User Simulation in Practice: Lessons Learned from Three Shared Tasks

Marcel Gohsen, Zahra Abbasiantaeb, Mohammad Aliannejadi, Krisztian Balog, Timo Breuer, Jeffrey Dalton, Maik Fröbe, Christin Katharina Kreutz, Andreas Kruff, Simon Lupart, Nailia Mirzakhmedova, Harrisen Scells, Philipp Schaer, Benno Stein, Johannes Kiesel [*]

Abstract

The Cranfield paradigm has been the dominant approach to evaluate information retrieval systems for decades, but—in its classical form—has clear limitations when it comes to conversational search systems, which synthesize unique outputs in a dynamic multi-turn interaction with the user. User simulation, i.e., the interaction of a computer program with a retrieval system instead of a human user to generate plausible conversations as a basis for evaluation, was proposed several years ago as a way to integrate the dynamics of conversational systems into an evaluation framework. Seen as a distant vision for years, the advent of large language models has propelled this idea forward. In 2025, there were the first three shared tasks in information retrieval where user simulation was used for evaluation or was the participants' goal. In this article, the organizers of these three shared tasks report on their specific evaluation approaches, highlight differences in setup, report on insights gained, and look to the future to discuss how user simulation can be integrated into a new evaluation paradigm.

1 Introduction

The evaluation of interactive and conversational retrieval systems has been a long-standing problem in the information retrieval community. To date, the de facto standard methods for evaluating such systems include Cranfield-style evaluations, user studies, and, more recently, "large language models (LLMs) as a judge" paradigms. All of these methods have their own disadvantages for evaluating conversational search: The Cranfield paradigm classifies only a fixed set from the almost infinite space of relevant responses as relevant [Penha and Hauff, 2020]. User studies are costly and time-consuming, and therefore difficult to scale [Gienapp et al., 2025]. LLMs can be biased towards their own content, can disagree with humans, limit reproducibility, and lack variance of opinion [Dietz et al., 2025], harming the reliability of the system evaluation.

User simulation, which can serve as an alternative to above-mentioned evaluation paradigms, was proposed decades ago. User simulation uses an "intelligent agent" to simulate how a user interacts with an interactive system [Balog and Zhai, 2024]; i.e., in the context of conversational

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

systems, a conversation between a user and the system is simulated. These conversations can be examined offline and analyzed both quantitatively and qualitatively to assess the utility of a system. For a long time, user simulation remained a theoretical idea, as there were no technologies capable of mimicking human conversation abilities with sufficient quality. With the advent of LLMs, however, the idea of user simulation has been revived and is currently heading towards becoming the next standard paradigm for evaluating conversational retrieval systems.

To advance research in this direction, we applied this paradigm to three complex shared tasks in the field of information retrieval, for the first time for these tasks: the Retrieval-Augmented Debating task at Touché [Kiesel et al., 2025], the Micro-Shared task at Sim4IA [Schaer et al., 2025b], and the Interactive Response Generation task at TREC iKAT. This article reports on the operationalization of the user simulation from the perspective of the task organizers. We provide an overview of the findings obtained from these field trials and formulate suggestions for the practical application of this new evaluation paradigm. Finally, we outline our vision of what user simulation-based evaluation of information retrieval systems could look like in the future.

2 Background

Balog and Zhai [2024] outline an extensive historical background on the origins of simulation technologies for the evaluation of information retrieval (IR) systems, from which we provide a brief summary. Using simulations for IR evaluation goes back to the 1960s, when, for example, Blunt [1965] employed simulations to measure the efficiency of a retrieval system. However, before the 2000s, only a handful of simulation based evaluations were conducted to simulate query generation [Cooper, 1973; Gordon, 1990; Griffiths, 1978], relevance feedback [Jones, 1979; Harman, 1992], or search processes [Tague et al., 1980].

In the early 2000s, a new interest for simulation was sparked in the research community of interactive information retrieval. Earlier lines of work such as the simulation of relevance feedback [Leuski, 2000; White et al., 2004, 2005; Keskustalo et al., 2008] and query generation [Jordan et al., 2006; Azzopardi and de Rijke, 2006], and later of employing user simulation for exploring efficiency aspects [Azzopardi, 2009, 2011; Baskaya et al., 2012], were continued. After this period, interest in user simulations waned for a while. With the increasing attention and technical advancements in conversational interfaces for information access, a new interest was triggered for user simulation in the 2020s. This resurgent engagement was made visible through the first workshop of Simulation for IR Evaluation (Sim4IR) as SIGIR'21 [Balog et al., 2021].

Nowadays, user simulation, with LLMs as its backbone, is gaining substantial traction for the evaluation of information access systems, particularly in the areas of ad hoc retrieval, conversational search and recommendation. In ad hoc retrieval, the primary goal is to generate queries for an information need [Breuer et al., 2022a], often extended by providing click and stop decisions in a search session [Engelmann et al., 2024]. Similarly, in conversational search, simulations are most concerned with producing the next user utterance in a conversation. For example, Wang et al. [2024] investigated simulation of user responses to clarifying questions, while Kiesel et al. [2024a] examined an LLM's ability to formulate follow-up questions in information-seeking conversations. The ability of a user simulator to interact with a system has enabled the evaluation of mixed-initiative systems that can ask clarifying questions to sharpen the understanding of an information need [Sekulić et al., 2022; Sekulic et al., 2024; Owoicho et al., 2023].

