LLM4Eval: Large Language Model for Evaluation in IR

Clemencia Siro
University of Amsterdam
Amsterdam, The Netherlands
c.n.siro@uva.nl

Hossein A. Rahmani
University College London
London, UK
hossein.rahmani.22@ucl.ac.uk

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Nick Craswell
Microsoft
Bellevue, US
nickcr@microsoft.com

Charles L. A. Clarke
University of Waterloo
Waterloo, Ontario, Canada
claclark@gmail.com

Guglielmo Faggioli
University of Padua
Padua, Italy
faggioli@dei.unipd.it

Bhaskar Mitra
Microsoft
Montréal, Canada
bmitra@microsoft.com

Paul Thomas
Microsoft
Adelaide, Australia
pathom@microsoft.com

Emine Yilmaz
University College London & Amazon
London, UK
emine.yilmaz@ucl.ac.uk

Abstract

Large language models (LLMs) have demonstrated increasing task-solving abilities not present in smaller models. Utilizing the capabilities and responsibilities of *LLMs for automated evaluation* (LLM4Eval) has recently attracted considerable attention in multiple research communities. Building on the success of previous workshops, which established foundations in automated judgments and RAG evaluation, this third iteration aims to address emerging challenges as IR systems become increasingly personalized and interactive. The main goal of the third LLM4Eval workshop is to bring together researchers from industry and academia to explore three critical areas: the evaluation of personalized IR systems while maintaining fairness, the boundaries between automated and human assessment in subjective scenarios, and evaluation methodologies for systems that combine multiple IR paradigms (search, recommendations, and dialogue). By examining these challenges, we seek to understand how evaluation approaches can evolve to match the sophistication of modern IR applications. The format of the workshop is interactive, including roundtable discussion sessions, fostering dialogue about the future of IR evaluation while avoiding one-sided discussions. This is the third iteration of the workshop series, following successful events at SIGIR 2024 and WSDM 2025, with the first iteration attracting over 50 participants.

CCS Concepts

• **Information systems → Evaluation of retrieval results**.

Keywords

Generative Models, Large Language Models, Automated Evaluation

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730367

ACM Reference Format:
Clemencia Siro, Hossein A. Rahmani, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2025. LLM4Eval: Large Language Model for Evaluation in IR. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3726302.3730367

1 Title & Motivation

Title. LLM4Eval @ SIGIR '25: The Third Workshop on Large Language Models (LLMs) for Evaluation in Information Retrieval.[1]

Motivation. The first LLM4Eval workshop at SIGIR 2024 [11] provided critical insights into the potential of LLMs for evaluation, particularly in search relevance assessment. With 22 accepted papers, over 50 participants, and the LLMJudge challenge [12], the workshop demonstrated that LLMs can generate relevance judgments closely aligned with human assessments, though their effectiveness depends on careful prompt design and systematic validation. It also highlighted challenges in evaluating retrieval-augmented generation (RAG) systems, including balancing retrieval accuracy with the quality and factuality of generated responses. Four major research priorities emerged: ensuring validity in LLM-based evaluations, addressing randomness from prompt and parameter variation, enhancing replicability and reproducibility, and understanding the interplay between human and LLM assessments. These findings highlight the need for further workshops to foster discussion and develop solutions to these pressing challenges.

Building on these findings, the second edition of LLM4Eval workshop [10] colocated with WSDM 2025 focused on addressing fundamental methodological questions raised during the first iteration, particularly in automated judgments and RAG evaluation. However, as information retrieval (IR) systems evolve, the scope of evaluation challenges extends far beyond basic relevance assessment.

[1] https://llm4eval.github.io/

Modern IR applications increasingly integrate elements of search, recommendations, and conversational interfaces while incorporating personalization. This evolution introduces fundamental evaluation challenges that traditional methodologies struggle to address. Search and recommendation systems now dynamically adapt to individual user characteristics, preferences, and interaction histories, creating multiple dimensions of personalization [1]. Systems modify rankings based on user context, tailor result presentations, and generate dynamic, personalized explanations [4, 15]. Each of these dimensions requires new approaches to evaluation: assessing the quality of personalized rankings, determining whether explanations align with user needs, and measuring the system's ability to adapt appropriately to evolving user preferences over time. These challenges underscore the need to rethink evaluation frameworks in ways that account for fairness, adaptability, and the diversity of user experiences.

