Evaluating Information Retrieval Models Along Time: The LongEval Lab at CLEF 2026

Timo Breuer[1], Matteo Cancellieri[2], Alaa El-Ebshihy[3], Maik Fröbe[4], Petra Galuščáková[5], Lorraine Goeuriot[6], Gabriel Iturra-Bocaz[5], Jüri Keller[1], Petr Knoth[2], Andreas Konstantin Kruff[1], Philippe Mulhem[6], Florina Piroi[3], David Pride[2], Philipp Schaer[1(✉)], and Didier Schwab[6]

[1] TH Köln - University of Applied Sciences, Cologne, Germany
philipp.schaer@th-koeln.de
[2] The Open University, Milton Keynes, UK
[3] TU Wien, Vienna, Austria
[4] Friedrich-Schiller-Universität Jena, Jena, Germany
[5] University of Stavanger, Stavanger, Norway
[6] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France

Abstract. Many components of information retrieval systems evolve over time. The LongEval Lab aims to provide a benchmark setting to the longitudinal evaluation of IR models. At its fourth edition, LongEval we focus on scholarly search and scholarly user models. We describe in this paper the tasks that are planned for the 2026 lab, the data necessary for each of the tasks, as well as the choice of evaluation activities.

Keywords: Longitudinal Evaluation · Continuous Evaluation · Temporal IR

1 Introduction

The majority of evaluation initiatives in information retrieval focus on evaluating systems at a fixed point in time. However, data and users' behavior and expectations evolve over time. Therefore, to maintain good performance, systems must adapt to these variations. In some cases, the performance of IR systems can drop over time as the patterns observed in data change, e.g., due to linguistic and societal changes [3]. The performance drop is more pronounced when the test data is further away in time from training data [12,22]. Similarly, it has been shown that a deep neural network-based IR is dependent on the consistency between the train and test data [24].

The aim of the LongEval lab (this year in its fourth edition [1,2,6]) is to develop models that mitigate performance drops over time. We provide participants with training data distant in time from testing and un-annotated data

Authors ordered alphabetically.

R. Campos et al. (Eds.): ECIR 2026, LNCS 16486, pp. 345–353, 2026.
https://doi.org/10.1007/978-3-032-21321-1_45

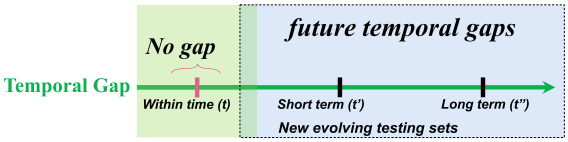

Fig. 1: Global framework for the LongEval Tasks.

from the testing time period. Our test scenarios include documents that evolve over the time, queries that are not known a priori, and relevance judgments that are non-binary and potentially not stable over time. We look at the temporal generalisability of IR systems trained and tested on data acquired at time t (Fig. 1) to their performance when operating on data acquired at time t' (shortly after time t), and at time t'' (a long period after time t).

The LongEval Lab is running since 2023, with a varying number of participants each year. In these years, the lab proposed two tasks per year, one involving classification, the other retrieval. For next year (2026), we plan a total of 4 tasks, widening the scope of long-term IR to new dynamics beyond documents, topics, and qrels, closer to evolving user behavior with user simulation tasks. Additionally, we newly experiment with longitudinal aspect in Retrieval Augmented Generation (RAG). Task 1 is, by now, a classical ad-hoc search task where systems are evaluated at varying points in time. Task 2 aims to support the derivation of Cranfield-style evaluation corpora from user behaviour observed over time by aiming to extract formalized topics from click logs. Task 3 is dedicated to modelling domain-specific user behaviour and different user types for use in user simulations. Task 4 is dedicated to the study of RAG behaviour when dealing with evolution of documents, focusing on evolving and conflicting information, still expecting the RAGs to generate relevant results according to the documents retrieved from different temporal collections.

