EVENT REPORT

Report on the Second Workshop on Simulations for Information Access (Sim4IA 2025) at SIGIR 2025

Philipp Schaer
TH Köln
Cologne, Germany
philipp.schaer@th-koeln.de

Christin Katharina Kreutz
TH Mittelhessen
Gießen, Germany
ckreutz@acm.org

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

Timo Breuer
TH Köln
Cologne, Germany
timobreuer@acm.org

Andreas Kruff
TH Köln
Cologne, Germany
andreas.kruff@th-koeln.de

Mohammad Aliannejadi, Christine Bauer, Nolwenn Bernard, Nicola Ferro, Marcel Gohsen, Nurul Lubis, Saber Zerhoudi *

Abstract

This paper is a report of the Workshop on Simulations for Information Access (Sim4IA) workshop at SIGIR 2025. We present the outcome of the keynote and panel discussion, the four invited tech talks, as well as the shared task overview and participants' presentations, along with the corresponding breakout discussions. We report on how we organized the workshop, provide a brief overview of what happened at the workshop, and summarize the main topics and findings of the workshop, as well as future work.

Date: 17 July 2025.

Website: https://sim4ia.org/sigir2025.

1 Introduction

The Workshop on Simulations for Information Access (Sim4IA) is a forum for researchers and practitioners from different domains and perspectives to discuss the many facets of simulations. After a successful first iteration at SIGIR 2024 [Breuer et al., 2024; Schaer et al., 2024], the second iteration of this workshop series took place at SIGIR 2025 in Padua.

In this current second iteration of the workshop, the goal was to dive deeper into the field of user simulations, grounded on recent developments like new available toolkits like SimIIR 3.0, tutorials on this exact topic at conferences like CIKM [Balog and Zhai, 2023], or similar endeavors

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

like the CLEF Touché Lab [Kiesel et al., 2025] or TREC iKAT, which both employed user simulations in the shared tasks' settings, demonstrating possibilities of LLM-based approaches. Sim4IA is meant to be a specific venue for presenting and discussing new and experimental approaches and evaluation settings. To understand how and whether the evaluation of information access technology can truly benefit from simulating user interactions, not only are tools and frameworks critical, but a multidisciplinary discussion and mutual understanding among the broad and sometimes conflicting perspectives are necessary.

Therefore, the goals of the workshop were as follows: (1) Continue our series of workshops to generate an open conversation about possible future scenarios, applications, and methods to include simulations in the evaluation of IA systems; (2) Provide a forum at SIGIR to discuss the pressing and emerging issues the IR community faces, and how simulations can help to overcome these; (3) Develop and advertise the idea of organizing a TREC/CLEF campaign that includes simulations as a core element; (4) Test an initial setting for two (micro) shared tasks designed around two information access use cases that might form the basis for the aforementioned TREC/CLEF campaign.

In this SIGIR Forum paper, we report on the Sim4IA[1] [Schaer et al., 2025] workshop at SIGIR 2025. We present the outcome of the keynote and panel discussion, the four invited tech talks, as well as the shared task overview and participants' presentations, along with the corresponding breakout discussions. We report on how we organized the workshop, provide a brief overview of what happened at the workshop, and summarize the main topics and findings of the workshop, as well as future work.

2 Workshop Overview

Sim4IA was a full-day workshop at SIGIR 2025, held in Padua, Italy, on 17 July 2025. The workshop attracted 35 participants who participated in an interactive setting where we continued our last year's approach of having a workshop instead of a typical "mini-conference" (see Table 1). After the keynote in the morning, we went over to a series of short but intensive tech talks on toolkits and infrastructures for user simulations, followed by a panel discussion. The SIGIR chairs organized a poster session for all workshops where we presented the design and preliminary outcomes of our shared task. These posters attracted quite some visitors, and we could attract some interested persons for the afternoon session. The afternoon was dedicated to the shared task we organized and two breakout sessions where we discussed validation measures for user simulations and the outline of an upcoming TREC track.

In contrast to last year's workshop, we decided to do an on-site only workshop, as the limited hybrid setting from last year was not as fruitful as we had hoped it to be. During the workshop, we fostered the usage of a new SIGIR Slack channel named `#usersim`.

