Preface

This volume contains the proceedings of the First Workshop on Large Language Models (LLMs) for Evaluation in Information Retrieval (LLM4Eval 2024) held on July 18th, 2024 in Washington D.C, USA, and co-located with The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024, July 14-18, 2024 Washington D.C., USA).

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 was 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.

The contributions to LLM4Eval 2024 mainly address the following relevant topics:

- 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.

- LLMs in specific IR tasks such as personalized search, conversational search, and multimodal retrieval.
- Challenges and future directions in LLM-based IR evaluation.

We received 21 submissions of original papers presenting new research results and 5 submissions of already published results. The program committee involved 24 researchers, highly diversified in background and geographical region. Three program committee members reviewed each submission. The reviewers looked at originality, technical depth, style of presentation, and impact. Finally, the committee accepted 18 original papers and all the previously published works for presentation at the workshop. Out of these, 7 papers were further published on the proceedings.

The workshop program included a booster session where the authors of the accepted paper presented their work, followed by a poster session, to allow a more detailed discussion between presenters and workshop participants. Furthermore, the workshop included a panel, with the following invited panellists: Charlie L. A. Clarke (University of Waterloo), Laura Dietz (University of New Hampshire), Michael D. Ekstrand (Drexel University), and Ian Soboroff (National Institute of Standards and Technology (NIST)). We also had two keynotes. The first was by Ian Soboroff (National Institute of Standards and Technology (NIST)), titled "A Brief History of Automatic Evaluation in IR", the second was by Donald Metzler (Google DeepMind), and was titled "LLMs as Rankers, Raters, and Rewarders".

The success of LLM4Eval 2024 would not have been possible without the considerable effort of several people including the Program Committee, and the participants who contribute their time and effort.
Thank you all very much!

July, 2024

Hossein A. Rahmani
Clemencia Siro
Mohammad Aliannejadi
Nick Craswell
Charles L. A. Clarke
Guglielmo Faggioli
Bhaskar Mitra
Paul Thomas
Emine Yilmaz

Program Committee

- Zahra Abbasiantaeb, University of Amsterdam
- Mofetoluwa Adeyemi, University of Waterloo
- Marwah Alaofi, RMIT University
- Negar Arabzadeh, University of Waterloo
- Shivangi Bithel, IIT Delhi
- Francesco Luigi De Faveri, University of Padua
- Yashar Deldjoo, Polytechnic University of Bari
- Gianluca Demartini, The University of Queensland
- Laura Dietz, University of New Hampshire
- Yue Feng, UCL
- Claudia Hauff, Spotify
- Bhawesh Kumar, Verily Life Sciences
- Yiqun Liu, Tsinghua University
- Sean MacAvaney, University of Glasgow
- James Mayfield, Johns Hopkins University
- Chuan Meng, University of Amsterdam
- Ipsita Mohanty, Carnegie Mellon University
- Mohammadmehdi Naghiaei, University of Southern California
- Pranoy Panda, Fujitsu Research
- Orion Weller, Johns Hopkins University
- Lu Wang, Microsoft
- Xi Wang, University of Sheffield
- Jheng-Hong Yang, University of Waterloo
- Oleg Zendel, RMIT University