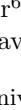


LongEval at CLEF 2025: Longitudinal Evaluation of IR Model Performance

Matteo Cancellieri¹ , Alaa El-Ebshihy^{2,3} , Tobias Fink^{2,3} ,
 Petra Galuščáková⁴ , Gabriela Gonzalez-Saez⁵ , Lorraine Goeuriot⁵ ,
 David Iommi² , Jüri Keller⁶ , Petr Knoth¹ , Philippe Mulhem⁵ ,
 Florina Piroi^{2,3} , David Pride¹ , and Philipp Schaefer⁶ 

¹ The Open University, Milton Keynes, UK***

² Research Studios Austria, Data Science Studio, Vienna, Austria

³ TU Wien, Austria

⁴ University of Stavanger, Stavanger, Norway

⁵ Univ. Grenoble Alpes, CNRS, Grenoble INP[†], LIG, Grenoble, France

⁶ TH Köln - University of Applied Sciences, Cologne, Germany

Abstract. This paper presents the third edition of the LongEval Lab, part of the CLEF 2025 conference, which continues to explore the challenges of temporal persistence in Information Retrieval (IR). The lab features two tasks designed to provide researchers with test data that reflect the evolving nature of user queries and document relevance over time. By evaluating how model performance degrades as test data diverge temporally from training data, LongEval seeks to advance the understanding of temporal dynamics in IR systems. The 2025 edition aims to engage the IR and NLP communities in addressing the development of adaptive models that can maintain retrieval quality over time in the domains of web search and scientific retrieval.

Keywords: Longitudinal Evaluation · Temporal Persistence · Temporal Generalisability · Temporal Change · Information Retrieval

1 Introduction

Information Retrieval (IR) systems are constantly challenged by the evolving search setting [7]. The foundational dataset is updated regularly, users develop new information needs, and their perception of relevance varies over time [1,13,14]. These temporal dynamics have strong implications for the aspired goal of maintaining high retrieval effectiveness over time. It is known that search is sensitive to temporal factors [8,12] and that incorporating information from previous points in time can be highly effective [2,10]. Additionally, in modern IR, the systems are updated or retrained often, making them a dynamic component themselves in the evolving search setting.

*** Authors ordered alphabetically

[†] Institute of Engineering Univ. Grenoble Alpes.

While these temporal factors strongly influence retrieval effectiveness, they are often overlooked or abstracted on purpose in conventional evaluations. The results from the previous iterations of the lab showed that the ranking of systems varies over time and that the most effective system is not necessarily also the system that performs the most consistently [3,4,9]. This shows how the experimental setup strongly influences the measured effectiveness.

In this third iteration, the LongEval Lab at CLEF (Conference and Labs of the Evaluation Forum¹) continues to explore the temporal dynamics in IR [3,4]. This includes the potential and limitations of temporal relevance signals for ranking, the temporal robustness of systems, and novel evaluation methods that factor in time. Thus, this lab sensitises researchers to uncertain and temporally limited validity of conventional evaluation results in IR. Considering the temporal dimension provides a new perspective on search and ultimately leads to a more holistic view on the retrieval problem.

This year, the lab provides a unique test bed comprising two evolving test collections. They cover the established retrieval scenarios of Web search and scientific retrieval, which have different goals and distinct dynamics. Participants are invited to submit retrieval runs to two tasks that address these dynamics.

2 Description of the LongEval 2025 Tasks

Until 2024, LongEval’s information retrieval tasks focused only on retrieving Web documents. In 2025, we enlarge the scope of LongEval as we want to study the potential differences, if any, between two retrieval contexts. The Web retrieval context is a classical Web case, in which very short queries are asked and the very top documents are considered. The scientific search contains potentially longer queries, and users are looking deeper in the result lists.

Both LongEval tasks use a sequence of datasets collected at different points in time. The (time) distances between two datasets are called “lags.” The IR systems that participants design are evaluated on the different lags, computing the differences in evaluation metrics between lags.