[1] https://sim4ia.org/sigir2025/

Time	Agenda
9:00–9:15	Welcome
9:15–10:00	Keynote by Christine Bauer
10:00–10:30	Invited tech talks on toolkits and infrastructure pt 1
10:30–11:00	Coffee break
11:00–12:00	Panel discussion
12:00–12:30	Invited tech talks on toolkits and infrastructure pt 2
12:30–14:00	Lunch break
13:00-14:30	Workshop Posters
14:30–15:00	Overview talk on shared tasks & Lightning talks on shared tasks
15:00–16:00	Breakout group discussions
16:00-16:30	Coffee break
16:30-17:30	Reports of the group discussions and closing

Table 1. Timeline of the Sim4IA workshop.

3 Keynote

Our keynote speaker, Christine Bauer (University of Salzburg), delivered her keynote titled *From toy models to tactics: What user simulation is good for* before taking questions from the audience (see the left side of Figure 1).

In her keynote, Christine explored the role of user simulation in advancing information access systems, with her special focus on recommender systems. She advocated moving from "toys" toward actionable "tactics," in the sense that many simulation tasks we look at today do not fully embrace the methodology and its potential. A central question was what insights simulated users can provide that real users cannot, and vice versa. Motivations for simulation include data sparsity, ethical concerns, and the high cost of user studies. Unlike traditional evaluations that provide hindsight, simulations enable foresight in system design.

She emphasized that user modeling is essential: the notion of an "average user" is misleading, as behavior is diverse and context-dependent. A key challenge lies in balancing abstraction and realism—trading off speed, interpretability, and generalizability against fidelity. Simulations can capture long-term dynamics such as session patterns, habit formation, and drift, and they help uncover non-obvious or paradoxical effects, as shown in recent studies [Ferraro et al., 2024]. Interestingly, variation introduced by algorithms often exceeds that from user choice models. Simulating extreme cases can also test robustness. Frameworks like CAFE [Bauer et al., 2024, 2025] were highlighted as ways to manage complexity. Finally, while large language models are increasingly used for simulation, the keynote urged caution: they are not valid representations of users, lack control, and carry significant biases. As a final remark, Christine concluded by framing simulations as nothing that is about perfect replication, but about probing possibilities, revealing hidden dynamics, and guiding system design.

Figure 1. Left: Keynote. Right: Panel discussion.

4 Panel Discussion

Between the keynote and talks, the workshop featured a panel discussion with four invited panelists, including our keynote speaker Christine Bauer, Nicola Ferro (University of Padua), Ian Soboroff (NIST), and Mohammad Aliannejadi (University of Amsterdam). The panelists shared short opening statements (see Table 2), then a structured debate followed, where panelists were assigned opposing sides of topics, and afterward one hypothetical user simulation scenario was discussed. The panel was moderated by Krisztian Balog (see the right side of Figure 1).

4.1 Structured Debate

User simulation can (soon) completely replace human evaluation in most large-scale evaluation campaigns.

Assigned pro: Nicola; assigned contra: Christine, Ian, Mohammad

There is growing skepticism towards the traditional Cranfield paradigm and a suggestion to remove human annotators from our evaluations. With the available technical means, this is no longer an unattainable future. While one could criticize Cranfield as being too short-sighted and use simulations to explore edge cases in evaluation, the human in the loop still holds importance. Especially for developing edge cases, research should be grounded in real users' behavior. The level at which humans are involved in evaluations is crucial: data-intense steps could be automated, similar to testing a car extensively before letting a human operate it.

It is crucial to underline that *simulation is simplification*. Simulation cannot fully capture human behavior and should not depict the complete reality, nor should it attempt to. Its value lies in generating useful insights and recognizing errors made for systematic reasons. The goal of evaluating interactive systems with simulations was questioned with two use cases mentioned: (1) Creating better data for machine learning and (2) supporting users in satisfying their information needs. In the first case, we run into the problem of models learning from learning output, leading to collapsing models. The second case sparks considerations on the difference between objective

