Second Workshop on Explainability in Information Retrieval

Catherine Chen
catherine_s_chen@brown.edu
Brown University
Providence, RI, USA

Maria Heuss
m.c.heuss@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Tanya Chowdhury
tchowdhury@cs.umass.edu
University of Massachusetts Amherst
Amherst, MA, USA

James Allan
allan@cs.umass.edu
University of Massachusetts Amherst
Amherst, MA, USA

Avishek Anand
Avishek.Anand@tudelft.nl
Delft Institute of Technology
Delft, The Netherlands

Carsten Eickhoff
c.eickhoff@acm.org
University of Tübingen
Tübingen, Germany

Suzan Verberne
s.verberne@liacs.leidenuniv.nl
Leiden University
Leiden, The Netherlands

Abstract

As models grow more complex and societal demands for transparency increase with emerging regulations, explainability has become an increasingly important research area. However, despite its recognized relevance, progress in explainability research in information retrieval (IR) has been slower than in related fields. This *full day* workshop aims to advance research in explainable IR by providing a more in-depth platform to reflect on recent developments and facilitate discussions across both new and persistent challenges. Building upon the first edition of the workshop, which was a great success in bringing together multiple perspectives on explainability in IR, this second edition will focus on synthesizing a common agenda for the research community. Through a set of interactive activities, the workshop will bring together a diverse group of researchers to build a shared understanding of key tasks and challenges, and to help shape future directions for explainable IR research. The workshop will have as concrete outcomes a roadmap document and a special issue proposal for a journal issue on explainability in IR.

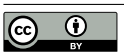
CCS Concepts

• **Information systems** → **Information retrieval**; *Users and interactive retrieval*; *Evaluation of retrieval results.*

Keywords

Explainable Search, Explainability, Interpretability, Transparency

ACM Reference Format:
Catherine Chen, Maria Heuss, Tanya Chowdhury, James Allan, Avishek Anand, Carsten Eickhoff, and Suzan Verberne. 2026. Second Workshop on Explainability in Information Retrieval . In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3805712.3808649



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808649

1 Motivation

Explainable AI research aims to explain the decision-making process of models in human-understandable terms. Explainability[1] is critical for ensuring safety in high-stakes domains such as finance, healthcare, and law. Additionally, it supports compliance with emerging AI regulations, such as the EU AI Act [2] and the Colorado AI Act [1], that stipulate a "right to explanation" for AI systems deployed in high-stakes situations.

Related communities, i.e., Natural Language Processing, Computer Vision, and Recommender Systems, have seen significant progress in explainability research, supported by dedicated workshops at top conferences. For example, *XAI in Action* at NeurIPS 2023 [9], the *Mechanistic Interpretability Workshop* at ICML 2024 [3] and NeurIPS 2025 [7], the *Actionable Interpretability Workshop* at ICML 2025 [5], and *BlackboxNLP* [4], which has been running at ACL/EMNLP for the past eight years and has had hundreds of attendees (up to 500 in 2023) and dozens of submissions. In addition, the *ACL and EMNLP conferences have attracted a growing number of submissions on interpretability, visible from the large number of sessions devoted to the topic.[2]

For IR, explainability has been listed as a relevant area under the SIGIR main conference call for papers since 2022, under the broader FATE Track (Fairness, Accountability, Transparency, Ethics, and Explainability). Despite IR being the backbone of AI systems in many high-stakes scenarios, however, we see slower progress in explainability research within IR, with a majority of work in the field focusing on recommender systems.

Additionally, the shift of the past 6 years toward more complex, highly parametric models (i.e., Transformer-based models) and towards multi-stage applications (such as RAG) has made it more challenging to interpret and explain retrieval results, in contrast to

[1] *Interpretability* is a closely related term, similarly used to describe efforts to provide insight into a model's internal processes. For simplicity, we do not distinguish between the two and use *explainability* throughout this proposal, encompassing any definitions of *interpretability* under its scope.
[2] 5 full paper sessions at EMNLP on interpretability: https://2025.emnlp.org/program/

more traditional methods (e.g., boolean search or BM25) that were more inherently explainable. Along with the emerging regulations, calling for explainability in high-stakes decision-making, there is a need for focused attention on explainability in IR systems.

Thus, **we propose a second, *full-day* workshop at SIGIR dedicated entirely to explainability in IR**. Building on the success of the inaugural workshop at SIGIR 2025 [6], this second edition aims to further showcase recent and ongoing research within this emerging community and to offer a platform for researchers to discuss a vision for the future of the field. In addition, this year we emphasize concrete outcomes, with plans to collaboratively develop a proposal for a special issue for the Information Retrieval Research Journal (IRRJ) and a research roadmap that can serve as a concrete starting point for the community to gain a shared understanding of the state of explainable information retrieval.