Several frameworks have been proposed for simulation-based evaluation of conversational IR systems. In the domain of conversational recommender systems (CRS), Afzali et al. [2023] introduced UserSimCRS, a toolkit for evaluating conversational recommender systems using agenda-based user simulation and context modeling. Bernard and Balog [2025] later updated this framework to include LLM-based simulators, "LLM-as-a-judge" evaluation methods, and support for standard benchmark datasets. Bernard et al. [2025] proposed SimLab, a cloud-based platform which includes pre-built user simulators, CRS systems, datasets, and evaluation metrics, enabling researchers to benchmark their systems, be it user simulators or recommender systems, in a controlled and reproducible manner. In addition to these CRS tools, Kiesel et al. [2024b] proposed GenIRSim, an LLM-based framework for evaluating Generative Information Retrieval (Gen-IR) systems, which allows for exploration of the user simulation parameter space and analysis of their impact on system evaluation. Similarly, SimIIR 3 [Azzopardi et al., 2024] supports the creation of LLM-based user simulators for interactive information retrieval tasks, which can respond to system outputs and interact with a conversational search result page.

3 The 2025 Shared Tasks Featuring Simulated Users

User simulation for the evaluation of Information Retrieval was employed or addressed in three shared tasks in 2025. While the Retrieval-Augmented Debating task at Touché (Section 3.1) and the Interactive Response task at TREC iKAT (Section 3.3) used simulated users in order to evaluate participants' systems, the Micro-Shared task at Sim4IA (Section 3.2) made the development of user simulators the objective of the participants and focused on the evaluation of the simulation systems itself. In the following sections, we describe these shared tasks with respect to how user simulation was used, what impact the simulation had, and what lessons were learned from employing user simulation in practice. Table 1 provides a comparison of the task setups and contrasts them with the widely known Cranfield paradigm.

3.1 Retrieval-Augmented Debating at Touché at CLEF

The Touché Lab at CLEF is concerned with computational approaches to argumentation. In particular, participants in the Retrieval-Augmented Debating (RAD) task of the sixth edition of Touché were asked to contribute conversational systems that allow users to engage in debates to improve their argumentation skills or to form or falsify an opinion on a topic of interest. The debate systems should respond to a user's initial claim, take the opposite stance, and try to persuade its user by supporting the system's stance with arguments or by refuting the user's stance with counterarguments. To retrieve and incorporate relevant real-world arguments, the debate systems had access to an indexed collection of arguments obtained from the ClaimRev [Skitalinskaya et al., 2021] dataset. More information on this task is provided in the overview paper [Kiesel et al., 2025] and on the task's web page.[1]

[1] https://touche.webis.de/clef25/touche25-web/retrieval-augmented-debating.html

Aspect	"Cranfield"	Touché	Sim4IA	iKAT
Collection	Any document collection	300 000 arguments from ClaimRev	160 sessions from CORE	116M passages from ClueWeb22-B
User model	Topics + relevance judgments (Qrels)	Retrieval-augmented (same collection) Llama 3.1 with custom prompts	Utterances + system responses; queries + SERPs + interacted with results	Personas as natural language statements + rubrics for topic progression, GPT-4.1 with custom prompts
Simulator Evaluation	-	-	Semantic similarity, SERP overlap, redundancy, rank-diversity score	-
System Evaluation	F-score, MAP, nDCG, ...	Grice's maxims of conversation	-	nDCG, Precision, Recall, MAP, manual assessment

Table 1. Comparative overview of setup aspects of the different shared tasks and a "typical Cranfield style evaluation."

User Simulation at Touché

To obtain debates for the evaluation of the participants' systems, we let simulated users interact with the system using a test dataset of claims. The simulated user was supposed to provide the first turn of the debate on the basis of the given claim and its description that clarifies the stance the user takes. The debate system and the simulated user took turns arguing against the stance of each other up to a total number of five turns. One debate was simulated for each claim and debate system, which yielded the pool of debates that were evaluated manually and automatically in order to obtain the final scores for the participants' systems.

The user simulation was build around Llama 3.1 [Dubey et al., 2024] and made use of the same argument collection as participants' systems. The simulated user employed a dense retrieval pipeline to find arguments in support of its own claims and for the attack of the debate system's stance. To this end, the simulated user used its own prior utterances as queries for supporting and the system's utterances as queries for attacking arguments, the relevance of which was regularized by the recency of the utterance in the conversation.

Impact of User Simulation on the Results of Touché

As the task for the user simulation system at Touché RAD was to discuss, the LLM-based implementation adapted its vocabulary to the style of a political debate. Since the user simulator provided the first utterance and the participant's systems were also LLM-based, most debate systems continued the conversation in a rather grandiloquent language. This phenomenon was especially present in debates concerning philosophical questions such as, for example, whether an objective morality exists in our society. This stylistic choice was often complemented by a verbose style and repetition on the part of both interlocutors. Although this domain adaptation can be considered thematic, it made the manual evaluation of the debates considerably more difficult, as the arguments put forward were not always comprehensible to the assessors.

