LLM4Eval: Large Language Model for Evaluation in IR

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

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

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

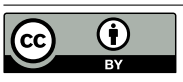
Large language models (LLMs) have demonstrated increasing task-solving abilities not present in smaller models. Utilizing the capabilities and responsibilities of *LLMs for automated evaluation* (LLM4Eval) has recently attracted considerable attention in multiple research communities. For instance, LLM4Eval models have been studied in the context of automated judgments, natural language generation, and retrieval augmented generation systems. We believe that the information retrieval community can significantly contribute to this growing research area by designing, implementing, analyzing, and evaluating various aspects of LLMs with applications to LLM4Eval tasks. The main goal of LLM4Eval workshop is to bring together researchers from industry and academia to discuss various aspects of LLMs for evaluation in information retrieval, including automated judgments, retrieval-augmented generation pipeline evaluation, altering human evaluation, robustness, and trustworthiness of LLMs for evaluation in addition to their impact on real-world applications. We also plan to run an automated judgment challenge prior to the workshop, where participants will be asked to generate labels for a given dataset while maximising correlation with human judgments. The format of the workshop is interactive, including roundtable and keynote sessions and tends to avoid the one-sided dialogue of a mini-conference.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Generative Models, Large Language Models, Automated Evaluation



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

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657992

ACM Reference Format:
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 (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626772.3657992

1 TITLE

LLM4Eval @ SIGIR '24: The First Workshop on Large Language Models (LLMs) for Evaluation in Information Retrieval.[1]

2 MOTIVATION

Large language models (LLMs), like GPT-4 [13], have demonstrated increasing effectiveness, such that a larger model performs well enough to be usable on a task where a smaller model was unusable. Recently, LLMs have been actively explored for various kinds of evaluation among other tasks.

In information retrieval (IR), among other applications, LLMs are being actively explored for estimating query-document relevance, both for ranking [10, 15] as well as for label generation [5, 17]. The latter can be subsequently used for training and evaluating other less powerful but more efficient rankers. For instance, HELM benchmark [10] simultaneously generates query-document relevance labels and a ranking score. The input of the prompt is a query and document. The generated output is a yes/no label. The ranking is carried out based on the generation probabilities of the yes/no tokens. Some later work [15] includes the HELM pointwise approach and adds more ranking methods.

More interestingly, LLMs are currently being employed for relevance labelling in the industry [17]. The evaluation methodologies can apply a wider range of LLMs and prompts to the labeling problem, and potentially address a wider range of potential quality problems.

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

More recently, with the advancement of LLMs, retrieval-augmented generation (RAG) systems, a class of LLM applications that use external data to augment the LLM's context, have received significant attention [8, 9]. Basically, a RAG system consists of a retriever and a downstream language model. Given a user question, the retriever finds relevant passages from a corpus (e.g., a company's internal knowledge base) and the language model uses these passages to generate a response. This formulation admits a multitude of choices: what retrieval model to use, how to divide the documents into retrieval chunks, and how to prompt or fine-tune the language model to use the retrieved information, to name only a few of the simplest design decisions. Recent LLM-based evaluation has emerged as a cheap and automatic strategy to evaluate the overall quality and the components of a RAG system [4, 16].

In natural language processing (NLP), some recent work showed that LLMs can be used as reference-free evaluators for text generation [7, 18]. The idea involves employing LLMs to assess the candidate output by considering its generation probability without relying on a reference target. This approach assumes that LLMs are able to assign higher probabilities to texts that are of high quality and fluency. Studies [2, 3, 11] have shown that LLMs can be perfect alternatives to human evaluation on NLG tasks. Some other work [11] showed that the way of prompting (so-called "prompt engineering") can enhance the LLM evaluation quality, with their proposed chain-of-thought (CoT) prompts outperforming various traditional evaluators [19, 20] by a large margin in terms of correlation with human evaluations.

Motivated by these observations, we believe that a workshop on evaluation strategies in the world of LLMs will question whether IR and NLP are truly facing a paradigm shift in evaluation strategies. Therefore, we have organized this workshop to provide a fresh perspective on LLM-based evaluation through an information retrieval lens. This workshop also provides a way to reflect on LLM-based evaluation benefits and challenges in academia and industry. Finally, we will encourage submissions and discussions on further evaluation topics and models, where existing literature is scarce, such as recommender systems, learning to rank, and diffusion models.

3 THEME AND SCOPE

The workshop focuses on models, techniques, data collections, and methodologies for information retrieval evaluation in the era of LLMs. These include but are not limited to:

- LLM-based evaluation metrics for traditional IR and generative IR
- Agreement between human and LLM labels
- Effectiveness and/or efficiency of LLMs to produce robust relevance labels
- Investigating LLM-based relevance estimators for potential systemic biases
- Automated evaluation of text generation systems
- End-to-end evaluation of Retrieval Augmented Generation systems
- Trustworthiness in the world of LLMs evaluation
- Prompt engineering in LLMs evaluation
- Effectiveness and/or efficiency of LLMs as ranking models

4 FORMAT AND PLANNED ACTIVITIES

In addition to the actual workshop at SIGIR, we plan to hold a challenge as a pre-workshop activity. Below we describe the details of these pre-workshop and workshop activities and our planned agenda for the actual workshop at SIGIR.

