate students, and anyone intrigued by the intersection of IR and GML are encouraged to participate. The workshop will be promoted through various strategic channels to ensure broad outreach. Our promotion strategy includes leveraging established academic networks, such as research institutions and pertinent academic mailing lists. Additionally, we plan to disseminate workshop announcements via targeted emails to researchers and practitioners in the field, encouraging them to participate and submit their contributions. Furthermore, our outreach strategy will encompass engagement through various social media platforms, including Twitter, LinkedIn, and other relevant networks. Additionally, a dedicated workshop website will serve as a centralized hub for disseminating comprehensive information regarding the workshop. This website will feature workshop details, submission guidelines, schedules, keynote speaker announcements, and other pertinent information, ensuring easy access to interested participants seeking detailed insights into the workshop's objectives and proceedings. By employing a multifaceted promotional approach, we aim to attract a diverse and engaged audience keen on exploring the evolving landscape of IR and GML integration.

10 RELATED WORKSHOPS

Generative IR Workshops: Similarly, workshops like "Generative IR" have concentrated on exploring generative approaches in Information Retrieval (IR). To clearly explain the difference, one might borrow former US President Kennedy's famous antimetabole: Generative IR asks what generation can do for IR; we ask what IR can do for generation. In practice, they focused on techniques like Differentiable Search Index (DSI) or similar. We focus not solely on generative methods in IR but rather on enhancing Information Retrieval techniques to augment the capabilities of Generative Machine Learning models. Retrieval-Enhanced Machine Learning (REML @ SIGIR 2023): This workshop primarily focuses on broader discussions involving machine learning, while our workshop distinctly concentrates on the pivotal role of retrieval in enhancing Generative Machine Learning models. Unlike approaches where models improve merely by utilizing machine learning techniques, our workshop uniquely emphasizes the necessity of retrieval within the context of generative models. Given these distinctions and the existing gap in workshops that explicitly focus on Retrieval-Augmented Generative (RAG) models, the need for our specialized workshop becomes evident. While other workshops touch on related themes, none concentrate on the indispensable role of retrieval in refining generative models' functionality.

ACKNOWLEDGMENT

This project was supported by the projects FAIR (PE0000013), SERICS (PE00000014), and IR0000013-SoBigData.it PNRR and the NEREO (Neural Reasoning over Open Data) project PRIN Grant no. 2022AEFHAZ.

REFERENCES

[1] Andrea Bacciu, Florin Cuconasu, Federico Siciliano, Fabrizio Silvestri, Nicola Tonellotto, and Giovanni Trappolini. 2023. RRAML: Reinforced Retrieval Augmented Machine Learning. In *Proceedings of the Discussion Papers - 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023 DP) co-located with 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), Rome, Italy, November 6-9, 2023 (CEUR Workshop Proceedings, Vol. 3537)*. CEUR-WS.org, 29–37. https://ceur-ws.org/Vol-3537/paper4.pdf

[2] Andrea Bacciu, Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, and Fabrizio Silvestri. 2023. Fauno: The Italian Large Language Model that will leave you senza parole!. In *Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023), Pisa, Italy, June 8-9, 2023 (CEUR Workshop Proceedings, Vol. 3448)*. CEUR-WS.org, 9–17. https://ceur-ws.org/Vol-3448/paper-24.pdf

[3] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems* 35 (2022), 31668–31683.

[4] Manish Bhatt, Sahana Chennabasappa, Cyrus Nikolaidis, Shengye Wan, Ivan Evtimov, Dominik Gabi, Daniel Song, Faizan Ahmad, Cornelius Aschermann, Lorenzo Fontana, et al. 2023. Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models.

[5] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. , 2206–2240 pages.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. *arXiv preprint arXiv:2401.14887* (2024).

[8] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

[9] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations.

[10] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B.

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[12] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

[13] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey.

[14] Fabio Pinelli, Gabriele Tolomei, and Giovanni Trappolini. 2023. FLIRT: Federated Learning for Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3472–3475.

[15] Artsiom Sauchuk, James Thorne, Alon Halevy, Nicola Tonellotto, and Fabrizio Silvestri. 2022. On the Role of Relevance in Natural Language Processing Tasks. , 1785–1789 pages.

[16] Wang-Chiew Tan, Jane Dwivedi-Yu, Yuliang Li, Lambert Mathias, Marzieh Saeidi, Jing Nathan Yan, and Alon Y Halevy. 2023. TimelineQA: A Benchmark for Question Answering over Timelines.

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models.

[18] Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, Alon Halevy, and Fabrizio Silvestri. 2023. Multimodal Neural Databases. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2619–2628. https://doi.org/10.1145/3539618.3591930