2 LongEval Data

The underlying dataset from where each task will extract queries, topics, and relevance assessments, is acquired from the CORE[1] [21] collection of scholarly documents. To our knowledge, CORE [20] is currently the largest aggregated collection of Open Access full text scholarly documents. CORE Search provides a web UI for users to query the entire database of scholarly documents with over a million searches per month. We will use the CORE Apache server logs to create datasets necessary for the tasks (i.e. documents, queries, qrels, and user sessions). These logs contain search and click information:

Search Information includes i) `search_id`–a unique identifiers for the search; ii) `query`–search query; iii) `serp`–documents identifiers of returned results[2] (example 1.1);

[1] CORE (COnnecting REpositories) https://core.ac.uk/.
[2] For LongEval we only consider the first ten results.

Click Information records from the user interaction with the CORE web-interface, for each click: i) `uid`–a unique (anonymous) identifier for individual user session; ii) the `trackId`–unique identifier of clicked document; iii) `type`–click type (e.g. author information, download PDF article, example 1.2).

Example 1.1: *Search query and results*

```
{"date": "2025-01-10␣00:00:05",
 "search_id": "83ed8653d71",
 "serp": [ 7233367, 85590132, 145375620, 23053234],
 "query": "Methodology␣definition␣and␣importance␣in␣research␣
    methodology"}
```

Example 1.2: *Search query and results*

```
{"date": "2025-01-10␣00:00:05",
 "search_id": "83ed8653d71",
 "serp": [ 7233367, 85590132, 145375620, 23053234],
 "query": "Methodology␣definition␣and␣importance␣in␣research␣
    methodology"}
```

3 Task Descriptions

In the following we give details on the four LongEval Lab tasls at CLEF 2026. We explain the general idea and objective of the task, the corresponding artefacts (e.g., queries, qrels, etc.) and the evaluation strategy for the specific task.

Task 1. LongEval-Sci: Ad-Hoc Scientific Retrieval

Description: The task aims to further encourage the development of IR systems able to handle temporal data evolution. The IR systems are expected to be persistent in their retrieval efficiency over time, as the test collection evolves. The participants will collect retrieval runs on various snapshots of the test collection.
Data: The data for this task is a set of scientific documents and queries from the CORE search engine (Sect. 2), following a similar acquisition process as for the previous LongEval iterations [1,2,6,14]. We compute discrete relevance assessments using a simplified Dynamic Bayesian Network (sDBN) Click Model [8,9]. As in previous years, we additionally provide convenient access to the dataset through the ir_datasets_longeval extension [18]. As of now, the data available to the participants is as follows:

1. Two training snapshots from the 2025 lab iteration (queries, documents, qrels, time intervals 2024-11 and 2025-01).
2. Two test snapshots: one acquired during a time interval t' shortly after t for evaluating the **short-term persistence**, and one acquired long after t during a time interval t'', for **long-term persistence** evaluation.

In total, the data contains about 4 million documents and one thousand queries. **Evaluation:** Submissions will be evaluated in terms of: *Effectiveness* per test snapshot, and *Robustness* across snapshots. *Effectiveness* will be measured with nDCG scores. *Robustness* will be assessed with: (1) Relative Improvement (RI) – measures relative change in effectiveness compared to the first snapshot, (2) Delta Relative Improvement (DRI) – measures relative improvement in relation to a reference system (e.g. BM25), and (3) Effect Ratio (ER) – measures degree of reproduction of the effectiveness [4,6,15,16].

