

R³AG: First Workshop on Refined and Reliable Retrieval Augmented Generation

Zihan Wang

University of Amsterdam
Amsterdam, The Netherlands
zihanwang.sdu@gmail.com

Haitao Yu

University of Tsukuba
Tsukuba, Japan
yuhaitao@slis.tsukuba.ac.jp

Xuri Ge

University of Glasgow
Glasgow, United Kingdom
x.ge.2@research.gla.ac.uk

Joemon M. Jose

University of Glasgow
Glasgow, United Kingdom
joemon.jose@glasgow.ac.uk

Weizhi Ma

Tsinghua University
Beijing, China
mawz12@hotmail.com

Zhaochun Ren

Leiden University
Leiden, The Netherlands
z.ren@liacs.leidenuniv.nl

Xin Xin

Shandong University
Shandong, China
xinxin@sdu.edu.cn

Abstract

Retrieval-augmented generation (RAG) has gained wide attention as the key component to improve generative models with external knowledge augmentation from information retrieval. It has shown great prominence in enhancing the functionality and performance of large language model (LLM)-based applications. However, with the comprehensive application of RAG, more and more problems and limitations have been identified, thus urgently requiring further fundamental exploration to improve current RAG frameworks. This workshop aims to explore in depth how to conduct refined and reliable RAG for downstream AI tasks.

To this end, we propose to organize the first R³AG workshop at SIGIR-AP 2024 to call for participants to re-examine and formulate the basic principles and practical implementation of refined and reliable RAG. The workshop serves as a platform for both academia and industry researchers to conduct discussions, share insights, and foster research to build the next generation of RAG systems. Participants will engage in discussions and presentations focusing on fundamental challenges, cutting-edge research, and potential pathways to improve RAG. At the end of the workshop, we aim to have a clearer understanding of how to improve the reliability and applicability of RAG with more robust information retrieval and language generation.

CCS Concepts

- Information systems → Information retrieval; • Computing methodologies → Natural language generation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR-AP '24, December 9–12, 2024, Tokyo, Japan

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0724-7/24/12
<https://doi.org/10.1145/3673791.3698435>

Keywords

Retrieval-Augmented Generation, Information Retrieval, Large Language Models, Reliability

ACM Reference Format:

Zihan Wang, Xuri Ge, Joemon M. Jose, Haitao Yu, Weizhi Ma, Zhaochun Ren, and Xin Xin. 2024. R³AG: First Workshop on Refined and Reliable Retrieval Augmented Generation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '24), December 9–12, 2024, Tokyo, Japan*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3673791.3698435>

1 BACKGROUND AND MOTIVATIONS

RAG (Retrieval-Augmented Generation) has emerged as a new paradigm for using information retrieval (IR) to improve the generated response from large language models (LLMs). On the one hand, traditional IR systems may encounter difficulties to handle more and more complex information seeking queries. On the other hand, LLM has shown notable natural language understanding capability while suffering from fictitious or inaccurate generation, also known as the hallucination problem. To this end, RAG has emerged to combine the best of IR and LLM generation. RAG has made significant progress in improving response quality by first retrieving relevant knowledge from external knowledge base and then generating responses based on the knowledge retrieval. RAG's advantages in handling information queries include but not limited to enhanced user experience, enriched information return, improved response accuracy, and multi-round conversational search for complex queries. Through integrating IR and language generation, RAG has become the keystone for various AI applications.

Existing RAG techniques [7, 14] focused on enhancing language models by integrating additional textual knowledge from external knowledge database. Transformer-based [18] language models have shown great promise in language generation, leading to notable LLMs like GPTs. LLMs owe their success to advanced architectures with billions of parameters, pre-trained on vast corpora from diverse sources, enabling remarkable generalization across

various AI applications [4, 8, 12]. However, LLMs also suffer from model hallucination [13] and difficulty in handling dynamic knowledge updates [19]. RAG alleviates the hallucination problem by providing LLMs with relevant knowledge using IR techniques to retrieve from external databases, achieving more accurate responses for knowledge-intensive generation. The decoupled database also supports more lightweight knowledge dynamic updates. For example, [15] integrates a RAG pipeline in an end-to-end generation system to improve the factual correctness of LLMs for domain-specific and time-sensitive queries.

While RAG has achieved great advancement, we recognize that there still exists challenges to conduct refined and reliable RAG (R^3AG). A typical RAG pipeline often includes user intent comprehension, knowledge parsing, knowledge retrieval, and response generation. Each pipeline step has specific challenges and plays an essential role to accomplish user queries. For example, how to understand user query intention under long and complex dialogue context; how to parse complex knowledge documents including tables and figures; how to conduct reliable knowledge retrieval; and how to refine the generated response. The workshop is expected to help researchers to conduct further investigation on R^3AG .