4.1 Pre-workshop Activities: Challenge

The goal of the challenge is to attract the attention of the community towards using LMs for evaluation and to release datasets that can later be used to enhance research in this area. In IR, among other applications, LLMs are being actively explored for estimating query-document relevance, both for ranking [10, 14] as well as for label generation [6, 17]. The latter can be subsequently used for training and evaluating other less powerful but more efficient rankers. The proposed challenge aims to study the effectiveness of LLMs in generating relevance labels on IR tasks. The goal of the proposed challenge is to evaluate LLMs on label generation.

The challenge will reuse the MS MARCO datasets [12] as our primary benchmark. The test queries will be a mix of previous years' TREC 2023 Deep Learning Track (TREC DL '23) test sets and a new set of heldout MS MARCO queries that have never been released before. Participants will be given a set of ⟨`query`, `document`⟩ pairs and will be asked to generate a relevance label, as well as a real-valued score for each of those pairs.

Participants will need to submit their exact prompt together with the predicted labels and predicted scores for the documents. When submitting prompts, participants will also be able to indicate the exact LLM model and parameters they employed to generate the run, which could be used to reproduce it. By allowing participants to submit their prompts, we can further analyze how these prompts may work across a variety of different LLM models.

In order to evaluate the quality of the generated labels, we plan to have the participants and the organizers audit the labels produced. We may also obtain some annotations via crowdsourcing as additional noisy ground truth. All the labels and the audits will be released as a shared dataset.

4.2 Synchronous Workshop

We plan to organize a full-day workshop, with the tentative schedule presented in Table 1.

5 CHALLENGE

The goal of the challenge is to attract the attention of the community towards using LMs for evaluation and to release datasets that can later be used to enhance research in this area. In IR, among other applications, LLMs are being actively explored for estimating query-document relevance, both for ranking [10, 14] as well as for label generation [6, 17]. The latter can be subsequently used for training and evaluating other less powerful but more efficient rankers. The proposed challenge aims to study the effectiveness of LLMs in generating relevance labels on IR tasks. The goal of the proposed challenge is to evaluate LLMs on label generation.

The challenge will reuse the MS MARCO datasets [12] as our primary benchmark. The test queries will be a mix of previous years' TREC 2023 Deep Learning Track (TREC DL '23) test sets and a new set of heldout MS MARCO queries that have never

Table 1: Planned Schedule for the LLM4Eval Workshop at SIGIR 2024.

Time	Agenda	Comment
9 - 9:15	Opening	-
9:15 - 10	Keynote 1	Keynote speaker: **Ian Soboroff, NIST**
10 - 10:30	Two invited talks	Invited talks from LLM4Eval articles published at major conferences
10:30 - 11	*Coffee break*	
11 - 12:30	Paper presentations	See Section 7 for more details on paper selection.
12:30 - 1:30	*Lunch break*	
1:30 - 2:15	Keynote 2	Keynote speaker: **Donald Metzler, Google**
2:15 - 3	Discussion panel	Topic: The role of IR community in LLM4Eval research, including 15 minutes Q&A
3 - 3:30	*Coffee break*	
3:30 - 4:15	Breakout sessions	Asking participants to form small groups to discuss the challenge of LLM4Eval
4:15 - 5	Poster presentation or spotlight talks	See Section 7 for more details on paper selection.

been released before. Participants will be given a set of ⟨query, document⟩ pairs and will be asked to generate a relevance label, as well as a real-valued score for each of those pairs.

Participants will need to submit their exact prompt together with the predicted labels and predicted scores for the documents. When submitting prompts, participants will also be able to indicate the exact LLM model and parameters they employed to generate the run, which could be used to reproduce it. By allowing participants to submit their prompts, we can further analyze how these prompts may work across a variety of different LLM models.

In order to evaluate the quality of the generated labels, we plan to have the participants and the organizers audit the labels produced. We may also obtain some annotations via crowdsourcing as additional noisy ground truth. All the labels and the audits will be released as a shared dataset.

6 ORGANIZERS

The organization team consists of active IR and NLP researchers from both academia and industry with recent experience in relevance judgments and ranking using LLMs research.

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). He also works as a part-time Applied Research Scientist at Thomson Reuters.

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

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 been involved in coordinating 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).

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 (NeuIR @ 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).

7 SELECTION PROCESS

We invited submission of papers up to six 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 burgeoning 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 channels such as LinkedIn, Twitter, Mastodon, and Slack (*e.g.*, SIGIR and TREC channels), as well as through mailing lists like SIGIR-List and CorporaList, in addition to dedicated website.

9 RELATED WORKSHOP

To the best of our knowledge, there have not been related workshops held previously at SIGIR or other conferences. The most indirectly relevant workshop to LLM4Eval is the recent SIGIR 2023 Workshop on **Generative Information Retrieval** [1].[2] Unlike Gen-IR, which mostly focused on generative IR techniques like document retrieval and direct response generation, LLM4Eval offers a venue for the discussion and exploration of 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] 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.

[2] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

[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] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated Evaluation of Retrieval Augmented Generation. *arXiv preprint arXiv:2309.15217* (2023).

[5] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on large language models for relevance judgment. arXiv:2304.09161 [cs.IR]

[6] Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*. 39–50.

[7] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166* (2023).

[8] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300* (2019).

[9] 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.

[10] 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).

[11] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[12] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *choice* 2640 (2016), 660.

[13] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[14] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563* (2023).

[15] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2023. Large language models are effective text rankers with pairwise ranking prompting. arXiv:2306.17563 [cs.IR]

[16] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *arXiv preprint arXiv:2311.09476* (2023).

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

[18] Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048* (2023).

[19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[20] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622* (2019).

[2] https://coda.io/@sigir/gen-ir