2.1 Task 1: LongEval-WebRetrieval

This task is a continuation of the Retrieval tasks from the previous two LongEval iterations. It uses evolving Web data to evaluate IR systems longitudinally: the systems are expected to be persistent in their retrieval effectiveness over time. The systems are evaluated on monthly several snapshots of documents and queries (lags), derived from real data acquired from a French Web search engine, Qwant². In this iteration, we evaluate the same IR systems on a sequence of test collections acquired after the last sample of the train collection.

¹ <https://www.clef-initiative.eu/>

² Most queries and documents are originally in French. However, English translations are additionally provided as well.

Lessons learned from the 2024 LongEval edition In 2024, the LongEval lab used two test environments, called Lag6 and Lag8. That is, we evaluated IR systems on test data that was 6 and 8 months newer than the data the systems were trained on. 28 teams registered for the second edition of the LongEval Retrieval task, and 14 teams submitted a total of 73 retrieval experiments. The number of teams that submitted is the same as the in the first edition, indicating that the task maintained its popularity. We observed the following [4]:

Approaches: Compared to 2023, some participants did use the temporal aspect of the LongEval test collection, incorporating past relevance signals as query reformulation [10]. The most effective approaches rely on multi-stage retrieval, using BM25 as a first-stage retrieval and neural-based or LLMs-based models for re-ranking.

Robustness: System rankings were computed with respect to retrieval performances on Lag6 and Lag8 (nDCG scores), and according to the changes in their performance between the two lags (Relative nDCG Drop, RnD, see section 3). The ranking correlation between the system rankings using nDCGs scores on Lag6 and Lag8 was high, while the ranking correlation between system rankings using RnD scores was low. This points to the fact that systems that are more robust to the evolution of the test collection were not the top-performing ones. This finding is consistent with the findings from Longeval 2023.

Data Preparations: As expected in an evolving collection of documents, there are large overlaps in documents and queries between the document snapshots across test collections and between test and train collections. This overlap was not easily identifiable with the released document and query IDs.

Based on these observations, in the LongEval 2025 lab we enlarge the training collection with additional snapshots that will allow fine-grained analysis of changes in the data collection from one snapshot to another. Similarly, the test environments will be composed of several consecutive snapshots, allowing for a deeper understanding of data evolution over time. Additionally, the LongEval test collection, in its totality, will be improved in terms of document and query identifiers, and its description.

Data. The 2025 dataset includes all data from the 2023 and 2024 editions, along with newly added, previously unreleased months. The training dataset consists of 18 million French documents (June 2022 - February 2023, translated to English) and 9,000 queries with computed relevance assessments based on a simplified Dynamic Bayesian Network (sDBN) Click Model [5,6], acquired from real users of the French Qwant search engine. The test collection spans 7 months of data (March 2023 - August 2023). Each month can be seen as a snapshot (lag). Each of these test collections is similar in structure to the train set, except that they do not contain any relevance assessments. Participants are expected to submit runs for each lag, using the same system trained only on the training dataset. Additionally, human-annotated data from the previous iteration will be used to

evaluate system performance. The total data for this task will be composed of 30 million documents and 15,000 queries, provided by Qwant³. Each document set will have a release time stamp, with the first set (in chronological order) being the training data.

2.2 Task 2: LongEval-SciRetrieval

The second task of the LongEval 2025 Lab is similar to the first task, and aims to examine how IR systems' effectiveness changes over time, when the underlying document collection changes, where the documents are scientific publications. The documents that will make the dataset for this task are acquired from the CORE⁴ collection of scholarly documents. To our knowledge, CORE [11] is currently the largest aggregated collection of Open Access full text scholarly documents. CORE provides a range of services built on top of this content and these services are currently used by over 30 million unique users each month. CORE Search provides a web UI for users to query the entire database of scholarly documents. This service registers over one million searches each month.

As can be seen from the sample results shown in Figure 1 for the query “open science”, the user has multiple options in terms of where to click for each individual search result: the PDF link (left hand image), the paper title (in brown color) and the author(s) name (underlined). Similarly to Task 1, we will use the click information to create relevance assessments for the test collection.