Each debate was deliberately initiated by the simulated user. As a consequence, the "quality" of the ensuing debate depended on the clarity of the initial utterance. In cases where the user simulator was unable to clearly state its viewpoint, the participants' debate systems had difficulty identifying the opposing stance and providing arguments consistent with that stance. In general, simulated users and debate systems were often inconsistent in their respective stances; they presented alleged counterarguments that actually supported the opposing side's stance. Assessors had to decide whether participants' systems were coherent with prior turns of the conversation, even when a stance switch occurred that was caused by the simulated user.

Lessons Learned from User Simulation at Touché

To employ user simulation for the evaluation of domain-specific IR applications (such as debate systems), the degree of complexity of the language should be controlled to increase the effectiveness of a subsequent manual assessment. Ideally, the complexity of the language is adapted to the expertise of the human assessors in the respective domain. We infer from our observations that if a simulated user initiates a conversation in a particular style, a generative IR system would copy that style when formulating its responses. Therefore, it is sufficient to control the output language of the user simulator to influence the style of the whole conversation. This language adaptation for LLM-based user simulators can most likely be achieved through in-context learning or lightweight fine-tuning with low-rank adaptation.

LLMs, especially those with fewer parameters, sometimes fail to follow instructions satisfactorily. Sometimes specific aspects of the prompt are ignored (lost in the middle) or sometimes the overall task is misunderstood. Based on our experience at Touché RAD, the more complex the task that the user simulator has to perform, the more errors can occur during evaluation. If feasible, we suggest repeating the conversation-generation process several times to reduce the impact of such simulation errors and ensure a fair evaluation across all assessed IR systems. Preferably, the simulation is repeated with different simulator implementations, changing input prompts, base models, or entire architectures. While this procedure increases the workload for human assessors, it simultaneously improves the validity of the evaluation process.

As part of the evaluation process, we discovered that the lack of feedback that the user simulator provided hurts the perceived authenticity of a conversation. For example, if the user simulator detected arguments that were incompatible with the system's stance, it should have pointed this out by questioning the validity of the argument in question. In most cases, the user simulator did not provide feedback on the system responses, which gave the impression that the two interlocutors were not listening to each other. Therefore, we advocate for explicit assessment and feedback mechanisms that a user simulator should be equipped with to produce authentic conversations.

3.2 Micro-Shared Task at Sim4IA at SIGIR

The Simulation for Information Access workshop (Sim4IA) was held two times—at SIGIR 2024 [Schaer et al., 2024; Breuer et al., 2025] and at SIGIR 2025 [Schaer et al., 2025b,a]. In addition to the workshop program that included a keynote, tech, and lightning talks of participants, the central element of the second edition of the workshop was the so-called Micro-Shared task on user simulations. The shared task concept was grounded in the fundamental design principle of validating user simulations rather than measuring system effectiveness. The organizers envisioned users interacting with specific information access systems, such as a traditional search engine or a conversational system. Participants were challenged to design and implement user simulators capable of replicating real user interactions with these systems with high fidelity. The workshop incorporated a simplified implementation of this concept, a Micro-Shared task. More information on this task is provided on the task's web page.[2]

User Simulation at Sim4IA

The shared task at Sim4IA was crafted around a set of interaction log files from the academic search engine CORE[3] [Knoth et al., 2023] and an extended version of the LongEval-Sci test collection [Cancellieri et al., 2025]. From the log files, we extracted interaction sessions including initial queries, reformulations, SERPs, and user clicks. The task for the participants was to build a user simulator capable of predicting the following search query of an interactive search session (Task A). Additionally, we extended this setting such that the system output of the system was not a SERP but a precomputed RAG-style LLM-generated response to which the user simulator should predict the next utterance (Task B). For each session, the task was to predict ten diverse candidate queries or utterances, ordered by their estimated likelihood of being the next user query. The participants were encouraged to submit the simulator itself, in addition to the generated queries/utterances, to allow further experimentation.

In total, we received the results of 62 different simulators (33 unique approaches): 6 (semi-)manual runs, 21 LLM-based runs without further finetuning (mostly using Gemini 2.5 Flash and GPT 4.1 nano), 5 LLM-based runs with a specific finetuning (using GPT-2 and LLama3.2 3B), as well as one rule-based run. Out of these submissions 48 (32 unique) were addressing Task A and 14 Task B.

The runs were compared against the real-world next query/utterance in the log files using cosine similarity, the SERP overlap, redundancy, and a newly proposed MMR-inspired measure that takes into account the diversity of the proposed candidates that we called Rank-Diversity Score (RDS).

Impact of User Simulation on the Results of Sim4IA

In contrast to the other shared tasks the main focus of Sim4IA was to extend the understanding of what constitutes a good simulator and how a simulator's performance should be evaluated. In previous discussions of a shared task on user simulation, the validation of simulators, due to the lack of established measures, remained a central open problem. To tackle this it would require two key components: (1) benchmark datasets that directly link real user interaction logs to simulated outputs, and (2) robust measures to quantify the similarity between simulated and real user behavior. Such datasets are not widely available and without a common ground for

[2] https://sim4ia.org/sigir2025/#micro-shared-task

[3] https://core.ac.uk/

comparison, it would be hard or even impossible to assess a new simulator and to understand the strengths and weaknesses of different simulation approaches.

