Team OpenWebSearch at LongEval: Large Language Model Relevance Judgments as Relevance Feedback for Scientific Search

Notebook for the LongEval Lab at CLEF 2026

Daria Alexander[1], Maik Fröbe[2], Gijs Hendriksen[1], Matthias Hagen[2], Djoerd Hiemstra1[1], Martin Potthast[3] and Arjen P. de Vries[1]

[1]*Radboud Universiteit Nijmegen*
[2]*Friedrich-Schiller-Universität Jena*
[3]*University of Kassel, hessian.AI, ScaDS.AI*

Abstract

We describe our participation in Task 1 on scientific search for LongEval 2026. The setup of Task 1 evaluates retrieval systems in a longitudinal setting where past click logs are available to evaluate the effectiveness of retrieval approaches on future snapshots. In this longitudinal scenario, we apply the same concepts as in our participation last year: we use explicit relevance feedback from the past to reformulate the queries on future snapshots so that documents that were relevant to the query in the past are also ranked highly in the result list on future evaluation snapshots. Our motivation for this approach is that when the intent of the query does not change, documents that have been relevant in the past also should stay relevant in the future, so that query reformulation against documents relevant in the past should also increase the effectiveness of the result ranking in the future. In the 2026 edition of the task, we observe that the click logs of past snapshots contain few relevant documents. Hence, (contrary to our submissions in previous years,) we do not use the click logs as explicit relevance feedback but instead use large language model relevance judgments. Besides changing this source for relevance feedback, our submission is conceptually identical to the previous editions, we use different approaches, like the concept of keyqueries, to ensure that documents labeled as relevant to a query on the past snapshot remain ranked in the top positions in future snapshots.

Keywords

Large Language Model Relevance Judgments, Keyquery, Counterfactual Query Rewriting

1. Introduction

Retrieval in the scientific domain has different longitudinal aspects than generic web search. In generic retrieval domains such as web search, documents might be deleted, modified, or created over time [1]. Our participations in the last iterations of Longeval evolved to specifically handle cases when documents have been observed as relevant in the past but might have been deleted or modified [2, 3, 4, 5]. In this scenarios, we used so called keyqueries [6] and counterfactually assume a relevant document from the past does still exist and reformulate the query so that documents relevant from the past appear at the top positions for a query. In the 2026 edition of the retrieval task of LongEval [7, 8, 9], the retrieval domain is scientific search. Consequently, the longitudinal aspects change substantially, for instance, scientific papers are rarely modified after publication, and also almost never get deleted. This changes the key aspect of our previous years' participation, as we do not need to counterfactually assume that the old document is still there, as the document is in almost all cases (maybe with retractions of papers as a counterexample) still in the corpus. An additional change is that this year we observed that only a few documents are relevant according to the click relevance judgments. For this reason, we decided to use large language models as relevance assessors. Overall, our two changes compared to our last year's submission is first, that we did not ned to counterfactually inject documents that were previously relevant into future snapshots (as the documents are still there and unmodified, due to the scientific

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

search domain), and, second, that we use large language models as relevance assessors for the explicit relevance feedback and not the click log data. We submit four runs, and our code is available online.[1]

2. Appraoches

We submit four approaches that are implemented using PyTerrier [10], ranx [11] and the `ir_datasets` integration specifically for LongEval [12, 13].

Large Language Model Relevance Judgments. We used large language models as relevance judges for all our approaches and used their labels as relevance feedback. We did not create the relevance judgments ourselves, but used the LLM relevance judgments published by Task 2 of LongEval (the LLMs are qwen3-32b, llama-3.1-8b, gpt-oss-120b, and gpt-oss-20b, each using the default UMBRELA [14] prompt). The LLM relevance judgments were on a scale from 0 (for not relevant) to 3 (for highly relevant). We used reciprocal rank fusion implemented in ranx to fuse the four LLM judgments into a single ordered list (from highly relevant by multiple assessors to not relevant). We then used this ranking as relevance feedback for all our four submissions.

kq-rm3-bm25 We used BM25 as retrieval model and used RM3 to get term candidates for reformulating the queries. The RM3 reformulation was done against the LLM relevance list fused from the four LLM relevance assessors.

kq-rm3-dirichlet We used Dirichlet as retrieval model and used RM3 to get term candidates for reformulating the queries. The RM3 reformulation was done against the LLM relevance list fused from the four LLM relevance assessors.

kq-rm3-pl2 We used PL2 as retrieval model and used RM3 to get term candidates for reformulating the queries. The RM3 reformulation was done against the LLM relevance list fused from the four LLM relevance assessors.

put-to-front We used BM25 as the basis retrieval model and moved all documents from the reciprocal rank fused LLM judgments to the top of the ranking.

