

# Sim4IA-Bench: A User Simulation Benchmark Suite for Next Query and Utterance Prediction

Andreas Konstantin Kruff<sup>1</sup>, Christin Katharina Kreutz<sup>2</sup>, Timo Breuer<sup>1</sup>,  
Philipp Schaer<sup>1</sup>, and Krisztian Balog<sup>3</sup>

<sup>1</sup> TH Köln - University of Applied Sciences, Germany  
{andreas.kruff,timo.breuer,philipp.schaer}@th-koeln.de

<sup>2</sup> TH Mittelhessen - University of Applied Sciences, Germany  
ckreutz@acm.org

<sup>3</sup> Stavanger University, Norway  
krisztian.balog@uis.no

**Abstract.** Validating user simulation is a difficult task due to the lack of established measures and benchmarks, which makes it challenging to assess whether a simulator accurately reflects real user behavior. As part of the Sim4IA Micro-Shared Task at the Sim4IA Workshop, SIGIR 2025, we present Sim4IA-Bench, a simulation benchmark suit for the prediction of the next queries and utterances, the first of its kind in the IR community. Our dataset as part of the suite comprises 160 real-world search sessions from the CORE search engine. For 70 of these sessions, up to 62 simulator runs are available, divided into Task A and Task B, in which different approaches predicted users’ next search queries or utterances. Sim4IA-Bench provides a basis for evaluating and comparing user simulation approaches and for developing new measures of simulator validity. Although modest in size, the suite represents the first publicly available benchmark that links real search sessions with simulated next-query predictions. In addition to serving as a testbed for next query prediction, it also enables exploratory studies on query reformulation behavior, intent drift, and interaction-aware retrieval evaluation. We also introduce a new measure for evaluating next-query predictions in this task. By making the suite publicly available, we aim to promote reproducible research and stimulate further work on realistic and explainable user simulation for information access: <https://github.com/irgroup/Sim4IA-Bench>.

**Keywords:** User simulation · Evaluation · Next Query Prediction

## 1 Introduction and Motivation

In recent years, user simulation has gained increasing attention within IR, as it provides a scalable and controllable method to study user behavior without large-scale user studies [4, 8, 39]. The advent of large language models (LLMs) dramatically lowered the barrier to entry, making it easier than ever to create simulators capable of generating human-like search queries and conversational utterances [2, 45, 58]. However, this rapid development outpaces our ability to

verify their performance. Consequently, there is little shared understanding of what constitutes a good simulator or how its performance should be evaluated [4].

This challenge arises from a fundamental gap in evaluation methodology. The validation of user simulators is an open problem that requires 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. Currently, there is a critical shortage of public resources dedicated to this task. Without a common ground for comparison, it is impossible to assess whether a new simulator is a true advancement or to understand the strengths and weaknesses of different simulation approaches.

To bridge this critical gap, this paper introduces Sim4IA-Bench, the first public benchmark resource specifically designed to evaluate user simulators. As the primary contribution of this work, we introduce a dataset derived from a recent user simulation initiative, the Sim4IA [43] Micro-Shared Task, tackling interactive IR simulation (Task A) and conversational session simulation (Task B). In addition to submissions from participating approaches, we complement this dataset with a proposed set of string-based and system-based similarity measures, offering a crucial starting point for the community to assess simulator quality. Importantly, our goal is not to measure a simulator’s success in a downstream retrieval task, but to directly address the more fundamental question of how well it reproduces authentic user behavior.

Beyond the methodological gap, a significant practical barrier has also hindered the wider adoption of user simulation: the substantial infrastructure and engineering effort required to build a simulator from scratch. To address this challenge, Sim4IA-Bench provides a comprehensive suite of practical resources designed to dramatically lower this barrier to entry. At the core, it includes a simulation toolkit that serves as a starting kit with baseline implementations, data loaders, and evaluation scripts. Sim4IA-Bench contains a rich collection of artifacts, such as prepared session logs, participants’ run files, and comprehensive documentation detailing data formats and evaluation protocols. By packaging these components together, we shift the focus from foundational engineering and enable researchers to concentrate on the core scientific challenges of simulator design. Sim4IA-Bench<sup>4</sup> is released under the MIT license, enabling both academic and industry researchers to access and use the resource.

**The Sim4IA-Bench Suite** In addition to two session datasets for typical IR and conversational search from the academic domain, Sim4IA-Bench provides a comprehensive set of artifacts to support experimentation and evaluation:

- Prepared session logs, including training and test sets.
- Submission run files (62) and corresponding lab notes from the three teams participating in the Sim4IA Micro-Shared Task.
- Benchmarking code for evaluating next-query prediction.
- A simulation toolkit, including Dockerized adaptations of SimIIR 3 [2].

<sup>4</sup> GitHub repository of Sim4IA-Bench: <https://github.com/irgroup/Sim4IA-Bench>

- Tutorials and detailed documentation with setup instructions and example workflows.

Furthermore, this work is intended to guide the development of future community-wide evaluation initiatives. The methodology, dataset structure, and experiences gained from organizing this shared task provide a valuable blueprint for establishing larger-scale, standardized evaluation campaigns at TREC or CLEF, and Sim4IA-Bench will be maintained as part of the User Simulation subtask (Task 3) in LongEval@CLEF’26 [6].

## 2 Related Work

This section covers current directions for the validation of user simulators before datasets for next query and utterance prediction are presented.

### 2.1 Validating User Simulators

There is a current trend towards relying on simulation-based evaluation, especially through usage of LLMs [39, 52, 53]. However, critical shortcomings can arise such as LLMs showcasing behavior that is unrealistic for humans [18, 54] or a lack of natural variation that is usually found in human interactions [49, 57].

