Information Retrieval Meets Large Language Models

Zheng Liu*
zhengliu1026@gmail.com
BAAI
Beijing, China

Yujia Zhou*
zhouyujia@ruc.edu.cn
Renmin University of China
Beijing, China

Yutao Zhu*
yutaozhu94@gmail.com
Renmin University of China
Beijing, China

Jianxun Lian
jianxun.lian@microsoft.com
Microsoft Research Asia
Beijing, China

Chaozhuo Li
cli@microsoft.com
Microsoft Research Asia
Beijing, China

Zhicheng Dou
dou@ruc.edu.cn
Renmin University of China
Beijing, China

Defu Lian
liandefu@ustc.edu.cn
University of Science and Technology of China
Hefei, China

Jian-Yun Nie
nie@iro.umontreal.ca
University of Montreal
Montreal, Canada

ABSTRACT

The advent of large language models (LLMs) presents both opportunities and challenges for the information retrieval (IR) community. On one hand, LLMs will revolutionize how people access information, meanwhile the retrieval techniques can play a crucial role in addressing many inherent limitations of LLMs. On the other hand, there are open problems regarding the collaboration of retrieval and generation, the potential risks of misinformation, and the concerns about cost-effectiveness. To seize the critical moment for development, it calls for the joint effort from academia and industry on many key issues, including identification of new research problems, proposal of new techniques, and creation of new evaluation protocols. It has been one year since the launch of ChatGPT in November last year, and the entire community is currently undergoing a profound transformation in techniques. Therefore, this workshop will be a timely venue to exchange ideas and forge collaborations. The organizers, committee members, and invited speakers are composed of a diverse group of researchers coming from leading institutions in the world. This event will be made up of multiple sessions, including invited talks, paper presentations, hands-on tutorials, and panel discussions. All the materials collected for this workshop will be archived and shared publicly, which will present a long-term value to the community.

CCS CONCEPTS

• **Information systems** → **Language models**.

*These authors contribute equally to the organization of this workshop.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05.
https://doi.org/10.1145/3589335.3641299

KEYWORDS

Information Retrieval; Large Language Models; Search; Ranking; Question Answering; Retrieval-Augmented Generation

ACM Reference Format:
Zheng Liu, Yujia Zhou, Yutao Zhu, Jianxun Lian, Chaozhuo Li, Zhicheng Dou, Defu Lian, and Jian-Yun Nie. 2024. Information Retrieval Meets Large Language Models. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3641299

1 INTRODUCTION

Information retrieval (IR) is the process of bridging people with their needed data and knowledge. Over the past few decades, the field of information retrieval has undergone significant developments, largely propelled by AI techniques. In recent years, deep learning become one of the major driving forces in this area. The introduction of word embeddings and deep neural networks marked a preliminary but important milestone. By transforming text representations from the lexical space into the latent space, it is enabled to model complex semantic relationship between query and document. Later on, the pre-trained language models further advanced the impact of deep neural representations. With language models pre-trained from massive corpora, the text understanding and generation quality have been substantially improved. Finally, the advent of large language models (LLMs) presents the most transformative moment in this technical trend. On one hand, LLMs can be used as a powerful and unified backbone to support every functionality within the conventional IR systems, such as first-stage retrieval, re-ranking, query reformulation, and answer generation. On the other hand, LLMs exhibit superior emergent capabilities in dealing with complicated reasoning and planning tasks. Therefore, it showcases the potential to build the IR-oriented autonomous agent, which will fully automate the information acquisition process.

Besides LLMs' propelling impact on IR, the retrieval techniques also play an important role to enhance LLMs through retrieval-augmented generation. For example, the factuality of LLMs can benefit from the knowledge retrieved from an external database.

The long-term memory of LLMs can be established by retrieving the off-loaded context in history. The retrieval and generation paradigm can also result in more cost-effectiveness architectures for LLMs. Finally, the retrieval-augmented generation is a common approach for LLMs to support domain-specific applications.

