Report on the Workshop on Explainability in Information Retrieval (WExIR) at SIGIR 2025

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

Catherine Chen
Brown University
USA
catherine_s_chen@brown.edu

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

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

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

Abstract

The first Workshop on Explainability in Information Retrieval (WExIR) took place at SIGIR 2025 in Padua, Italy. As chairs, we hoped to engage the sub-community working on transparency, interpretability and explainability in information retrieval, to identify the main challenges and an initial roadmap for research. The workshop was a success, not only in number of participants, but especially in the level of engagement from the whole room. Apart from keynotes and regular presentations and posters, we had two interactive panels with a 'hot seat' for anyone from the audience to join as well as two plenary discussion sessions. We worked on a shared document throughout the whole workshop to collect a shared research agenda. This report presents a summary of those notes. We plan to follow-up with the community in 2026 with another edition of the workshop, a survey paper, and the preparation of a special issue on the topic.

Date: 17 July 2025.

Website: https://xirworkshop.github.io.

1 Introduction

The *First Workshop on Explainable Information Retrieval (WExIR)* was held in conjunction with *SIGIR 2025*, advancing the community's agenda on transparency, accountability, and interpretability in information access systems. In this inaugural edition, the main themes went beyond the *need* for explainability toward questions of *evaluation*, *standardization*, and *deployment* in practice.[1] Explainability remains central to IR not only as a scientific challenge

[1] See workshop introduction and scope in the event report.

but also as a societal requirement in high–stakes domains such as healthcare, law, finance, and public policy. The rapid adoption of large language models (LLMs) intensifies these concerns, extending explainability beyond retrieval pipelines to hybrid retrieval–generation settings where faithfulness, plausibility, warranted trust, and user–intent alignment are critical. Currently, evolving regulatory frameworks (e.g., the EU AI Act) render explainability a compliance and governance imperative: explanations must be useful to different stakeholders (e.g., developers, auditors, and end users) and accompanied by reproducible and accountable evaluation protocols.

The workshop brought together researchers from academia and stakeholders from industry and end-user communities, to reflect on current challenges, share recent advances, and outline future directions. The format mixed invited talks, panels, lightning presentations, and extended discussion sessions.

This report synthesizes the main discussions, emerging themes, and future directions that surfaced at WExIR 2025, including (i) movement toward shared evaluation frameworks and taxonomies of explanation methods; (ii) user– and stakeholder–centric design principles; and (iii) opportunities for auditing–oriented adoption, community benchmarks, and reproducibility.

2 Accepted Extended Abstracts

During the workshop, speakers briefly pitched their work, followed by poster presentations. Accepted submissions ranged from already published work to work in progress to visions on future research. The extended abstracts, published (non-archival) on the workshop website[2], provide a good overview of the topics presented during the workshop. Across the workshop, three dominant themes emerge:

Model-Internal Interpretability

Several papers focus on understanding how neural IR models encode relevance internally, including attention-head specialization, layer-wise relevance emergence, and low-rank adaptation effects [Nijasure et al., 2025; Vast et al., 2025; Kareem et al., 2025]. These works align with mechanistic interpretability approaches adapted to IR and recommendation settings.

Explainability for Generative and Conversational IR

Multiple contributions address explainability in systems combining retrieval with generation or dialogue [Sudhi et al., 2025; Wang and Verberne, 2025; Łajewska et al., 2025]. Key questions include how to explain multi-stage pipelines, how to communicate uncertainty, and how to align explanations with conversational context. LLMs are explored not only as targets of explanation but also as tools for generating explanations themselves [Laksito and Stevenson, 2025]. This raises questions about explanation fidelity, evaluation, and the risk of explanation hallucination.

[2] https://xirworkshop.github.io/papers/

Domain-Specific Explainable IR

Domain context plays a critical role in explainability requirements. High-stakes settings such as auditing and news recommendation demand explanations that support verification, bias detection, and trust calibration [Frummet et al., 2025; Kareem et al., 2025].