Panelist	*The most promising use case for user simulation in my opinion is...*	*The biggest challenge we face with user simulation for evaluation is...*
Christine Bauer	to evaluate and compare the outcomes of extreme scenarios.	that we misuse them—relying on user simulators designed without sufficient thought; then, misinterpreting their outputs as reality, which leads to fundamentally flawed conclusions.
Ian Soboroff	where a simulation can reveal boundary cases and unexpected behavior in the system.	having a system learn the simulation and not the 'real problem."
Mohammad Aliannejadi	enabling evaluation and development of systems for corner-case and non-mainstream users.	how to handle and whether to model user biases and stereotypes.
Nicola Ferro	generating query variants, providing (pseudo-)relevance feedback, generating interaction patterns.	(1) connecting the content generation (aka LLM) with the interaction generation (aka Markov models) part and (2) evaluating and validating simulation.

Table 2. Panelists' opening statements for pre-defined questions.

relevance and perceived relevance, as well as balancing human annotators being imperfect with verifiable rewards that are not based on human judgment.

Ultimately, the open question remains where exactly in the loop a human should be employed.

The development of intelligent information access systems and user simulators should go hand-in-hand.

Assigned pro: Christine, Nicola; assigned contra: Mohammad; unassigned: Ian

Intelligent information access systems are typically developed and refined iteratively. One approach is to design a system and a simulator separately, based on requirements collected from potential users. Preventing an alignment between system and simulator allows for a more unbiased evaluation. Here, evaluation campaigns could help ensure comparability across simulators. However, a drawback of such disjoint development is the potentially limited usefulness of simulators in evaluation if they do not fit our systems. Integrating systems into real work environments or supporting legacy workflows should go hand-in-hand not only with simulators but also with real users.

Regardless of a joint or disjoint development of system and simulator, explicit and implicit assumptions on users' information needs shape their design. Simulators can be developed as data-driven or model-driven, but do not necessarily have to rely on data. Our systems can be considered as sets of hypotheses that should be critically evaluated. Simulation serves as a tool to disprove these hypotheses.

What users want and what they need are different and oftentimes even unknown to them. For instance, if a user searches for biased information, it is difficult to design a system that strikes a balance between serving user preferences and promoting information quality. One extreme could be to *not simulate, just build*, trusting that users will adapt to a system. This risks users employing known strategies to new systems, even if they are ill-fitting—"if users are given hammers, they think about their job as hammers and nails."

A more user-focused approach could be a system asking clarifying questions or allowing agents to take a more proactive role in a conversational search setting. Lastly, the definition of users' *success* of an information interaction is task-dependent. A more suitable evaluation measure would be more flexible than merely considering satisfaction.

Building an effective user simulator requires massive amounts of real user interaction logs.

Assigned pro: Nicola; assigned contra: Christine, Ian, Mohammad

Building upon massive amounts of user data is already a reality. Currently, access to such logs is only available for individuals working in companies that collect this data. Employing smaller interaction logs leads to problems of independence and collapsing models if these logs are used to develop both systems and simulators. Historical interactions might not contain all modes we would be interested in for simulation, and also come with the problem of limited applicability of the old data to new ideas. The question arose whether log data might not be the correct thing to use for building effective user simulators.

One possibility to overcome the lack of data could be an attempt to draw logs from different fields. Another possibility would be the conduction of user studies. As one downside, user studies typically lack the consideration of edge cases. These edge cases or extreme scenarios could be simulated by identifying gaps in logs and data and simulating the opposite of the observable behavior. In this case, one would come up with assumptions and hypotheses based on trusting and understanding the logs to define heuristics. Sometimes it is impossible to collect data through a user study for simulation, such as with persons who are not using a system. Other times, it is not necessary to collect data in order to be able to realistically simulate, such as the simulation of patients with different degrees of Alzheimer's. Here, we are facing a data-less simulator by defining characteristics of user groups. While a simulation needs data to run, a simulator only needs assumptions.

In the end, the question of the difference between using massive amounts of data with an abstraction of characteristics and using heuristics only arose. It is unclear when and if both become equivalent.

4.2 Hypothetical Scenario

Table 3 presents the hypothetical scenario that was given to the panelists.

For developing systems and simulations, we require good training and evaluation sets. As collecting training labels is expensive, this step can be outsourced. However, the collection of evaluation labels should not be outsourced; rather, it should remain in-house. Crowdsourcing this step would save money in annotation, but sanitizing would be expensive.