While simulators need to be validated against real human interactions [4], the specific requirements for simulators differ depending on what they are going to be used for, i.e. training vs. evaluation [5]. In general, such a comparison against human interactions may be performed at a distributional level, for example, by comparing (i) query characteristics (length, terms) [3], similarity [26], or retrieval performance and shared task utility [7] for traditional search or (ii) the distribution of dialogue acts or success rate for conversational agents [59]. Other approaches for validation include labeling specific instances according to different dimensions, like naturalness, usefulness, grammar for conversational utterances [44, 53, 61] or comparing entire conversations (human vs. simulated) in a side-by-side manner [53, 59]. Another method is the evaluation based on testers where testers are sets of IR systems over which a specific performance pattern can be expected that a simulator is trying to reproduce [29, 30].

While there are few resources dedicated to validating simulators, with Sim4IA-Bench we provide exactly this to bridge this gap.

### 2.2 Datasets

To the best of our knowledge, our introduced resources are the first to provide a common evaluation environment for both interactive IR as well as conversational search. To highlight the novelty of our resource, we surveyed existing log datasets and conversational resources containing session interaction logs. 1 provides an overview of publicly available session datasets at the time of our study.

There are several large-scale datasets that primarily originate from the web search domain or rather small-scale domain-specific datasets from academic,

Table 1: Comparison of Interactive IR and Conversational Search datasets for next-query or next-utterance prediction. **Size** refers to the number of sessions or conversations. **T** informs whether the dataset covers traditional IR, **C** informs whether the dataset covers conversations. **A** refers to additional public assets like system runs from a shared task. **Domain** indicates the topic of the dataset.

Dataset	Size	T	C	A	Domain
AOL [40]	283,207 (AOL17)	✓	✗	✗	Web search
SUSS [35]	484,449	✓	✗	✗	Academic search
Yandex [46]	797,867	✓	✗	✗	Web search
TREC Session [12]	1564	✓	✗	✓	Web search
ConvAI [34]	4750	✗	✓	✓	Human Chatting
TianGong-ST [14]	147,155	✓	✗	✓	Web search
ConvAI2 [20]	4406	✗	✓	✓	Human Chatting
TripClick [41]	1,602,648	✓	✗	✓	Health
ConvAI3 (ClariQ) [1]	1,596,757	✗	✓	✓	Human Chatting
Baidu-ULTR [64]	1.2 bil	✓	✗	✓	Web search
Webis-FUQ-24 [26]	18,980	✓	✗	✗	Web search, arguments, exhibitions, product search
Persona-Chat [60]	10,907	✗	✓	✗	Human chatting
Webis-CQR-2 [25]	284	✗	✓	✗	Arguments, books, news, trips
SoguoQ [47]	14,075,717	✗	✓	✓	Web search
TREC CAsT (2022) [38]	50	✗	✓	✓	Web search
LLM-REDIAL [31]	47,600	✗	✓	✗	Movies, books, sports
WildChat [62]	1,039,785	✗	✓	✗	Web search
LMSYS-CHAT-1M [63]	1,000,000	✗	✓	✗	Web search
Ours	160	✓	✓	✓	Academic search

health-related, and other fields. Most of these datasets serve as evaluation toolkits for different aspects of the user modeling in an interactive search setting with varying degrees of an explicit user simulation. For instance, some datasets allow a comprehensive evaluation of the different interactions in a simulated sessions, while others have a more specific focus like query suggestion, click modeling, or utterance prediction. Building on these datasets, prior work has examined related aspects such as query expansion and suggestion [16, 36], query and utterance prediction [23, 33, 51, 55], or session modeling [13, 17, 22, 56].

The additional assets (A) column in 1 denotes whether a dataset has previously been employed in a shared task or is included in a benchmark that enables systematic evaluation. Several of the listed datasets have served as complementary assets in this sense. For example, the Yandex dataset was used in a challenge focused on personalizing search results based on user context and search history [46]. The four TREC Session datasets were employed in a task designed to improve retrieval effectiveness through the use of historical queries, ranked result lists, and user interaction information [12]. Similarly, the Baidu-ULTR dataset was used in a task where participants developed feature-based re-ranking models

that exploited behavioral and display features to better capture user preferences [37]. The SogouQ dataset was used in a task addressing ambiguous queries and promoting ranking diversification to account for multiple possible user intents [48]. TianGong-ST has been applied in a task focused on ranking documents for the final query of a session, taking into account the complete preceding session context [15]. More recently, dialogue datasets such as WildChat [62] have been designed to evaluate large language models in realistic conversational settings [24, 32].

Taken together, these datasets exemplify how complementary assets facilitate structured comparison, either by providing directly comparable shared task runs or by being integrated into established benchmarks. Building on this, our resource is the first to provide complementary assets specifically for next query prediction in the context of user simulation, offering run files that enable reproducible system comparisons and systematic evaluation of measures under controlled experimental conditions.

In general, there is a trade-off between the desire to make as much user interaction data available for rigorous validation of user simulations and the requirements to keep users anonymous and respecting their privacy. Sim4IA-Bench enables validations of user simulations across multiple sessions in traditional IR and conversational search, while guaranteeing full user privacy based on rigorous anonymization measures.

### 3 Dataset

Here we present the dataset at the core of Sim4IA-Bench: its structure, contents, and key characteristics, highlighting the aspects that make it suitable for evaluating and developing user simulators in interactive IR.

#### 3.1 Task Descriptions and Dataset Contents

We provide session-based datasets for the Sim4IA Micro-Shared Task [43], designed to support the evaluation and development of user simulators in interactive IR. The resources include two sets of data corresponding to the two task variants: Task A (interactive IR simulation) and Task B (conversational session simulation). Each task has a training set of 45 sessions and a test set of 35 sessions. The test sets differ from the training sets in that the final query of each session is withheld, allowing participants to evaluate simulator predictions without overfitting. In all tasks ten next queries or utterances are to be predicted. For a concise overview of the two tasks and their respective workflows, see 1.

