SimIIR 3

A Framework for the Simulation of Interactive and Conversational Information Retrieval

Azzopardi, Leif; Breuer, Timo; Engelmann, Björn; Kreutz, Christin; MacAvaney, Sean; Maxwell, David; Parry, Andrew; Roegiest, Adam; Wang, Xi; Zerhoudi, Saber

DOI
10.1145/3673791.3698427

Publication date
2024

Document Version
Final published version

Published in
SIGIR-AP 2024 - Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region

Citation (APA)
Azzopardi, L., Breuer, T., Engelmann, B., Kreutz, C., MacAvaney, S., Maxwell, D., Parry, A., Roegiest, A., Wang, X., & Zerhoudi, S. (2024). SimIIR 3: A Framework for the Simulation of Interactive and Conversational Information Retrieval. In *SIGIR-AP 2024 - Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (pp. 197-202). (SIGIR-AP 2024 - Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region). ACM. https://doi.org/10.1145/3673791.3698427

SimIIR 3: A Framework for the Simulation of Interactive and Conversational Information Retrieval

Leif Azzopardi
University of Strathclyde
Glasgow, United Kingdom
leifos@acm.org

Timo Breuer
TH Köln - University of Applied Sciences
Cologne, Germany
timobreuer@acm.org

Björn Engelmann
TH Köln - University of Applied Sciences
Cologne, Germany
bjoern.engelmann@th-koeln.de

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

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

David Maxwell
Delft University of Technology
Delft, Netherlands
maxwelld90@acm.org

Andrew Parry
University of Glasgow
Glasgow, United Kingdom
a.parry.1@research.gla.ac.uk

Adam Roegiest
Triangle Lab
Toronto, Canada
adam@roegiest.com

Xi Wang
University of Sheffield
Sheffield, United Kingdom
xi.wang@sheffield.ac.uk

Saber Zerhoudi
University of Passau
Passau, Germany
saber.zerhoudi@uni-passau.de

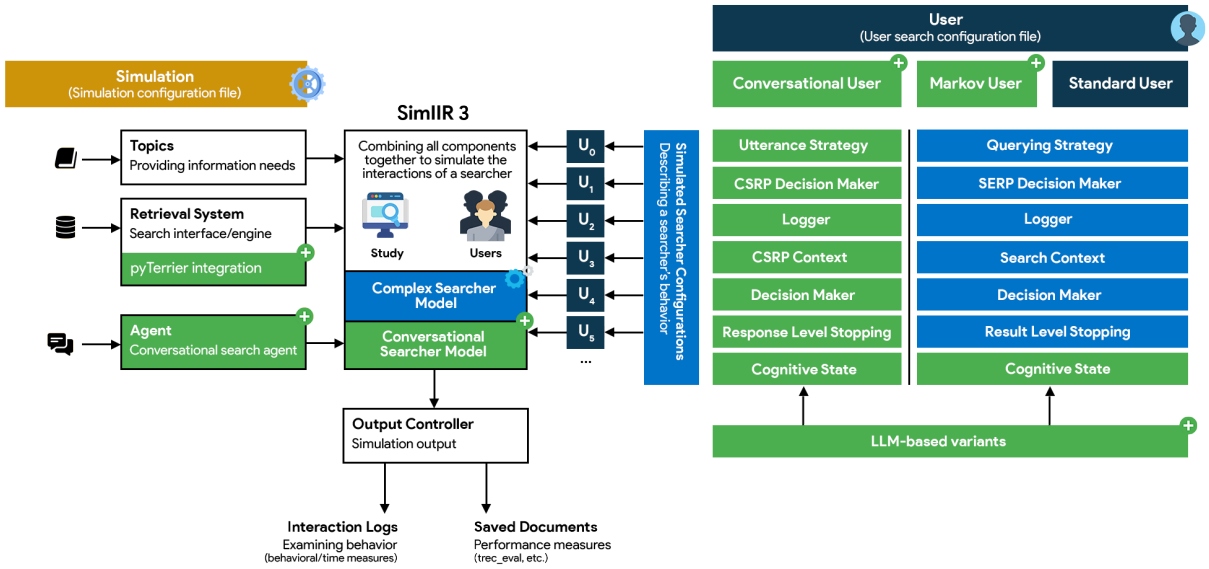

Figure 1: Updated SimIIR 3.0 Framework – the green boxes denote the new components added in order to support conversational search simulations, simulated users with cognitive states, Markovian users, search systems powered by pyTerrier, and users powered by Large Language Models.

