Report on the Collab-a-thon at ECIR 2024

Sean MacAvaney
University of Glasgow
United Kingdom
sean.macavaney@glasgow.ac.uk

Adam Roegiest, Aldo Lipani, Andrew Parry, Björn Engelmann, Christin Katharina Kreutz, Chuan Meng, Erlend Frayling, Eugene Yang, Ferdinand Schlatt, Guglielmo Faggioli, Harrisen Scells, Iana Atanassova, Jana Friese, Janek Bevendorff, Javier Sanz-Cruzado, Johanne Trippas, Kanaad Pathak, Kaustubh Dhole, Leif Azzopardi, Maik Fröbe, Marc Bertin, Nishchal Prasad, Saber Zerhoudi, Shuai Wang, Shubham Chatterjee, Thomas Jaenich, Udo Kruschwitz, Xi Wang, Zijun Long *

Abstract

We present a report on the Collab-a-thon, a series of sessions at the European Conference on Information Retrieval (ECIR) 2024 designed to help foster new collaborations during a conference. This report presents the motivation and design of the Collab-a-thon, a summary of the discussions covered at each session, and a set of recommendations for conducting similar events in the future. The event is set to run again at ECIR 2025 and planning is underway to pilot the event in a different community at the IEEE International Conference on Distributed Computing Systems (ICDCS) 2025.

Date: 25–27 March 2024.

Website: https://www.ecir2024.org/collab-a-thon/.

1 Introduction

Conferences provide a platform for sharing scientific findings and building new connections within a community. While they typically allocate time for information exchange and networking, many meaningful connections are built informally outside the formal conference activities. Some members of the community, particularly newcomers such as students, may find it challenging to engage in this manner. To address this issue, ECIR 2024 introduced a new initiative called the "Collab-a-thon." This series of events ran in parallel with the main conference sessions and aimed to give conference delegates dedicated time to meet peers interested in the same topics and form new

*Affiliation not shown for all authors due to space limitations (see Appendix A for details).

Figure 1. Solving fairness in IR, one sentence at a time. (Photo courtesy of `@andreas_chari` via X.)

potential collaborations. By providing some structure around these activities, we aimed to engage with more of the community and help form connections that may have otherwise been missed.

A total of 34 conference delegates, some of whom appear in Figure 1, participated in one or more Collab-a-thon session during the conference, accounting for approximately 8% of registrants. Each session sparked lively discussions on various topics, with several resulting in concrete ideas for future collaboration. These included plans to work together on an open-source project, a research endeavor, and the development of a web application.

The rest of this event report is structured as follows. Section 2 discusses the event format and the motivations behind its design. Section 3 offers a brief overview of the discussions and results from each Collab-a-thon session. Finally, Section 4 concludes by sharing lessons learned to guide future events of a similar nature.

2 Format

The Collab-a-thon took place as a set of sessions largely tied to each regular conference session. As scheduling allowed, each Collab-a-thon session was conducted parallel to the following conference activity. An excerpt of the ECIR schedule with the Collab-a-thon schedule is shown in Figure 2, showing how the main conference session on "Entities and ML" took place on Monday at 11:00–12:30, while the corresponding Collab-a-thon session 13:30–14:00 (after lunch and in parallel with other main conference sessions).

In each session, a moderator (the Collab-a-thon Chair) facilitated discussions by assisting with introductions and offering prompts if the conversation lagged (such as, "Could we merge the idea from X's paper with Y's?"). The main aim was to generate new research ideas that

build on everyone's skills and expertise. The moderator, often with help from another conference volunteer, took notes and recorded attendees, which contributed to this report. After the session, participants were encouraged to keep the discussions going throughout the conference and beyond (e.g., multiple channels were created in the SIGIR Slack).

Halfway through the conference, there was also a "Midterm Report" presented to the larger conference. This gave the participants a chance to share what they had discussed and encourage further participation.

Linking each Collab-a-thon session to the main conference sessions had multiple benefits. Firstly, it ensured that the topics were detailed enough to attract the interest of several conference attendees. If a delegate presented a paper on or attended a session about a specific topic, there is a high likelihood that it is something they are particularly interested in and would like to discuss further. Moreover, the topic would already be on the delegates' minds. Providing a dedicated opportunity to shape these ideas while they're fresh is likely to be more valuable than postponing it.