In the shared task at Sim4IA we therefore introduced a ready-to-use testbed in the form of the Sim4IA-Bench Suite [Kruff et al., 2025]. This public benchmark resource is specifically designed for the evaluation of user simulators. It consists of (1) the prepared session logs, including training and test sets, (2) all submission run files and corresponding lab notes from the teams participating in the Sim4IA shared task, (3) the benchmarking code for evaluating the predictions, (4) and a simulation toolkit, including a dockerized adaptation of the SimIIR 3 [Azzopardi et al., 2024] toolkit.

Lessons Learned from User Simulation at Sim4IA

One key element of the Sim4IA shared task was to evaluate how good and authentic simulators could reproduce real user behavior. The performance or success in a down-stream retrieval task was explicitly not addressed. Measuring this was not trivial as different perspectives on the simulators' outcomes produced different insights. While, for example, the similarity between the generated queries and the resulting SERPs from the rule-based approach was the highest, an in-depth analysis of this approach revealed that the generated query candidates were mostly all the same. In our new RDS measure, we therefore penalized this behavior to obtain a more nuanced view of the results. Overall, we could see that persona-based simulators that used LLMs fluctuated the most, producing the most diverse outcomes. Compared to the manual runs, most automatic runs produced a comparable similarity ($\bar{RD} \approx 0.7$), which we interpret as an overall sustainable vocabulary overlap of simulated and real queries/utterances.

Further experiments on the question of how to compare and validate simulators' outcomes still remain future work and mark a research gap that needs to be addressed in subsequent shared tasks.

3.3 Interactive Response Generation Task at iKAT at TREC

The goal of the Interactive Knowledge Assistant Track (iKAT) at TREC is to advance research in the field of personalized, conversational information access systems. The task for the participants of iKAT is to develop conversational systems that retrieve relevant passages and synthesize informative answers to a user's information need, taking into account the user's preferences and characteristics (persona). In the 2025 edition of iKAT, participants received a test dataset of conversations on 17 different topics between nine users with distinct personas and an imagined conversational system, which was created manually by the organizers. These user personas are encoded as a personal text knowledge base (PTKB), which are short sentences in natural language that describe the user (e.g., "I want to increase my protein intake"). Participants should generate the next system turn based on prior turns and the PTKB of the user by identifying relevant PTKB statements to the utterance, retrieving relevant passages, and generating a relevant response. More information on this task is provided on the task's web page.[4]

[4] https://trecikat.com

User Simulation at iKAT

In year three of TREC iKAT (2025), a novel interactive task was offered to the participants in which their systems had to respond to user utterances in real time. These user utterances were generated by a user simulator to produce conversations that were evaluated to assess participants' systems. To interact with the user simulator, participants should call an HTTP-API that was developed by the organizers. The API allowed participants to debug their system with a baseline user simulator and make submissions by interacting with the task-specific user simulators.

For a fair assessment across all systems, the organizers opted for a user simulation paradigm that generates a comparable conversation progression for each topic in the test dataset. To achieve this, conversations from the test dataset were broken down into sequences of simple underlying questions (rubrics) that were addressed by the user in the conversation. For example, on the topic of "how to make good coffee" one of the rubrics was "what is the impact of roasting on coffee taste?". The user simulator was informed about the sequence of rubrics for a topic, which was used as guidance to condition the utterance generation process. Further information for the control of the user simulator was the user's PTKB and the sequence of the previous turns.

Impact of User Simulation on the Results of iKAT

The essential requirement for the design of the user simulator at iKAT was reproducibility. This was achieved by enforcing comparable conversation progressions for each information need independent of what system the user simulator was talking to. We observed semantically congruent conversations with only a few topical deviations produced by the simulator. Some of these deviations are desirable, such as providing feedback to clarifying questions or recommendations. Others, however, are less desirable, for example, if the user simulator has ignored important instructions or meta information. Since the simulator had a mechanism for detecting these deviations, another turn was generated on the topic, so that the negative effects of the deviations were marginal.

The interactivity of the task due to the user simulation provoked many participants to build systems that frequently ask for clarification (e.g., "Are you interested in group activities or solo pursuits?"), recommend follow-up topics (e.g., "Do you want information on free or discounted swim sessions for families?"), or use information scaffolding (e.g., "Would you like more jazz artist or playlist recommendations?") to avoid overloading the user with unnecessary details. This mixed initiative paradigm was considered in the evaluation, and systems that implemented this had an advantage in the evaluation. Although the user simulator had strong control mechanisms to stay on topic, questions from the system to the user were sufficiently answered in most cases.

Getting an LLM to behave like a user can sometimes fail, resulting in unnatural interaction patterns. This phenomenon was particularly prevalent in expressions of acknowledgments and farewells where, for example, the simulated user offered to answer further questions of the system or thanked the system for its curiosity. Since the user simulator and the systems were LLM based, such utterances can cause the system to become confused about its role in the conversation as well. Luckily, this phenomenon occurred mostly at the end of the conversation, but still has an impact on the perceived "naturalness" or coherence of a conversation.

