# Extended Overview of the CLEF 2026 LongEval Lab on Longitudinal Evaluation of Model Performance*

LongEval at CLEF 2026

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

[1]*Univ. Grenoble Alpes, CNRS, Grenoble INP, FR*
[2]*TH Köln - University of Applied Sciences, Cologne, DE*
[3]*The Open University, Milton Keynes, UK*
[4]*TU Wien, AT*
[5]*Friedrich-Schiller-Universität Jena, DE*
[6]*University of Stavanger, NO*

**Abstract**

The LongEval lab is about the evaluation of information retrieval systems over time. This year, we collaborate with the academic search platform CORE that provides documents, queries, and real-world user interactions. The participants are provided with different snapshots derived from this data. In the 2026 edition of the lab, four different tasks around temporal aspects of search are proposed: (1) ad-hoc scientific retrieval, (2) topic extraction from the query logs, (3) the simulation of user interactions in the form of query (re-)formulations, and (4) a RAG task involving the evolving document collection. We present an overview of this year's tasks, the participating systems, and results. A total of 104 runs were submitted, which were evaluated with respect to their effectiveness, but also robustness measures that quantify changes in retrieval effectiveness over time.

**Keywords**

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

## 1. Introduction

The CLEF 2026 LongEval lab investigates how information retrieval (IR) systems perform and remain robust as document collections evolve over time. The lab closely cooperates with the COnnecting REpositories (CORE) search platform for open-access academic papers, which provides real-world queries, documents, and interactions from the search logs. The evolving corpus is represented through time-ordered snapshots, enabling longitudinal analyses of effectiveness and stability as the collection changes. In 2026, LongEval offered four tasks addressing challenges relevant to temporal IR: 1) ad-hoc scientific retrieval, 2) topic extraction from query logs, 3) user simulation via query (re-)formulation, and 4) retrieval-augmented generation (RAG) over an evolving scientific corpus. The underlying dataset for all these tasks, the SciRetrieval dataset, extends prior work with three cumulative snapshots (March–May, June–August, and September–November 2025), constructed from CORE logs, to which we added the top 1000 search results from the January 2026 CORE ranking. The corpus sizes grew, thus, from roughly 870K to 1.7M documents across the snapshots.

In Task 1, for training and evaluation, LongEval provides raw and Document Click-Through Rate (DCTR) click-based qrels, and includes synthetic LLM-generated relevance judgments using GPT-OSS-120b as an additional baseline for alternative evaluations. Across 20 submitted retrieval runs from eight teams, pipelines consistently include a lexical first stage (often BM25), with many multi-stage variants and re-ranking strategies leveraging temporal and citation-derived signals. Effectiveness was

*CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany*
* Authors are given in alphabetical order.
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

**Table 1**
Teams participating in the LongEval 2026 lab.

| Task | Team | Runs | Team size (# persons) | Location | Notebook |
|---|---|---|---|---|---|
| 1 | glhf | 5 | 2 | United States | [1] |
| 1 | Lexora | 3 | 4 | Chennai, India | [2] |
| 1 | OpenWebSearch | 4 | 7 | Nijmegen, The Netherlands; Jena and Kassel, Germany | [3] |
| 1 | Time4Citation | 3 | 4 | Cologne, Germany | [4] |
| 1 | dense-fusion | 1 | 3 | Cologne, Germany | [4] |
| 1 | air-benders | 2 | - | Vienna, Austria | - |
| 1 | air-02 | 1 | - | Vienna, Austria | - |
| 1 | nlp hashiras | 1 | - | Ankara, Turkey | - |
| 2 | CIR | 17 | 4 | Cologne, Germany | [5] |
| 2 | JOINorDIE | 6 | 3 | Cologne, Germany | [4] |
| 2 | VerbaNexAI | 1 | 4 | Cartagena de Indias, Colombia | [6] |
| 3 | JOINorDIE | 18 | 3 | Cologne, Germany | [4] |
| 3 | Promptly | 6 | 2 | Cologne, Germany | [4] |
| 3 | Split n' Simulate | 1 | 2 | Cologne, Germany | [4] |
| 3 | VerbaNexAI | 4 | 4 | Cartagena de Indias, Colombia | [7] |
| 4 | VerbaNexAI | 2 | 4 | Cartagena de Indias, Colombia | [8] |
| 4 | glhf | 9 | 2 | United States | [9] |
| 4 | DS-GT | 4 | 2 | Atlanta, United Stated | [10] |
| 4 | LabIndividual | 4 | 3 | Foshan, China | [11] |
| 4 | ockham | 6 | - | - | - |

measured using the nDCG@10 metric, and results show marked variability in system order and a general decline in effectiveness over time, underscoring the need for temporally aware features and timeline-specific judgments to ensure reliable longitudinal evaluation.