We also considered an alternative format: hosting the Collab-a-thon as a standalone, all-day event, akin to a workshop or tutorial. While this format had some advantages —like providing participants with dedicated time to collaborate, akin to a traditional hackathon— it also had significant drawbacks that led us to choose the parallel approach instead. Firstly, we were concerned that an all-day event might clash with other activities scheduled for the day, such as workshops, tutorials, Industry Day, and IR for Good Day. Organizers and attendees of these events would have had to decide between them, potentially dividing attendance. Given the diverse offerings, convincing people to prioritize the Collab-a-thon over other events could prove challenging. Furthermore, if successful, it could detract from the attendance of these other activities, which is less than ideal. Additionally, organizers and presenters of these events would have been unable to participate in the Collab-a-thon. Secondly, in an all-day event format, forming teams would have been complicated. Unlike the proposed parallel approach, which ensures common interests among attendees, an all-day format might have resulted in disparate groups with little overlap in interests.

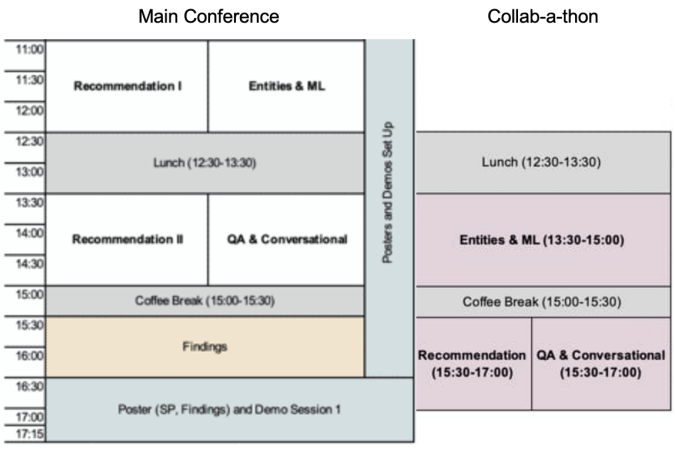

Figure 2. Excerpt of the ECIR 2024 schedule showing the parallel Collab-a-thon sessions.

3 Sessions

This section summarizes topics discussed and the outcomes from each Collab-a-thon session chronologically.

3.1 Entities and Machine Learning

The first session was topically aligned with the "Entities and Machine Learning" main conference session. Two ECIR delegates joined this session, both with a particular interest in entities, so the conversation largely revolved around this topic. A variety of problems in the area of entities were discussed, including how the term "entity" is a bit overloaded, problems in effectively representing entities semantically, and how entity representations often do not transfer well between domains. The discussion then shifted to the topic of how entities and their relationships evolve over time. We discussed the possibility of building a new benchmark dataset to evaluate this change. A potential feature of this dataset would be different granularities in time for different types of entities (this is something that humans pick up on rather easily but is difficult to model with current systems).

The session wrapped up with a discussion about a possible application of entities to help build personalized conference timetables. In particular, a conference could be modeled as a graph of users, papers, and interests. Path recommendation could be applied to help match attendees to sessions they should attend and other delegates they should meet.

3.2 Question Answering and Conversational Search

The second session aligned with the "Question Answering and Conversational Search" main conference track, and attracted four conference delegates. The conversation started on the topic of evaluating these systems, which remains a challenge when the answers are generated. Though an LLM (Large Language Model)-based evaluation is likely more reliable than BLEU/ROUGE/BERTScore evaluations, the attendees expressed concerns about using LLMs to evaluate LLM output. The conversation then shifted to a variety of open problems, such as how to separate out the "semantics" from the "idea" of responses, moving away from reference-based evaluations, and how to ensure that evaluation remains simple and repeatable.