3. Evaluation

The evaluation setup of LongEval has a focus on longitudinal effectiveness and robustness of the retrieval system over time. For instance, in the best case, a retrieval system can, with more user interactions, better understand in which scenario it is used so that it remains effective or even increases its effectiveness (but it ideally should not become less effective). We follow the official LongEval evaluation that incorporates diverse modifications of standard IR effectiveness evaluation towards longitudinal evaluation [1, 15, 16] and use nDCG@10 on relevance judgments derived via the discretized Document Click-Through Rate (DCTR) as relevance labels. Table 1 shows the evaluation results for the average retrieval performance (ARP; the effectiveness in terms of nDCG for the three temporal snapshots), relative change (RC [17, 18, 19]; measures the effectiveness delta between future and past snapshots, normalized by the past snapshot), and the delta relative improvement (DRI; measures the relative change in effectiveness against the official baseline system).

In Table 1, we observe that there is a general trend that retrieval gets less effective over time, e.g., the BM25 baseline has an nDCG@10 of 0.279 at timestamp 1, 0.254 at timestamp 2, and 0.150 at timestamp 3. The LLM judgments seem to make the retrieval less effective in all cases; none of our approaches is better than BM25.

[1] https://github.com/OpenWebSearch/LONGEVAL-26

Table 1
The evalution of our four submitted approaches and two baselines evaluated via the average retrieval performance (ARP), relative change (RC), and delta relative improvement (DRI).

run	dctr						
	ARP			RC		DRI	
	nDCG@10			nDCG@10		nDCG@10	
	1	2	3	2	3	2	3
baseline-bm25	0.279	0.254	0.150	0.089	0.462	0.000	0.000
baseline-qwen3-4b	0.234	0.201	0.133	0.139	0.432	0.047	-0.047
kq-rm3-bm25	0.273	0.219	0.139	0.198	0.492	0.117	0.054
kq-rm3-dirichlet	0.181	0.123	0.079	0.322	0.564	0.167	0.124
kq-rm3-pl2	0.276	0.233	0.142	0.154	0.484	0.071	0.040
put-to-front	0.261	0.217	0.132	0.169	0.496	0.082	0.060

4. Conclusion

We participated with four submissions in Task 1 of LongEval. We used LLM relevance judgments as relevance feedback for query reformulations so that documents judged as relevant by an LLM appear at the top positions of the ranking. We observed that this decreased the effectiveness. Potential reasons for this reduction might be that the LLM judgments are not aligned with the user behavior of the production search engine, or that LLMs would put documents at a very high position that are not retrieved by the production search engine so that they are not relevant in the evaluation. For future work, we think it could be interesting to study how the alignment between LLMs and click logs can be increased.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] 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.

[2] J. Keller, M. Fröbe, G. Hendriksen, D. Alexander, M. Potthast, M. Hagen, P. Schaer, Counterfactual query rewriting to use historical relevance feedback, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part III, volume 15574 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 138–147. URL: https://doi.org/10.1007/978-3-031-88714-7_11. doi:10.1007/978-3-031-88714-7_11.

[3] D. Alexander, M. Fröbe, G. Hendriksen, M. Hagen, D. Hiemstra, M. Potthast, A. P. de Vries, Team openwebsearch at longeval: Using historical data for scientific search, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3316–3326. URL: https://ceur-ws.org/Vol-4038/paper_266.pdf.

[4] D. Alexander, M. Fröbe, G. Hendriksen, F. Schlatt, M. Hagen, D. Hiemstra, M. Potthast, A. P. de Vries, Team openwebsearch at CLEF 2024: Longeval, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2304–2313. URL: https://ceur-ws.org/Vol-3740/paper-215.pdf.

[5] M. Fröbe, G. Hendriksen, A. P. de Vries, M. Potthast, Open web search at longeval 2023: Reciprocal rank fusion on automatically generated query variants, in: M. Aliannejadi, G. Faggioli, N. Ferro,

M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2432–2440. URL: https://ceur-ws.org/Vol-3497/paper-195.pdf.

[6] T. Gollub, M. Hagen, M. Michel, B. Stein, From keywords to keyqueries: content descriptors for the web, in: G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, T. Sakai (Eds.), The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013, ACM, 2013, pp. 981–984. URL: https://doi.org/10.1145/2484028.2484181. doi:10.1145/2484028.2484181.