**Task A: Interactive IR Simulation** Each session in Task A contains queries, the corresponding retrieved SERPs, and the documents clicked by users. Clicks are categorized into three types: clicks on authors, clicks on the work itself, and clicks on the “Download PDF” option. Timestamps for queries and clicks are included, enabling simulators to account for temporal aspects of interactions

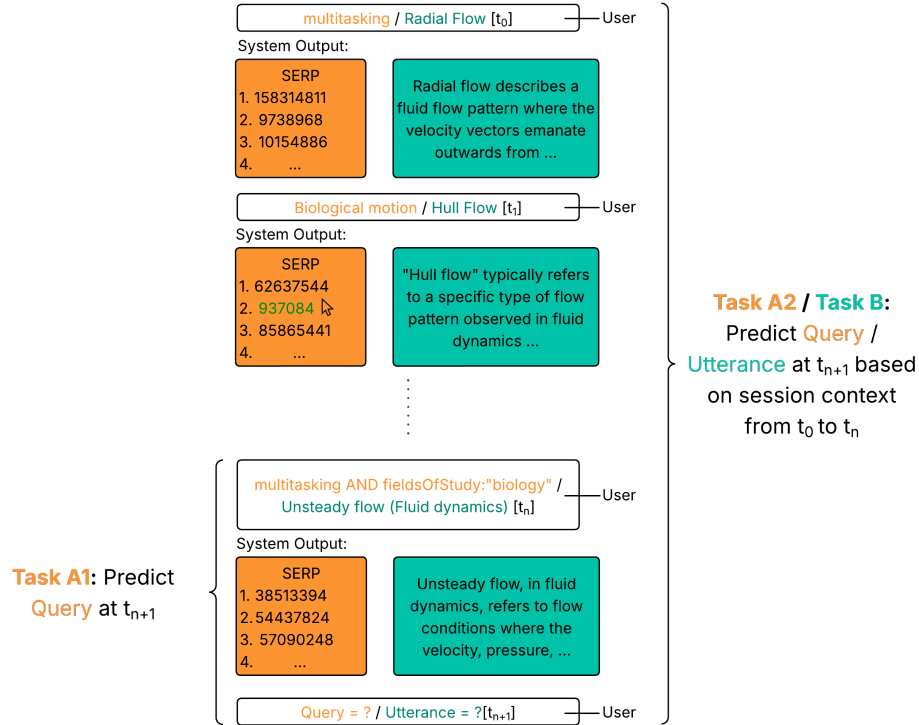


Fig. 1: Overview of Task A (interactive IR simulation, in orange) and Task B (conversational session simulation, in teal). Document IDs highlighted in green represent documents a real user interacted with.

and to model different types of user behavior. We composed two variants of Task A using the same data: Task A1 only considers the last query and corresponding SERP information, while in Task A2, the whole session was allowed to be used. For Task A, the dataset contains an average session length of 5.20 queries (4.20 reformulations on average), with 1.49 clicks per query.

**Task B: Conversational Session Simulation** Task B sessions contain only utterance-response pairs. Responses were generated using Google’s Gemma 3 12B model served on Ollama. For each query, the model was provided with the top three retrieved documents, the preceding response, and all upcoming utterances. The prompt (see 2) was designed to ensure that the model focused solely on the current utterance while steering towards the upcoming utterances, producing coherent and plausible conversational flows. For Task B, the average session length is 4.85 queries (3.85 reformulations on average).

Utterance: **What are the main symptoms of diabetes?**

Relevant Documents:

Title: **Diabetes Overview**

Abstract: **Diabetes is a chronic condition characterized by high blood sugar levels. Common symptoms include frequent urination, excessive thirst, fatigue, and blurred vision.**

Title: **Type 2 Diabetes Symptoms**

Abstract: **Type 2 diabetes often develops gradually. Symptoms include increased hunger, frequent infections, and slow-healing sores.**

Previous Response: **Diabetes is a condition affecting the body’s ability to regulate blood sugar.**

Upcoming User Utterances:

Utterance: **How can diabetes be managed effectively?**

Utterance: **What lifestyle changes are recommended for patients?**

Instruction:

- Answer the query using the relevant documents and the previous response.
- Act like a RAG system: provide relevant, informative, and context-aware responses tailored to the user’s query intent.
- Try to only answer the current utterance and explicitly do not provide answers in the response, that might answer the provided upcoming utterances.
- Limit your answer to **no more than 150 words**.
- Focus on key points and avoid unnecessary repetition.
- Please do answer formally and don’t use phrases like ‘Okay, here’s a response to “Diabetes”’ acting as a RAG system.

Answer:

Fig. 2: Prompt template for generating the synthetic responses of the conversational system for Task B. Black parts show the unchanging structural and instructional parts, purple components depict the variables inserted depending on sessions and utterances.

### 3.2 Session Extraction from CORE Logs

Sessions were reconstructed from CORE [27] log files using a heuristic approach to group queries into sessions: starting from the last query in a potential session, queries occurring within a time window of -10 to +5 minutes were considered candidates for inclusion. Queries were added to a session if their cosine similarity, calculated using embeddings generated by the SentenceTransformer model all-MiniLM-L6-v2, with the current session queries, was greater than or equal to 0.1. All reconstructed sessions were independently manually reviewed by two people to ensure coherence and plausibility. Query sessions that contained fewer than three reformulations or did not originate from a plausible information need were excluded.

Table 2: Distribution of submitted run types for tasks by groups from CIR, Webis and THM.

	(semi-) manual	persona	prompting & tuning	other LLM	rule-based	$\Sigma$
Task A1	3	6	2 + 3 = 5	3 + 12 = 15	1	30
Task A2	2 + 3 = 5	6	3	4		18
Task B	4	6		4		14

### 3.3 Usage Notes

The datasets and artifacts are designed to be fully accessible and ready for use. Example scripts and tools facilitate loading, processing, and analysis of the sessions. For Task A, the combination of training and test sets allows evaluation of next-query predictions at multiple levels, from string similarity to semantic and SERP-based metrics. Task-specific features, such as click categories in Task A and semi-synthetic responses in Task B, enable more nuanced simulator evaluations and the exploration of interaction dynamics in both interactive and conversational settings.