Lessons Learned from User Simulation at iKAT

With the objective of reproducible evaluations in mind, we learned that size does matter with respect to the LLM that is used for the utterance generation of the simulator. With the amount of control that is necessary to guide the conversation, smaller LLMs with fewer parameters are not yet up to the task, since important details in the prompts can be overlooked. This is especially true for long contexts like we had at iKAT. The prompt had to accommodate for general instructions, instructions specifically for the next utterance (e.g., topical guidance with rubrics), PTKB statements of the modeled user, and the full dialog history (up to 15 turns). In pilot experiments, we experimented with recent open-weight models with up to seven billion parameters, with and without quantization of parameters, and found that these models tend to deviate too often from the topic, confuse their roles, do not follow instructions, or do not infuse their utterances with personal information. Therefore, the task was conducted with a simulator based on GPT-4.1 (with presumably hundreds of billions of parameters). Interesting directions to make smaller models feasible for reproducible user simulation can include techniques such as prompt compression, more efficient encoding of user models, or stronger control mechanisms.

At iKAT, user models were mostly defined as a set of personal interests, preferences, and demographics. As a result, the language style or search behavior was homogeneous across all nine personas that were opted for to imitate. This fact is most likely not representative of a comparable population of real humans. To vary these attributes, explicit interventions and control mechanisms are needed as off-the-shelf LLMs only reflect persona properties on a surface level. However, we think that modeling different search strategies for the evaluation of information retrieval system is important to get a more sophisticated idea of a system's performance.

4 An Evaluation Paradigm for the Future?

Our experience of organizing shared tasks featuring user simulation, which we shared in Section 3, inspired us to think about the future of evaluating conversational retrieval systems. We believe that user simulation will eventually become a crucial step in the development process of information retrieval systems before real users become involved.

We understand user simulation as the next evolutionary step of the Cranfield paradigm. In discussions, we identified many parallels between these two paradigms. In a Cranfield experiment, a retrieval system is evaluated using a metric calculated on the basis of a set of search queries and the corresponding relevant documents (known as "qrels"), whereas in a user simulation, a retrieval system is evaluated by calculating metrics based on simulated search queries (or utterances) from simulated users. User simulation shares many of the desirable properties of a Cranfield-based evaluation, such as the ability to be conducted fully automatically, to be repeatable, and to be executed offline. Furthermore, the user simulation paradigm requires near identical "ingredients" such as a test collection with predefined information needs, a set of metrics, and at least one user model. A user model of the Cranfield paradigm is defined by a set of qrels and metrics, where the qrels define what a user finds relevant while the metrics approximate a user's search behavior. In contrast, a user model for user simulation-based evaluation can take many forms, such as a set of character traits, user knowledge, or past conversations. From a theoretical standpoint, there are no limitations to user modeling in the user simulation paradigm.

User simulation aims to relax some of the overly simplified assumptions underlying the Cranfield experiment. One of these assumptions is that the set of relevant documents in a test collection for a specific information need is representative of all documents that the entire user population considers relevant [Voorhees, 2001]. However, relevance has an inherent subjective component [Cosijn and Ingwersen, 2000], which renders a single user model representing the entire user population an unrealistic simplification. The ability to create more complex user profiles in simulation-based evaluations, for example by defining multiple user profiles that differ in their assessment of the relevance of aspects of the same topic, is a step towards a more realistic evaluation. However, in the future, user simulation research has to address how many and which user models must be included to obtain a population that is representative of the population of real users.

However, a user simulator is not required to be perfect in order to be useful. A simulated user has the advantage that it can serve as a crash test dummy before a retrieval system is released to real users. Simulated users can invoke a system's responses to extreme user behavior, similar to performing unit tests. With the analogy to unit testing in mind, in the future it may be possible to develop retrieval systems in a simulator-driven fashion, in which simulators are created first and then executed and evaluated whenever major changes are made to the system.

Despite this promise, to fully realize simulator-driven development, the community must address two glaring open issues: the validation of user simulators and the lack of standardized resources and methodology. Without addressing these, it is difficult to determine if a simulator is realistic or if results are comparable across different studies. We propose addressing these issues through both long-term community initiatives and immediate best practices.

A TREC track for Standardizing Methodology and Resources

To address these challenges in the long term, a new User Simulation track has been proposed and accepted to run at TREC 2026.[5] This track explicitly addresses both the validation of simulators and the standardization of resources. Just as `trec_eval` became the standard tool for the Cranfield paradigm, the community requires a standardized `sim_eval` tool to ensure rigor in simulation-based evaluation. This track will serve as the venue to develop these standardized metrics, protocols, and validation frameworks.

Immediate Best-Practice Recommendations

While we await the emergence of a standardized methodology and tools, it is critical to maintain high documentation standards now to enable reproducibility. To make these simulators reusable in the immediate term, public information would be required, for example, about implementation details, training data, or communication protocols. In the domain of interactive information retrieval, an exemplary set of principles to enable reusable research resources was established [Gäde et al., 2021], from which we want to draw inspiration to outline comparable principles for reusable simulators. In particular, the documentation principle will be the most important feature of reusable user simulators.

We believe that publishing the following information about a user simulator (e.g., in the form of a model card or README file) is a prerequisite for establishing reusable simulators.