Scenario: Suspicion of Simulated Annotations
You are running an evaluation campaign for a conversational information access task, where you require annotations at both the turn level and the conversation level. You outsourced relevance annotations to a third-party provider. Later, you begin to suspect that some annotations may have been generated by simulated users rather than real human annotators.

What evidence would you look for to support this suspicion?
(How would you identify annotations that are most likely machine-generated, and what strategies might you use to validate or flag them?)

Table 3. Hypothetical scenario.

The reason why we assume annotations have been artificially generated can lead to hypotheses on hints in the data to confirm these suspicions. Depending on the type of data that has been annotated, strange annotations raising suspicion should lead to looking into more detail. Evidence on artificially generated annotations could focus on coherence: A topical ill-fit of a response to a specific domain, such as an answer to a technical task not containing any technical terms, as the LLM defaulted to the material it has been trained on. Other evidence could be found in consistency, such as specific patterns found in responses. Examples, especially telling in long texts, could be recurring numbers of turns in conversations and the number of sentences. Some AI detection tools developed for textual data flag the usage of specific linguistic features. Nevertheless, such indicators are no definite proof as the flagged annotations could still stem from humans.

Besides detecting artificially generated annotations automatically, one could include adversarial examples or honeypots for systems that humans would not pick up on. Such an approach could be useful for a limited time until the next generation of systems can pass these tests.

Another perspective on the issue of generated annotations is to accept them as valid if they are high-quality data. Concrete ideas on how to incentivize humans to produce such high-quality data in collaboration with LLMs are still to be developed.

5 Invited Tech Talks

A total of four invited talks on toolkits and infrastructure were given, spread out over two designated sessions. The time frame for each lighting talk was 10 minutes plus some time for questions. Table 4 holds the titles and presenters of the four talks.

A Step Towards the Interactive Shared Task

Marcel Gohsen introduced the user simulation infrastructure at the Touché Retrieval-Augmented Debating (RAD) at CLEF '25 and the Interactive Knowledge Assistance Track (iKAT) at TREC '25 shared tasks.[2,3]

[2] https://touche.webis.de/clef25/touche25-web/retrieval-augmented-debating.html
[3] https://www.trecikat.com

Authors	Title
Marcel Gohsen **Saber Zerhoudi**	A Step Towards the Interactive Shared Task Simulators Meet Agents: Towards a Hybrid User Search Simulation
Nurul Lubis	ConvLab-3: A Flexible Dialogue System Toolkit Based on a Unified Data Format
Nolwenn Bernard, **Krisztian Balog**	UserSimCRS v2

Table 4. List of invited tech talks. Presenters in bold.

At RAD, participant submissions consisted of a dockerized debate system that was executed inside TIRA [Fröbe et al., 2023]. Participants had to implement their system against a predefined HTTP-API to ensure successful communication with a secret simulator on a secret set of debate topics. This procedure lead to low software stack limitations and no hardware requirements for participants coupled with high developmental complexity and the risk of errors in the execution of the approaches on the unseen topics. For organizers this submission setup lead to significant efforts in support and maintenance but produced reusable software artifacts, reproducibility of runs on different datasets and simulators. Additionally, metadata about the run executions (e.g., runtimes, energy consumption) could be collected with this setup.

At iKAT, participant submissions consisted of either run files in the offline track or interactions with an API in the interactive track. The submissions are assessed on conversations with a secret simulator on topics from a public test set. This approach lead to hardware requirements for participants but did not imply any software stack limitations leading to low development complexity, the possibility of manual runs, and no issues due to formatting or run failures. For organizers this setup lead to no reusable software artifacts and the inability to use a post-hoc simulator or change of datasets. On the plus side, this approach lead to a lower level of required support and the possibility to measure run metadata. The setup of interacting with a submission API would also allow for secret test datasets.

Simulators Meet Agents: Towards a Hybrid User Search Simulation

Saber Zerhoudi identified one of the pain points of simulation-based evaluation as experimental overhead; the difficulty in scaling experiments to increase their scope and granularity. A simulation framework currently in development, UXSim [Zerhoudi and Granitzer, 2025], is being engineered to directly address this challenge. It enables the simplified construction, running, and sharing of simulations. This approach lowers the barrier to entry for complex and high-fidelity simulation.