Task 2. LongEval-TopEx: Topic Extraction From Query Logs

Description: Retrieval systems infer which documents are relevant to an information need either from (1) usage data or (2) expert judgments [7]. In search engine production deployments these variants can complement each other, e.g., when retrieval systems are first evaluated with available expert judgments so that only promising candidates are then evaluated under usage data [19]. Our goal for Task 2 is to also enable similar synergies for LongEval submissions.
Data: The input is the training set from time interval t. Participants can use the relevance judgments derived from the observed usage data for each query to create a TREC-style description and narrative for each training query. The extracted topic consists of the training query and the generated description and narrative. The query describes what users submit to the search engine, whereas the description formalizes what the users actually did mean and the narrative formalizes what makes documents relevant respectively not relevant. To obtain relevance judgments, we will build a top-10 judgment pool for all retrieval runs submitted for the test sets for Task 1 that overlap with the training queries. We then will annotate this top-10 pool for all test sets with multiple large language model relevance assessors [10,11,25,26] prompted with the query and the generated description and narrative that formalize what the search engine users did mean and what will be considered relevant or not. Thereby, we obtain one alternative set of qrels per extracted topic and LLM relevance assessor.
Evaluation: Our evaluation aims to detect "good" extracted topics that are (1) aligned with observed usage data, (2) allow to distinguish retrieval systems, and (3) provide a clear formalization of the information need to derive what documents will be relevant or not. Each of the three dimensions is needed to ensure that subsequent evaluations are meaningful. We will assess the quality of each extracted topic for each of the three dimensions using the LLM-generated relevance assessments and the retrieval runs submitted to Task 1:

1. Alignment: We calculate the annotator-agreement between the qrels from the click logs and the LLM-generated qrels from the extracted topic for each test time interval.
2. Distinguishability: We calculate how well nDCG scores created from the LLM-generated qrels of the extracted topic can distinguish the runs submitted to Task 1.

3. Clarity: We will calculate the annotator agreement when (potentially different) large language models repeatedly create alternative qrels for the same extracted topic.

Similarly to Task 1, we compute the relative drop on the three dimensions for the short- and long-term test sets. Topics extracted in Task 2 are Task 3 inputs.

Task 3. LongEval-USim: User Simulations

Description: The CORE-based dataset in LongEval differs from typical test collections not only by the temporal characteristics described earlier, but also by the inclusion of user interactions in the academic search task. In these search tasks, we observe patterns that are less common in other (web) search tasks, such as journal runs, numerous Boolean block searches, or faceted search. These characteristics may lead to interaction behaviors that have not been previously explored enabling a wide range of interesting scenarios.

In this task, we invite participants to model different user behaviors and types for user simulations, with a particular focus on query formulations. This approach was explored in the SIGIR 2025 Workshop on Simulations in Information Access (Sim4IA[3]) [5], in the form of a micro-shared task, where the evaluation of query (re-)formulation simulations was tested out.

From the original CORE search log files we extract a set of interaction sessions, similar to the ones in Sim4IA. With the `search_id`, a fingerprint `uid`, and heuristics, we extract sessions from the search logs. These sessions enable us to track user interactions over (short) time spans, including queries, SERPs, and clicks. Participants in this task will get a pre-filtered session logs excerpts. From here, they should predict the next query in the sequence of interactions. The participants can detect and train specific user models on historic interaction logs and test those models on the test data in the final LongEval round.

To simplify this task, we provide a pre-configured SimIIR v3 environment that will allow off-the-shelf user simulations by configuring user models or simple interaction steps. Later, these simulators can be extended to a scenario that includes click or stopping decisions. Participants can also utilize the synthesized topics from Task 2, as these can serve as additional input for this task. Instead of relying solely on session logs, the topic and encoded context within these can provide more information on user intentions compared to queries alone.

Relevance Computation and Evaluation Metrics: As participants are to submit queries, we will compare the submitted queries to the queries from the session logs. We can look at two main aspects: (1) the inability to distinguish between simulated and real data, and (2) the performance prediction capabilities of a simulation. We measure the differences in the generated queries themselves or the differences in the output results by comparing them to the interactions in the CORE log files. Using reproducibility measures [17], we identify the simulators that are closest (or best reproduce) the original set of interactions. In the

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

LongEval setting, we can additionally compare the interactions generated by the simulator based on the development of topics over time. A simulator performing well at one point in time does not necessarily perform well at another.

Task 4. LongEval-RAG: Retrieval Augmented Generation (RAG)

Description and Data: This task aims to evaluate the ability of RAG systems to retrieve correct information when this information evolves over the time. In such case, the connection between the retrieval step of a RAG and its generation step should incorporate temporal awareness of the retrieved documents.