The session wrapped up with a discussion on the idea of a "worst-case" evaluation of these systems. Instead of focusing on average cases, you could sample a system multiple times and evaluate it based on the worst performance, which establishes a form of a lower-bound. Interesting research questions exist here, such as whether the worst-case performance aligns with average cases. Understanding the worst case might be especially important in high-stakes settings like medical or financial question answering.

3.3 Multi-modal IR

The session on "Multi-modal IR" was attended by four delegates. The conversation kicked off with a discussion on the challenges of aligning representations across different modalities. It was noted that similar challenges also arise when considering entities as another modality. A significant portion of the discussion revolved around the TREC AToMiC track Yang et al. [2023],

particularly its relevance to conversational question answering and its potential as a foundation for more advanced multi-modal *conversational* search, building upon ideas from the previous session. Towards the end, participants delved into the observation that many current systems process different modalities independently, contrasting with how humans naturally process multiple modalities simultaneously, with each modality enhancing the understanding of the others. This led to a suggestion that adopting a processing model more akin to human cognition could be beneficial moving forward.

3.4 Long Sequences

During the Multi-modal IR session, two of the attendees decided that they would rather discuss the topic of modeling long sequences instead, so a separate session was split out. The discussion started with how handling long sequences is challenging with transformer-based models regardless of the modality. The particular challenge in all the transformer-based models lies in the efficiency with long sequence, owing to the computation complexity of attention and feed forward network with the increase in the sequence length. This problem persists across a variety of domains and tasks utilizing any transformer-based models (such as LLM), including legal text processing [Prasad et al., 2024], book summarizing, long question answering, long text generation, continuous video/audio processing, etc. The discussion moved on to some strategies utilizing augmenting the attention with sparse rank computations, low rank approximations and strategies such as chunking without any architectural changes which requires enabling efficient attention between chunks. This discussion highlighted some recent advancements in this domain, the importance of long sequence processing and the need to ameliorate this computation complexity when using any transformer-based model/LLM.

3.5 Fairness and Privacy

The "Fairness and Privacy" session attracted ten delegates. The group discussed the definition of fairness from the perspective of content creators and searchers and how they are related to the fairness hypothesis [Wilkie and Azzopardi, 2014]. The discussion led to interesting debates on how fairness should and could be defined in ranking and its impact on designing search engines. A central point of discussion revolved around how the literature refers to fairness and bias in many different ways using many different criteria – and these can lead to competing, conflicting and confusing notions of fairness. Each definition in turn can have very different implications and impacts on the ranking and design. For example, the notion of "equality of opportunity" vs "equality of outcome" was heavily discussed. The former suggests all documents/individuals/groups should have a similar opportunity to be retrieved, while the latter suggests that all documents/individuals/groups should be retrieved equally. The former means that some documents/individuals/groups would be retrieved more often than other (which may give the impression that the system is biased), while the latter would mean all documents/individuals/groups are equally likely to be retrieved regardless of relevance (which means the system may be considered less biased towards the documents/individuals/groups, but at the expense of relevance). However, bias does not necessarily mean discrimination; whether it is fair or not depends on the criteria used, such as equality, statistical parity, or another standard.

Another highlight of the discussion was whether and how systems should provide flexibility to select the type of fairness criteria to employ. Or whether it was right for retrieval systems to be imposing their notion of fairness on users. For example, certain notions of fairness may impose greater costs on a majority of users, in order to be more representative. This highlighted the trade-off between relevance and fairness – and if it is fair not to retrieve relevant documents because of some ideological principle or supposed ideal.

Discussion related to evaluation also sparked interesting problems along with fairness. The group discussed whether the typical Cranfield Paradigm is the right tool for evaluating fairness and how its assumptions may conflict with the quality that we want to measure for fairness. Existing IR evaluation collections usually contain one set of relevance judgments for each topic, which may or may not reflect the diversity of the opinion different searchers may possess.

Despite having fewer discussions on privacy, the group also recognized that many fairness questions are based on or lead to privacy problems. The group is interested in continuing the conversation and could extend it as a workshop or tutorial in future conferences.