The platform is designed for accessibility. It provides a "plug-and-run" Python backend and a web-based interface that supports real-world user interface interaction and simplifies the entire research workflow: from configuration and execution to the visual analysis of results. The framework's core is the hybrid orchestration of different simulation components. It integrates traditional rule-based simulators, like those in SimIIR [Zerhoudi et al., 2022; Azzopardi et al., 2024], with new LLM-based components, emphasizing grounded LLM reasoning as well as explainability. This structure allows researchers to quickly conduct baseline experiments using traditional, rule-based

simulators to determine if a simulation-based approach is a viable answer to their research question before committing to more complex models. To enhance explainability, agents are able to reveal their reasoning for decisions directly in the web UI.

ConvLab-3: A Flexible Dialogue System Toolkit Based on a Unified Data Format

Nurul Lubis presented ConvLab-3 [Zhu et al., 2023], an open-source toolkit built for researchers to perform experiments across various corpora, developers to construct task-oriented dialogue systems and community contributors to share their models and datasets. The framework provides a unified format for dialogue corpora and system modules. It streamlines training, enables in-depth evaluation and contains rich user simulators. A demonstration[4] of the toolkit showcased its straightforward use.

UserSimCRS v2

Krisztian Balog presented the UserSimCRS v2 toolkit[5] [Bernard and Balog, 2025] as an extension of the first version [Afzali et al., 2023]. Main features of the toolkit include agenda-based user simulation with a modular architecture typical of task-oriented dialogue systems and a component for user modeling with options to modify preferences, personality traits and environmental context. Extensions to the toolkit include the inclusion of LLMs for natural language understanding and generation as well as information need-guided creation and update of agendas. Additionally, LLM-based user simulations utilizing different prompting strategies are integrated.

6 Micro Shared Task

This year's iteration of our workshop featured a micro shared task with three subtasks on CORE log files: (A1) next query prediction based on the previous query, (A2) next query prediction based on the entire previous session and (B) next utterance prediction in a conversational setting with a given session.

Andreas Kruff first introduced the micro shared task and the submitted runs, before he, Marcel Gohsen, and Christin Kreutz gave short presentations of their groups' submissions. All results and all information on the different approaches, the corresponding artifacts, including posters, presentations, and run files, are publicly available in our Zenodo community[6].

In the following, we provide a brief overview of the shared task. The Micro-Shared Task itself was conducted using an adapted version of the SimIIR 3 framework. Participants were provided with training and test session files derived from the CORE Search Engine prior to the conference in order to submit their contributions. While the training set included complete sessions, to allow participants to validate their current approaches, the test set withheld the last query or utterance of each session to prevent overfitting to the given dataset. For all 35 test sessions, participants were required to submit ten candidate queries for the respective last query. Different evaluation measures were tested beforehand and served as the basis for further discussion in the workshop's

[4] https://convlab.github.io
[5] https://github.com/iai-group/UserSimCRS
[6] https://zenodo.org/communities/sim4ia_workshop_2025/

breakout groups. These included cosine similarity, used to assess the semantic indistinguishability of queries; SERP overlap, which evaluates similarity at the system-response level (without the need of relevance judgments); and a newly proposed MMR-inspired measure, the Rank-Diversity Score (RDS), designed to reward simulators capable of generating multiple diverse yet semantically related query candidates.

We hope that the artifacts resulting from the Micro-Shared Task will help sharpen the understanding of how a shared task on user simulation can be designed and inspire ideas for validating simulators in terms of their ability to replicate real user behavior. To establish a common ground, we released all artifacts related to the workshop and the Micro-Shared Task as part of our submission to the ECIR 2026 Resource Track. A more detailed description of the experimental setup and the proposed measures is available as an arXiv preprint [Kruff et al., 2025]. For further insights into the expected impact and the lessons learned from this and other user simulation Shared Task, see the corresponding parallel paper in this issue of the SIGIR Forum [Gohsen et al., 2025]. For additional details on the Micro-Shared Task, we refer to the official website[7] and the corresponding repository[8].