Topic. The topic of this workshop includes interesting points related to the RAG pipelines, i.e., user intent comprehension, knowledge parsing, knowledge retrieval, and response generation. The detailed topics are described in section 2.

Relevance. On the one hand, generative LLMs are hot research topics in the IR communities. The capability of LLMs enriches the scope of IR and has changed the process of information seeking to a large extend. On the other hand, RAG is one of the most important applications of IR in LLMs. IR plays a new and essential role in LLM-based downstream tasks. To this end, this workshop of R^3AG is highly relevant to the SIGIR-AP conference.

Motivation. Recently, the potential of RAG has been verified in various AI applications. However, there still exists fundamental research challenges to further improve current RAG methods. SIGIR-AP is one of the leading conferences with numerous recognized research works focusing on IR-related research and applications. We believe that organizing the R^3AG workshop with SIGIR-AP now can (i) stimulate interesting research to meet complex information seeking demands; (ii) encourage the community to conduct in-depth research and practical applications on refined and reliable RAG; (iii) expand the impact of SIGIR-AP conferences and the workshop.

2 FUNDAMENTAL CHALLENGES AND TOPICS

This year we will focus more on fundamental challenges in this field and expect thorough discussions during the R^3AG workshop.

User Intent Comprehension. User intent comprehension directly affects the following retrieval and final response generation. However, user intent comprehension could be challenging especially under long dialogue context. R^3AG is expected to help users clarify their information needs during the interaction process. Besides, R^3AG should have the capability to split complex queries into simple queries to accomplish user demands. Related methods include

query expansion, which introduces hypothetical answer generation from LLMs into the retrieval process to improve the retrieval relevance, query summarization [11], query rewrite [16], etc.

Query & Knowledge Encoding. The effectiveness of RAG depends heavily on the representation learning of both queries and knowledge. Mainstream methods use pre-trained feature extractors [10, 17] to encode them, while there could be a misalignment between diverse knowledge and user queries. Currently, fine-tuning based query-knowledge encoding plays an important role in RAG. However, there still lacks sufficient exploration to conduct more effective query & knowledge encoding for R^3AG , especially in terms of efficient fine-tuning to support knowledge updates.

RAG for Complex Documents. RAG systems aim to deliver knowledge chunks to improve LLM generation. However, existing RAG methods could encounter difficulties to parse complex documents with embedded tables and figures. How to parse complex tabular knowledge is still an open research question. Besides, the chunks division strategy also affects response generation. Too large chunks introduce irrelevant information while small chunks could lead to incomplete knowledge. In addition, multi-hop document retrieval is also a challenge. The above three issues need to be further investigated to conduct R^3AG .

Reliable Retrieval for RAG. The inclusion of noisy or contradictory information during retrieval can significantly impair the performance of existing RAG models. There has been growing interest in improving the resilience of RAG against harmful or counterfactual inputs, which is now considered a crucial performance indicator in recent studies [5, 20, 21]. Meanwhile, current research [9] reveals that including irrelevant documents can unexpectedly increase accuracy by more than 30%, challenging the initial belief that it would degrade quality. These findings inspire us to organize R^3AG for developing specialized strategies for effective retrieval and to emphasize the ongoing need for further investigation into RAG's robustness and reliability.

Response Evaluation and Refinement. While RAG enhances LLMs by providing additional information, LLMs may still face challenges with unreliable or inaccurate response generation. These challenges may arise from incorrect information in the context [6] or issues with hallucinations [2]. Unfortunately, there is a noticeable gap in understanding how these challenges impact the output quality of RAG, as well as in developing strategies for models to mitigate these issues and refine their responses. As such, R^3AG is dedicated to comprehensively evaluating and enhancing the response quality of RAG-based LLMs across multiple dimensions, such as relevance, faithfulness, negative rejection, information integration, creative generation, and error correction.

Multimodal R^3AG . Most current research focuses on textual RAG, despite the need for multimodality in many applications. Recent large-scale models like Flamingo [3], and GPT-4 [1] show significant multimodal capability when scaled to tens of billions of parameters and trained on extensive multimodal corpora. These large multimodal models require RAG even more than unimodal textual LLMs for external knowledge support. Therefore, the workshop encourage discussions regarding R^3AG for large multimodal models.

3 PROGRAM SKETCH

3.1 Workshop Format

The workshop is planned to be hosted for half a day, including 2 invited talks and 4 oral research talks. There are two encouraging types of invited talks: (i) academic talks on fundamental research on the adaptability and reliability of RAG techniques; and (ii) industrial talks on the practice of designing or applying refined and reliable RAG techniques for real-world applications, including information retrieval and generative systems.

Each talk should be delivered as a slide-based lecture. A Q&A session will follow the conclusion of each talk.