[5] https://trec.usersim.ai

- **Interaction Modes.** The interaction capabilities of simulators and systems should be explicitly described. Possible interaction modes include (but are not limited to) queries and utterances, mouse movements or clicks, gaze, relevance feedback, or voice commands.
- **Interaction Protocol.** With a protocol, we define how a simulated session between systems and users consisting of different interactions is structured. Questions about initiative, start and end of a session, or common interaction patterns (e.g., query→browse→click) should be addressed to ensure compatibility between simulators and systems.
- **Session Participants.** Sessions typically consist of interactions between one simulated user and one system but could be extended to multiple systems or users in collaborative settings.
- **User Model.** As user simulation aims to mimic the behavior of a specific user or population, the modeled behavior should be described. Furthermore, the initial state of the simulated user, including user knowledge, preferences, or other persona properties should be made available.
- **Training Data.** When a simulator is trained on, for example, user-specific behavioral data or interaction logs, the data should be described and preferably made available to the public.
- **User-System Interface.** As a user simulator can consist of complex interactions, the interface through which the simulated user can interact with a system needs to be well documented for easy integration. These APIs can be designed based on the offered interaction modes and can range from HTTP-APIs to screen monitoring.
- **Benchmarks.** For the release of a new dataset, typically state-of-the-art system performances are provided to give baseline readings of the systems on the given dataset. For user simulation, we would expect similar disclosures. The benchmark details should consist of the assessed task, (at least one) metric, a dataset of information needs, and the performances of the state-of-the-art systems when interacting with the given simulator.
- **Benchmark Metadata.** Analogous to `ir_metadata` [Breuer et al., 2022b], metadata readings and parameters should be provided in the documentation of the user simulators. This metadata may contain hardware information, experiment parameters, or profiling and energy consumption measurements.

5 Conclusion

In three shared tasks in 2025, we gained hands-on experience and insights into the practical application of user simulations for evaluating conversational retrieval systems. Across Touché RAD and TREC iKAT, we learned about the importance of controllability and guidance of user simulators and the lack of these capabilities of current simulators based on LLMs with few parameters. The Sim4IA micro-shared task highlighted the difficulty of assessing the quality and authenticity of user simulators which emphasizes the lack of standardized resources and methodology. Generally, we understand user simulation to become an extension to the Cranfield paradigm for the evaluation of interactive retrieval systems. However, to get there, the establishment of standardized resources, methodologies, and metrics is crucial. With this paper, we proposed a first attempt on a list of requirements for user simulators to become standard tools in the tool box of IR evaluation, but more research in this direction is needed until commonly accepted test suites will become available. To unite forces, we will organize the User Simulation track at TREC 2026 which has the goal to establish standard evaluation methodologies for user simulation and their application for the evaluation of conversational information retrieval.

A Authors and Affiliations

Touché RAD Organizers

- Marcel Gohsen; Bauhaus-Universität Weimar; Weimar, Germany; marcel.gohsen@uni-weimar.de
- Nailia Mirzakhmedova; Bauhaus-Universität Weimar; Weimar, Germany; nailia.mirzakhmedova@uni-weimar.de
- Harriscen Scells; University of Tübingen; Tübingen, Germany; harriscen.scells@uni-tuebingen.de
- Mohammad Aliannejadi; University of Amsterdam; Amsterdam, The Netherlands; m.aliannejadi@uva.nl
- Maik Fröbe; Friedrich-Schiller-Universität Jena; Jena, Germany; maik.froebe@uni-jena.de
- Johannes Kiesel; GESIS - Leibniz Institute for the Social Sciences; Cologne, Germany; johannes.kiesel@gesis.org
- Benno Stein; Bauhaus-Universität Weimar; Weimar, Germany; benno.stein@uni-weimar.de

Sim4IA 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
- Timo Breuer; TH Köln; Cologne, Germany; timo.breuer@th-koeln.de
- Andreas Kruff; TH Köln; Cologne, Germany; andreas.kruff@th-koeln.de

TREC iKAT Organizers

- Mohammad Aliannejadi; University of Amsterdam; Amsterdam, The Netherlands; m.aliannejadi@uva.nl
- Simon Lupart; University of Amsterdam; Amsterdam, The Netherlands; s.c.lupart@uva.nl
- Marcel Gohsen; Bauhaus-Universität Weimar; Weimar, Germany; marcel.gohsen@uni-weimar.de
- Zahra Abbasiantaeb; University of Amsterdam; Amsterdam, The Netherlands; z.abbasiantaeb@uva.nl

- Nailia Mirzakhmedova; Bauhaus-Universität Weimar; Weimar, Germany; nailia.mirzakhmedova@uni-weimar.de
- Johannes Kiesel; GESIS - Leibniz Institute for the Social Sciences; Cologne, Germany; johannes.kiesel@gesis.org
- Jeffrey Dalton; University of Edinburgh; Edinburgh, Scotland, UK; jeff.dalton@ed.ac.uk

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, pages 1160–1163, 2023.

Leif Azzopardi. Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 556–563. ACM, 2009. doi: 10.1145/1571941.1572037.

Leif Azzopardi. The Economics in Interactive Information Retrieval. In Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft, editors, *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 15–24. ACM, 2011. doi: 10.1145/2009916.2009923.

