

SynIRgy: Synthetic Data and Simulation Synergy for Information Retrieval

Manel Slokom¹[0000–0002–9048–1906], Alejandro Bellogin²[0000–0001–6368–2510],
Mayank Singh³[0009–0005–5279–5049], and Andrea Barraza-Urbina⁴

¹ Vrije Universiteit & CWI, Amsterdam m.slokom@vu.nl

² Universidad Autónoma de Madrid, Spain alejandro.bellogin@uam.es

³ advicehub.ai, mayanksingh@advicehub.ai

⁴ Grubhub, apbarraza@gmail.com

Abstract. The increasing complexity of information access systems necessitates innovative evaluation methods beyond real-world user data, which is often biased, incomplete, or difficult to obtain due to privacy concerns. Simulation and synthetic data methods provide controlled and reproducible environments for training, evaluating, refining, and benchmarking algorithms in information retrieval and personalization.

This workshop aims to bring together researchers and practitioners to share insights, discuss best practices, and address pressing open problems by exploring synergies in the use of synthetic data and simulation for Information Retrieval (IR). Building on prior work in the community, we welcome methodological advances, ethical considerations, and practical applications of these approaches. The workshop also aims to produce a report summarizing key discussions, challenges, and future directions, serving as a reference point for researchers and practitioners working at the intersection of simulation, synthetic data, and IR. The workshop is intended to be an in-person event with some possibility of online presentations.

The workshop website is: <https://synirgy-workshop.github.io/>

Keywords: Synthetic data · Simulation · Search · Recommendation · GenAI · Algorithms · Evaluation.

1 Motivation and relevance to ECIR

While traditional IR research has long relied on static benchmark collections such as TREC and fixed recommendation datasets such as MovieLens and the Netflix Prize, these resources capture only a limited portion of the dynamics found in real deployment settings. In practice, information access systems, including search and recommendation, are interactive and evolve over time as users and systems adapt to each other [6]. As a result, static evaluation fails to reflect user adaptation, long-term effects, and phenomena such as filter bubbles, fairness, and causal feedback loops. Synthetic Data and Simulation Synergy for Information Retrieval (SynIRgy) aims to explore the increasing role of simulation and synthetic data in information retrieval (IR). The workshop aims to bridge this

gap by bringing together researchers and practitioners to discuss best practices, emerging methodologies, and open challenges in simulation-based IR research.

This workshop addresses recent breakthroughs in generative AI, reinforcement learning, and synthetic data generation. The rise of large language models has enabled new approaches for generating synthetic data, simulated user reviews, and interactive retrieval evaluation [4]. However, these advancements also raise concerns about data/evaluation reliability, biases [3,9], privacy [8], and ethical implications [2], which we aim to examine critically.

The SynIRgy Workshop: Synthetic Data and Simulation Synergy for Information Retrieval is designed to complement, rather than duplicate, the main conference themes by focusing specifically on the role of simulation and synthetic data in the development, evaluation, and deployment of information retrieval technologies. While synthetic data and simulation-based evaluation are becoming increasingly relevant across IR research, they remain underrepresented in major conferences. The European Conference on Information Retrieval (ECIR), for instance, explicitly lists as one of its core topics of interest: “*User aspects in IR, including information interaction, human-AI collaboration, contextualisation, personalisation, simulation, characterisation, and behaviours.*” This growing recognition points to the need for a dedicated space to bring together researchers who are developing, applying, or critically evaluating these methods. SynIRgy aims to be that space—a community hub where researchers can share methodologies, datasets, challenges, and results, while also discussing standards and open questions around the use of synthetic data and simulation in IR.

2 Workshop goals, vision, and expected outcomes

We outline below a set of key research questions that the call for papers aims to explore (but is not limited to), highlighting how the workshop topics *complement* the main conference:

1. *User behavior & interaction*
 - How can we model realistic user interactions in simulated IR, including the use of LLMs to generate behavior that reflects real-world users?
 - How do synthetic user profiles compare to real-world data, and how can their validity be assessed?
 - How can LLM-generated interactions help bridge the gap between simulated and real user behavior?
2. *Synthetic data generation*
 - What metrics can assess similarity between real and synthetic data?
 - How can we generate synthetic datasets that preserve utility while protecting privacy?
 - What trade-offs arise between privacy preservation and retrieval model accuracy?
 - How can LLMs be adapted to create high-quality synthetic datasets, and how can these be validated?

- Which aspects of real-world performance in information access systems can be evaluated using synthetic data?
- 3. *Algorithm development & evaluation*
 - What conditions must a simulation environment satisfy to support algorithm development?
 - What role can simulation play in reinforcement learning for interactive information access?
 - In which scenarios is synthetic data suitable for evaluation and algorithm development?
 - How can retrieval-augmented generation pipelines be simulated and evaluated?
- 4. *Bias and ethics*
 - How can synthetic data mitigate bias in information retrieval systems, and what trade-offs arise between fairness, utility, and privacy?
 - What ethical issues arise when using LLMs to generate synthetic data, such as transparency, accountability, and misuse?
 - How can user personas support diversity and inclusion in simulations?
 - What methods can limit the reinforcement of existing biases?
- 5. *Reproducibility & infrastructure*
 - How should reproducibility be defined for synthetic data and simulations?
 - How can FAIR principles be applied to improve transparency and comparability?
 - What role do open-source tools and cloud platforms play in large-scale simulation?