SIGIR-AP '24, December 9–12, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-0724-7/24/12
https://doi.org/10.1145/3673791.3698427

Abstract

Evaluating the interactions between users and systems presents many challenges. Simulation offers a reliable, re-usable, and repeatable methodology to explore how different users, user behaviours and/or retrieval systems impact performance. With Large Language Models and Generative AI now widely available and accessible, new affordances are possible. These allow researchers to create more "realistic" simulated users that can generate queries and judge items like humans, and to develop new retrieval systems where responses and interactions are conversational and based on retrieval augmented generation. This resource paper presents a community-led initiative to update the Simulation of Interactive Information Retrieval (SimIIR) Framework to enable the simulation of conversational search using LLMs. The largest update provides a conversational search workflow which involves a number of new possible interactions with a search system or agent – enabling a host of new development and evaluation opportunities. Other developments include the Markovian Users, Cognitive States, LLM-based components for assessing snippets/documents/responses, generating queries, deciding on when to stop/continue, and PyTerrier integration. This paper aims to mark the release of SimIIR 3.0 and invites the community to build, extend, and use the resource.

CCS Concepts

• **Information systems** → **Retrieval effectiveness**; **Users and interactive retrieval**; **Search interfaces**; *Task models*; *Retrieval tasks and goals*; • **Human-centered computing** → *User models*; **HCI theory, concepts and models**; *Graphical user interfaces*; **HCI design and evaluation methods**.

Keywords

Interactive IR, Conversational IR, Simulation, Interaction, Open Source Framework

ACM Reference Format:

Leif Azzopardi, Timo Breuer, Björn Engelmann, Christin Kreutz, Sean MacAvaney, David Maxwell, Andrew Parry, Adam Roegiest, Xi Wang, and Saber Zerhoudi. 2024. 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 '24), December 9–12, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3673791.3698427

1 Introduction

Modelling how users interact with search engines (and now search agents) is key in assessing user experience, predicting user behaviours, evaluating system performance, and optimising user interfaces [3, 25, 27, 36]. Standard static test collections that follow the Cranfield / TREC paradigm [16, 40], however, fail to account for the interactive and iterative nature of search – where performance, experience and behaviour are influenced by the almost infinite variety of ways in which people can interact with search systems [12]. While large-scale A/B testing and controlled user studies offer more comprehensive evaluations, they also come with their own set of limitations: (1) A/B tests require a fully operational working system with a substantial user base to yield meaningful results [28] and (2) controlled lab-based user studies are time-consuming, costly, difficult to reproduce, and restricted to very specific tasks/settings [9, 24, 27]. This is where simulation emerges as a powerful tool to repeatably and reliably conduct controlled experimentation to examine and explore the behaviour and performance of different user querying styles, search strategies, goals, costs, and so on given different search systems, interfaces and agents [4, 5, 10, 36]. As such, simulation provides a flexible and cost-effective means to analyse search behaviour beyond the constraints of traditional methodologies. By simulating user interactions, researchers and developers can conduct virtual A/B tests on various back-end configurations and user interface designs. This approach allows for the exploration of different user types and information needs, offering insights that would be challenging to obtain through other methods.

Simulation bridges the gap between controlled laboratory experiments and real-world user interactions, providing a scalable and reproducible way to evaluate complex retrieval systems and interactive search scenarios. As search technologies continue to evolve, simulation will play an increasingly vital role in shaping the future evaluations of information retrieval agents and systems, ensuring they meet the diverse and dynamic needs of users. Given the potential for simulation, in 2016, the SimIIR framework was released to enable researchers to conduct simulated interactive information retrieval experiments. Since 2016, several variants of the SimIIR framework have been introduced [15, 22, 23, 29, 30, 43]. To consolidate these efforts, this paper brings together the community's efforts into a single, unified framework with the goal of providing a base for simulations that use LLMs and generative models.