### 3.4 Submitted Runs

For the Micro-Shared Task we obtained 62 runs over all tasks in total and three accompanying lab notes detailing information on these runs by groups from CIR [10], Webis [21] and THM [19]. In general, we distinguish between runs that were composed (semi-) manually, through an LLM, as well as rule-based ones. We further differentiate LLM-based runs as ones composed by prompting the LLM to behave as a persona alone, through a combination of tuning and prompting and others. 2 provides the number of runs of different types from each group for the respective tasks.

## 4 Assessing Simulator Fidelity

Evaluating the quality and validity of a simulator in query prediction tasks is challenging, and the choice of suitable measures remains an open question. In this work, our methodology shifts the focus from a simulator’s effectiveness in retrieval tasks to its *reproduction quality*—that is, how well it replicates the authentic user behavior observed in real interaction logs. To assess this, we employ a complementary set of string-based and system-based similarity measures. Each measure is designed to capture a different aspect of simulator fidelity by quantifying the degree to which simulated queries match actual user inputs. In addition to the following measures presented in this paper, a broader overview of suitable measures was conducted in the work of Kruff et al in an additional comprehensive study [28].

#### 4.1 Measures

**Semantic Similarity.** Semantic similarity assesses how close the predicted queries are in meaning to the original user queries. We computed the average cosine similarity between sentence embeddings of the original query  $q_i^{true}$  and the  $Q$  candidate queries  $q_{i,j}$  over  $N$  sessions. We then bounded the similarity values from  $[-1, 1]$  to  $[0, 1]$  for better visual comparability with the other measures.

$$\bar{S} = \frac{1}{|N|} \sum_{i=1}^N \left( \frac{1}{Q} \sum_{j=1}^Q \text{cosine}(q_i^{true}, q_{i,j}) \right)$$

This measure ensures that candidates are not only syntactically similar but also semantically aligned with the user’s intent.

**Redundancy.** To evaluate novelty and diversity, we introduced a Redundancy measure, which calculates the Jaccard similarity between all  $Q$  candidate queries or utterances within a session averaged across all  $N$  sessions:

$$\bar{R} = \frac{1}{|N|} \sum_{i=1}^N \left( \frac{2}{Q(Q-1)} \sum_{1 \leq j < k \leq Q} \text{Jaccard}(q_{i,j}, q_{i,k}) \right)$$

Low redundancy indicates that a simulator produces multiple distinct candidates that remain semantically similar to the original query, whereas high redundancy signals minimal variation. Redundancy is not intended as a stand-alone metric; rather, it is defined for its role in the Rank-Diversity Score described below. It serves to penalize simulators that produce candidates with only minor variations, while giving simulators the opportunity to also predict queries exhibiting a degree of in-session topic drift.

**SERP Overlap.** To capture system-level effects, we measured the average overlap between the search engine results retrieved for the original query and the  $Q$  candidates over the  $N$  sessions:

$$\bar{O} = \frac{1}{|N|} \sum_{i=1}^N \left( \frac{1}{Q} \sum_{j=1}^Q \text{overlap}(q_i^{true}, q_{i,j}) \right)$$

Overlap denotes the fraction of shared documents in the top-10 ranked results by a fixed retrieval system. In our experiments, queries were executed using BM25. This metric assumes access to a search engine and corpus on which both original and candidate queries can be run, and is therefore dependent on the underlying retrieval model and index. The cutoff was set with reference to typical precision scores and the typical length of a webpage in the absence of pagination.

Accordingly, SERP overlap should be interpreted as a system-dependent indicator of how query reformulation affects downstream retrieval behavior, complementing string- and meaning-based similarity metrics rather than serving as a standalone effectiveness measure.

**Rank-Diversity Score.** While the above measures capture overall candidate quality, they do not consider the ordering of candidates. To address this,

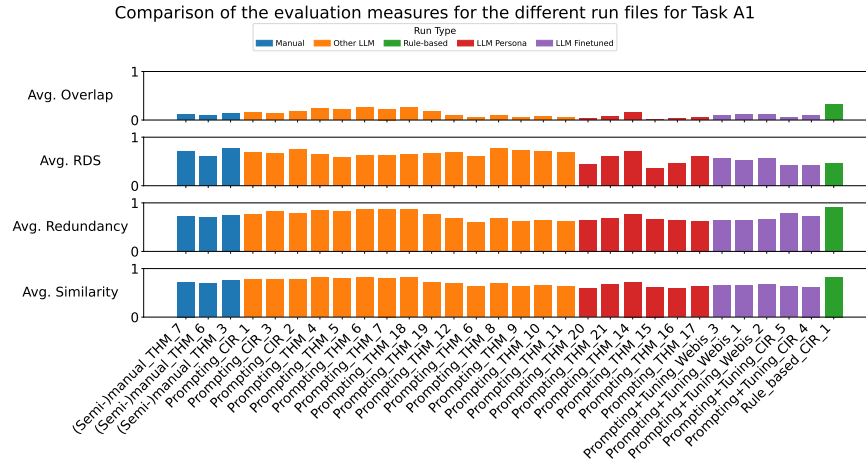


Fig. 3: Exemplary run results for Task A1.

we designed an MMR-inspired Rank-Diversity Score (RDS) that combines rank-based evaluation with redundancy:

$$\bar{RD} = \text{RDS}_{\cos \geq 0.7} \cdot (1 - \bar{R})$$

$$\text{RDS}_{\cos \geq 0.7} = \frac{1}{|N|} \sum_{i \in N} \left( \sum_{q \in Q_i} \frac{1}{\text{rank}_q^{\cos \geq 0.7}} \right)$$

RDS is computed over all sessions  $N$  and the list of candidates  $Q_i$  generated for each session  $i$ . It rewards simulators that rank high-quality, diverse candidate queries or utterances at the top while penalizing poor ordering and runs submitting fewer than the required ten candidates, as this lowers their potential multiplier. Originally introduced by Carbonell and Goldstein [11], MMR provides a conceptual foundation for our measure.