The workshop brings together researchers working on search, recommender systems, LLMs, and related information access technologies. It supports community building and the development of a shared vocabulary, while also identifying key challenges and open problems. A short report summarizes the discussions and outlines future directions. These outcomes provide a basis for later editions and related activities such as tutorials, survey papers, or a special issue.

Accepted papers are published in the Communications in Computer and Information Science (CCIS) proceedings series by Springer.

3 Format, Schedule and Structure

The SynIRgy workshop runs as a **half-day in-person event**, combining peer-reviewed research presentations with informal networking and focused discussion on synthetic data and simulation for information retrieval.

Schedule: We start the workshop with a brief welcome and introduction that situates synthetic data and simulation within current IR research. The program is organized into two main sessions: (1) simulation methods and (2) synthetic data and evaluation. Each session includes a keynote and oral presentations of full

paper contributions. Short papers, demos, negative results, and position papers are presented in a poster session. These formats support both mature research and early-stage ideas, allowing participants to engage with methodological proposals, empirical studies, and critical perspectives. Posters remain on display during the break to support continued discussion and interaction. The workshop closes with a summary of the main insights and concluding remarks.

Review process: Submissions are evaluated using a double-blind peer-review process. Each paper receives at least two expert reviews. Evaluation criteria include relevance to synthetic data, simulation, and IR (cf. Section 2), ability to stimulate discussion and collaboration, and clarity and quality of presentation.

4 Related workshops

SynIRgy builds on the momentum of prior workshops such as Sim4IA and Sim4IR [1,7], held as part of the SIGIR conference series, and SimuRec [5], held at ACM RecSys 2021. These events demonstrate sustained community interest in simulation based evaluation, user modeling, and synthetic data generation.

SynIRgy continues this line of inquiry with a more focused emphasis on information retrieval, large language models, user personas, and their interaction with synthetic data, algorithm design, and evaluation.

5 Intended audience

The workshop is open to all attendees of ECIR. We anticipate around 30–50 participants, primarily drawn from researchers and practitioners working in evaluation, synthetic data generation, simulation of short and long-term effects in user behavior, and related fields. To ensure broad visibility, the workshop is promoted via research mailing lists (e.g., SIGIR-announce, RecSys-list, UM, etc.), social media, and targeted outreach to researchers active in simulation and synthetic data.

6 Organizers

This section introduces the four organizers behind the SynIRgy workshop. The organizers are actively involved in the preparation and execution of the workshop at the conference.

- Manel Slokom is a postdoctoral researcher at the human-centered data analytics group at CWI and the quantitative data analytics group at Vrije Universiteit Amsterdam.
- Alejandro Bellogin is an Associate Professor at the Department of Computer Engineering of Universidad Autónoma de Madrid (Spain).
- Mayank Singh is Founder/CEO at advicehub.ai.
- Andrea Barraza-Urbina is a Senior Data Scientist at Grubhub.

Acknowledgments. This work has received funding by grant PID2022-139131NB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by “ERDF A way of making Europe.”

Manel’s research is supported by the AI, Media & Democracy Lab. For more information about the lab and its further activities, visit <https://www.aim4dem.nl/>.

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

References

1. Balog, K., Maxwell, D., Thomas, P., Zhang, S.: Sim4ir: The SIGIR 2021 workshop on simulation for information retrieval evaluation. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 2697–2698. ACM (2021). <https://doi.org/10.1145/3404835.3462821>, <https://doi.org/10.1145/3404835.3462821>
2. Bellogín, A., Grau, O., Larsson, S., Schimpf, G., Sengupta, B., Solmaz, G.: The eu ai act and the wager on trustworthy ai. *Communications of the ACM* **67**(12), 58–65 (2024). <https://doi.org/10.1145/3665322>, <https://doi.org/10.1145/3665322>
3. Burke, R., Kontny, J., Sonboli, N.: Synthetic attribute data for evaluating consumer-side fairness (2018)
4. Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., Pu, Y., Lei, Y., Chen, X., Wang, X., et al.: When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web* **27**(4), 42 (2024)
5. Ekstrand, M.D., Chaney, A., Castells, P., Burke, R., Rohde, D., Slokom, M.: Simurec: Workshop on synthetic data and simulation methods for recommender systems research. In: *Proceedings of the 15th ACM Conference on Recommender Systems*. p. 803–805. RecSys ’21 (2021). <https://doi.org/10.1145/3460231.3470938>
6. Gunawardana, A., Shani, G., Yogev, S.: Evaluating recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) *Recommender Systems Handbook*, pp. 547–601. Springer US (2022). https://doi.org/10.1007/978-1-0716-2197-4_15, https://doi.org/10.1007/978-1-0716-2197-4_15
7. Schaer, P., Kreutz, C.K., Balog, K., Breuer, T., Kruff, A.K.: Second SIGIR workshop on simulations for information access (sim4ia 2025). In: Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S. (eds.) *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*. pp. 4172–4175. ACM (2025). <https://doi.org/10.1145/3726302.3730363>, <https://doi.org/10.1145/3726302.3730363>
8. Slokom, M.: Comparing recommender systems using synthetic data. In: *Proceedings of the 12th ACM Conference on Recommender Systems*. p. 548–552. RecSys ’18, Association for Computing Machinery (2018). <https://doi.org/10.1145/3240323.3240325>
9. Tian, M., Ekstrand, M.D.: Estimating error and bias in offline evaluation results. In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. pp. 392–396. CHIIR ’20, Association for Computing Machinery (2020). <https://doi.org/10.1145/3343413.3378004>