2 SimIIR Background

The framework for the Simulation of Interactive Information Retrieval, called SimIIR, was developed by Maxwell and Azzopardi [35] as an initial attempt to provide the basis for a simulation platform. This framework provided a mechanism to create simulated users – which are defined by a querying strategy, a document and snippet assessment strategy, a stopping strategy, and a goal/cost model. Then for a given search system (defined by a search interface, search engine, and test collection), and a given task, the SimIIR framework would simulate how the simulated user would interact with the system. The user model provided by the framework followed the Complex Search Model (CSM) [33, 37]. The CSM essentially defines the actions and decisions that a user will perform and consider during their search session. Given a topic, the process is as follows: (1) the user generates a query (or list of queries given the topic and any past interaction), (2) the user then selects a query to issue to the search system, (3) the system returns the search engine result page (SERP), and the user views the SERP, (4) if the SERP appears useful, they then examine a result snippet, (5) if the result snippet looks attractive, they examine the document, else (6) they either examine another snippet or stop examining snippets and considering querying again, (7) if they examine a document, they judge it as relevant or not, and return to the SERP to consider further snippets or consider querying again. If they run out of queries, meet their goal, or run out of time, they stop.

As an extension [34], the framework also models a user's cognitive state in the User State Model (USM). It contains users' background knowledge, information needs, the history of interactions, as well as a series of models for querying, the attractiveness of snippets, and relevance decisions. Additionally, CSM within SimIIR has been extended to support the notion of information scent through a SERP-level stopping decision point [36].

The framework has been subsequently used to conduct a variety of experiments by various researchers. Some examples are noted below. Labhishetty and Zhai [29, 30] use SimIIR to quantify the reliability of testers as a way to evaluate user simulators.

Câmara et al. [15] focused on learning-oriented search tasks. They modified SimIIR to support a searcher model based on CSM which considers subtopics of users' information needs. Four parameters describe user behaviours: the speed of which a user integrates new terms into their vocabulary, the willingness of users to explore a subtopic, the willingness of users to click result snippets, and the strategy of switching subtopics.

Zerhoudi et al. [43] present SimIIR 2.0 as an extension of SimIIR by introducing components enabling the training of Markov models from real users' logged sessions. Similar to Câmara et al. [15] they also extend the CSM such that new seen terms are integrated into the vocabulary used for query generation. Additionally, they enable grouping of user characteristics.

Engelmann et al. [23] use SimIIR 2.0 to explore query reformulation strategies in the context of web table retrieval. They adapt CSM and include a Doc2Query-based query generation. In more recent work [22], they additionally implement LLM-based query generation and integrate users' background information in this process.

Given these recent developments, various branches of the SimIIR framework have been created, leading to a fractured code base and development path. At a recent community event, there was impetus to bring together the different branches and to update the framework. Moreover, with recent advances in Large Language Models affording the creation of conversational search systems and agents, there was great interest in implementing conversational search models to enable the simulation of conversational search. This resource paper outlines the combined, community-based effort to coordinate and update the framework so that LLM-based and conversational search simulations and experiments can be conducted, shared and reproduced.

3 SimIIR 3.0 Framework

Figure 1 presents an overview of the SimIIR 3.0 framework. The components shown in white and blue represent the existing infrastructure developed in [35, 43]. The green represents the new additions to the framework, which include the following:

Large Language Model Support: In order to facilitate easier experimentation with different ways to incorporate LLMs into the CSM, a generic wrapper around the LangChain library was implemented that takes care of interacting with specific LLMs (e.g., OpenAI models, Google models, local models – such as Llama-3 and Mistral). In doing so, the wrapper manages issues around retries and output formatting to avoid boilerplate duplication in specific use cases. Each of the basic CSM "decision" components in Figure 1 (called Decision Maker (DM)) have had simple implementations that can serve as the basis for more complex interactions. Each component expects a prompt template and information about which LLM is to be used and how it should be used (i.e., temperature). The individual components expose different data to the template as follows:

- **Querying Strategy:** This component exposes topic title and topic description to the prompt template. The selected LLM uses the information to formulate a ranked list of queries that the simulated user can then issue.
- **SERP Decision Maker:** This component exposes topic title, topic description, previously viewed relevant snippets, and the current SERP's snippets to the prompt template. Using this information, the LLM is instructed to decide whether the SERP is attractive enough to be examined, or not.
- **Result Decision Maker:** This component exposes the topic title and topic description to the prompt template. For snippets, it exposes the document title and associated snippet. For a document, it exposes the document title and its entire contents. The LLM-based Result DM decides whether the simulated user examines the snippet or considers the document relevant, or not.
- **Result Level Stopping Decision Maker:** This component exposes the topic title, topic description, previously viewed snippets, and the current snippet being decided upon to the prompt template. The LLM-based Result Level Stopping DM then decides whether the user should continue examining results, or not.

