LLM4Eval@WSDM 2025: Large Language Model for Evaluation in Information Retrieval

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. This is the second iteration of the workshop. The first version was held in conjunction with SIGIR 2024, attracting over 50 participants.

CCS Concepts

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

WSDM '25, March 10–14, 2025, Hannover, Germany

ACM ISBN 979-8-4007-1329-3/25/03
https://doi.org/10.1145/3701551.3705706

Keywords

Generative Models, Large Language Models, Automated Evaluation

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. 2025. LLM4Eval@WSDM 2025: Large Language Model for Evaluation in Information Retrieval. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25), March 10–14, 2025, Hannover, Germany*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3701551.3705706

1 Title

`LLM4Eval` @ WSDM '25: The Second Workshop on Large Language Models (LLMs) for Evaluation in Information Retrieval.[1]

2 Motivation

Large language models (LLMs), like ChatGPT, 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 [4] as well as for label generation [3, 7]. The latter can be subsequently used for training and evaluating other less powerful but more efficient rankers. More interestingly, LLMs are currently being employed for relevance labelling in the industry [11]. 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.

In natural language processing (NLP), some recent work showed that LLMs can be used as reference-free evaluators for text generation [12]. 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.

[1] https://llm4eval.github.io/WSDM2025/

Studies [2, 5] have shown that LLMs can be perfect alternatives to human evaluation on NLG tasks. Some other work [5] 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 [13] by a large margin in terms of correlation with human evaluations.

The first edition of the LLM4Eval workshop was organised at SIGIR 2024 [9]. The workshop was popular at SIGIR with more than 50 in-person participants, 18 papers accepted; all presented in a poster session. We also organised LLMJudge challenge [10] as part of the workshop. The LLMJudge challenge was focused on evaluating the capabilities of LLMs for relevance judgment prediction task. In total, LLMJudge had 39 submissions (i.e., the 39 labelers) from 7 groups of universities and industries. The workshop also included a breakout discussion on various aspects of evaluation in the era of LLMs and a panel discussions. After the workshop, a report [8] was published. The report discussed the keynote speakers, accepted papers, and highlighted takeaways linked to potential research avenues. Among them, we can cite (i) *Evaluation validity*, how to validate of the evaluation using LLMs, (ii) *intrinsic randomness of the LLMs*, some operations that are becoming more and more common when operating with an LLM, such as prompt engineering or parameter tuning, induce randomness in the generation, (iii) *replicability and reproducibility*, and (iv) *the parallelism between human and LLMs assessment*. In the second edition of the LLM4Eval workshop, we propose to reflect on these hot-takes via publications and discussions; assess whether some have been addressed by researchers and take note of any additional hot-takes in discussions at the workshop.

3 Format and Planned Activities

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

3.1 Pre-workshop: LLMJudge Challenge

We plan to reorganize the LLMJudge challenge, originally conducted for the LLM4Eval workshop at SIGIR 2024 [8, 9]. The proposed challenge aims to study the effectiveness of LLMs in generating relevance labels on IR tasks. The challenge will reuse the LLMJudge challenge dataset [10]. Full details of the challenge, including submission requirements for participants, are available on our workshop website[2] and the GitHub repository[3].

3.2 Synchronous Workshop

We plan to organize a full-day workshop, with the schedule presented on our workshop website[4].

4 Related Workshop

The most indirectly relevant workshop to LLM4Eval are the recent **Information Retrieval Meets Large Language Models (IRLLM)** [6] at TheWebConf 2024 and SIGIR 2024 Workshop on **Generative Information Retrieval (Gen-IR)** [1]. Unlike IRLLM and Gen-IR, LLM4Eval offers a venue for the discussion and exploration of how LLMs can be applied for evaluation in information retrieval systems.

[2]https://llm4eval.github.io/WSDM2025/challenge/
[3]https://github.com/llm4eval/LLMJudge
[4]https://llm4eval.github.io/WSDM2025/program/

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] 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]

[4] 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).

[5] 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

[6] 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* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1586–1589. https://doi.org/10.1145/3589335.3641299

[7] Hossein A Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024. Synthetic test collections for retrieval evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2647–2651.

[8] Hossein A Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles LA Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. Report on the 1st Workshop on Large Language Model for Evaluation in Information Retrieval (LLM4Eval 2024) at SIGIR 2024. *arXiv preprint arXiv:2408.05388* (2024).

[9] 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

[10] 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).

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

[12] 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).

[13] 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).