For compiling the dataset for the LongEval-SciRetrieval task, we create a specific pipeline for capturing user queries, the results of these queries, and the user interactions with the displayed results from the CORE Apache server logs. We then process the data to remove any traffic that was generated by bots so only searches conducted by human users remain. Using this pipeline, the dataset for this task is extracted and consists of two main components that contain both the search and click information:

Search Information (see example 1.1) includes i) unique (anonymous) identifiers for individual user session; ii) search query; iii) returned results⁵ ;

Click Information (see example 1.2) records, for each click, i) a unique (anonymous) identifier for individual user session; ii) the link that was clicked in the results list; iii) the position of clicked link in results list.

Since this is the first time this task is organized, the number of dataset lags is lower than those used in the first task. We aim to release two training datasets and one or two test datasets.

³ Qwant search engine: <https://www.qwant.com/>

⁴ CORE (COnnecting REpositories) <https://core.ac.uk/>

⁵ For LongEval we only consider the first ten results.

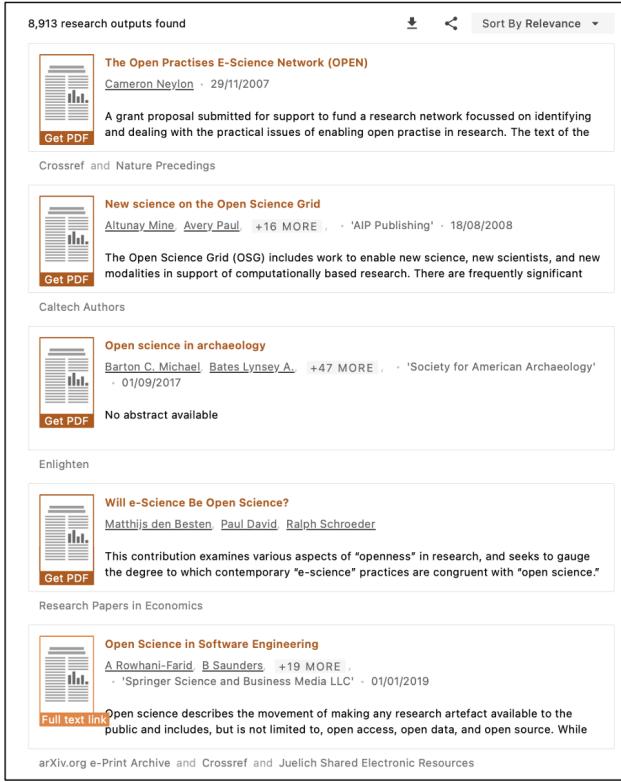


Fig. 1: Sample result from CORE for a search for “open science”

Example 1.1: Search query and returned results

```
{
  "search_id": "e5385afdaedcc11f9a6ba092cb613f27",
  "query": "inclusive\u2022codesign",
  "serp": [
    "https://core.ac.uk/works/156973302",
    "https://core.ac.uk/works/8080300",
    "https://core.ac.uk/works/45177790",
    "https://core.ac.uk/works/148924361",
    "https://core.ac.uk/works/149405922",
    "https://core.ac.uk/works/149554048",
    "https://core.ac.uk/works/15034633",
    "https://core.ac.uk/works/150382029",
  ],
  "date": "2024-12-03T07:11:11.000Z"
}
```

Example 1.2: *Unique click information*

```
{
  "search_id": "e5385afdaedcc11f9a6ba092cb613f27",
  "url": "https://core.ac.uk/works/156973302",
  "serp": "0",
  "date": "2024-12-03T07:11:11.000Z"
}
```

3 Evaluation

Since the two retrieval tasks are very similar in design, differing in the type of data provided to the users (Web documents vs. scientific publications), the evaluation is, conceptually the same. Namely, the submitted runs will be mainly evaluated in two ways:

1. **nDCG** scores calculated on each lag test set provided for the sub-tasks. Such a classical evaluation measure is consistent with Web search, for which the discount emphasises the ordering of the top results.
2. **Relative nDCG Drop (RnD)** measured by computing the difference between nDCG values between different lag datasets. Such values will allow to check the robustness of systems against the evolution of the data.