3 Keynote

The keynote was delivered by Dr. Oana Inel, who discussed fostering human reflection through explanations. Acknowledging the diversity of users' opinions, perspectives, and backgrounds, the talk explored the increasingly common role of explanations in recommender and information retrieval systems. Their effectiveness, however, depends greatly on the trustworthiness of the underlying data, the characteristics of the target users, and the contexts in which the explanations are consumed. Drawing on lessons from data annotation practices and the evaluation of explanations in recommender systems, Dr. Inel illustrated why it is essential to consider both the users for whom explanations are designed and the data being explained. The second part of the talk took a practical perspective, placing users back in control and enabling them to make sense of distributed information spaces through human-centered and empowering solutions. In particular, explanations were presented as key enablers of transparency in decision-support systems and as tools for encouraging human reflection on online content. The talk concluded with various approaches to designing explanations that promote user-driven diversity, transparency, and reflection, emphasizing the importance of thorough design and evaluation processes.

4 Panels and Plenary Discussion

The workshop had two rounds of panel discussion with hot seat followed by a plenary discussion of 45 minutes each. The first round, centered around the evaluation of explanations in IR, included panelists James Allan, Gineke Wiggers and Avishek Anand. The second round, concerned with the needs and challenges of stakeholders and practitioners in IR, featured Oana Inel, Debasis Ganguly and Carsten Eickhoff.

Here we give a summary of some of the main points that were raised during those discussion sessions.

4.1 Evaluation of Explanations

Next to some concrete suggestions for the evaluation of explanations such as self-consistency metrics or building organic tasks that allow you to observe the user engagement with the explanations, the following points stood out during the discussions:

Evaluating for whom?

The user plays a central role for the definition and evaluation of any IR task, and similarly should do so for the evaluation of explanations. It should be made explicit whether the

evaluation of an explanation focuses on the model side (e.g., faithfulness) or on the user side (e.g,. plausibility, trust) and who are the system's stakeholders.

The need for a taxonomy of explanation methods

A recurring theme during the discussion was the lack of clear definitions that are applicable for different tasks and types of explanations. One goal might be defining explainability as a pipeline that allows clear communication which processing step the explanation addresses. A taxonomy of explanation approaches in the field was suggested as a possible solution to this issue. Where currently for each task and explanation type different evaluation schemes are applied, a shared taxonomy might allow for a more uniform evaluation framework. Dimensions of such taxonomy might include: task, scenario, goal of the explanation, stakeholder, rubric of evaluation items, type of user, and style of explanation.

Taking inspiration from other fields

During the workshop, several related fields and resources were brought up that could aid in defining standardized evaluation frameworks in the future, among them an existing taxonomy on explainable AI [Schwalbe and Finzel, 2024; Anand et al., 2022], a workshop on actionable XAI at ICML25 [3], and explanations in NLP [Madsen et al., 2022].

Proposed next steps

During the discussion sessions we identified a need for further conversation and alignment between different researchers and practitioners in the field. Reproducibility tracks were mentioned as a way to test hypotheses and approaches in different contexts and with different evaluation techniques. Other future steps might include a (Dagstuhl) workshop on the topic and a white paper detailing the discussed taxonomy further. A special issue in the Information Retrieval Research Journal (IRRJ) might help to further raise interest in these topics and work towards a solution as a community. Lastly, a shared task might help to build a basis for an evaluation framework that the community can iterate on.

4.2 The Needs and Challenges of Stakeholders and Practitioners in IR

Explaining for whom?

The discussion highlighted the need to actively account for the needs of stakeholders and practitioners in the design process of explainability methods. Level of interest, time and maturity/skill level might all play into the ability of the user to interact with explanations. Ways to address those different needs might include different explanation modalities and the possibility to chose a level of explanation complexity.

[3] https://actionable-interpretability.github.io

Sparse adoption of explainability methods in practical IR applications

A recurring question during the discussions asked why existing methods are not applied more frequently and why there is not more active interest from stakeholders and practitioners in explaining their algorithms to the user.

In reality for some real world applications we can even see a step back in terms of transparency, from presenting ranked lists towards AI generated summaries. Also, users do not seem to use explanations much once they are implemented. Reasons discussed ranged from a lack in available tooling, over possibility that explanations are more a novelty item, but would be overshadowed by utility overtime to the lack of actionable insights provided by explanations.