To conclude, there come two distinct but interrelated research directions for IR in the era of LLMs, namely **LLM-For-IR**, where people take advantage of LLMs to realize more precise, interactive, and automated IR process, and **IR-For-LLM**, where retrieval techniques help to address the inherent limitations of LLMs.

1.1 Statement of Objectives

It has been one year since ChatGPT was launched in last November. The entire industry has been experiencing profound changes since that time, where unprecedented opportunities come forth in this historical moment. Many conventional IR functionalities and products are being dramatically shaped by the new powerful foundation models. Meanwhile, the IR tools are intensively integrated into the LLM-based applications. Besides, there are also numerous challenges to be addressed. For example, it is needed to explore the effective collaboration mechanism between retrieval and generation. The retrieval-augmented generation is still prone to hallucination phenomenons, whose result can be even harder to discriminate. It remains to work out the alignment strategy for IR-oriented LLMs where the generation result can best satisfy people's information need. Given the above context, our workshop will contribute to the following objectives.

- **Problem identification**. First and foremost, we aim to identify and define the new research problems, which are of high importance in the coming stage.
- **Knowledge sharing**. The invited speakers will be encouraged to present their research insights and their hands-on developing experiences in this field.
- **Resource Publishing**. We also expect the introduction of new resources in this event, including training data, benchmarks, and evaluation protocols.
- **Cross-domain Collaboration**. The workshop will be oriented to researchers and developers in both academia and industry. The presence of both types audience will facilitate the collaborations between the two communities.

1.2 Description of Format

We plan to organize the full-day workshop so as to accommodate four different sessions. The first session will comprise 3 invited talks from the leading researchers in this area. The second session will be paper presentations, where 3 quality research papers will be selected to present. The third session will be hands-on tutorial, where experienced researchers will be asked to share their developing experiences on LLMs and retrieval-augmented generation. The last session will be the panel discussion, which will focus on the exchange of visions and ideas about the future research.

1.3 Topic and Themes

Yutao and Yujia: Please put down some of your interested topic and themes in this place.

1.3.1 Application of LLMs for information retrieval.

- **LLMs for Query Understanding and Reformulation:**
  - Using LLMs to understand the intent behind ambiguous queries.
  - LLMs for query expansion with semantic understanding.
  - The role of context in LLM-powered query reformulation.
- **LLMs for User Behavior Understanding:**
  - Predicting user satisfaction using LLMs in search sessions.
  - Personalization of search results through LLM analysis of historical user data.
- **Personalized Techniques Based on LLMs:**
  - Creating user profiles using LLMs to enhance search result relevance.
  - Personalized knowledge graphs constructed with the help of LLMs.
- **LLM-driven Conversational Search:**
  - Dialogue systems for search powered by LLMs.
  - Continuous learning from user interactions in conversational IR.
- **LLM-based Differential Index:**
  - Building a model-based indexing system with LLMs for generative retrieval.
  - Index pruning and optimization through LLMs' semantic understanding.
  - LLMs for creating abstract representations of documents for quicker retrieval.
- **LLMs for Ranking and Matching:**
  - LLMs for contextual ranking of search results.
  - Semantic matching of queries to documents using LLMs.
  - The use of LLMs in multi-modal search result ranking.
- **Evaluation Metrics Based on LLMs:**
  - Developing new IR evaluation metrics using LLMs' language understanding capabilities.
  - Using LLMs to automate relevance judgement for IR evaluation.
  - LLMs in the emulation of user satisfaction for search result testing.
- **LLM-based Data Augmentation for Information Retrieval:**
  - Generating synthetic queries for IR system training using LLMs.
  - Enhancing corpora diversity for IR evaluation with LLM-generated content.