While LLMs have proven effective at generating relevance judgments and correlating well with human assessments [7], recent studies reveal important limitations in their ability to evaluate subjective aspects of IR systems [3, 14]. For example, conversational interfaces require evaluation of subtle interaction qualities such as coherence, naturalness, and contextual relevance, often demanding nuanced human judgment. Similarly, recommender systems necessitate assessments of explanation quality, long-term engagement, and trustworthiness – factors that are subjective [5, 9]. This raises critical questions about the boundaries between human and automated evaluation. Identifying where LLMs excels and where human insight is indispensable will be key to designing hybrid evaluation frameworks that combine the strengths of both.

Additionally, modern applications no longer operate within isolated domains; instead, they blend traditional search capabilities with conversational interfaces, personalized recommendations, and RAG [6, 13]. This integration creates complex user interaction patterns where a single session might involve initiating a search query, refining it through dialogue, receiving personalized recommendations, and consuming generated explanations. Evaluating such systems requires a holistic approach that not only measures the effectiveness of individual components but also assesses how these components interact and contribute to overall user satisfaction and task completion. Beyond relevance, evaluation must account for inter-component synergy, user engagement across different interaction modes, and the system's ability to deliver coherent and meaningful experiences.

The third iteration of `LLM4Eval` aims to foster critical dialogue about the future of IR evaluation in an era of increasingly sophisticated systems. By bringing together researchers working across different domains - from search to recommendations to conversational systems - we seek to explore challenges, share insights, and identify promising directions for evaluation methodology. Our goal is to advance the conversation about how evaluation approaches can evolve alongside the systems they assess, helping the community chart a course for future research in this rapidly developing field.

2 Theme and Purpose

The third iteration of `LLM4Eval` focuses on "IR Evaluation for Complex, Personalized, and Interactive Systems with LLMs." Building upon the success of previous `LLM4Eval` workshops, we aim to deepen understanding of core IR evaluation challenges and expand into emerging directions.

2.1 Areas of Particular Interest

We especially encourage discussions that:

- Bridge traditional IR evaluation with emerging approaches
- Examine the interplay between search, recommendations, and dialogue
- Address challenges in evaluating complex, multi-component IR systems
- Consider user diversity and fairness in evaluation methodology
- Investigate trade-offs between automation and human judgment
- Study evaluation approaches for novel IR applications
- Focus on risks related to using personalized LLMs as assessors, such as privacy concerns and biases that they might induce or reinforce.

2.2 Distinction from Main Conference Topics

While the main SIGIR conference typically focuses on IR algorithms, systems, and their evaluation, this workshop specifically addresses the evolving nature of evaluation methodology itself. We examine emerging challenges that arise from:

- The intersection of automated and human evaluation approaches
- The need for personalized evaluation frameworks
- Cross-system evaluation challenges spanning search, recommendations, and dialogue
- The evolution of evaluation metrics for complex IR systems

These topics complement the main conference by focusing on methodological challenges that arise as IR systems become more complex and personalized, providing fresh perspectives on how we assess the next generation of IR systems.

3 Format and Planned Activities

3.1 Format and Schedule

We will organize a full-day physical workshop, following the tentative schedule in Table 1.

3.2 Planned Interaction and Engagement

The workshop combines various formats to encourage active participation:

- **Lightning talks:** Brief presentations highlighting key ideas and challenges
- **Interactive poster session:** Extended discussions of presented work
- **Roundtable discussions:** Focused small-group exploration of specific themes

Each session is designed to maximize participant interaction and idea exchange, moving beyond traditional presentation formats to foster genuine dialogue and collaboration.

Table 1: Tentative Schedule for the LLMEval Workshop at SIGIR 2025.

Time	Agenda	Description
9:00 - 9:15	Opening Remarks	Workshop themes and goals
9:15 - 10:00	Keynote Talk	Invited keynote (speaker and topic to be confirmed)
10:00 - 10:30	Paper Presentations (Short Session)	Short session with a few paper presentations
10:30 - 11:00	*Coffee break + Poster Presentations*	Informal poster viewing and discussions
11:00 - 12:30	Paper Presentations (Long Session)	Longer session with multiple paper presentations
12:30 - 13:30	*Lunch break + Poster Presentations*	Networking over lunch and posters
13:30 - 14:00	Poster Session (continued)	Final opportunity to view posters
14:00 - 15:30	Breakout Discussion	Small group discussions on topics decided during the workshop
15:30 - 16:00	*Coffee break*	
16:00 - 16:45	Breakout Discussion Reports	Synthesis and sharing of group outcomes
16:45 - 17:00	Closing Remarks	Summary and next steps

4 Special Requirements

Three or more of the organizers will organize the workshop in person. The only requirement is poster stands during the second half of the workshop.