3.2 Tentative Workshop Schedule

The workshop schedule is planned with one half-day session:

- 9:00 - 9:10 Welcome & opening
- 9:10 - 9:50 Academic invited talk
- 9:50 - 10:30 Industrial invited talk
- 10:30 - 10:50 Coffee break
- 10:50 - 11:40 Oral paper talks
- 11:40 - 12:00 Panel discussion

Tentative Speakers. The tentative speakers include academia researchers, such as Dr. Xiangyu Zhao from the City University of Hong Kong, and Prof. Hideo Joho from University of Tsukuba and industrial staff, such as Dr. Alexandros Karatzoglou from Amazon.

Panel Discussion. We are also considering hosting a panel discussion as the final part of the workshop. The decision to include this session will depend on the availability of panelists attending SIGIR-AP and the number of accepted research papers.

Contingency Plan. To ensure a successful workshop, our contingency plan identifies key requirements (venue, speakers, materials) and assesses risks. We will have backup venues, alternate speakers, and a ready team for onsite support. A designated lead will oversee a communication plan, with specific roles for last-minute organizers to manage registration, technical support, and logistics.

3.3 Selection Process

Each invited speaker should be highly esteemed within the community. Invitations should be agreed upon by all organizers without any disagreement. The workshop accepts paper submissions via a standard peer-review process, expects 6~10 paper submissions, and accepts 3~4 papers. Each submission is evaluated by at least two members of the program committee. The senior PCs or workshop organizers will make the final decision. Authors will receive detailed review comments and a notification letter.

Tentative Program Committee. In addition to the current 7 organizers, we also plan to invite the following tentative PC members: (1) Dr. Chao Huang from University of Hong Kong, (2) Dr. Xiangyu Zhao from the City University of Hong Kong, (3) Dr. Andrew Yates from University of Amsterdam, and (4) Dr. Alexandros Karatzoglou from Amazon.

3.4 Online Materials

A website for the R³AG workshop will be made available online. All relevant materials, including talk information, presentation slides,

referred papers, speaker details, and related open-source projects, will be accessible on this website.

3.5 Workshop Advertisement

The R³AG workshop will be promoted on various social media platforms to increase visibility and encourage paper submissions. These platforms include, but are not limited to, Twitter, Facebook, and WeChat. Additionally, the organizers will send personalized emails to further advertise the workshop.

4 RELATED WORKSHOPS

List of related workshops:

- Information Retrieval's Role in RAG Systems (SIGIR 2024¹)
- Multimodal Representation and Retrieval (SIGIR 2024²)
- Information Retrieval Meets Large Language Models (WWW 2024³)
- Large Knowledge-Enhanced Models (IJCAI 2024⁴)
- Knowledge Retrieval and Language Models (ICML 2022⁵)

The Information Retrieval's Role in RAG Systems workshop at SIGIR (2024) explored retrieval's integral role in RAG frameworks. As multimodal LLMs and RAG gain traction, the Multimodal Representation and Retrieval workshop at SIGIR (2024) introduced the challenge of multimodal queries and documents. The Information Retrieval Meets LLMs workshop at WWW (2024) addressed issues like retrieval-generation collaboration and hallucination. The Large Knowledge-Enhanced Models workshop at IJCAI (2024) discussed integrating LLMs with symbolic knowledge, while the Knowledge Retrieval and Language Models workshop at ICML (2022) highlighted the limitations of knowledge retrieval. R³AG is the first workshop to focus on refined and reliable RAG techniques. It will feature invited talks, paper presentations, and the release of real datasets and code for future practice.

5 ORGANIZERS INFORMATION

Prof. Joemon M. Jose is a Professor at the School of Computing Science, University of Glasgow, Scotland and a member of the Information Retrieval group. His research focuses on the following three themes: (i) Social Media Analytics; (ii) Multi-modal LLMs for information retrieval; (iii) Multimedia mining and search. He has published over 300 papers with more than 10,000 Google Scholar citations, and an h-index of 51. He leads the GAIR Lab investigating research issues related to the above themes. He has been serving as the program committee chair and member for numerous top international conferences (e.g., SIGIR, WWW, and ECIR). He also serves as a PC chair for SIGIR-AP 2024.

¹<https://coda.io/@rstless-group/ir-rag-sigir24>

²<https://mrr-workshop.github.io/>

³<https://irmeestsllm.github.io/>

⁴<https://lkm2024.openkg.org/>

⁵<https://knowledge-retrieval-workshop.github.io/>

Dr. Zhaochun Ren is an Associate Professor at Leiden University. His research interests focus on joint research in IR and natural language processing, with an emphasis on conversational information seeking, question-answering, and recommender systems. He aims to develop intelligent systems that can address complex user requests and solve core challenges in both information retrieval and natural language processing towards that goal. In addition to his academic experience, he worked on e-commerce search and recommendation at JD.com for 2+ years. He has co-organized workshops at SIGIR (2020), WSDM (2019, 2020), and ECIR 2025.