## 4.2 Case Study: Analysis of Task A1

Using the runs from Task A1 as an example, we illustrate how the applied measures provide complementary perspectives on simulator performance. Cosine similarity captures the semantic closeness between simulated and actual user queries, showing that manually created runs and large, out-of-the-box LLMs perform comparably well in reproducing the linguistic and conceptual intent of users. Persona-based and fine-tuned simulators, by contrast, show slightly lower alignment, while rule-based approaches reach high similarity scores due to their strong adherence to the original queries.

SERP overlap extends this perspective to the system level, reflecting whether different simulators lead to comparable retrieval outcomes. Here, the rule-based

approach again performs strongest, as its low deviation from the original queries results in highly similar retrieval results. However, this measure also highlights that approaches generating more diverse queries, such as manual, persona-based, or fine-tuned runs, tend to diverge in retrieval outcomes, which suggests a trade-off between fidelity and behavioral variability.

Finally, the redundancy-based analysis, incorporated into the newly proposed measure, reveals a key limitation of several simulator types. Their tendency to produce highly similar candidate queries with little internal variation becomes particularly apparent for rule-based and some prompting-based approaches as well as for fine-tuned models that closely follow prior query patterns. In contrast, manual and large LLM runs benefit from this measure, as they generate more diverse candidate sets while maintaining overall semantic coherence.

Overall, while cosine similarity and SERP overlap are well suited for assessing how closely simulators mirror real user intent and retrieval outcomes, the redundancy-oriented measure complements them by exposing a lack of diversity across candidate queries. Together, these measures provide a multifaceted understanding of simulator behavior, revealing the strengths of faithful reproduction and the weaknesses in behavioral variability across approaches.

### 4.3 Reuse Value of the Benchmark Artifacts

Ultimately, Sim4IA-Bench establishes a foundation for developing future validation measures for user simulations. The corresponding process would follow standard practices used in developing IR evaluation metrics. For example, methods such as swap counting [50], stability tests [9], or bootstrapping [42] could be applied to determine whether a new validation measure aligns with existing ones, including those introduced in this work, or opens up complementary evaluation perspectives. In this context, the focus of the validation would shift from ranking systems to ranking user simulators instead.

## 5 Reflections on the Shared Task

The considerable effort invested in preparing the study environment, including comprehensive setup instructions and video guides, proved successful. None of the participating teams required substantial support from the organizers. The smooth onboarding process was a key logistical success, ensuring that the setup can be easily reproduced in future studies.

The shared task and its resulting dataset provide a valuable first step toward systematically exploring user simulation in interactive search. However, the process also highlighted several key challenges and limitations. A significant finding, particularly for Task B (conversational session simulation), was that the simulated utterances remained largely “query-like” and lacked the natural verbosity typical of this setting. This points to a major area for future improvement in simulator design. Furthermore, the semi-synthetic nature of the dataset

highlighted both possibilities and constraints, offering opportunities for experimentation while revealing areas where realism could be improved. Finally, the task underscored that selecting appropriate evaluation measures is still an open and underexplored challenge for the community.

## 6 Conclusion and Outlook

This work addresses the critical need for standardized evaluation of user simulators in IR. We introduce Sim4IA-Bench, a comprehensive suite of resources derived from the Sim4IA [43] shared task, which includes datasets, task definitions, a baseline toolkit, tutorials, and documentation. This suite represents an initial yet significant step toward advancing the systematic study and evaluation of user simulators. By providing a structured, reusable framework, they enable reproducible experimentation and support methodological exploration in interactive IR. It also facilitates the testing of alternative simulation approaches and evaluation measures, inspiring the development of new tasks, datasets, or experimental designs.

This resource is not a static endpoint but a foundation for future community efforts. As an immediate next step, the methodology and toolkit will be maintained and expanded as part of the User Simulation subtask (Task 3) in LongEval@CLEF'26. This transition from a micro-shared task to a full, recurring shared task will facilitate the collection of new datasets (starting with new data for Task A) and the evaluation of new measures, providing further insights into simulation-based evaluation of next-query prediction.

Looking ahead, the limitations identified in our reflections (Section

**Acknowledgments.** This work is partially funded by Deutsche Forschungsgemeinschaft (DFG) under grant number 509543643 and within the funding programme FH-Personal (PLan CV, reference number 03FHP109) by the German Federal Ministry of Education and Research (BMBF) and Joint Science Conference (GWK).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## Bibliography

- [1] Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.: Building and evaluating open-domain dialogue corpora with clarifying questions. In: EMNLP '21', pp. 4473–4484 (2021), <https://doi.org/10.18653/v1/2021.emnlp-main.367>
- [2] Azzopardi, L., Breuer, T., Engelmann, B., Kreutz, C., MacAvaney, S., Maxwell, D., Parry, A., Roegiest, A., Wang, X., Zerhoubi, S.: SimIIR 3: A framework for the simulation of interactive and conversational information retrieval. In: SIGIR-AP '24, pp. 197–202 (2024), <https://doi.org/10.1145/3673791.3698427>