Nevertheless, the discussion concluded that we should still care about providing good explanations to build warranted user trust especially in high stakes scenarios and to allow the user to refer to the explanation in long tail cases, when results do not align with the user's expectations. Moreover, an increased interest in the future was expected, stemming from the auditing requirements stipulated in the European AI Act.

Proposed next steps

Throughout the discussions, several steps were proposed that might increase interest in applying explanations and hence building more transparent and trustworthy systems. Both advertisement of events such as future iterations of this workshop as well as active engagement within events and conferences from relevant communities and regulators might lead to increased awareness from practitioners and stakeholders. Furthermore, incoming regulations such as the EU AI act might offer new opportunities to engage with stakeholders through the auditing domain.

5 Conclusions

The discussions at WExIR 2025 converged on three natural conclusions that outline the next steps for advancing explainable information retrieval.

Towards a Shared Evaluation Framework

There was broad agreement that the field requires a common vocabulary and evaluation taxonomy. Future work should prioritize the design of a standardized evaluation pipeline that explicitly distinguishes between *faithfulness* — the model-side validity of explanations — and *plausibility* — their user-side usefulness. Community-wide *shared tasks*, *benchmark datasets*, and *reproducibility tracks* were proposed to enable consistent comparison and iterative refinement across methods and application domains.

Bridging Research and Real-World Adoption

Despite notable scientific progress, explainability methods are rarely deployed in operational IR systems. Participants emphasized the need for usable toolkits, actionable explanation

formats, and interdisciplinary collaboration to close this gap. Engagement with practitioners, auditors, and policymakers is essential to ensure that explainability tools address concrete needs rather than remaining proof-of-concept prototypes.

Regulation as an Opportunity

Evolving regulatory frameworks, such as the *EU AI Act*, were recognized as potential catalysts for wider adoption. Explainable IR can play a central role in meeting emerging requirements for transparency, accountability, and auditing. Aligning future research agendas with these governance needs could transform explainability from an optional feature into a foundational property of trustworthy and verifiable information access systems.

Together, these conclusions reflect a maturing research area — one that is moving from conceptual exploration toward shared standards, deployment readiness, and measurable societal impact.

6 Final Note

The first Workshop on Explainability in Information Retrieval made a first step towards bringing the community together to discuss the current state of research in the field and obstacles that might slow progress in the future. Through lively panel and plenary discussions, a few potential blockers of the field could be identified. Participants as well as organizers have voiced interest in starting a collaboration working on some these issues together.

References

Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. Explainable information retrieval: A survey. *arXiv preprint arXiv:2211.02405*, 2022.

Alexander Frummet, Emanuel Slany, Jonas Amling, Moritz Lang, and Stephan Scheele. Explainable information retrieval in the audit domain. In *First Workshop on Explainability in Information Retrieval (WExIR)*, 2025.

Jasmin Kareem, Martijn C. Willemsen, and Maarten de Rijke. Mechanisms of trending news in recommender systems. In *First Workshop on Explainability in Information Retrieval (WExIR)*, 2025.

Arif Laksito and Mark Stevenson. Generating search explanations using large language models. In *First Workshop on Explainability in Information Retrieval (WExIR)*, 2025.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42, 2022.

Atharva Nijasure, Tanya Chowdhury, and James Allan. How relevance emerges: Interpreting lora fine-tuning in reranking llms. In *First Workshop on Explainability in Information Retrieval (WExIR)*, 2025.

Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101, 2024.

Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, Roman Teucher, and Nicolas Flores-Herr. Towards end-to-end model-agnostic explanations for rag systems. In *First Workshop on Explainability in Information Retrieval (WExIR)*, 2025.

Mathias Vast, Basile Van Cooten, Laure Soulier, and Benjamin Piwowarski. Understanding matching mechanisms in cross-encoders. In *First Workshop on Explainability in Information Retrieval (WExIR)*, 2025.

Yumeng Wang and Suzan Verberne. Towards user-centric explainability in conversational information retrieval. In *First Workshop on Explainability in Information Retrieval (WExIR)*, 2025.

Weronika Łajewska, Damiano Spina, Johanne R. Trippas, and Krisztian Balog. Explainability for transparent conversational information-seeking. In *First Workshop on Explainability in Information Retrieval (WExIR)*, 2025.