Leif Azzopardi and Maarten de Rijke. Automatic Construction of Known-Item Finding Test Beds. In Efthimis N. Efthimiadis, Susan T. Dumais, David Hawking, and Kalervo Järvelin, editors, *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 603–604. ACM, 2006. doi: 10.1145/1148170.1148276.

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 Tetsuya Sakai, Emi Ishita, Hiroaki Ohshima, Faegheh Hasibi, Jiaxin Mao, and Joemon M. Jose, editors, *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2024, Tokyo, Japan, December 9-12, 2024*, pages 197–202. ACM, 2024. doi: 10.1145/3673791.3698427.

Krisztian Balog and ChengXiang Zhai. User Simulation for Evaluating Information Access Systems, 2024.

Krisztian Balog, David Maxwell, Paul Thomas, and Shuo Zhang. Sim4IR: The SIGIR 2021 Workshop on Simulation for Information Retrieval Evaluation. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2697–2698. ACM, 2021. doi: 10.1145/3404835.3462821.

Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–114. ACM, 2012. doi: 10.1145/2348283.2348301.

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

Nolwenn Bernard, Sharath Chandra Etagi Suresh, Krisztian Balog, and ChengXiang Zhai. SimLab: A Platform for Simulation-based Evaluation of Conversational Information Access Systems, 2025.

Charles R. Blunt. An Information Retrieval System Model. Technical Report HRB-352.14-R-1, January 1965.

Timo Breuer, Norbert Fuhr, and Philipp Schaer. Validating Simulations of User Query Variants. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, pages 80–94. Springer, 2022a. doi: 10.1007/978-3-030-99736-6_6.

Timo Breuer, Jüri Keller, and Philipp Schaer. Ir_metadata: An Extensible Metadata Schema for IR Experiments. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3078–3089. ACM, 2022b. doi: 10.1145/3477495.3531738.

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 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, March 2025. ISSN 0163-5840. doi: 10.1145/3722449.3722460. URL https://doi.org/10.1145/3722449.3722460.

Matteo Cancellieri, Alaa El-Ebshihy, Tobias Fink, Maik Fröbe, Petra Galuscáková, Gabriela González Sáez, Lorraine Goeuriot, David Iommi, Jüri Keller, Petr Knoth, Philippe Mulhem, Florina Piroi, David Pride, and Philipp Schaer. LongEval at CLEF 2025: Longitudinal Evaluation of IR Systems on Web and Scientific Data. In Jorge Carrillo-de-Albornoz, Alba García Seco de Herrera, Julio Gonzalo, Laura Plaza, 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, Madrid, Spain, September 9-12, 2025, Proceedings*, volume 16089 of *Lecture Notes in Computer Science*, pages 363–387. Springer, 2025. doi: 10.1007/978-3-032-04354-2_20.

Michael D. Cooper. A Simulation Model of an Information Retrieval System. *Inf. Storage Retr.*, 9(1):13–32, 1973. doi: 10.1016/0020-0271(73)90004-1.

Erica Cosijn and Peter Ingwersen. Dimensions of Relevance. 36(4):533–550, 2000. doi: 10.1016/S0306-4573(99)00072-2.

Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. Principles and Guidelines for the Use of LLM Judges. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, pages 218–229, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 979-8-4007-1861-8. doi: 10.1145/3731120.3744588.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and et al. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783.

Björn Engelmann, Timo Breuer, Jana Isabelle Friese, Philipp Schaer, and Norbert Fuhr. Context-Driven Interactive Query Simulations Based on Generative Large Language Models. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II*, pages 173–188, Berlin, Heidelberg, March 2024. Springer-Verlag. ISBN 978-3-031-56059-0. doi: 10.1007/978-3-031-56060-6_12.

Maria Gäde, Marijn Koolen, Mark M. Hall, Toine Bogers, and Vivien Petras. A Manifesto on Resource Re-Use in Interactive Information Retrieval. In Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith, editors, *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021*, pages 141–149. ACM, 2021. doi: 10.1145/3406522.3446056.

Lukas Gienapp, Tim Hagen, Maik Fröbe, Matthias Hagen, Benno Stein, Martin Potthast, and Harrisen Scells. The Viability of Crowdsourcing for RAG Evaluation. In Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne, editors, *48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025)*, pages 159–169, New York, July 2025. ACM. ISBN 979-8-4007-1592-1/2025/07. doi: 10.1145/3726302.3730093.

Michael D. Gordon. Evaluating the Effectiveness of Information Retrieval Systems using Simulated Queries. *Journal of the American Society for Information Science*, 41(5):313–323, 1990. doi: 10.1002/(SICI)1097-4571(199007)41:5\⟨313::AID-ASI1\⟩3.0.CO;2-G.

José-Marie Griffiths. *The Computer Simulation of Information Retrieval Systems*. PhD thesis, University of London, 1978.

Donna Harman. Relevance Feedback Revisited. In Nicholas J. Belkin, Peter Ingwersen, and Annelise Mark Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992*, pages 1–10. ACM, 1992. doi: 10.1145/133160.133167.

Karen Sparck Jones. Search Term Relevance Weighting given Little Relevance Information. *J. Documentation*, 35(1):30–48, 1979. doi: 10.1108/EB026672.