- [3] Azzopardi, L., de Rijke, M., Balog, K.: Building simulated queries for known-item topics: An analysis using six european languages. In: SIGIR '07, pp. 455–462 (2007), <https://doi.org/10.1145/1277741.1277820>
- [4] Balog, K., Zhai, C.: User simulation for evaluating information access systems. *Foundations and Trends® in Information Retrieval* **18**(1-2), 1–261 (2024), <https://doi.org/10.1561/15000000098>
- [5] Bernard, N., Balog, K.: Towards a formal characterization of user simulation objectives in conversational information access. In: ICTIR '24, pp. 185–193 (2024), <https://doi.org/10.1145/3664190.3672529>
- [6] Breuer, T., Cancellieri, M., El-Ebshihy, A., Fröbe, M., Galuščáková, P., Goeuriot, L., Iturra-Bocaz, G., Keller, J., Knoth, P., Kruff, A., Mulhem, P., Piroi, F., Pride, D., Schaer, P., Schwab, D.: Evaluating information retrieval models along time: The longeval lab. In: Anand, A., Ren, Z., Verberne, S., Jatowt, A., Campos, R., Lan, Y., MacAvaney, S., Aliannejadi, M., Bauer, C., Mansoury, M. (eds.) *Advances in Information Retrieval, 48th European Conference on Information Retrieval, ECIR 2026* (2026)
- [7] Breuer, T., Fuhr, N., Schaer, P.: Validating simulations of user query variants. In: Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørnvåg, K., Setty, V. (eds.) *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, pp. 80–94, Springer (2022), [https://doi.org/10.1007/978-3-030-99736-6\\_6](https://doi.org/10.1007/978-3-030-99736-6_6), URL [https://doi.org/10.1007/978-3-030-99736-6\\_6](https://doi.org/10.1007/978-3-030-99736-6_6)
- [8] Breuer, T., Maistro, M.: Toward evaluating the reproducibility of information retrieval systems with simulated users. In: ACM-REP '24, pp. 25–29 (2024), <https://doi.org/10.1145/3641525.3663619>
- [9] Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: SIGIR '00, pp. 33–40 (2000), <https://doi.org/10.1145/345508.345543>
- [10] Busch, T., El Ghadioui, M., Knippenberg, P., Mörsheim, M.H.: CIR@Sim4IA: Lab Note Submission for Team 1 and Team 2 for Subtask A1 (2025), <https://doi.org/10.5281/zenodo.16909638>
- [11] Carbonell, J.G., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: SIGIR '98, pp. 335–336 (1998), <https://doi.org/10.1145/290941.291025>
- [12] Carterette, B., Clough, P., Hall, M., Kanoulas, E., Sanderson, M.: Evaluating retrieval over sessions: The trec session track 2011-2014. In: SIGIR '16, pp. 685–688 (2016), <https://doi.org/10.1145/2911451.2914675>
- [13] Chen, H., Dou, Z., Zhu, Y., Cao, Z., Cheng, X., Wen, J.: Enhancing user behavior sequence modeling by generative tasks for session search. In: CIKM '22, pp. 180–190 (2022), <https://doi.org/10.1145/3511808.3557310>
- [14] Chen, J., Mao, J., Liu, Y., Zhang, M., Ma, S.: Tiangong-st: A new dataset with large-scale refined real-world web search sessions. In: CIKM '19, pp. 2485–2488 (2019), <https://doi.org/10.1145/3357384.3358158>
- [15] Chen, J., Wu, W., Mao, J., Wang, B., Zhang, F., Liu, Y.: Overview of the NTCIR-16 session search (SS) task. In: NT-

- CIR '22 (2022), URL <https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/01-NTCIR16-0V-SS-ChenJ.pdf>
- [16] Chen, W., Cai, F., Chen, H., de Rijke, M.: Attention-based hierarchical neural query suggestion. In: SIGIR '18, pp. 1093–1096 (2018), <https://doi.org/10.1145/3209978.3210079>
- [17] Cheng, Q., Ren, Z., Lin, Y., Ren, P., Chen, Z., Liu, X., de Rijke, M.: Long short-term session search: Joint personalized reranking and next query prediction. In: WWW '21, pp. 239–248 (2021), <https://doi.org/10.1145/3442381.3449941>
- [18] Davidson, S., Romeo, S., Shu, R., Gung, J., Gupta, A., Mansour, S., Zhang, Y.: User simulation with large language models for evaluating task-oriented dialogue. CoRR **abs/2309.13233** (2023), <https://doi.org/10.48550/ARXIV.2309.13233>
- [19] Dietzler, N.O., Hofmann, N., Dauenhauer, J., Idahor, I.D., Kreutz, C.K.: THM@Sim4IA: Manual and automated next query prediction for user simulation (2025), <https://doi.org/10.5281/zenodo.17386068>
- [20] Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., Prabhunoye, S., Black, A.W., Rudnicky, A., Williams, J., Pineau, J., Burtsev, M., Weston, J.: The Second Conversational Intelligence Challenge (ConvAI2). CoRR **abs/1902.00098** (2019), URL <http://arxiv.org/abs/1902.00098>
- [21] Gohsen, M., Hagen, M., Stein, B.: Webis at Sim4IA 2025: Prediction of Next User Queries as a Sequence-To-Sequence Problem (2025), <https://doi.org/10.5281/zenodo.16909542>
- [22] Günther, S., Göttert, P., Hagen, M.: Exploring lstms for simulating search sessions in digital libraries. In: TPD L '22, pp. 469–473 (2022), [https://doi.org/10.1007/978-3-031-16802-4\\_47](https://doi.org/10.1007/978-3-031-16802-4_47)
- [23] Ivey, J., Kumar, S., Liu, J., Shen, H., Rakshit, S., Raju, R., Zhang, H., Ananthasubramaniam, A., Kim, J., Yi, B., Wright, D., Israeli, A., Möller, A.G., Zhang, L., Jurgens, D.: Real or Robotic? Assessing Whether LLMs Accurately Simulate Qualities of Human Responses in Dialogue. CoRR **abs/2409.08330** (2024), <https://doi.org/10.48550/ARXIV.2409.08330>
- [24] Joko, H., Amirshahi, S., Clarke, C.L.A., Hasibi, F.: WildClaims: Information Access Conversations in the Wild(Chat). CoRR **abs/2509.17442** (2025), <https://doi.org/10.48550/ARXIV.2509.17442>
- [25] Kiesel, J., Cai, X., Baff, R.E., Stein, B., Hagen, M.: Toward conversational query reformulation. In: DESIRES '21, pp. 91–101 (2021), URL <https://ceur-ws.org/Vol-2950/paper-12.pdf>
- [26] Kiesel, J., Gohsen, M., Mirzakhmedova, N., Hagen, M., Stein, B.: Simulating follow-up questions in conversational search. In: ECIR '24, pp. 382–398 (2024), [https://doi.org/10.1007/978-3-031-56060-6\\_25](https://doi.org/10.1007/978-3-031-56060-6_25)
- [27] Knoth, P., Herrmannova, D., Cancellieri, M., Anastasiou, L., Pontika, N., Pearce, S., Gyawali, B., Pride, D.: Core: a global aggregation service for open access papers. Scientific Data **10**(1), 366 (2023), <https://doi.org/10.1038/s41597-023-02208-w>