7 Summary of Breakout Groups

The afternoon session was complemented by two breakout group discussions, one on evaluation measures and one on the outline of a new simulation-based shared task.

7.1 Group Discussion on Measures

This breakout session was seven people strong. The discussion focused on how to design and validate measures for evaluating user simulations, with particular attention to the shared task on how users formulate and reformulate queries. A key question was how to handle cases where more than one query is used in the evaluation and what the appropriate comparison baseline should be. The consensus was that simulated interactions should be compared to original human interactions rather than system outcomes, since factors outside the query simulation itself can influence outcomes.

Participants emphasized the importance of validity, arguing that measures must be compared against human behavior to ensure realism and that comparisons should consider both individual users and aggregate (average) behavior. When multiple human users are available, simulated behaviors can be compared to the overall distribution of real user actions, including not only queries but also clicks and other types of interactions. The group also discussed the challenge of topic drift, which should be assessed over multiple query suggestions, and the difficulty of evaluation without explicit topic descriptions—though these can potentially be generated or sourced from datasets such as TREC. While LLMs might be used as judges, concerns have been raised about their reproducibility, making human or crowd-sourced evaluations, such as those conducted by Turkers, more reliable alternatives. Finally, the importance of defining specific user types in advance was highlighted, as comparing across heterogeneous groups (such as adults versus children)

[7] https://sim4ia.org/sigir2025/
[8] https://github.com/sim4ia/sigir2025-shared-task

risks producing meaningless averages. Overall, the measures discussed were considered appropriate for the problem, provided that comparisons remain grounded in human behavior and clearly defined user contexts.

7.2 Group Discussion on a New Shared Task

The breakout group focused on defining a new shared task that bridges the human-centered and system-centered IR communities. The discussion was structured around three key pillars: the information access problem, the evaluation setup, and the assessment of submissions.

The group aimed to define a task that moves beyond standard ad hoc search by incorporating exploratory, multi-turn interactions. We identified a specific information access problem: researchers searching for datasets with the help of a conversational agent. In this scenario, the user must converse with a system to identify suitable datasets for their research. The proposed training data is sourced from a user study recording researchers' interactions with two LLM-based assistants. This dataset is uniquely suited for user simulation as it is being extended to include transcripts of researchers' thought processes (think-alouds) alongside standard click data.

A major topic of discussion was the infrastructure required for interactive evaluation. The group noted that requiring participants to conduct their own user studies or collect interaction data would create a prohibitive barrier to entry. To address this, we reached a consensus on an API-based setup. In this framework, participants operate user simulators that interact with the hosted conversational agents via a shared API. This ensures standardized evaluation, while enabling participants to build simulators using their choice of tooling.

Finally, we determined that the evaluation burden must lie with the organizers rather than NIST or the participants. The creation of a high-quality test dataset—including the transcription of user intent and thought processes—will be centralized. This ensures consistent quality checks and allows participant submissions to be evaluated directly against ground-truth human behaviors.

These discussions provided critical input for the formulation of the User Simulation track proposal, which has been submitted (and accepted) for TREC 2026.

8 Conclusion

The second iteration of the Sim4IA workshop demonstrated a high level of interest from the participants while shedding light on the open issues and unsolved problems of user simulations. The combination of a keynote, a panel discussion, invited tech talks, and the shared task highlighted the many facets of the workshop's topic and the challenges ahead in the near future.

The shared task was the first of its kind and introduced a new perspective on how to design and evaluate user simulations, showing the potential but also the limitations of such an endeavor. As one of the workshop's goals, we discussed and outlined a new shared task for TREC, which was well-received and addressed during the breakout sessions. We already recognized a strong interest from the community last year, but we concluded that the discussion at a workshop like Sim4IA needs to be more focused and streamlined. With the micro shared task, we did exactly this. This resulted in a clear plan for the next steps. The valuable input and in-depth discussion with experienced shared task organizers and current/future participants were channeled into a

shared task proposal for the TREC 2026 User Simulation track that was accepted just days before submitting this manuscript. We will keep you posted.

In the meantime, we invite you to visit the `usersim.ai` website[9] and subscribe to the mailing list. We envision this portal as the central hub for fostering collaboration and building a dedicated community around user simulation.