3.6 Special Session: Midterm Report

At the end of the second day of the event, the Collab-a-thon Midterm Report was presented in the main conference ballroom. A reminder of the event's format was presented, followed by individual pitches from members of each session. Most sessions involved a spokesperson from each group presenting a single slide that summarized the points of discussion and ideas. The "fairness" session (shown in Figure 1) was particularly noteworthy for featuring several attendees, each of whom shared a one-sentence takeaway from the discussion (in a random order for fairness, of course).

3.7 SimIIR

An organic collaborative effort emerged during ECIR when delegates from various institutes realized that they were independently working on the open-source framework called SimIIR [Maxwell and Azzopardi, 2016; Zerhoudi et al., 2022]. This framework enables the simulation of interactive information retrieval experiments – but has become fragmented through various forks. So the Collab-a-thon was seen as an ideal setting to bring the community together and coordinate development efforts on improving and enhancing the framework. This spawned an 'ad-hoc' Collab-a-thon session to bring participants together.

A lively discussion took place, both introducing newcomers to simulation in IR and discussing practical matters for the next steps with the SimIIR project. A new GitHub Organization[1] was created, as well as guidelines for how to contribute. A variety of new features were proposed including, new user workflows to simulate conversational search and to support for LLM-based search agents, along with quality-of-life improvements (new output formats), new potential directions for research (e.g., simulated mood, engagement or emotion tracking), and modernization (e.g., integrating PyTerrier [Macdonald et al., 2021] to take advantage of techniques like neural IR and dense retrieval). A channel `#simiir`[2] on ACM SIGIR's Slack was also created to help facilitate the ongoing collaboration where all are welcome to join and participate.

[1] https://github.com/simint-ai

[2] https://acmsigir.slack.com/archives/C06RC0A31EX

3.8 Neural IR

The session on "Neural IR" discussed a variety of current problems in neural IR, particularly with respect to training and evaluation. Given how rapidly progress had been made, there was concern from the group about whether we made meaningful progress in the past several years, or whether we are just over-fitting the types of queries and labels present in frequently used test collections like MS MARCO [Nguyen et al., 2016] respectively benchmark suites like BEIR [Thakur et al., 2021]. Answering this question is interesting for future research projects, as we might need to switch to new datasets for developing future algorithms if we already overfitted our frequently used datasets. The discussions took inspiration from Norbert Fuhr's Keynote at SIGIR 2020 [Fuhr, 2020].

The group coalesced around a collaborative plan to make a first step to answering this by testing whether we are overfitting the particular relevance assessments in the TREC Deep Learning datasets [Craswell et al., 2019, 2020]. A new set of assessments, potentially also including user-specific "narratives", plan to be annotated by the group and a variety of systems will be compared with TREC's assessments vs the new assessments. A Slack channel was created to help coordinate these efforts.

3.9 Domain-Specific IR

The final Collab-a-thon session was on Domain-Specific IR. The discussion started on the topic of whether, in the age of LLMs, specific considerations need to be made to handle domains anymore at all. One participant suggested that the CLEF 2024 Monster Track [Ferro et al., 2024] might be able to help us answer this question. Meanwhile, the general consensus of the group was that there will likely always be room for specialization; at the very least, a specialist model could likely be smaller and more efficient than a generalist model. Further, from a system design perspective, it can be beneficial for individual components (potentially targeted towards a specific domain) to be tested, deployed, and updated independently. The conversation then shifted to the relationship between modalities and domains, with the proposition that certain modalities are more important in some domains than others. Especially with respect to certain modalities (like text), there is the potential for different domain-specific models to learn from one another.

Although there were no concrete outcomes from this session, the participants expressed that it was valuable for them to unpack and discuss the ideas and problems in the area.

3.10 Canceled Sessions

There were also sessions scheduled on "Recommendation" and "IR & NLP", which were aligned with main conference sessions. Both these sessions were canceled due to lack of interest, however. The "IR & NLP" was scheduled opposite the SimIR session, demonstrating the potential challenges in running parallel Collab-a-thons.

4 Conclusions and Lessons Learned

We believe the Collab-a-thon was a success and can serve as a model for similar events in future conferences. It facilitated new connections among delegates and led to concrete plans for ongoing

collaboration. However, we have noted some key observations that we think would be valuable to share with the community to improve future events.