Taken together, it is possible to fully instantiate a simulated search user that makes all decisions using Search User-based LLMs powered components.

Conversational Search User: The conversational search user extends the user class, which was augmented to handle a number of new actions:

- **Utterance Strategy**: Given the topic, and user state, the utterance strategy decides what the user will utter to the Conversational Search Interface (CSI); this is akin to the Query Strategy.
- **CSRP Decision Maker**: Given the response from the CSI, this component decides whether the user finds the Conversational Search Result Page (CSRP), attractive enough to continue to inspect it; this is akin to the SERP Decision Maker.
- **Response Decision Maker:** This component decides whether the response shown in the CSI is relevant, or not.
- **Response Level Stopping Decision Maker:** This component decides whether the user should stop given the response, or not.

Conversational Searcher User Model: An update to the architecture supports a variety of simulations based on different user models. Currently, the framework supports two distinct user models: (a) the Complex Searcher Model, where a user interacts with a search interface (via querying, examining snippets, assessing documents, etc.) and (b) the new Conversational Searcher Model, where a Conversational Search User interacts with a CSI (via uttering requests, examining and assessing responses, etc.).

Conversational Search Interface: Under the standard search interface, users interact with a SERP, which is typically configured to be 10 results per page, while for the CSI, by default, we assume that the CSRP is a response (with links to documents).

Conversational Search Agent: To provide responses to user utterances given the CSI, a conversational agent/system/engine is instantiated to handle such requests (note that the user does not directly interact with the back-end – only via the CSI – so existing conversational agents/system APIs are wrapped to provide the CSI).

Taken together it is possible to instantiate a Conversational Search User, that interacts with a CSI given the basic Conversational Searcher Model. As we develop the framework, we will include more actions, and include more complex user models – for example if the CSI returns both a response and a set of links, or a response and set of results, then the user could examine both results and responses, before making another utterance.

Markovian Searcher User Model: To perform simulations based on a Markov Model as proposed in [43], we created a Markovian Search User – where the probability of performing the different actions within the CSM could be provided to seed the simulation. The SimIIR framework has been extended to incorporate four main variants of Markovian Search Users: Basic, Contextual, Query-based, and Conditional. The Basic model represents user interactions as probabilistic transitions between states, with each state corresponding to a user action. The Contextual model improves upon this by categorising users based on their search behaviour and context [42]. The Query-Based model analyses query changes to represent user search behaviour [41], while the Conditional model incorporates dependencies between user actions and preceding queries [7]. These models are implemented as modular components in SimIIR, configurable through the simulation configuration file, allowing for easy comparison and evaluation of different user behaviours in search scenarios.

PyTerrier Search Engine Integration: To allow the use of broader search systems, we provide an interface to Terrier search components [31]. Using either a reference to a PyTerrier index [32] or a dataset from which an index can be retrieved from the Terrier data repository, the interface provides both retrieval and access to meta-data such as document texts and titles. Additionally, we provide a dense retrieval interface which accepts a PyTerrier flex index structure[1] and any standard bi-encoder model. Finally, the Terrier engine accepts an optional pipeline which overrides standard search components solely employing an index for meta-data.

Other components being added included **cognitive states and personas** – such that user biases, profession, and personality traits, connected to their simulated historical search and interactions, can be available to the simulated user — so that components can then draw upon such information influencing decisions. For example, an LLM-based Result DM could use a simulated role, such as a lawyer, and any disclosed biases (e.g., confirmation bias) in the prompt when making decisions. Also, new **querying strategies** are being developed that focus on reformulation using a Doc2Query approach [22]. These reformulations support the integration for relevance feedback of a simulated user by favouring keywords from relevant documents.

[1]Requires the pyterrier_dr plugin.

The **SimIIR 3.0 Framework** is available at **https://github.com/simint-ai/simiir-3**. A Slack channel in ACM SIGIR community Slack has been set up for the community to propose changes and make contributions to the framework. This resource paper serves as an advertisement to the community to provide feedback on its developments, as well as an invitation to contribute.