Acknowledgments

This workshop has been partially funded by the project PLan_CV. Within the funding programme FH-Personal, the project PLan_CV (reference number 03FHP109) is funded by the German Federal Ministry of Education and Research (BMBF) and Joint Science Conference (GWK). The German Research Foundation (DFG) also partially funded this event under project number 509543643.

We would also like to thank our panelist, Ian Soborrof (NIST), for making the panel discussion the most active and best-attended part of the workshop.

A Authors and Affiliations

Workshop organizers:

- Philipp Schaer; TH Köln, Cologne, Germany; philipp.schaer@th-koeln.de
- Christin Katharina Kreutz; TH Mittelhessen, Gießen, Germany; ckreutz@acm.org
- Krisztian Balog; University of Stavanger, Stavanger, Norway; krisztian.balog@uis.no
- Andreas Kruff; TH Köln, Cologne, Germany; andreas.kruff@th-koeln.de
- Timo Breuer; TH Köln, Cologne, Germany; timobreuer@acm.org

Keynote speaker and panelists:

- Christine Bauer; University of Salzburg, Salzburg, Austria; christine.bauer@plus.ac.at
- Mohammad Aliannejadi; University of Amsterdam, Amsterdam, The Netherlands; m.aliannejadi@uva.nl
- Nicola Ferro; University of Padua, Padua, Italy; ferro@dei.unipd.it

Other authors:

- Nolwenn Bernard; TH Köln, Cologne, Germany; nolwenn.bernard@th-koeln.de
- Marcel Gohsen; Bauhaus-Universität Weimar, Weimar, Germany; marcel.gohsen@uni-weimar.de
- Nurul Lubis; Heinrich Heine University Düsseldorf, Düsseldorf, Germany; lubis@hhu.de
- Saber Zerhoudi; University of Passau, Passau, Germany; saber.zerhoudi@uni-passau.de

[9] https://usersim.ai

References

Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. Usersimcrs: A user simulation toolkit for evaluating conversational recommender systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1160–1163, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3573029. URL https://doi.org/10.1145/3539597.3573029.

Leif Azzopardi, Timo Breuer, Björn Engelmann, Christin Kreutz, Sean MacAvaney, David Maxwell, Andrew Parry, Adam Roegiest, Xi Wang, and Saber Zerhoudi. Simiir 3: A framework for the simulation of interactive and conversational information retrieval. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, page 197–202, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707247. doi: 10.1145/3673791.3698427. URL https://doi.org/10.1145/3673791.3698427.

Krisztian Balog and ChengXiang Zhai. Tutorial on user simulation for evaluating information access systems. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos, editors, *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 5200–5203. ACM, 2023. doi: 10.1145/3583780.3615296. URL https://doi.org/10.1145/3583780.3615296.

Christine Bauer, Li Chen, Nicola Ferro, and Norbert Fuhr. Conversational agents: A framework for evaluation (CAFE) (Dagstuhl Perspectives Workshop 24352). *Dagstuhl Reports*, 14(8):53–58, 2024. doi: 10.4230/DAGREP.14.8.53. URL https://doi.org/10.4230/DagRep.14.8.53.

Christine Bauer, Li Chen, Nicola Ferro, Norbert Fuhr, Avishek Anand, Timo Breuer, Guglielmo Faggioli, Ophir Frieder, Hideo Joho, Jussi Karlgren, Johannes Kiesel, Bart P. Knijnenburg, Aldo Lipani, Lien Michiels, Andrea Papenmeier, Maria Soledad Pera, Mark Sanderson, Scott Sanner, Benno Stein, Johanne R. Trippas, Karin Verspoor, and Martijn C. Willemsen. Manifesto from dagstuhl perspectives workshop 24352 – conversational agents: A framework for evaluation (CAFE). *CoRR*, abs/2506.11112, 2025. doi: 10.48550/arXiv.2506.11112. URL https://doi.org/10.48550/arXiv.2506.11112.

Nolwenn Bernard and Krisztian Balog. UserSimCRS v2: Simulation-based evaluation for conversational recommender systems. *arXiv*, cs.IR/2512.04588, 2025. URL https://arxiv.org/abs/2512.04588.