5 Organizers

The organization team consists of active IR and NLP researchers from both academia and industry.

Clemencia Siro is a fourth-year PhD Student at the University of Amsterdam. Her research focuses on the evaluation of conversational systems from user interactions and user-centric evaluation of and with LLMs. She has previously co-organized workshops at ICLR (2023, 2024) and SIGIR 2024.

Hossein A. Rahmani is a second-year PhD student at the University College London (UCL) advised by Prof. Emine Yilmaz and Nick Craswell. His PhD research focuses on utilizing LLMs to generate synthetic data and labels in information retrieval. He previously co-organised the TREC Deep Learning Track (2023), LLM4Eval, and LLMJudge.

Mohammad Aliannejadi is an Assistant Professor at the University of Amsterdam, the Netherlands. His main research interests are conversational information seeking and recommendation, user simulation, and data augmentation using large language models. Mohammad has organized several workshops and data challenges on various topics, including conversational search and cross-market recommendation at NeurIPS, EMNLP, TREC, WSDM, and ECIR.

Nick Craswell is a Principal Applied Scientist at Microsoft in Redmond Washington, working on enhancing search, recommendation, and other information access methods, for personal and enterprise data such as email, chat, and shared files. This includes work on developing and evaluating generative AI solutions to such problems. He has coordinated multiple past TREC tracks including Web Track, Enterprise Track, Tasks Track, and Deep Learning Track.

Charles Clarke is a Professor in the School of Computer Science at the University of Waterloo, Canada. His research focuses on data-intensive tasks and efficiency, including search, ranking, question answering, and other problems involving human language data at scale. He has previously co-organized workshops at ECIR (2024, 2014, 2011), SIGIR (2016, 2015, 2013, 2012), WSDM (2012), and CHIIR (2023, 2020).

Guglielmo Faggioli is a Post-Doc researcher at the University of Padua (UNIPD), Italy. His main research interests regard Information Retrieval focusing on evaluation, performance modeling, query performance prediction, conversational search systems, and privacy-preserving IR. He contributed as co-editor to the Proceedings of CLEF (2021, 2022, 2023, 2024).

Bhaskar Mitra is a Principal Researcher at Microsoft Research. His research focuses on AI-mediated information and knowledge access and questions of fairness and ethics in the context of these sociotechnical systems. He co-organized several workshops (Neu-IR @ SIGIR 2016-2017, HIPstIR 2019, and Search Futures @ ECIR 2024), shared evaluation tasks (TREC Deep Learning Track 2019-2023, TREC Tip-of-the-Tongue Track 2023-2024, and MS MARCO ranking leaderboards), and tutorials (WSDM 2017-2018, SIGIR 2017, ECIR 2018, and AFIRM 2019-2020).

Paul Thomas is a Senior Applied Scientist at Microsoft. His research is in information retrieval: particularly in how people use web search systems and how we should evaluate these systems, including evaluation with and of large language models. He has previously co-organized the CHIIR and ADCS conferences, various tracks at SIGIR, and TREC tracks.

Emine Yilmaz is a Professor and Turing Fellow at University College London, Department of Computer Science. She also works as an Amazon Scholar as part of the Amazon Alexa team. Her research mainly focuses on retrieval evaluation, task-based information retrieval, misinformation detection, and fairness in machine learning. She has previously organized workshops at various conferences, including ECIR, CIKM, CSCW, WSDM, and NeurIPS. She also co-organized the TREC Tasks Track (2015-2017) and the TREC Deep Learning Track (2019-2023).