1.3.2 Incorporation of information retrieval to LLMs.

- **LLM Pre-training with Retrieval:**
  - Combining traditional IR techniques with LLM pre-training for improved domain adaptation.
  - Retrieval-enhanced pre-training strategies for LLMs.
  - The impact of incorporating IR tasks during the pre-training of LLMs.
- **Enhancing LLM by Retrieval Adapters:**
  - Modular retrieval adapters for LLMs to fine-tune them for specific IR tasks.
  - Customizable IR features in LLMs using retrieval adapters.
  - Improving LLM transfer learning with retrieval-focused adapters.

- **Knowledge-Enriched LLMs for IR:**
  - Integrating external knowledge bases with LLMs for IR.
  - Using IR to provide real-time data for LLM responses.
  - Enhancing LLMs' factual accuracy and timeliness with dynamic retrieval methods.
- **Retrieval Augmented Generation (RAG) for LLMs:**
  - Leveraging document retrieval to enhance LLM responses.
  - Comparing RAG with end-to-end LLMs in terms of information accuracy.
  - The complex reasoning strategies of retrieval-augmented LLMs.
  - Benchmarking the performance of retrieval-augmented LLMs with long-form answers.
- **Hybrid Models Combining LLMs and Classic IR Models:**
  - Case studies of hybrid models' performance in specialized search domains such as legal documents, medical records, and academic research.
  - The role of LLMs in enhancing the feature extraction capabilities of classic IR models, thereby improving the semantic matching of queries and documents.
  - Strategies for maintaining the interpretability and explainability of IR systems when incorporating black-box LLM components.

1.3.3 Training and Reasoning Strategies of LLMs in IR scenarios.

- **Incorporating Feedback Loops in LLM Training:**
  - Implementing reinforcement learning from user interactions.
  - Utilizing click-through rates and other implicit feedback for model fine-tuning.
  - Feedback-based curriculum learning for progressive LLM training.
- **Multi-task Learning for LLMs in IR:**
  - Combining query understanding with relevance feedback in a multi-task framework.
  - Shared representations for different IR tasks in LLMs.
  - Cross-lingual and cross-domain multi-task learning strategies.
- **Meta-learning and Few-shot Learning:**
  - Applying meta-learning for rapid adaptation to new IR tasks.
  - Few-shot learning techniques for query classification and result ranking.
  - LLMs and the challenges of sparse data in specialized IR domains.
- **Explainable AI in IR:**
  - Techniques for interpreting LLM predictions in IR.
  - Building trust through explainable LLM-enhanced search results.
  - Visualizing LLM reasoning for IR professionals and end-users.
- **Transfer Learning and Domain Adaptation:**
  - Utilizing transfer learning from large-scale datasets to IR applications.
  - Domain adaptation techniques for LLMs in changing information landscapes.
  - Cross-domain knowledge transfer and its implications for IR.
- **Scalability and Efficiency in LLM Training:**
  - Efficient training strategies for LLMs with large-scale IR datasets.
  - Scaling up LLMs for enterprise-level IR applications.

Extensions.

- **Application in multi-lingual scenarios:**
  - Investigating the capacity of LLMs to facilitate cross-lingual information retrieval, enabling users to query in one language and retrieve relevant information in another.
  - Enhancing LLMs with multi-lingual corpora to improve the accuracy and coverage of IR across diverse languages.
- **Application in multi-modal scenarios:**
  - Expanding LLM capabilities to interpret and index multi-modal data, such as images, videos, and audio, alongside textual information for comprehensive IR.
  - Integrating LLMs with computer vision and audio processing techniques to create unified multi-modal search platforms.

1.4 Targeted Audience

This workshop is mainly oriented to all levels of researchers in IR and LLM communities. The shared content will be beneficial to a wide spectrum of scholars in other related areas, such as natural language processing, multi-modality research, data mining, and general artificial intelligence. The event will emphasize on detailed technical practice; therefore, the practitioners and developers will find it useful as well.

1.5 Tentative Program Committee List

The program committee will be composed of a diverse group of researchers working in IR, NLP, and AI, who are from both industrial labs and universities. The tentative members are listed as follows.