Consideration of physical location is crucial. Due to limitations in venue space, the Collab-a-thon took place in three different rooms throughout the conference. The third room, situated between the two main session rooms, was the most effective. Its central placement allowed delegates to conveniently drop by, observe ongoing activities, and choose to participate if interested. In contrast, the other two locations were less accessible, missing out on such spontaneous interactions. We suggest deliberate and consistent room placement in future events to encourage these serendipitous encounters.

Offer plans, but stay open to change. As famously attributed to US President Dwight D. Eisenhower, "Plans are useless, but planning is indispensable." While the topics from the main conference sessions served as a good starting point for many discussions, they did not always fit every need. For instance, during the Multimodal session, two attendees veered off to discuss Long Sequences instead. Similarly, there was interest in initiating a Collab-a-thon on SimIIR, which was then included in the official schedule. We recommend remaining flexible to such changes to allow discussions and collaborations to evolve naturally.

Promoting Collab-a-thon during main conference sessions. Since Collab-a-thon sessions usually aligned with the topics of the main conference sessions, it could have been beneficial for the session chairs to encourage attendees to join Collab-a-thon discussions for further exploration of the subject matter. It might even be advantageous for session chairs to lead the corresponding Collab-a-thon sessions in future events. We recommend organizers to explore this approach to boost participation.

Although the first Collab-a-thon is wrapped up, the collaborations are just getting started! Meanwhile, the event is scheduled to run again at ECIR 2025 and planning is underway to pilot the event at the IEEE International Conference on Distributed Computing Systems (ICDCS) 2025.

Acknowledgments

We thank the ECIR 2024 General Chairs—Graham McDonald, Craig Macdonald, and Iadh Ounis—for their help in the conception of the event and their flexibility in its organization and format. We also thank the Program Chairs—Nicola Tonellotto and Nazli Goharian—for the inclusion of the "Collab-a-thon Midterm Report" session in the main program. Finally, we thank Andreas Chari for taking the photograph in Figure 1.

A Authors and Affiliations

Organizer

- Sean MacAvaney, University of Glasgow, United Kingdom, sean.macavaney@glasgow.ac.uk

Other authors

- Adam Roegiest, Zuva, Canada, adam@roegiest.com
- Aldo Lipani, University College London, United Kingdom, aldo.lipani@ucl.ac.uk
- Andrew Parry, University of Glasgow, United Kingdom, a.parry.1@research.gla.ac.uk

- Björn Engelmann, TH Köln, Germany, bjoern.engelmann@th-koeln.de
- Christin Katharina Kreutz, TH Mittelhessen, Germany, christin.kreutz@th-koeln.de
- Chuan Meng, University of Amsterdam, The Netherlands, c.meng@uva.nl
- Erlend Frayling, University of Glasgow, United Kingdom, Erlend.Frayling@glasgow.ac.uk
- Eugene Yang, Johns Hopkins University, United States, eugene.yang@jhu.edu
- Ferdinand Schlatt, Friedrich-Schiller-Universität Jena, Germany, ferdinand.schlatt@uni-jena.de
- Guglielmo Faggioli, University of Padova, Italy, guglielmo.faggioli@unipd.it
- Harrisen Scells, Leipzig University, Germany, harry.scells@uni-leipzig.de
- Iana Atanassova, Université de Bourgogne Franche-Comté, France, iana.atanassova@univ-fcomte.fr
- Jana Friese, Universität Duisburg-Essen, Germany, jana.friese@uni-due.de
- Janek Bevendorff, Leipzig University, Germany, janek.bevendorff@uni-weimar.de
- Javier Sanz-Cruzado, University of Glasgow, United Kingdom, javier.sanz-cruzadopuig@glasgow.ac.uk
- Johanne Trippas, RMIT University, Australia, j.trippas@rmit.edu.au
- Kanaad Pathak, University of Strathclyde, United Kingdom, kanaad.pathak@strath.ac.uk
- Kaustubh Dhole, Emory University, Atlanta, USA, kaustubh.dhole@emory.edu
- Leif Azzopardi, University of Strathclyde, United Kingdom, leifos@acm.org
- Maik Fröbe, Friedrich-Schiller-Universität Jena, Germany, maik.froebe@uni-jena.de
- Marc Bertin, Université Claude Bernard Lyon 1, France, marc.bertin@univ-lyon1.fr
- Nishchal Prasad, Institut de Recherche en Informatique de Toulouse, France, Nishchal.Prasad@irit.fr
- Saber Zerhoudi, Universität Passau, Germany, saber.zerhoudi@uni-passau.de
- Shuai Wang, The University of Queensland, Australia, shuai.wang5@uq.net.au
- Shubham Chatterjee, University of Edinburgh, United Kingdom, shubham.chatterjee@ed.ac.uk
- Thomas Jaenich, University of Glasgow, United Kingdom, t.jaenich.1@research.gla.ac.uk
- Udo Kruschwitz, University of Regensburg, Germany, Udo.Kruschwitz@ur.de
- Xi Wang, The University of Sheffield, United Kingdom, xi.wang@sheffield.ac.uk
- Zijun Long, University of Glasgow, United Kingdom, Zijun.Long@glasgow.ac.uk