Timo Breuer, Christin Katharina Kreutz, Norbert Fuhr, Krisztian Balog, Philipp Schaer, Nolwenn Bernard, Ingo Frommholz, Marcel Gohsen, Kaixin Ji, Gareth J. F. Jones, Jüri Keller, Jiqun Liu, Martin Mladenov, Gabriella Pasi, Johanne R. Trippas, Xi Wang, Saber Zerhoudi, and ChengXiang Zhai. Report on the 1st workshop on simulations for information access (sim4ia 2024) at SIGIR 2024. *SIGIR Forum*, 58(2):1–14, 2024. doi: 10.1145/3722449.3722460. URL https://doi.org/10.1145/3722449.3722460.

Andres Ferraro, Michael D. Ekstrand, and Christine Bauer. It's not you, it's me: The impact of choice models and ranking strategies on gender imbalance in music recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys 2024, pages 884–889. ACM, 2024. doi: 10.1145/3640457.3688163. URL https://doi.org/10.1145/3640457.3688163.

Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, pages 236–241, Berlin Heidelberg New York, April 2023. Springer. doi: 10.1007/978-3-031-28241-6_20. URL https://doi.org/10.1007/978-3-031-28241-6_20.

Marcel Gohsen, Zahra Abbasiantaeb, Mohammad Aliannejadi, Krisztian Balog, Timo Breuer, Jeffrey Dalton, Maik Fröbe, Christin Katharina Kreutz, Andreas Kruff, Simon Lupart, Nailia Mirzakhmedova, Harriscen Scells, Philipp Schaer, Benno Stein, and Johannes Kiesel. User simulation in practice: Lessons learned from three shared tasks. *SIGIR Forum*, 59(2), 2025.

Johannes Kiesel, Çağrı Çöltekin, Marcel Gohsen, Sebastian Heineking, Maximilian Heinrich, Maik Fröbe, Tim Hagen, Mohammad Aliannejadi, Sharat Anand, Tomaž Erjavec, Matthias Hagen, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevičius, Harrisen Scells, Moritz Wolter, Ines Zelch, Martin Potthast, and Benno Stein. Overview of Touché 2025: Argumentation Systems. In Jorge Carrillo de Albornoz, Julio Gonzalo, Laura Plaza, Alba García Seco de Herrera, Josiane Mothe, Florina Piroi, Paolo Rosso, Damiano Spina, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025)*, Lecture Notes in Computer Science, pages 486–508, Berlin Heidelberg New York, September 2025. Springer. doi: 10.1007/978-3-032-04354-2_25. URL https://doi.org/10.1007/978-3-032-04354-2_25.

Andreas Konstantin Kruff, Christin Katharina Kreutz, Timo Breuer, Philipp Schaer, and Krisztian Balog. Sim4IA-Bench: A User Simulation Benchmark Suite for Next Query and Utterance Prediction. *arXiv*, cs.IR/2511.09329, 2025. URL https://arxiv.org/abs/2511.09329.

Philipp Schaer, Christin Katharina Kreutz, Krisztian Balog, Timo Breuer, and Norbert Fuhr. Sigir 2024 workshop on simulations for information access (sim4ia 2024). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 3058–3061, New York, NY, USA, 7 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657991. URL https://doi.org/10.1145/3626772.3657991.

Philipp Schaer, Christin Katharina Kreutz, Krisztian Balog, Timo Breuer, and Andreas Konstantin Kruff. Second SIGIR workshop on simulations for information access (sim4ia 2025). In Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan

Verberne, editors, *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 4172–4175. ACM, 2025. doi: 10.1145/3726302.3730363. URL https://doi.org/10.1145/3726302.3730363.

Saber Zerhoudi and Michael Granitzer. Uxsim: Towards a hybrid user search simulation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea*. ACM, 2025. doi: 10.1145/3746252.3761640. URL https://doi.org/10.1145/3746252.3761640.

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 *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 4661–4666, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557711. URL https://doi.org/10.1145/3511808.3557711.

Qi Zhu, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Shutong Feng, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gasic, and Minlie Huang. ConvLab-3: A flexible dialogue system toolkit based on a unified data format. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 106–123, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.9. URL https://aclanthology.org/2023.emnlp-demo.9/.