- **Xing Xie**, senior principal researcher in Microsoft Research Asia.
- **Yeyun Gong**, principal researcher in Microsoft Research Asia.
- **Qi Chen**, principal researcher in Microsoft Research Asia.
- **Prabhat Agarwal**, senior machine learning engineer in the applied science team at Pinterest.
- **Ruiming Tang**, research manager in Huawei Noah's Ark Lab.
- **Bo Zhao**, researcher in Beijing Academy of Artificial Intelligence.
- **Qingyao Ai**, assistant professor in Tsinghua University.
- **Ting Bai**, assistant professor in Beijing University of Posts and Telecommunications.
- **Jiongnan Liu**, Ph.D. student in Renmin University of China.
- **Shuting Wang**, Ph.D. student in Renmin University of China.
- **Fengran Mo**, Ph.D. student in University of Montreal.

1.6 Workshop Organizer

The organizers of this workshops are introduced as follows.

- **Zheng Liu**. Zheng Liu is currently a researcher with Beijing Academy of Artificial Intelligence (BAAI). He received his Ph.D. degree from Hong Kong University of Science and Technology (HKUST) in 2018. He worked as a senior researcher in Microsoft Research Asia (MSRA) and technical specialist in Huawei 2012

Labs. His research interests include information retrieval, natural language model processing, and recommender systems.

• **Yujia Zhou** received the BE degree in computer science and technology from School of Information, Renmin University of China, in 2019. And he is studying for PhD in the School of Information, Renmin University of China. He won the best student paper award in CCIR 2018. His research interests include personalized search, model-based information retrieval, and retrieval-augmented generation.

• **Yutao Zhu** received the B.S. and M.S. degree from Renmin University of China, and the Ph.D. degree from the University of Montreal. He is currently a postdoc at Renmin University of China. His current research interests are Large Language Models and Information Retrieval. He received the Best Paper Award from CCIR 2021 and the Google Scholarship for UdeM in 2019. He served as the PC member of several top-tier conferences, such as ACL, SIGIR, SIGKDD, AAAI, EMNLP, etc.

• **Zhicheng Dou** is currently a professor at Renmin University of China. He received his Ph.D. and B.S. degrees in computer science and technology from Nankai University in 2008 and 2003, respectively. He worked at Microsoft Research Asia from July 2008 to September 2014. His current research interests are Information Retrieval, Natural Language Processing, and Big Data Analysis. His homepage is http://playbigdata.ruc.edu.cn/dou.

• **Jianxun Lian** is now a senior researcher at Microsoft Research Asia. He received his Ph.D. degree from University of Science and Technology of China in 2018. His research interests include recommender systems and LLM-based agents. He has published 30+ academic papers on top-tier international conferences such as KDD, ACL, WWW, and SIGIR. He actively contributes as a program committee member for prestigious conferences such as AAAI, WWW, and IJCAI.

• **Chaozhuo Li** received his Ph.D. degree in computer software and theory from School of Computer Science and Engineering, Beihang University, Beijing, China, in 2020. He is currently a senior researcher at Microsoft Research Asia, Beijing, China. He has published over 70 papers, such as NeurIPS, AAAI, SIGIR, KDD and CIKM. His research interests include data mining and social network analysis.

• **Defu Lian** received the PhD degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2014. He is currently a professor with the School of Computer Science and Technology, USTC. He has published prolifically in referred journals and conference proceedings, such as TIST, TKDE, AAAI, KDD, SIGIR, IJCAI, WWW. His current research interests include spatial data mining, recommender systems, and learning to hash.

• **Jian-Yun Nie** is a professor with the University of Montreal, Canada. He has published more than 150 papers in information retrieval and natural language processing in journals and conferences. He served as a general co-chair of the ACM SIGIR Conference in 2011. He is currently on the editorial board of seven international journals. He has been an invited professor and researcher at several universities and companies.

ACKNOWLEDGEMENT

The workshop is supported by grants from the National Key R&D Program of China (No. 2021ZD0111801). We also thank all PC members for their valuable assistance of this workshop.