Extended Abstract of LongEval at CLEF 2025: Longitudinal Evaluation of IR Systems on Web and Scientific Data

Notebook for the LongEval Lab at CLEF 2025

Matteo Cancellieri[1], Alaa El-Ebshihy[2,3], Tobias Fink[2,3], Maik Fröbe[4], Petra Galuščáková[5], Gabriela Gonzalez-Saez[6], Lorraine Goeuriot[6], David Iommi[2], Jüri Keller[7], Petr Knoth[1], Philippe Mulhem[6], Florina Piroi[2,3], David Pride[1] and Philipp Schaer[7]

[1]*The Open University, Milton Keynes, UK*[1]
[2]*Research Studios Austria, Data Science Studio, Vienna, Austria*
[3]*TU Wien, Austria*
[4]*Friedrich-Schiller-Universität Jena, Germany*
[5]*University of Stavanger, Stavanger, Norway*
[6]*Univ. Grenoble Alpes, CNRS, Grenoble INP*[2]*, LIG, Grenoble, France*
[7]*TH Köln - University of Applied Sciences, Cologne, Germany*

Abstract

The LongEval lab focuses on the evaluation of information retrieval systems over time. Two datasets are provided that capture evolving search scenarios with changing documents, queries, and relevance assessments. Systems are assessed from a temporal perspective—that is, evaluating retrieval effectiveness as the data they operate on changes. In its third edition, LongEval featured two retrieval tasks: one in the area of ad-hoc web retrieval, and another focusing on scientific article retrieval. We present an overview of this year's tasks and datasets, as well as the participating systems. A total of 19 teams submitted their approaches, which we evaluated using nDCG and a variety of measures that quantify changes in retrieval effectiveness over time.

Keywords

Longitudinal Evaluation, Temporal Persistence, Temporal Generalisability, Temporal Change, Information Retrieval

1. Introduction

Information Retrieval (IR) systems are challenged by evolving search settings, where document collections, user needs, and relevance judgments evolve continuously [1, 2, 3, 4]. However, most evaluation test collections are static, ignoring the impact of temporal changes. LongEval addresses this gap by introducing evolving test collections and measuring performance over time. Previous editions of the lab showed that retrieval effectiveness can vary across time, and that the most effective system is not always the most consistent one [5, 6, 7]. The lab's main goals are to (i) assess how the performance of retrieval systems changes over time as test collections evolve, and (ii) propose methods that mitigate performance drop by making models more robust over time.

[1]Authors ordered alphabetically
[2]Institute of Engineering Univ. Grenoble Alpes.
CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
0000-0002-9558-9772 (M. Cancellieri); 0000-0001-6644-2360 (A. El-Ebshihy); 0000-0002-1045-8352 (T. Fink); 0000-0002-1003-981X (M. Fröbe); 0000-0001-6328-7131 (P. Galuščáková); 0000-0003-0878-5263 (G. Gonzalez-Saez); 0000-0001-7491-1980 (L. Goeuriot); 0000-0002-4270-5709 (D. Iommi); 0000-0002-9392-8646 (J. Keller); 0000-0003-1161-7359 (P. Knoth); 0000-0002-3245-6462 (P. Mulhem); 0000-0001-7584-6439 (F. Piroi); 0000-0002-7162-7252 (D. Pride); 0000-0002-8817-4632 (P. Schaer)

2. Tasks and Datasets

In 2025, LongEval featured two retrieval tasks:
WebRetrieval: Based on monthly snapshots from the French search engine Qwant[1], this task evaluates how systems trained on earlier data perform on future web collections. The dataset includes: (1) a training set of 19 million documents and 119,341 queries (June 2022–February 2023), and (2) a test set of 14 million documents and 63,416 queries (March–August 2023).
SciRetrieval: A new task using snapshots from the CORE[2] search engine of open-access scholarly articles. The dataset includes: (1) a training set of 2 million documents and 393 queries (mid-November to mid-December 2024), and (2) a test set with two parts: 2 million documents and 492 queries collected in January 2025, and 99 held-out queries from the training snapshot.

In both tasks, systems were trained once and evaluated on future snapshots without retraining, enabling analysis of short- and long-term performance shifts.

3. Participation and Systems

We received 45 runs for WebRetrieval and 23 for SciRetrieval from 19 participating teams. Submitted approaches ranged from classic BM25 pipelines [8, 9, 10, 11, 12] to advanced neural methods involving reranking [13, 14, 9, 10, 15, 16], query expansion [17, 16, 12], use of historical signals [14, 18, 15], large language models (LLMs)[17, 19, 12], and clustering techniques[15].

4. Results and Discussion

Effectiveness was measured using nDCG@10 and nDCG@1000. For the WebRetrieval task, systems generally showed stable performance over short time spans (March to May 2023), but drops over longer spans (March to August 2023). This trend aligns with document overlap data: earlier snapshots share fewer documents with later ones, indicating distributional shifts.

In SciRetrieval, the smaller number of snapshots limited long-term analysis, but some systems still showed robustness, especially those using query clustering and re-ranking. System rankings varied more across snapshots than in the web task.

To better capture temporal behavior, we used additional metrics: Relative Improvement (RI), Delta RI (DRI), and Effect Ratio (ER). These showed that only a few systems maintained or improved effectiveness long-term, and that standard effectiveness metrics alone may not capture persistence.