Chris Jordan, Carolyn R. Watters, and Qigang Gao. Using Controlled Query Generation to Evaluate Blind Relevance Feedback Algorithms. In Gary Marchionini, Michael L. Nelson, and Catherine C. Marshall, editors, *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2006, Chapel Hill, NC, USA, June 11-15, 2006, Proceedings*, pages 286–295. ACM, 2006. doi: 10.1145/1141753.1141818.

Heikki Keskustalo, Kalervo Järvelin, and Ari Pirkola. Evaluating the Effectiveness of Relevance Feedback Based on a User Simulation Model: Effects of a User Scenario on Cumulated Gain Value. 11(3):209–228, 2008. doi: 10.1007/S10791-007-9043-7.

Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, and Benno Stein. Simulating Follow-up Questions in Conversational Search. 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 IR Research (ECIR 2024)*, volume 14609 of *Lecture Notes in Computer Science*, pages 382–398, Berlin Heidelberg New York, March 2024a. Springer. doi: 10.1007/978-3-031-56060-6_25.

Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, and Benno Stein. Who Will Evaluate the Evaluators? Exploring the Gen-IR User Simulation Space. In Lorraine Goeuriot, Philippe Mulhem, Georges Quénot, Didier Schwab, Giorgio Maria Di Nunzio, Laure Soulier, Petra Galuscakova, Alba García Seco de Herrera, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, volume 14958 of *Lecture Notes in Computer Science*, pages 166–171, Berlin Heidelberg New York, September 2024b. Springer. doi: 10.1007/978-3-031-71736-9_11.

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, Berlin Heidelberg New York, September 2025. Springer.

Petr Knoth, Drahomira Herrmannova, Matteo Cancellieri, Lucas Anastasiou, Nancy Pontika, Samuel Pearce, Bikash Gyawali, and David Pride. CORE: A Global Aggregation Service for

Open Access Papers. *Scientific Data*, 10(1):366, June 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02208-w.

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, 2025.

Anton Leuski. Relevance and Reinforcement in Interactive Browsing. In *Proceedings of the 2000 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 6-11, 2000*, pages 119–126. ACM, 2000. doi: 10.1145/354756.354809.

Paul Owoicho, Ivan Sekulic, Mohammad Aliannejadi, Jeffrey Dalton, and Fabio Crestani. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*, SIGIR '23, pages 632–642, New York, NY, USA, 2023. ACM. ISBN 978-1-4503-9408-6. doi: 10.1145/3539618.3591683.

Gustavo Penha and Claudia Hauff. Challenges in the Evaluation of Conversational Search Systems. In Giuseppe Di Fabbrizio, Surya Kallumadi, Utkarsh Porwal, and Thrivikrama Taula, editors, *Proceedings of the KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption Co-Located with the 26TH ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2020), Virtual Workshop, August 24, 2020*, volume 2666 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

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, 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, , Andreas Kruff, Mohammad Aliannejadi, Christine Bauer, Nolwenn Bernard, Nicola Ferro, Marcel Gohsen, Nurul Lubis, and Saber Zerhoudi. Report on the Second Workshop on Simulations for Information Access (Sim4IA 2025) at SIGIR 2025. *SIGIR Forum*, 59(2):1–15, December 2025a.

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, 2025b. doi: 10.1145/3726302.3730363.

Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. Evaluating Mixed-initiative Conversational Search Systems via User Simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pages 888–896, New

York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9132-0. doi: 10.1145/3488560.3498440.

Ivan Sekulic, Mohammad Aliannejadi, and Fabio Crestani. Analysing Utterances in LLM-Based User Simulation for Conversational Search. *ACM Transactions on Intelligent Systems and Technology*, 15(3):62:1–62:22, 2024. doi: 10.1145/3650041.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1718–1729. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EACL-MAIN.147.

Jean Tague, Michael J. Nelson, and Harry Wu. Problems in the Simulation of Bibliographic Retrieval Systems. In Robert N. Oddy, Stephen E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, *Information Retrieval Research, Proc. Joint ACM/BCS Symposium in Information Storage and Retrieval, Cambridge, UK, June 1980*, pages 236–255. Butterworths, 1980.

Ellen M. Voorhees. The Philosophy of Information Retrieval Evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2001. doi: 10.1007/3-540-45691-0_34.

Zhenduo Wang, Zhichao Xu, Vivek Srikumar, and Qingyao Ai. An in-depth investigation of user response simulation for conversational search. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pages 1407–1418, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645447. URL https://doi.org/10.1145/3589334.3645447.

Ryen W. White, Joemon M. Jose, C. J. van Rijsbergen, and Ian Ruthven. A Simulated Study of Implicit Feedback Models. In Sharon McDonald and John Tait, editors, *Advances in Information Retrieval, 26th European Conference on IR Research, ECIR 2004, Sunderland, UK, April 5-7, 2004, Proceedings*, volume 2997 of *Lecture Notes in Computer Science*, pages 311–326. Springer, 2004. doi: 10.1007/978-3-540-24752-4_23.

Ryen W. White, Ian Ruthven, Joemon M. Jose, and C. J. van Rijsbergen. Evaluating Implicit Feedback Models Using Searcher Simulations. *ACM Trans. Inf. Syst.*, 23(3):325–361, 2005. doi: 10.1145/1080343.1080347.