2 Theme and Purpose of the Workshop

Overall, this workshop series aims to address challenges of the field by providing a platform with specific interaction and engagement activities beyond what is possible in the dedicated main conference session. Building upon the momentum from the workshop's first edition, our overall objective is to establish and further a shared understanding of explainability concepts, methods, and challenges specific to IR, and to outline a full research agenda to help guide future research.

Specifically, the second edition of this workshop centers on the theme of *"Explainability Beyond the Classical IR Pipeline."* This theme emerged from discussions at the inaugural workshop at SIGIR 2025, which brought together diverse perspectives and surfaced open questions around the evaluation, standardization, and deployment of explainable IR systems. Building on this foundation, the current edition aims to develop these ideas more concretely.

Thus, we have three primary goals:

(1) **Community Building:** Bring together researchers interested in XIR to foster discussions and identify key challenges in the field. This includes both established experts in the field as well as new researchers who are interested in working on XIR in the future.
(2) **Showcase Recent Progress:** Provide a platform for sharing and discussing XIR research, including work in progress and recently published studies, to increase exposure and encourage discussion on previously faced challenges in the field.
(3) **Identify Future Directions for XIR:** Consolidate perspectives on key notions and definitions, establish best practices, and identify challenges of explainability to outline a shared trajectory for future research.

For a full list of topics of interest, see the Call for Papers, Section 3.2.

3 Workshop Format

We propose a **full-day** workshop with one keynote speaker and multiple interactive activities to encourage active discussions, idea exchange, and collaboration among participants throughout the day.

3.1 Proposed Activities

Keynote. The workshop will include one keynote talk intended to provide an opening perspective and offer context for the discussions that follow. We have invited Mor Vered, a senior lecturer at Monash University. Her research interests include topics in Explainable AI, such as human-centered XAI, XAI used in deceptive ways, and user study design for XAI research. She also works on generating explanations built on cognitive theories, considering that these explanations need to be consumed by people and therefore take into account human situation awareness models. This makes her perspective particularly relevant for reflecting on the importance of understanding the human in the explainability pipeline.

Author Lightning Talks and Posters. Instead of a traditional paper presentation session, we will invite authors of accepted contributions to our workshop to present a 2-3 minute *"lightning talk"* to give an overview of their work. This will provide authors an opportunity to promote their posters and allow other participants to more effectively decide which posters they may want to visit during the shared workshop poster session.

Discussion Panel. The workshop will feature a discussion panel in line with our workshop theme that aims to highlight the multiple perspectives on the explainability pipeline, from system to user. The panel will feature 2-3 invited panelists, along with one open "hot-seat", reserved for audience members who wish to contribute perspectives or respond to specific questions. This format was used successfully in the first edition of the workshop and led to active and sustained audience participation. A moderator will guide the discussion using a set of prepared questions and encourage contributions from workshop attendees throughout the session. The organizers will take notes in a shared document. The panel will be followed by a plenary discussion aimed at consolidating key insights and converging on a shared research agenda, which will be further developed during a subsequent writing session.

Writing Session. We will organize a plenary discussion that will transition into a collaborative writing session in which participants break into groups of 5–10 to concretize ideas discussed throughout the day and work toward defining future research agendas. The goal of this session is to gather community input, surface key challenges facing the field, and connect researchers interested in taking concrete steps toward addressing these challenges. One or more groups will focus specifically on the workshop's main theme, *Explainability Beyond the Classical IR Pipeline*, with the option for additional groups to form around other topics of interest. For example, one group may build on discussions from the previous edition of the workshop, such as developing a taxonomy of explainability needs in IR.

Tentative Schedule of Events. The tentative schedule is shown in Table 1.

3.2 Call for Papers

To foster the growth of an active community, we aim to involve both senior and junior researchers in our workshop. To curate an insightful and engaging program, we will invite submissions of **extended**

Table 1: Tentative workshop schedule

09:00-09:45	Opening Remarks & Keynote
09:45-10:30	Discussion Panel: *Perspectives on the Explainability Pipeline*
10:30-11:00	Coffee Break
11:00-11:45	Plenary Discussion (panel)
11:45-12:30	Author Lightning Talks
12:30-13:30	Lunch Break
13:30-14:30	Workshop Posters
14:30-16:00	Plenary Discussion (writing brainstorming)
16:00-16:30	Coffee Break
16:30-17:15	Writing Session
17:15-17:30	Closing Remarks

abstracts up to 2 pages in length for research across various stages: already published work, ongoing work not yet submitted elsewhere, and finished work that has not yet been published. The workshop will also allow authors of in-progress work to receive constructive feedback from the community to help refine their papers. We will actively encourage people (in particular students) to submit ongoing work and preliminary results and reach out to authors of previously published relevant work to encourage submissions. Below is a list of potential topics for submissions:

- **Novel Methods and Models for Explainability:** Development of new algorithms, techniques, or models to explain IR systems.
- **Internal Analysis and Model Interpretability:** Novel techniques for analyzing the inner workings of IR systems, insights learned from analysis and interpretation.
- **Defining and Evaluating Explainability:** How can we unify different notions and definitions of explainability and achieve consensus over what constitutes a "good" explanation?
- **Human-Centered Explainability:** Work focused on identifying, investigating, or integrating users' explainability needs in IR systems.
- **Practical Applications of Explainability:** Explainability in high-stakes domains (e.g., healthcare, legal, finance, etc.), applications to personalized search/recommendation, bias mitigation and fairness.
- **Other:** Other topics not mentioned in this list but still pertaining to explainability in IR.

In addition to "traditional" topics in explainability, we will also open extended abstract submissions for *Perspectives on the Challenges of Explainability in IR*, which include but are not limited to the following topics:

- **Explainability Beyond the Classical IR Pipeline:** What challenges for explainability arise with the emergence of new technologies in the IR pipeline, such as LLMs, agentic IR, generative IR, and how should we begin to solve them?
- **Insights from Other Fields:** What takeaways or solutions from fields outside of IR can be used to advance XIR research?
- **Interdisciplinary XIR:** How do ethical concerns, regulatory requirements (e.g., GDPR), and domain-specific constraints (e.g., health, legal, finance, etc.) affect explainability research?

3.3 Special Requirements and Contingency Plans

AV equipment for speakers and panel discussions will be important for the workshop but as this equipment is typically already provided as part of the main conference, we do not expect this to be an issue.

To facilitate the plenary discussion and breakout writing sessions, it would be nice to have flexibility with the seating arrangements. However, this is not a must-have and we could easily break out into different spaces that still foster engaging discussions.

In the event that our invited keynote speaker is unable to attend, we have identified several alternative candidates who are likely to be present at SIGIR, increasing the likelihood of accepting an invitation, including Anja Reusch, Krisztian Balog, and Procheta Sen.

Finally, with at least three organizers expected to be on-site, the workshop is well-prepared to proceed smoothly even in a situation where someone encounters last-minute conflicts.

3.4 Envisioned Workshop Output

Following the workshop, we aim to produce three artifacts resulting from dedicated writing session with participating attendees: (1) a SIGIR Forum report summarizing the event, (2) a proposal to the IRRJ[3] for a special issue on explainability, and (3) a whitepaper that outlines a roadmap of future research directions in explainable IR.

4 Workshop Organizers

At least 3 organizers will attend in person, while the remaining organizers will support workshop preparation and lead follow-up community engagement and content activities.

Catherine Chen (likely to attend onsite) is a final-year PhD Candidate at Brown University, supervised by Carsten Eickhoff. Her research centers on explainable IR, particularly developing methods for explainable search, and she has published papers in leading IR and NLP venues.

Maria Heuss is an assistant professor at the IRLab, University of Amsterdam. Her research focuses on responsible advice-giving systems, with a particular interest in explainability and fairness. She has co-organized the IR-for-Good track at ECIR, organized and given tutorials on fairness and explainability topics and published at leading IR venues.

Tanya Chowdhury (potentially onsite) is a final-year PhD candidate at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, supervised by Prof James Allan. Recently, she has been particularly interested in mechanistically understanding the functioning of neural retrievers and reranking models. Her previous work has been published in ICLR, EMNLP, IJCAI, SIGIR, The Web Conf, ICTIR, etc.

James Allan is a Distinguished Professor and the director of the Center for Intelligent Information Retrieval (CIIR) in the Manning

[3] https://irrj.org/

College of Information and Computer Sciences at the University of Massachusetts Amherst. He is broadly interested in explainability for information access, with current research focusing on developing and evaluating mechanistic interpretability approaches for ranking systems.

Avishek Anand is an associate professor at the Delft University of Technology. His research focuses on explainable IR and other topics around Question Answering and information retrieval. He has published, organized, and given tutorials in leading IR venues.

Carsten Eickhoff (likely to attend onsite) is a professor at the University of Tübingen and holds a visiting appointment at Brown University. His research focuses on transparency and explainability of language models. His lab pursues these topics both in the context of information retrieval as well as natural language processing, and frequently explores applications in biomedicine.

Suzan Verberne (likely to attend onsite) is a professor of Natural Language Processing at Leiden University. Her recent work centers around interactive information access for low-resource contexts. She has a strong interest in the interplay between search engines and LLMs and the role of the user in both. She currently supervises three PhD projects on explainability. Suzan is highly active in the NLP and IR communities, holding chairing positions in the large world-wide conferences.