- [28] Kruff, A.K., Bernard, N., Schaer, P.: Validating search query simulations: A taxonomy of measures. In: Anand, A., Ren, Z., Verberne, S., Jatowt, A., Campos, R., Lan, Y., MacAvaney, S., Aliannejadi, M., Bauer, C., Mansoury, M. (eds.) *Advances in Information Retrieval, 48th European Conference on Information Retrieval, ECIR 2026* (2026)
- [29] Labhishetty, S., Zhai, C.: An exploration of tester-based evaluation of user simulators for comparing interactive retrieval systems. In: *SIGIR '21*, pp. 1598–1602 (2021), <https://doi.org/10.1145/3404835.3463091>
- [30] Labhishetty, S., Zhai, C.: RATE: A reliability-aware tester-based evaluation framework of user simulators. In: *ECIR '22*, pp. 336–350 (2022), [https://doi.org/10.1007/978-3-030-99736-6\\_23](https://doi.org/10.1007/978-3-030-99736-6_23)
- [31] Liang, T., Jin, C., Wang, L., Fan, W., Xia, C., Chen, K., Yin, Y.: LLM-REDIAL: A large-scale dataset for conversational recommender systems created from user behaviors with llms. In: *Findings of the ACL '24*, pp. 8926–8939 (2024), <https://doi.org/10.18653/v1/2024.FINDINGS-ACL.529>
- [32] Lin, B.Y., Deng, Y., Chandu, K.R., Ravichander, A., Pyatkin, V., Dziri, N., Bras, R.L., Choi, Y.: WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild. In: *ICLR '25* (2025), URL <https://openreview.net/forum?id=MKEHCx25xp>
- [33] Liu, Q., Chen, Y., Chen, B., Lou, J.G., Chen, Z., Zhou, B., Zhang, D.: You impress me: Dialogue generation via mutual persona perception. In: *ACL '20* (2020), URL <https://api.semanticscholar.org/CorpusID:215745354>
- [34] Logacheva, V., Burtsev, M., Malykh, V., Polulyakh, V., Seliverstov, A.: Convai dataset of topic-oriented human-to-chatbot dialogues. In: *The NIPS '17 Competition: Building Intelligent Systems*, pp. 47–57 (2018), [https://doi.org/10.1007/978-3-319-94042-7\\_3](https://doi.org/10.1007/978-3-319-94042-7_3)
- [35] Mayr, P., Kacem, A.: A complete year of user retrieval sessions in a social sciences academic search engine. In: *TPDL '17*, pp. 560–565 (2017), [https://doi.org/10.1007/978-3-319-67008-9\\_46](https://doi.org/10.1007/978-3-319-67008-9_46)
- [36] Mitra, B.: Exploring session context using distributed representations of queries and reformulations. In: *SIGIR '15*, pp. 3–12 (2015), <https://doi.org/10.1145/2766462.2767702>
- [37] Niu, Z., Mao, J., Ai, Q., Zou, L., Wang, S., Yin, D.: Overview of the ntcir-17 unbiased learning to rank evaluation 2 (ultre-2) task. In: *NTCIR '23* (2023), <https://doi.org/10.20736/0002001320>
- [38] Owoicho, P., Dalton, J., Aliannejadi, M., Azzopardi, L., Trippas, J.R., Vakulenko, S.: TREC CAsT 2022: Going beyond user ask and system retrieve with initiative and response generation. In: *TREC '22* (2022), URL <https://api.semanticscholar.org/CorpusID:261288646>
- [39] Owoicho, P., Sekulic, I., Aliannejadi, M., Dalton, J., Crestani, F.: Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In: *SIGIR '23*, pp. 632–642 (2023), <https://doi.org/10.1145/3539618.3591683>
- [40] Pass, G., Chowdhury, A., Torgeson, C.: A picture of search. In: *InfoScale '06*, p. 1 (2006), <https://doi.org/10.1145/1146847.1146848>