The LongEval RAG task will provide a set of queries, for which the correctness of the answer depends on the temporal aspect. Queries will be selected such that the relevant documents either changed over time or are documents from different temporal collections with conflicting information. Typical examples of queries are related to evolution and/or contradictions of specific research topics, e.g. "What are the main technical evolutions of attention models between the short-term and the long-term datasets?" The participants will create RAG systems that use both short- and long-term collections from the CORE data set (explained in Task 1) and submit, for each query, a generated response along with the list of support documents or passages. Time information is provided explicitly in the metadata of the documents (creation, publication, update times). We follow the line of exploring time-sensitive information in RAG [13,23,27] focusing on specific topics of the scientific literature domain and on shorter time frames where more complex information needs are the case.
Evaluation: We will evaluate the submissions (i.e. the generated answers) manually using Answer Relevancy and Faithfulness metrics. The Answer Relevancy measures to what extent the generated results addresses the user's query. The Faithfulness metric measures how accurately the answer reflects the information in the source documents. Both dimensions will be weighted equally to assess to which extent good results benefit from the retrieved documents.

4 Conclusion

The proposed next iteration of the LongEval Lab at CLEF in 2026 aims to evaluate information access systems within dynamic, temporal environments. Task 1 will continue the scientific search task, evaluated at multiple points in time. Task 2 aims to generate TREC-style topics from interaction logs to facilitate automatic relevance judgments. Task 3 focuses on predicting the next query in a session based on a simulated users. Task 4 invites participants to test their RAG systems in the evolving environment on temporal depended queries. Through these tasks we aim to improve the state of knowledge on and the robustness of information access systems in dynamic, temporal settings. All information, including further details such as the anticipated schedule will be made available on the LongEval website[4].

[4] https://clef-longeval.github.io/.

Acknowledgments. This work is partially funded by Deutsche Forschungsgemeinschaft (DFG) under grant numbers 509543643 and 407518790, 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.

References

1. Alkhalifa, R., et al.: Overview of the CLEF-2023 LongEval Lab on Longitudinal Evaluation of Model Performance. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proc. of the 14th International Conference of the CLEF Association (CLEF 2023). LNCS. Springer (2023)
2. Alkhalifa, R., et al.: Overview of the CLEF 2024 LongEval Lab on Longitudinal Evaluation of Model Performance. In: Proc. of the 15th International Conference of the CLEF Association (CLEF 2024) (2024)
3. Alkhalifa, R., Zubiaga, A.: Capturing stance dynamics in social media: open challenges and research directions. Int. J. Digital Humanities, 1–21 (2022)
4. Breuer, T., Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Schaer, P., Soboroff, I.: How to measure the reproducibility of system-oriented ir experiments. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 349–358. SIGIR '20. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3397271.3401036
5. Breuer, T., et al.: Report on the 1st workshop on simulations for information access (sim4ia 2024) at SIGIR 2024. ACM SIGIR Forum **58**(2) (December 2024)
6. Cancellieri, M., et al.: Longeval at clef 2025: Longitudinal evaluation of ir systems on web and scientific data. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9–12, 2025, Proceedings. p. 363–387. Springer, Heidelberg (2025). https://doi.org/10.1007/978-3-032-04354-2_20
7. Chapelle, O., Joachims, T., Radlinski, F., Yue, Y.: Large-scale validation and analysis of interleaved search evaluation. ACM Trans. Inf. Syst. **30**(1), 6:1–6:41 (2012). https://doi.org/10.1145/2094072.2094078
8. Chapelle, O., Zhang, Y.: A dynamic bayesian network click model for web search ranking. In: Proceedings of the 18th International Conference on World Wide Web, WWW '09, pp. 1–10. Association for Computing Machinery, New York, April 2009. https://doi.org/10.1145/1526709.1526711
9. Chuklin, A., Markov, I., Rijke, M.d.: Click models for web search. Synthesis Lectures on Information Concepts, Retrieval, and Services **7**(3), 1–115 (2015). https://doi.org/10.2200/S00654ED1V01Y201507ICR043
10. Faggioli, G., et al.: Perspectives on large language models for relevance judgment. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR (2023). https://doi.org/10.1145/3578337.3605136
11. Faggioli, G., Dietz, L., Clarke, C.L.A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., Wachsmuth, H.: Who determines what is relevant? humans or AI? why not both? Commun. ACM **67**(4), 31–34 (2024). https://doi.org/10.1145/3624730