5 PC Members and Selection Process

The workshop is open for participation for everyone interested. Selection of extended abstracts (2 pages ACM format) will be mainly based on relevance to the workshop. We invite submissions from different stages of research, ranging from perspective papers over work in progress to previously published work, and welcome submissions from students. Reviewing will be done by the organizers themselves, as well as a number of PC members.

6 Target Audience

As previously mentioned, the target audience for our workshop is both junior and senior researchers who currently work on XIR or are interested in working on XIR in the future. Due to the highly interdisciplinary nature of explainability, we will encourage participation not only from researchers in academia, but from stakeholders and practitioners who have a vested interest in deploying explainable systems for real-world applications. Thus, SIGIR is an appropriate venue to attract such an audience, as the conference attracts a diverse group of academics, as well as a wide variety of industry and government specialists that may rely on explanations of IR systems to inform high-stakes decisions and outcomes.

7 Related Workshops

This workshop will be a follow-up to the first workshop on explainable information retrieval, which was held at SIGIR 2025 [6]. This workshop had 11 abstract submissions of which 9 were accepted and had active discussion sessions on the evaluation of explanations in IR and on the needs and challenges of practitioners and stakeholders, with 30+ participants. With this year's iteration, we want to focus on challenges that come with the emergence of novel approaches within IR and pick up the momentum created last year to start a shared community effort to tackle previously discussed challenges.

Prior to the first edition of this workshop, most related workshop activity on explainability has taken place within the recommender systems community, such as the *International Workshop on ExplainAble Recommendation and Search (EARS)* (SIGIR 2018-2020) [10] and *Measuring the Quality of Explanations in Recommender Systems (QUARE)* (SIGIR 2022 and RecSys 2023) [8].

Our proposed workshop builds on the foundation of the topics covered by EARS, investigating the progress in recent years while introducing new focuses: (1) tackling open challenges such as consistent definitions and evaluation of explanations, (2) integrating insights from related fields such as NLP and general ML, and (3) focusing on different IR use cases and their needs for explanations. While progress in explainability for recommender systems has gained traction, explainability in IR systems has not received the same focused attention, and we hope this workshop series will continue to help the community identify a shared trajectory for future exploration and advance research in this area.

Acknowledgments

Suzan Verberne acknowledges that this publication is part of the project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO).

References

[1] 2024. Colorado Senate Bill 24-205: Concerning the Regulation of Artificial Intelligence. Colorado General Assembly. https://leg.colorado.gov/bills/sb24-205 Will go into effect on February 1, 2026.

[2] 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[3] Fazl Barez, Lawrence Chan, Mor Geva, Kayo Yin, Neel Nanda, and Max Tegmark. 2024. Workshop on Mechanistic Interpretability @ ICML 2024. https://icml.cc/virtual/2024/workshop/29953

[4] Yonatan Belinkov, Aaron Mueller, Najoung Kim, Hosein Mohebbi, Hanjie Chen, Dana Arad, and Gabriele Sarti (Eds.). 2025. *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

[5] Tal Haklay, Hadas Orgad, Anja Reusch, Marius Mosbach, Sarah Wiegreffe, Ian Tenney, and Mor Geva. 2025. Workshop on Actionable Interpretability @ ICML 2025.

[6] Maria Heuss, Catherine Chen, Avishek Anand, Carsten Eickhoff, and Suzan Verberne. 2025. Workshop on Explainability in Information Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[7] Neel Nanda, Andy Arditi, Stefan Heimersheim, Anna Soligo, Andrew Lee, Martin Wattenberg, Sarah Wiegreffe, Atticus Geiger, Julius Adebayo, Kayo Yin, Fazl Barez, and Lawrence Chan. 2025. Workshop on Mechanistic Interpretability @ NeurIPS 2025. https://neurips.cc/virtual/2025/loc/san-diego/workshop/109547

[8] Alessandro Piscopo, Oana Inel, Sanne Vrijenhoek, Martijn Millecamp, and Krisztian Balog. 2023. Report on the 1st Workshop on Measuring the Quality of Explanations in Recommender Systems (QUARE 2022) at SIGIR 2022. In *ACM SIGIR Forum*, Vol. 56.

[9] FChhavi Yadav, Michal Moshkovitz, Nave Frost, Suraj Srinivas, Bingqing Chen, Valentyn Boreiko, Himabindu Lakkaraju, J. Zico Kolter, Dotan Di Castro, and Kamalika Chaudhuri. 2023. XAI in Action: Past, Present, and Future Applications. Workshop at Advances in Neural Information Processing Systems (NeurIPS). https://neurips.cc/virtual/2023/workshop/66529

[10] Yongfeng Zhang, Yi Zhang, and Min Zhang. 2019. Report on EARS'18: 1st International Workshop on ExplainAble Recommendation and Search. In *ACM SIGIR Forum*, Vol. 52.