6 Program Committee

Below is the list of current PC members:

- Amit Jaspal, Meta
- Hossein A. Rahmani, University College London
- James Mayfield, Johns Hopkins University
- Marwah Alaofi, RMIT University

- Paul Thomas, Microsoft
- Ipsita Mohanty, Carnegie Mellon University
- Zackary Rackauckas, Columbia University
- Yiqun Liu, Tsinghua University
- Senjuti Dutta, Self
- Haolun Wu, Stanford University, Mila - Quebec AI Institute
- Eugene Yang, Johns Hopkins University
- Mahdi Dehghan, Shahid Beheshti University
- Bhashithe Abeysinghe, American Institutes for Research
- Guglielmo Faggioli, University of Padua
- Sean MacAvaney, University of Glasgow
- Karin Sevegnani, Heriot-Watt University
- Yue Feng, University of Birmingham
- Arthur Câmara, Zeta Alpha Vector
- Xi Wang, University of Sheffield

7 Selection Process

We invited submission of papers up to nine pages plus additional space for the references and appendices. Each submission was reviewed by at least three reviewers, evaluating their originality, presentation, clarity, relevance to workshop scopes, and technical soundness. We anticipate a variety of submissions, such as early research findings, reports on original research, resources or toolkits for evaluation, and position papers. The most compelling papers will be selected for oral presentation, while the remaining papers will be presented in a poster session or through brief spotlight presentations. The proceedings of the `LLM4Eval` workshop are non-archival, and authors can resubmit their work to other peer-reviewed venues.

8 Target Audience

With the growing interest in LLMs, especially retrieval-augmented models, we anticipate a diverse audience comprising researchers from both industry and academia engaged in information retrieval and natural language processing research and engineering. We intend to advertise the workshop across various platforms, including social media platforms used by the IR community and Slack (*e.g.*, SIGIR and TREC channels), direct outreach to participants from previous `LLM4Eval` workshop, as well as through mailing lists like SIGIR-List and CorporaList, in addition to a dedicated website.

9 Related Workshop

The most indirectly relevant workshop to `LLM4Eval` is the recent **Information Retrieval Meets Large Language Models (IRLLM)** [8] at TheWebConf 2024 and SIGIR 2024 Workshop on **Generative Information Retrieval (Gen-IR)** [2]. Unlike IRLLM and Gen-IR, `LLM4Eval` offers a venue for discussing and exploring how LLMs can be applied for evaluation in information retrieval systems.

Acknowledgments

This research is supported by the Engineering and Physical Sciences Research Council [EP/S021566/1], the EPSRC Fellowship titled “Task Based Information Retrieval” [EP/P024289/1], CAMEO, PRIN 2022 n. 2022ZLL7MW and by the Dreams Lab, a collaboration between Huawei Finland, the University of Amsterdam, and the Vrije Universiteit Amsterdam.

References

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Augusto Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Model. User Adapt. Interact.* 30, 1 (2020), 127–158. https://doi.org/10.1007/S11257-019-09256-1

[2] Garbiel Bénédict, Ruqing Zhang, and Donald Metzler. 2023. Gen-ir@ sigir 2023: The first workshop on generative information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3460–3463.

[3] Cheng-Han Chiang and Hung-yi Lee. 2023. A Closer Look into Using Large Language Models for Automatic Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8928–8942. https://doi.org/10.18653/v1/2023.findings-emnlp.599

[4] Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2023. Towards Explainable Conversational Recommender Systems. 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, 2786–2795. https://doi.org/10.1145/3539618.3591884

[5] Dietmar Jannach. 2023. Evaluating conversational recommender systems. *Artif. Intell. Rev.* 56, 3 (2023), 2365–2400. https://doi.org/10.1007/S10462-022-10229-X

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[7] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

[8] 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 on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw (Eds.). ACM, 1586–1589. https://doi.org/10.1145/3589335.3641299

[9] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/2043932.2043962

[10] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. [n. d.]. LLM4Eval@WSDM 2025: Large Language Model for Evaluation in Information Retrieval. In *Proc. WSDM (WSDM '24)*.

[11] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. LLM4Eval: Large Language Model for Evaluation in IR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 3040–3043. https://doi.org/10.1145/3626772.3657992

[12] Hossein A Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles LA Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. 2024. LLMJudge: LLMs for Relevance Judgments. *arXiv preprint arXiv:2408.08896* (2024).

[13] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 338–354. https://doi.org/10.18653/v1/2024.naacl-long.20

[14] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Rethinking the Evaluation of Dialogue Systems: Effects of User Feedback on Crowdworkers and LLMs. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1952–1962. https://doi.org/10.1145/3626772.3657712

[15] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621* (2023).