[7] T. Breuer, M. Cancellieri, A. El-Ebshihy, M. Fröbe, P. Galuscáková, L. Goeuriot, G. Iturra-Bocaz, J. Keller, P. Knoth, A. K. Kruff, P. Mulhem, F. Piroi, D. Pride, P. Schaer, D. Schwab, Evaluating information retrieval models along time: The longeval lab at CLEF 2026, in: R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, A. Anand, Z. Ren, S. Verberne, N. Bai, M. Mansoury (Eds.), Advances in Information Retrieval - 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29 - April 2, 2026, Proceedings, Part IV, volume 16486 of *Lecture Notes in Computer Science*, Springer, 2026, pp. 345–353. URL: https://doi.org/10.1007/978-3-032-21321-1_45. doi:10.1007/978-3-032-21321-1_45.

[8] S. Bouaraba, T. Breuer, M. Cancellieri, A. El-Ebshihy, M. Fröbe, P. Galuscáková, L. Goeuriot, G. Iturra-Bocaz, J. Keller, P. Knoth, A. K. Kruff, P. Mulhem, F. Piroi, D. Pride, P. Schaer, D. Schwab, Extended overview of the clef 2026 longeval lab on longitudinal evaluation of model performance, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[9] S. Bouaraba, T. Breuer, M. Cancellieri, A. El-Ebshihy, M. Fröbe, P. Galuscáková, L. Goeuriot, G. Iturra-Bocaz, J. Keller, P. Knoth, A. K. Kruff, P. Mulhem, F. Piroi, D. Pride, P. Schaer, D. Schwab, Overview of the clef 2026 longeval lab on longitudinal evaluation of model performance, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science (LNCS), Springer, 2026.

[10] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using pyterrier, in: K. Balog, V. Setty, C. Lioma, Y. Liu, M. Zhang, K. Berberich (Eds.), ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020, ACM, 2020, pp. 161–168. URL: https://doi.org/10.1145/3409256.3409829. doi:10.1145/3409256.3409829.

[11] E. Bassani, L. Romelli, ranx.fuse: A python library for metasearch, in: M. A. Hasan, L. Xiong (Eds.), Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, ACM, 2022, pp. 4808–4812. URL: https://doi.org/10.1145/3511808.3557207. doi:10.1145/3511808.3557207.

[12] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified data wrangling with ir_datasets, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2429–2436. URL: https://doi.org/10.1145/3404835.3463254. doi:10.1145/3404835.3463254.

[13] J. Keller, M. Fröbe, G. Hendriksen, D. Alexander, M. Potthast, P. Schaer, Simplified longitudinal retrieval experiments: A case study on query expansion and document boosting, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2024, Madrid, Spain, September 9-12, 2025, Proceedings, Part I, Lecture Notes in Computer Science, Springer, 2025.

[14] S. Upadhyay, R. Pradeep, N. Thakur, N. Craswell, J. Lin, UMBRELA: umbrela is the (open-source reproduction of the) bing relevance assessor, CoRR abs/2406.06519 (2024). URL: https://doi.org/10.48550/arXiv.2406.06519. doi:10.48550/ARXIV.2406.06519. arXiv:2406.06519.

[15] J. Keller, T. Breuer, P. Schaer, Replicability measures for longitudinal information retrieval evaluation, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. D. Nunzio, L. Soulier, P. Galuscáková,

A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part I, volume 14958 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 215–226. URL: https://doi.org/10.1007/978-3-031-71736-9_16. doi:10.1007/978-3-031-71736-9_16.

[16] G. N. G. Sáez, P. Mulhem, L. Goeuriot, Towards the evaluation of information retrieval systems on evolving datasets with pivot systems, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 91–102. doi:10.1007/978-3-030-85251-1_8.

[17] R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. E. Anke, T. Fink, P. Galuscáková, G. G. Sáez, 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: CLEF (2), volume 14959 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 208–230.

[18] R. Alkhalifa, I. M. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. E. Anke, G. G. Sáez, P. Galuscáková, L. Goeuriot, E. Kochkina, M. Liakata, D. Loureiro, P. Mulhem, F. Piroi, M. Popel, C. Servan, H. T. Madabushi, A. Zubiaga, Overview of the CLEF-2023 longeval lab on longitudinal evaluation of model performance, in: CLEF, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 440–458.

[19] M. Cancellieri, A. El-Ebshihy, T. Fink, M. Fröbe, P. Galuscáková, G. G. Sáez, L. Goeuriot, D. Iommi, J. Keller, P. Knoth, P. Mulhem, F. Piroi, D. Pride, P. Schaer, Extended abstract of longeval at CLEF 2025: Longitudinal evaluation of IR systems on web and scientific data, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3296–3299. URL: https://ceur-ws.org/Vol-4038/paper_264.pdf.