12. Florio, K., Basile, V., Polignano, M., Basile, P., Patti, V.: Time of your hate: The challenge of time in hate speech detection on social media. Appl. Sci. **10**(12), 4180 (2020)
13. Gade, A., Jetcheva, J.: It's about time: Incorporating temporality in retrieval augmented language models. arXiv preprint arXiv:2401.13222 (2024)
14. Galuscáková, P., et al.: Longeval-retrieval: French-english dynamic test collection for continuous web search evaluation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3086–3094. SIGIR '23. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3539618.3591921
15. Keller, J., Breuer, T., Schaer, P.: Evaluation of temporal change in IR test collections. In: Oosterhuis, H., Bast, H., Xiong, C. (eds.) Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024, pp. 3–13. ACM (2024). https://doi.org/10.1145/3664190.3672530
16. Keller, J., Breuer, T., Schaer, P.: Leveraging prior relevance signals in web search. In: CLEF (Working Notes). CEUR Workshop Proceedings, vol. 3740, pp. 2396–2406. CEUR-WS.org (2024)
17. Keller, J., Breuer, T., Schaer, P.: Replicability measures for longitudinal information retrieval evaluation. In: Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Di Nunzio, G.M., Soulier, L., Galuščáková, P., García Seco de Herrera, A., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 215–226. Springer, Cham (2024)
18. Keller, J., Fröbe, M., Hendriksen, G., Alexander, D., Potthast, M., Schaer, P.: Simplified longitudinal retrieval experiments: A case study on query expansion and document boosting. In: Carrillo-de-Albornoz, J., de Herrera, A.G.S., Gonzalo, J., Plaza, L., Mothe, J., Piroi, F., Rosso, P., Spina, D., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings. LNCS, vol. 16089, pp. 117–127. Springer (2025). https://doi.org/10.1007/978-3-032-04354-2_8
19. Kharitonov, E.: Using interaction data for improving the offline and online evaluation of search engines. Ph.D. thesis, University of Glasgow, UK (2016), theses.gla.ac.uk/7750/
20. 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. Sci. Data **10**(1), 366 (2023). https://doi.org/10.1038/s41597-023-02208-w
21. Knoth, P., Zdrahal, Z.: CORE: three access levels to underpin Open Access. D-Lib Magazine **18**(11/12) (2012)
22. Lukes, J., Søgaard, A.: Sentiment analysis under temporal shift. In: Balahur, A., Mohammad, S.M., Hoste, V., Klinger, R. (eds.) Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 65–71. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-6210, https://aclanthology.org/W18-6210/
23. Piryani, B., Abdullah, A., Mozafari, J., Anand, A., Jatowt, A.: It's high time: A survey of temporal information retrieval and question answering. arXiv:2505.20243 (2025)
24. Ren, R., et al.: A thorough examination on zero-shot dense retrieval. In: Findings of the Association for Computational Linguistics: EMNLP (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.1057

25. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proc. of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1930–1940 (2024). https://doi.org/10.1145/3626772.3657707
26. Upadhyay, S., Pradeep, R., Thakur, N., Craswell, N., Lin, J.: UMBRELA: umbrela is the (open-source reproduction of the) bing relevance assessor. CoRR (2024)
27. Wu, F., et al.: Time-sensitve retrieval-augmented generation for question answering. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 2544–2553. CIKM '24. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3627673.3679800