4 Future Directions

The new framework enables a greater variety of new simulations to be performed – where researchers can ask many more questions, such as: When LLMs are used to simulate users, how well do they perform? Are some more efficient/effective at interacting with search engines than others? At the same time the framework also enables research on questions regarding conversational search – presenting a two sided problem of building "simulated LLM-based test users" to use "conversational search interfaces/agents": How good are our simulated LLM-based users at using our conversational search interfaces/agents? And, how good are our conversational search interfaces/agents, given our pool of simulated LLM-based users?

That being said, there are many more possible components to develop and different possible user models to support – which are not yet supported – but the framework is suitably flexible to enable more nuanced and complex simulation, if required, or of interest.

Many of the current LLM-based components in the latest version of the SimIIR framework operate independently of each other. This was done to facilitate ease of experimentation and explore whether different LLMs may tackle different components better than others. Moreover, as the issue of context length is not yet a completely solved problem, though less of one than it has been historically, the independent nature means that the SimIIR framework user does not have to try to explicitly manage the interaction "memory" with the LLMs. This unified approach, where all previous interactions are accessible to the simulated user (even those generated by different LLMs), is a promising direction to explore. This would lead to a more realistic simulation, mimicking how humans conduct searches. By having access to the full context of prior interactions, the simulated user would be able use the entire process to complete the next task — similar to how a human would.

In its current state, the framework allows the prompt-based generation of queries with LLMs. Following earlier work by Alaofi et al. [1], the model is guided by additional topical information like the description as provided in the topic file of an ad hoc test collection. However, for some testbeds and experimental setups, this additional context information may not always be available. For instance, the popular MS MARCO test collection [20] does not have topic files that describe the information need in a structured format with a title, description, narrative. Similarly, in real search sessions, the users' information needs are not explicit, and the query is the single reference point. In this regard, we identified several interesting methods that will be integrated into the framework. We envision the query generation component of SimIIR 3.0 as a *zoo* of diverse state-of-the-art LLM-based query generation techniques that can be compared alongside others in a standardised manner using different simulated users. For instance, Jagerman et al. [26] propose an effective query expansion technique that prompts the

LLM to generate an answer for a given query, which is then appended to the initial query and used to retrieve the ranking. This approach could be used to *revitalise* the idea of the known-item searcher as proposed by Azzopardi et al. [6], where the simulated user has a concrete understanding of what is searched. Likewise, the methodology by Bonifacio et al. [14], where the LLM generates queries based on a given relevant document, is an additional viable method.

Another interesting future direction is the use of existing query logs or user query variants to guide the models with a *chain-of-thought* approach. Dai et al. [21] guide the model with samples of different query-document pairs to generate effective queries for retrieval. The UQV datasets by Bailey et al. [8] and Benham and Culpepper [13] provide excellent resources that can be used as samples of user queries that are combined with documents of TREC Web 2013/14 [17, 18] and TREC Common Core 2017 [2] or Robust 2004/05 [38, 39]. Likewise, ORCAS [19] — a companion resource to MS MARCO with clicked query-document pairs — should be considered for these endeavours.

Lastly, we aim for more interactive query generations that happen on the fly during the simulation. Doing so would better consider the context of the simulated sessions. For instance, earlier issued queries and the text contents of clicked documents could be added to the prompt to inform the model of what was seen before and is already familiar to the simulated user. This approach has been recently explored by Zhang et al. [44], who use the instruction-tuned LLM to run the user simulation in its entirety. While this approach has several merits, it sacrifices control over the behaviour of the simulated users. The SimIIR 3.0 framework provides the ability to understand these kinds of trade-offs, we would have to face when relying on the LLM entirely. Suppose that LLMs would generate more human-like queries than principled query generation methods [11]. In that case, SimIIR 3.0 could be used to explore and analyse the trade-offs between *fidelity versus controllability*.

In summary, the SimIIR 3.0 framework provides the basis for exploring and examining a whole range of different research questions in a controlled and repeatable environment. With the new updates, we hope this resource is a useful contribution to the community which can be further built upon and experimented with.