References

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the TREC 2019 deep learning track. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*, volume 1250 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2019. URL https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.PM.pdf.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the TREC 2020 deep learning track. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2020. URL https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.DL.pdf.

Nicola Ferro, Julio Gonzalo, Jussi Karlgren, and Henning Müller. The CLEF 2024 monster track: One lab to rule them all. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part VI*, volume 14613 of *Lecture Notes in Computer Science*, pages 11–18. Springer, 2024. doi: 10.1007/978-3-031-56072-9_2. URL https://doi.org/10.1007/978-3-031-56072-9_2.

Norbert Fuhr. Proof by experimentation?: towards better IR research. *SIGIR Forum*, 54(2):2:1–2:4, 2020. doi: 10.1145/3483382.3483385. URL https://doi.org/10.1145/3483382.3483385.

Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. Pyterrier: Declarative experimentation in python from BM25 to dense retrieval. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 4526–4533. ACM, 2021. doi: 10.1145/3459637.3482013. URL https://doi.org/10.1145/3459637.3482013.

David Maxwell and Leif Azzopardi. Simulating interactive information retrieval: Simiir: A framework for the simulation of interaction. In Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel, editors, *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 1141–1144. ACM, 2016. doi: 10.1145/2911451.2911469. URL https://doi.org/10.1145/2911451.2911469.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne, editors, *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Nishchal Prasad, Mohand Boughanem, and Taoufiq Dkaki. Exploring large language models and hierarchical frameworks for classification of large unstructured legal documents. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II*, volume 14609 of *Lecture Notes in Computer Science*, pages 221–237. Springer, 2024. doi: 10.1007/978-3-031-56060-6_15. URL https://doi.org/10.1007/978-3-031-56060-6_15.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December*

2021, virtual, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

Colin Wilkie and Leif Azzopardi. Best and fairest: An empirical analysis of retrieval system bias. In Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, volume 8416 of *Lecture Notes in Computer Science*, pages 13–25. Springer, 2014. doi: 10.1007/978-3-319-06028-6_2. URL https://doi.org/10.1007/978-3-319-06028-6_2.

Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio de Rezende, Krishna Srinivasan, Miriam Redi, Stéphane Clinchant, and Jimmy Lin. TREC2023 AToMiC overview. In Ian Soboroff and Angela Ellis, editors, *Proceedings of the Thirty-Second Text REtrieval Conference, TREC 2023*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2023. URL https://trec.nist.gov/pubs/trec32/papers/Overview_atomic.pdf.

Saber Zerhoudi, Sebastian Günther, Kim Plassmeier, Timo Borst, Christin Seifert, Matthias Hagen, and Michael Granitzer. The simiir 2.0 framework: User types, markov model-based interaction simulation, and advanced query generation. In Mohammad Al Hasan and Li Xiong, editors, *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 4661–4666. ACM, 2022. doi: 10.1145/3511808.3557711. URL https://doi.org/10.1145/3511808.3557711.