- [41] Rekabsaz, N., Lesota, O., Schedl, M., Brassey, J., Eickhoff, C.: TripClick: The Log Files of a Large Health Web Search Engine. In: SIGIR '21, pp. 2507–2513 (2021), <https://doi.org/10.1145/3404835.3463242>
- [42] Sakai, T.: Evaluating evaluation metrics based on the bootstrap. In: SIGIR '06, pp. 525–532 (2006), <https://doi.org/10.1145/1148170.1148261>
- [43] Schaer, P., Kreutz, C.K., Balog, K., Breuer, T., Kruff, A.K.: Second SIGIR Workshop on Simulations for Information Access (Sim4IA 2025). In: SIGIR '25, pp. 4172–4175 (2025), <https://doi.org/10.1145/3726302.3730363>
- [44] Sekulić, I., Aliannejadi, M., Crestani, F.: Evaluating mixed-initiative conversational search systems via user simulation. In: WSDM '22, pp. 888–896 (2022), <https://doi.org/10.1145/3488560.3498440>
- [45] Sekulić, I., Alinannejadi, M., Crestani, F.: Analysing utterances in llm-based user simulation for conversational search. *ACM Trans. Intell. Syst. Technol.* **15**(3) (2024), <https://doi.org/10.1145/3650041>
- [46] Serdyukov, P., Dupret, G., Craswell, N.: Log-based personalization: the 4th web search click data (wscd) workshop. In: WSDM '14, pp. 685–686 (2014), <https://doi.org/10.1145/2556195.2556207>
- [47] Song, R., Zhang, M., Luo, C., Sakai, T., Liu, Y., Dou, Z.: SogouQ: The First Large-Scale Test Collection with Click Streams Used in a Shared-Task Evaluation, pp. 143–150 (2021), [https://doi.org/10.1007/978-981-15-5554-1\\_10](https://doi.org/10.1007/978-981-15-5554-1_10)
- [48] Song, R., Zhang, M., Sakai, T., Kato, M.P., Liu, Y., Sugimoto, M., Wang, Q., Oriei, N.: Overview of the ntcir-9 intent task. In: NTCIR '11 (2011), URL <https://api.semanticscholar.org/CorpusID:18551696>
- [49] Terragni, S., Filipavicius, M., Khau, N., Guedes, B., Manso, A.F., Mathis, R.: In-context learning user simulators for task-oriented dialog systems. *CoRR* **abs/2306.00774** (2023), <https://doi.org/10.48550/ARXIV.2306.00774>
- [50] Voorhees, E.M., Buckley, C.: The effect of topic set size on retrieval experiment error. In: SIGIR '02, pp. 316–323, ACM (2002), <https://doi.org/10.1145/564376.564432>
- [51] Wang, K., Li, X., Yang, S., Zhou, L., Jiang, F., Li, H.: Know you first and be you better: Modeling human-like user simulators via implicit profiles. In: ACL '25, pp. 21082–21107 (2025), <https://doi.org/10.18653/v1/2025.acl-long.1025>
- [52] Wang, L., Zhang, J., Yang, H., Chen, Z.Y., Tang, J., Zhang, Z., Chen, X., Lin, Y., Sun, H., Song, R., Zhao, X., Xu, J., Dou, Z., Wang, J., Wen, J.R.: User behavior simulation with large language model-based agents. *ACM Trans. Inf. Syst.* **43**(2) (2025), <https://doi.org/10.1145/3708985>
- [53] Wang, X., Tang, X., Zhao, X., Wang, J., Wen, J.R.: Rethinking the evaluation for conversational recommendation in the era of large language models. In: EMNLP '23, pp. 10052–10065 (2023), <https://doi.org/10.18653/v1/2023.emnlp-main.621>
- [54] Wang, Z., Xu, Z., Srikumar, V., Ai, Q.: An in-depth investigation of user response simulation for conversational search. In: Proceedings of the ACM on Web Conference 2024, pp. 1407–1418, WWW '24 (2024), <https://doi.org/10.1145/3589334.3645447>

- [55] Yang, D., Zhang, Y., Fang, H.: Zero-shot query reformulation for conversational search. In: ICTIR '23, pp. 257–63 (2023), <https://doi.org/10.1145/3578337.3605143>
- [56] Ye, Y., Li, Z., Dou, Z., Zhu, Y., Zhang, C., Wu, S., Cao, Z.: Learning from the wisdom of crowds: Exploiting similar sessions for session search. In: IAAI '23, pp. 4818–4826 (2023), <https://doi.org/10.1609/AAAI.V37I4.25607>
- [57] Yoon, S.e., He, Z., Echterhoff, J., McAuley, J.: Evaluating large language models as generative user simulators for conversational recommendation. In: NAACL '24, pp. 1490–1504 (2024), <https://doi.org/10.18653/V1/2024.NAACL-LONG.83>
- [58] Zhang, E., Wang, X., Gong, P., Yang, Z., Mao, J.: Exploring human-like thinking in search simulations with large language models. In: SIGIR '25, pp. 2669–2673 (2025), <https://doi.org/10.1145/3726302.3730193>
- [59] Zhang, S., Balog, K.: Evaluating conversational recommender systems via user simulation. In: KDD '20, pp. 1512–1520 (2020), <https://doi.org/10.1145/3394486.3403202>
- [60] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? In: ACL '18, pp. 2204–2213 (2018), <https://doi.org/10.18653/v1/P18-1205>
- [61] Zhang, S., Wang, M.C., Balog, K.: Analyzing and simulating user utterance reformulation in conversational recommender systems. In: SIGIR '22, pp. 133–143 (2022), <https://doi.org/10.1145/3477495.3531936>
- [62] Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., Deng, Y.: WildChat: 1M ChatGPT Interaction Logs in the Wild. In: ICLR '24 (2024), URL <https://openreview.net/forum?id=B18u7ZR1bM>
- [63] Zheng, L., Chiang, W., Sheng, Y., Li, T., Zhuang, S., Wu, Z., Zhuang, Y., Li, Z., Lin, Z., Xing, E.P., Gonzalez, J.E., Stoica, I., Zhang, H.: LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In: ICLR '24 (2024), URL <https://openreview.net/forum?id=B0fDKxfwt0>
- [64] Zou, L., Mao, H., Chu, X., Tang, J., Ye, W., Wang, S., Yin, D.: A large scale search dataset for unbiased learning to rank. In: NeurIPS '22, pp. 1127–1139 (2022), URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/07f560092a0edceabf55af32a40eae3-Paper-Datasets\\_and\\_Benchmarks.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/07f560092a0edceabf55af32a40eae3-Paper-Datasets_and_Benchmarks.pdf)