References

[1] Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 1869–1873. https://doi.org/10.1145/3539618.3591960

[2] James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen M. Voorhees. 2017. TREC 2017 Common Core Track Overview. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017 (NIST Special Publication, Vol. 500-324)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec26/papers/Overview-CC.pdf

[3] Leif Azzopardi. 2014. Modelling interaction with economic models of search. In *SIGIR 2014 - Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 3–12. https://doi.org/10.1145/2600428.2602298

[4] Leif Azzopardi. 2016. Simulation of Interaction: Modelling and Simulating User Interaction and Search Behaviour. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 1227–1230. https://doi.org/10.1145/2911451.2914799

[5] L. Azzopardi and M. De Rijke. 2006. Automatic construction of known-item finding test beds. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Vol. 2006.

[6] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. 2007. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando (Eds.). ACM, 455–462. https://doi.org/10.1145/1277741.1277820

[7] Ricardo A. Baeza-Yates, Carlos A. Hurtado, Marcelo Mendoza, and Georges Dupret. 2005. Modeling User Search Behavior. In *Third Latin American Web Congress (LA-Web 2005), 1 October - 2 November 2005, Buenos Aires, Argentina*. IEEE Computer Society, 242–251. https://doi.org/10.1109/LAWEB.2005.23

[8] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 725–728. https://doi.org/10.1145/2911451.2914671

[9] Krisztian Balog and ChengXiang Zhai. 2024. User Simulation for Evaluating Information Access Systems. *Found. Trends Inf. Retr.* 18, 1-2 (2024), 1–261. https://doi.org/10.1561/1500000098

[10] Krisztian Balog and ChengXiang Zhai. 2024. User Simulation for Evaluating Information Access Systems. arXiv:2306.08550 [cs.HC] https://arxiv.org/abs/2306.08550

[11] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2013. Modeling behavioral factors ininteractive information retrieval. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi (Eds.). ACM, 2297–2302. https://doi.org/10.1145/2505515.2505660

[12] Nicholas J. Belkin. 2008. Some(what) grand challenges for information retrieval. *SIGIR Forum* 42, 1 (jun 2008), 47–54. https://doi.org/10.1145/1394251.1394261

[13] Rodger Benham and J. Shane Culpepper. 2017. Risk-Reward Trade-offs in Rank Fusion. In *Proceedings of the 22nd Australasian Document Computing Symposium, ADCS 2017, Brisbane, QLD, Australia, December 7-8, 2017*, Bevan Koopman, Guido Zuccon, and Mark James Carman (Eds.). ACM, 1:1–1:8. https://doi.org/10.1145/3166072.3166084

[14] Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, and Rodrigo Frassetto Nogueira. 2022. InPars: Unsupervised Dataset Generation for Information Retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2387–2392. https://doi.org/10.1145/3477495.3531863

[15] Arthur Câmara, David Maxwell, and Claudia Hauff. 2022. Searching, Learning, and Subtopic Ordering: A Simulation-Based Analysis. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 142–156.

[16] Cyril W Cleverdon, Jack Mills, and E Michael Keen. 1966. Factors determining the performance of indexing systems,(Volume 1: Design). *Cranfield: College of Aeronautics* 28 (1966).

[17] Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz, Charlie Clarke, and Ellen M. Voorhees. 2013. TREC 2013 Web Track Overview. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013 (NIST Special Publication, Vol. 500-302)*, Ellen M. Voorhees (Ed.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec22/papers/WEB.OVERVIEW.pdf

[18] Kevyn Collins-Thompson, Craig Macdonald, Paul N. Bennett, Fernando Diaz, and Ellen M. Voorhees. 2014. TREC 2014 Web Track Overview. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014 (NIST Special Publication, Vol. 500-308)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec23/papers/overview-web.pdf

[19] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 20 Million Clicked Query-Document Pairs for Analyzing Search. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 2983–2989. https://doi.org/10.1145/3340531.3412779

[20] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. MS MARCO: Benchmarking Ranking Models in the Large-Data Regime. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1566–1576. https://doi.org/10.1145/3404835.3462804

[21] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot Dense Retrieval From 8 Examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=gmL46YMpu2J

[22] Björn Engelmann, Timo Breuer, Jana Isabelle Friese, Philipp Schaer, and Norbert Fuhr. 2024. 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 (Lecture Notes in Computer Science, Vol. 14609)*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer, 173–188. https://doi.org/10.1007/978-3-031-56060-6_12

[23] Björn Engelmann, Timo Breuer, and Philipp Schaer. 2023. Simulating Users in Interactive Web Table Retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 3875–3879. https://doi.org/10.1145/3583780.3615187

[24] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online Evaluation for Information Retrieval. *Found. Trends Inf. Retr.* 10, 1 (2016), 1–117. https://doi.org/10.1561/1500000051

[25] Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc.