These measures assess the quality of systems and also their robustness against the data (queries/documents) evolution along time: a system that has good results using nDCG, and also good results according to the RnD measure is considered to be able to cope with the evolution over time of the Information Retrieval collection.

4 LongEval Timeline

Information and updates about the LongEval Lab, and the submission guidelines, will be communicated mainly through the lab's website⁶. The training data release for both tasks is scheduled for February 2025, and the test data for end of March 2025. In concordance with the CLEF schedule, the participant submission deadline is planned for May 2025, with the evaluation results to be released in June 2025. As in the previous iterations, we invite participants to the LongEval workshop to be organized as part of the CLEF 2025 conference. The workshop is open to researchers interested in the temporal persistence of IR models, and we welcome submissions that are not part of the shared task but deal with this topic.

⁶ <https://clef-longeval.github.io>

Acknowledgements

This work is supported by the ANR Kodicare bi-lateral project, grant ANR-19-CE23-0029 of the French Agence Nationale de la Recherche, and by the Austrian Science Fund (FWF, grant I4471-N). This work has been using services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

1. Adar, E., Teevan, J., Dumais, S.T., Elsas, J.L.: The web changes everything: understanding the dynamics of web content. In: WSDM. pp. 282–291. ACM (2009)
2. Alexander, D., Fröbe, M., Hendriksen, G., Schlatt, F., Hagen, M., Hiemstra, D., Potthast, M., de Vries, A.P.: Team openwebsearch at CLEF 2024: Longeval. In: CLEF (Working Notes). CEUR Workshop Proceedings, vol. 3740, pp. 2304–2313. CEUR-WS.org (2024)
3. Alkhalifa, R., Bilal, I.M., Borkakoty, H., Camacho-Collados, J., Deveaud, R., El-Ebshihy, A., Anke, L.E., Sáez, G.G., Galuščáková, P., Goeuriot, L., Kochkina, E., Liakata, M., Loureiro, D., Mulhem, P., Piroi, F., Popel, M., Servan, C., Madabushi, H.T., Zubiaga, A.: Overview of the CLEF-2023 longeval lab on longitudinal evaluation of model performance. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14163, pp. 440–458. Springer (2023)
4. Alkhalifa, R., Borkakoty, H., Deveaud, R., El-Ebshihy, A., Espinosa-Anke, L., Fink, T., Galuščáková, P., Gonzalez-Saez, G., Goeuriot, L., Iommi, D., Liakata, M., Madabushi, H.T., Medina-Alias, P., Mulhem, P., Piroi, F., Popel, M., Zubiaga, A.: Overview of the clef 2024 longeval lab on longitudinal evaluation of model performance. 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. 208–230. Springer Nature Switzerland, Cham (2024)
5. 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. pp. 1–10. WWW '09, Association for Computing Machinery, New York, NY, USA (Apr 2009)
6. Chuklin, A., Markov, I., Rijke, M.d.: Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* **7**(3), 1–115 (Jul 2015)
7. Dumais, S.T.: Putting searchers into search. In: SIGIR. pp. 1–2. ACM (2014)
8. Kanhabua, N., Blanco, R., Nørvåg, K.: Temporal information retrieval. *Found. Trends Inf. Retr.* **9**(2), 91–208 (2015)
9. 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)

10. 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)
11. 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)
12. Liu, Y.A., Zhang, R., Guo, J., de Rijke, M., Fan, Y., Cheng, X.: Robust neural information retrieval: An adversarial and out-of-distribution perspective (2024), <https://arxiv.org/abs/2407.06992>
13. Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E., Wang, L.L., Hersh, W.R.: Searching for scientific evidence in a pandemic: An overview of trec-covid. *Journal of Biomedical Informatics* **121**, 103865 (2021)
14. Tikhonov, A., Bogatyy, I., Burangulov, P., Ostroumova, L., Koshelev, V., Gusev, G.: Studying page life patterns in dynamical web. In: SIGIR. pp. 905–908. ACM (2013)