Task 2 focused on extracting formalized information needs from query logs. The aim is to enable reliable Cranfield-style evaluations when only production clicks exist over time. Three teams submitted 23 topic sets, which were judged by four LLM assessors. We note that, consistent with prior observations, the LLM assessors tend to over-label relevance. Among the formalized topic variants, the manual ones achieved the best system-ranking agreement (Tau-AP), while a query-augmented method yielded the highest inter-LLM agreement (Cohen's Kappa) at the document level.

Task 3 is aimed at simulating user behaviour where the inputs are session-level query sequences, SERPs, and clicks. Participants were asked to predict the next query. Four teams submitted a total of 29 experiments. It appears that, using more recent context (e.g., the last visible query) and contrastive prompt enrichment improves performance, while complex reasoning prompts (e.g., chain-of-thought, self-refinement) generally do not. Persona-based methods produced mixed outcomes, with effectiveness highly sensitive to implementation and prompting.

Task 4 studies RAG under temporal evolution by providing 47 questions and, for each, a set of 10 candidate documents derived from temporally-aware clustering and paper-pair selection using dense scientific embeddings. Participants report cited sources and generated answers, which are evaluated by Precision, ROUGE-L, BERTScore, nugget coverage, and an LLM-as-a-judge. Results show that stronger retrieval alone does not guarantee better generation quality, and generation-only systems can give good results when the candidate set is small.

Overall, LongEval 2026 collected 97 runs (Table 1) across its four tasks, evaluating not only effectiveness but also robustness over time, and highlights the importance of temporally grounded datasets, signals, and evaluation protocols for advancing IR in evolving scholarly environments. In the following sections, we describe the details of each task including: data preparation, submitted runs, results, discussions, and conclusions.

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

The LongEval-Sci task is a continuation of the LongEval SciRetrieval task from last year's edition [12]. It is a traditional ad-hoc search task in the scientific domain. We evaluate the longitudinal effectiveness and robustness of Information Retrieval (IR) systems in retrieving scholarly publications from an evolving corpus. Each evolved state of the corpus is represented by a snapshot at a specific time point. In this edition, we provide participants with three new snapshots.

### 2.1. Datasets

In this year's edition of the LongEval lab, we continued to use the CORE search engine[1] to construct a dynamic test collection. We gathered user queries, search engine results pages (SERPs), and click feedback over the year 2025, and created a dynamic test collection with three snapshots, each spanning three months: Snapshot 1 (March to May), Snapshot 2 (June to August), and Snapshot 3 (September to November). This dataset continues the 2025 collection [12] chronologically, but the creation process differs from the previous one, so the two collections are not necessarily comparable. Namely, the snapshots span longer time periods, documents are only added to the snapshots but never removed, and qrels are only constructed using selected click types. We introduced these changes for two reasons: (1) we have user queries spanning a full year, which was a limitation last year with only a short period of two months, and (2) we wanted to simulate the real scenario of a scientific collection, where publications are added but almost never removed. The dataset is available from the TU Wien Research Data Repository [13, 14] and through the LongEval `ir_datasets` integration [15, 16][2].

We sampled real user queries from the CORE search logs, selecting ad-hoc queries with at least 30 clicks in any snapshot and manually filtering inappropriate or harmful queries. This resulted in 434 queries, to which we added the 47 questions from Task 4. These candidate queries are manually filtered to remove inappropriate or harmful queries.

Based on these queries, we construct the document collection for each snapshot by selecting a sub-sample of the CORE collection of scholarly documents. More concretely, for the queries we selected above, we use the CORE API[3] to fetch the documents that appear in the query's click log, including logged SERPS, as well as the top 1,000 documents in the CORE ranking from January 2026. To reflect the temporal dynamics of scholarly literature, the collection snapshots are strictly cumulative. Each successive snapshot is a superset of the preceding one, augmented with documents newly appearing in the system rankings. Since these documents are likely to be relevant, a proportional number of randomly selected documents is added as well. This yielded corpus sizes of approximately 87K, 1.3M, and 1.7M documents for Snapshot 1, Snapshot 2, and Snapshot 3, respectively.