[26] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query Expansion by Prompting Large Language Models. *CoRR* abs/2305.03653 (2023). https://doi.org/10.48550/ARXIV.2305.03653 arXiv:2305.03653

[27] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends Inf. Retr.* 3, 1-2 (2009), 1–224. https://doi.org/10.1561/1500000012

[28] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.* 18, 1 (2009), 140–181. https://doi.org/10.1007/S10618-008-0114-1

[29] Sahiti Labhishetty and Chengxiang Zhai. 2021. An Exploration of Tester-based Evaluation of User Simulators for Comparing Interactive Retrieval Systems. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1598–1602. https://doi.org/10.1145/3404835.3463091

[30] Sahiti Labhishetty and ChengXiang Zhai. 2022. RATE: A Reliability-Aware Tester-Based Evaluation Framework of User Simulators. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 336–350. https://doi.org/10.1007/978-3-030-99736-6_23

[31] Craig Macdonald, Richard McCreadie, Rodrygo L. T. Santos, and Iadh Ounis. 2012. From Puppy to Maturity: Experiences in Developing Terrier. In *OSIR@SIGIR*. https://api.semanticscholar.org/CorpusID:13725528

[32] Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation inInformation Retrieval using PyTerrier. In *Proceedings of ICTIR 2020*.

[33] David Maxwell. 2019. *Modelling search and stopping in interactive information retrieval*. Ph. D. Dissertation. University of Glasgow, UK. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.775861

[34] David Maxwell and Leif Azzopardi. 2016. Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi (Eds.). ACM, 731–740. https://doi.org/10.1145/2983323.2983805

[35] David Maxwell and Leif Azzopardi. 2016. Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 1141–1144. https://doi.org/10.1145/2911451.2911469

[36] David Maxwell and Leif Azzopardi. 2018. Information Scent, Searching and Stopping - Modelling SERP Level Stopping Behaviour. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 10772)*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer, 210–222. https://doi.org/10.1007/978-3-319-76941-7_16

[37] Paul Thomas, Alistair Moffat, Peter Bailey, and Falk Scholer. 2014. Modeling decision points in user search behavior. In *Proceedings of the 5th Information Interaction in Context Symposium* (Regensburg, Germany) *(IliX '14)*. Association for Computing Machinery, New York, NY, USA, 239–242. https://doi.org/10.1145/2637002.2637032

[38] Ellen M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004 (NIST Special Publication, Vol. 500-261)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf

[39] Ellen M. Voorhees. 2005. Overview of the TREC 2005 Robust Retrieval Track. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005 (NIST Special Publication, Vol. 500-266)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec14/papers/ROBUST.OVERVIEW.pdf

[40] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.

[41] Saber Zerhoudi, Michael Granitzer, Jörg Schlötterer, and Christin Seifert. 2021. Query Change as a Contextual Markov Model for Simulating User Search Behaviour. In *FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, India, December 13 - 17, 2021*, Debasis Ganguly, Surupendu Gangopadhyay, Mandar Mitra, and Prasenjit Majumder (Eds.). ACM, 43–51. https://doi.org/10.1145/3503162.3503165

[42] Saber Zerhoudi, Michael Granitzer, Christin Seifert, and Jörg Schlötterer. 2022. Simulating User Interaction and Search Behaviour in Digital Libraries. In *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event) (CEUR Workshop Proceedings, Vol. 3160)*, Giorgio Maria Di Nunzio, Beatrice Portelli, Domenico Redavid, and Gianmaria Silvello (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-3160/paper8.pdf

[43] Saber Zerhoudi, Sebastian Günther, Kim Plassmeier, Timo Borst, Christin Seifert, Matthias Hagen, and Michael Granitzer. 2022. 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, Atlanta, GA, USA, October 17-21, 2022*, Mohammad Al Hasan and Li Xiong (Eds.). ACM, 4661–4666. https://doi.org/10.1145/3511808.3557711

[44] Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. 2024. USimAgent: Large Language Models for Simulating Search Users. *CoRR* abs/2403.09142 (2024). https://doi.org/10.48550/ARXIV.2403.09142 arXiv:2403.09142