Dr. Haitao Yu is a Tenured Associate Professor at University of Tsukuba and leading the Information Intelligence research group. His research focuses on Information Retrieval, Knowledge Graph, and Machine Learning. He has published numerous papers on top international conferences (e.g., WSDM, CIKM, SIGIR, WWW, ECIR, and AAAI) and journals (e.g., Information Processing and Management, and Information Retrieval Journal). He is the co-organizer of the NTCIR tasks of Temporalia-2 and AKG. He has been serving as the program committee member for numerous top international conferences (e.g., WSDM, CIKM, SIGIR, and ECIR).

Dr. Xin Xin is a Tenure-Track Assistant Professor at the School of Computer Science and Technology of Shandong University. Before that, he earned his Ph.D. degree from the University of Glasgow. His current research interests include information retrieval, natural language processing, and reinforcement learning. He has published more than 40 papers in top conferences (e.g., WWW, SIGIR, ACL, WSDM) and journals (e.g., TOIS, TKDE), and received the Best Paper Honor Mention at WSDM 2024. He has organized the DRL4IR workshop in SIGIR and KEIR workshop in ECIR.

Dr. Weizhi Ma is a Research Assistant Professor at the Institute for AI Industry Research (AIR) at Tsinghua University. He earned his B.E. and Ph.D. in the Department of Computer Science and Technology from Tsinghua University. Dr. Ma's research focuses on information retrieval, recommender systems, and LLM-powered agents. He has published over 70 papers in leading international conferences and journals, including TOIS, TKDE, SIGIR, KDD, AAAI, IJCAI, WSDM, CIKM, etc. His accolades include the Best Paper Honorable Mention Award at SIGIR 2020 and several other paper awards. In addition to his research, Dr. Ma is the Secretary of the Youth Working Committee of the Chinese Information Processing Society of China and serves as an Assistant Editor for ACM TOIS. He is a recipient of the Young Elite Scientists Sponsorship Program by CAST and the Shuimu Tsinghua Scholar Program.

Zihan Wang is a fourth-year PhD student at the University of Amsterdam (UvA). He has published over ten papers in prestigious conferences including KDD, SIGIR, CCS, and WSDM. He received the Best Student Paper Award at WSDM 2018. His current research focuses on information extraction, knowledge graph embedding, and the reliability of large language models.

Xuri Ge is a fourth-year PhD student at the University of Glasgow (UofG). He has published over 15 papers in prestigious conferences and journals, including ACM MM, SIGIR, NeurIPS, IP&M, CIKM, ICME and ACM TIST, etc. His current research focuses on information retrieval, efficient multimodal representation learning,

and multimodal large language models. Xuri has organized 3DMM workshop in IEEE ICME2024.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Vaibhav Adlakha, Parishad BehnamChader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. *Trans. Assoc. Comput. Linguistics* 12 (2024), 681–699.
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aidi Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*.
- [4] Mohammad Alkhafal, Ping Yu, Mengyang Yin, and Chao Deng. 2024. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics* (2024), 104662.
- [5] Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park, and Sung Ju Hwang. 2023. Knowledge-Augmented Language Model Verification. In *EMNLP*. 1720–1736.
- [6] Ning Bian, Hongyu Lin, Peilin Liu, Yaojie Lu, Chunkang Zhang, Ben He, Xianpei Han, and Le Sun. 2023. Influence of external information on large language models mirrors social cognitive patterns. *arXiv preprint arXiv:2305.04812* (2023).
- [7] 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. In *ICML*. 2206–2240.
- [8] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *AAAI*, Vol. 38. 17754–17762.
- [9] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellootto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. *arXiv preprint arXiv:2401.14887* (2024).
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [11] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).
- [12] Sefika Efeoglu and Adrian Paschke. 2024. Retrieval-Augmented Generation-based Relation Extraction. *arXiv preprint arXiv:2404.13397* (2024).
- [13] 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. *arXiv preprint arXiv:2311.05232* (2023).
- [14] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*.
- [15] Jiaru Li, Ye Yuan, and Zehui Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446* (2024).
- [16] Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. RaFe: Ranking Feedback Improves Query Rewriting for RAG. *arXiv preprint arXiv:2405.14431* (2024).
- [17] Alex Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. 8748–8763.
- [18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.
- [19] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *EMNLP*.
- [20] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558* (2023).

- [21] Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented

language models. *arXiv preprint arXiv:2311.09210* (2023).

This figure "acm-jdslogo.png" is available in "png" format from:

<http://arxiv.org/ps/2410.20598v2>

This figure "sample-franklin.png" is available in "png" format from:

<http://arxiv.org/ps/2410.20598v2>