Two query relevance label sets (qrels) are constructed directly from the click interactions. The raw qrels mark every document ever clicked as relevant. Since a click cannot be directly translated into a relevance signal, the Document Click Through Rate (DCTR) is determined and discretized into the grades not relevant, relevant, and highly relevant. Since not all queries have enough clicks in all snapshots, only a query subset can be evaluated by the DCTR in each snapshot. Snapshot 1 contains 149 queries, Snapshot 2 100, and Snapshot 3 136. 12 queries are present in all snapshots. For the evaluation of the retrieval task, we have a connection to Task 2 on topic formalization, as the formalized topics are used to generate relevance judgments with large language models in Task 2. Those relevance judgments allow for alternative evaluations of the retrieval effectiveness. Task 2 yielded 94 sets of LLM-generated relevance judgments (using top-10 pooling). For the evaluation of Task 1, we only include qrels generated with GPT-OSS-120b [17] on the query only, without description and narrative, which is also used as a baseline in Task 2 and somewhat established practice for LLM-as-a-judge [18].

We provided participants with training and test datasets. The training dataset is composed of documents from Snapshot 1, 100 queries, 2,275 raw relevance assessments, and 10,406 DCTR relevance

[1] https://core.ac.uk/
[2] https://github.com/clef-longeval/ir-datasets-longeval
[3] https://core.ac.uk/services/api

**Table 2**
Overview of the collected ir_metadata according to (a) implemented retrieval paradigms and (b) frequently used software libraries.

(b) Retrieval paradigms

| **Paradigm** | **Used** | |
|---|---|---|
| | Yes | No |
| Lexical | 20 | 0 |
| Deep Neural | 3 | 17 |
| Sparse Neural | 0 | 20 |
| Dense Neural | 3 | 17 |
| Single Stage | 5 | 15 |

(a) Most popular libraries

| Library | Teams |
|---|---|
| python-terrier | 8 |
| ir-datasets-longeval | 4 |
| numpy | 4 |
| scikit-learn | 3 |
| tirex-tracker | 3 |
| torch | 3 |
| pandas | 3 |
| sentence-transformers | 3 |

assessments. The test datasets are composed of documents from Snapshot 2 and Snapshot 3, and 381 queries. Participants submitted runs using the test queries for each of the 3 snapshots, using the same system trained on the training dataset. No overlap between this train split, the test split, or the LongEval 2025 collection exists. This is different from previous years and does not allow for the exploitation of this overlap.

## 2.2. Submissions and Approaches

In total, we received 20 runs, from eight teams (cf. Table 1). Compared to last year, two more teams participated in this task, but submitted three fewer runs. Of those eight teams, five teams also submitted a notebook paper (compared to two teams in 2025). The submissions were done through the TIRA platform [19], where participants could upload runs and containerized approaches. Additionally, we collected metadata, including short descriptions of each run, which participants had to provide in the `ir_metadata` specification format [20]. Table 2 provides an overview of the metadata of the submitted approaches.

Table 2 provides an overview of the run metadata submitted by the participating teams. All 20 submitted runs use a lexical ranker at some point in the retrieval pipeline. Three runs use a deep neural approach, and three runs use a dense neural component. Five runs are classified as single-stage, but 15, the majority of the runs, use a multi-stage approach. BM25 [21] is used most often for the first-stage ranking. Re-ranking approaches varied from classical boosting approaches to specific temporal signals to Learning-To-Rank on sparse feature sets or deep neural cross-encoders. Temporal signals are created, for example, from the publishing date (teams GLHF and dense-fusion), by incorporating a retrieval signal from previous snapshots (team Lexora), or from the newest citation as a combined proxy for recency and quality (teams GLHF and Time4Citation). Further approaches included the use of query expansion using RM3 [22] and Reciprocal Rank Fusion [23] approaches (teams GLHF and Lexora).

Most teams made the source code publicly available on GitHub; two teams kept their code private. All eight teams used the `PyTerrier` [24] library to implement their approaches. Four teams used the `IR_datasets` integration [15], `NumPy` [25] followed by three teams using `sklearn` [26], `PyTorch` [27] and `SentenceTransformers` [28]. The OpenWebSearch team additionally provided metadata on the resource consumption tracked with the TirexTracker [29]. In the following, we give an overview of the provided baselines and the submitted runs by participants.

**Baselines** We provided two baselines as a reference point for the evaluations and as first-stage rankers for reranking approaches. The first baseline is BM25 [21] with default parameters over the abstract texts. It was implemented using `PyTerrier` [24], with the default parameters. The second baseline also used the abstract collection which is embedded with Qwen3 4B [30] at the default temperature. It was implemented with `PyTerrier-Dense` and a `vLLM` backed [31].

**GLHF** The GLHF team experimented with different first-stage rankers based on the full texts that are also combined using rank fusion, query expansion with RM3, and a pre-trained MiniLM cross-encoder. Additionally, they use bibliometric and temporal boosts to rerank the documents. The findings indicate that using the full-text improved the results. The temporal signal also improved the effectiveness and robustness when calibrated correctly [1].