5. Conclusion

The third edition of LongEval expanded the study of temporal robustness to a new domain and received strong engagement from the community. While many systems performed well initially, maintaining effectiveness over time remains challenging.

Acknowledgments

This work was supported by the ANR Kodicare project (ANR-19-CE23-0029), the Austrian Science Fund (FWF, I4471-N), the UKRI/EPSRC Turing AI Fellowship to Maria Liakata (EP/V030302/1), the German Research Foundation (DFG, 407518790), and the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062, LINDAT/CLARIAH-CZ). The work also used services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz).

[1] https://www.qwant.com/
[2] https://core.ac.uk/

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o in order to: check grammar and spelling. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. T. Dumais, Putting searchers into search, in: SIGIR, ACM, 2014, pp. 1–2.

[2] E. Adar, J. Teevan, S. T. Dumais, J. L. Elsas, The web changes everything: understanding the dynamics of web content, in: WSDM, ACM, 2009, pp. 282–291.

[3] K. Roberts, T. Alam, S. Bedrick, D. Demner-Fushman, K. Lo, I. Soboroff, E. Voorhees, L. L. Wang, W. R. Hersh, Searching for scientific evidence in a pandemic: An overview of trec-covid, Journal of Biomedical Informatics 121 (2021) 103865.

[4] A. Tikhonov, I. Bogatyy, P. Burangulov, L. Ostroumova, V. Koshelev, G. Gusev, Studying page life patterns in dynamical web, in: SIGIR, ACM, 2013, pp. 905–908.

[5] R. Alkhalifa, I. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, G. Gonzalez-Saez, P. Galuščáková, L. Goeuriot, E. Kochkina, M. Liakata, D. Loureiro, H. T. Madabushi, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Overview of the clef-2023 longeval lab on longitudinal evaluation of model performance, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science (LNCS), Springer, Thessaliniki, Greece, 2023.

[6] R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, T. Fink, P. Galuščáková, G. Gonzalez-Saez, L. Goeuriot, D. Iommi, M. Liakata, H. T. Madabushi, P. Medina-Alias, P. Mulhem, F. Piroi, M. Popel, A. Zubiaga, Overview of the CLEF 2024 LongEval Lab on Longitudinal Evaluation of Model Performance, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2024.

[7] J. Keller, T. Breuer, P. Schaer, Evaluation of temporal change in IR test collections, in: H. Oosterhuis, H. Bast, C. Xiong (Eds.), Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024, ACM, 2024, pp. 3–13. URL: https://doi.org/10.1145/3664190.3672530. doi:10.1145/3664190.3672530.

[8] A. Bruttomesso, D. Cavazza, A. Corrò, S. Peraro, D. Seghetto, N. Ferro, SEUPD@CLEF: Team 3DS2A on Performance Evaluation over Time of IR Systems with Proximity Search and Reranking Components, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[9] A. Mukhtar, P. Leonardo, F. Zaccarin, Z. Shen, N. Ferro, SEUPD@CLEF: Team [DataHunter] on Temporal Stability Analysis of Boolean and CamemBERT-Based Retrieval Systems, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[10] J. Stryszewski, W. Prosowicz, T. Kawiak, A. Jaśkowiec, Improving Scientific Information Retrieval with Dense Representations and Cross-Encoder Re-ranking, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[11] G. Amato, N. Brentel, A. Demo, S. Laghetto, F. Pivotto, I. Toporov, N. Ferro, Team RAND at LongEval 2025: Composable Information Retrieval with Semantic and Language-Aware Components, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[12] D. Caon, R. D. Maschio, A. Disarò, S. Maule, N. Ferro, SARD at LongEval 2025: On Longitudinal Evaluation of IR Systems by Using Query Rewriting and Hybrid Queries, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[13] A. Bottari, L. Croce, F. M. H. Abadi, N. Ferro, SEUPD@CLEF: Team BASETTE on an IR system for basic hardware, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[14] F. Braun, T. Busch, M. S. Coban, M. E. Ghadioui, D. Hovhannisyan, K. Jonina, A. Large, F. Z. Y. Lin, E. Loewenstein, L. Maaßen, N. Maron, M. H. Mörsheim, J. A. N. Ofunim, V. Romanovskis, A. Simon, J. Witalla, M. Wollenberg, J. Keller, P. Schaer, CIR at LongEval 2025: Exploring Temporal Sensitivity in Web Retrieval, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[15] D. Alexander, M. Fröbe, G. Hendriksen, M. Hagen, D. Hiemstra, M. Potthast, A. de Vries, Team OpenWebSearch at CLEF 2025: LongEval, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[16] G. Gaio, F. Mazzarotto, M. Meneghin, E. Saro, F. Visonà, SEUPD2425-RACOON at LongEval 2025: A novel approach to Information Retrieval with LLM-based query expansion and temporal relevance feedback techniques, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[17] A. Miyaguchi, I. Afrulbasha, A. Pramov, DS@GT at LongEval: Evaluating Temporal Performance in Web Search Systems and Topics with Two-Stage Retrieval, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[18] A. M. Ndiema, J. Keller, P. Schaer, LongEval: CIR_cluster at LongEval 2025: Clustering Query Variants for Temporal Generalization, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[19] D. Furlan, G. Gibellato, S. S. Nazirialhashem, E. Pase, A. Pasqualetto, F. Tiberio, N. Ferro, SEUPD@CLEF: Team RISE on Improving Search by Crafting Titles and Matching URLs, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.