**Lexora** The Lexora team submitted three approaches. The first approach uses RM3 query expansion with BM25. The second approach combines RM3 with BM25 and the retrieval score from the previous snapshot. The third approach uses a hybrid dense–sparse pipeline combining BM25 with RM3, and Qwen3 embeddings via Reciprocal Rank Fusion. Their findings show that considering temporal signal improves the stability of submitted systems [2].

**OpenWebSearch** The OpenWebSearch team continued to experiment with keyqueries [32]. A keyquery is a perfect query reformulation that finds a set of target documents. They employed synthetic relevance labels, the same as used for task 2, to create the set of relevant target documents [3].

**Time4Citation** The Time4Citation team created three relevance signals, based on the citations of the documents, that were used to rerank the BM25 baseline. The first signal models the recency of documents based on the last citation and emphasizes more recent documents. The second approach boosts documents with more citations, and the third signal treats the availability of any citation as a relevance indicator [4].

**dense-fusion** The dense-fusion team trained a Learning-To-Rank (LTR) model to rerank the top 500 ranked documents. They used a diverse feature set, including the document age, the log-scaled publication counts, and embedding representations of the abstracts. The final ranking is determined by a linear combination of the BM25 score and the new LTR score [4].

## 2.3. Results

The submitted approaches are evaluated for effectiveness per snapshot and robustness across snapshots. We contrast the results with both baselines and compare them to Snapshot 1. Our primary relevance labels are based on the DCTR, which we supplement with the synthetic relevance labels.

### 2.3.1. Effectiveness

The test collection is designed for a scientific ad-hoc search task with short, general, and unstructured keyword queries, and we rely on the click interactions for relevance estimation. Therefore, we select nDCG@10 as our effectiveness measure and average the results across many queries. We look for the most effective retrieval approach per snapshot, first for the DCTR qrels and then also for the synthetic labels generated with GPT-OSS-120b. The effectiveness leaderboard per snapshot and approach is presented in Table 3, and the results are also visualized in Figure 1.

Based on the DCTR labels, the top three approaches for the first snapshot are `Dense-Fusion-Run` (0.326), `custom-lexicalft-tc` and `custom-lexical-FT-T` (both 0.3), and `bm25-ft-additive` (0.291). On the second snapshot, `custom-lexicalft-tc` and `custom-lexical-FT-T` (both 0.284) outperform `Dense-Fusion-Run` (0.274). The third best system is `temporal-fusion` (0.268). The third snapshot is led again by `custom-lexicalft-tc` (0.195) followed by the approaches `Log-boost-03` and `Dense-Fusion-Run`(both 0.182). Investigating the top three systems per snapshot on the DCTR qrels indicates variability in the system order and effectiveness. On average, the effectiveness drops from the first to the second snapshot and increases again, almost to the level of Snapshot 1, in the third snapshot.

For the synthetic relevance labels (generated with GPT-OSS-120b), we find that the best approaches on the first snapshots are `put-to-front` (0.914), `bm25-dense-rerank` (0.846), and

**Table 3**
Effectiveness across qrels and snapshots, measured by nDCG@10. The approaches are sorted by the snapshot 3 results on the DCTR qrels, the baselines are gray and the best result per snapshot is **bold**.

| **Qrels** | **dctr** | | | **gpt-oss-120b** | | |
|---|---|---|---|---|---|---|
| **Run / Snapshot** | 1 | 2 | 3 | 1 | 2 | 3 |
| custom-lexicalft-tc [1] | 0.300 | **0.284** | **0.195** | 0.698 | 0.709 | 0.713 |
| custom-lexical-FT-T[1] | 0.300 | **0.284** | **0.195** | 0.698 | 0.709 | 0.713 |
| Log-boost-03 [4] | 0.263 | 0.253 | 0.182 | 0.648 | 0.711 | 0.747 |
| Dense-Fusion-Run [4] | **0.326** | 0.274 | 0.182 | 0.746 | 0.752 | 0.756 |
| binary-boost-09 [4] | 0.257 | 0.241 | 0.168 | 0.646 | 0.712 | 0.748 |
| custom-lexical-FT [1] | 0.290 | 0.256 | 0.166 | 0.706 | 0.714 | 0.717 |
| bm25-ft-additive [1] | 0.291 | 0.257 | 0.166 | 0.706 | 0.714 | 0.717 |
| bm25-ft-router [1] | 0.290 | 0.256 | 0.166 | 0.707 | 0.714 | 0.717 |
| temporal-fusion [2] | 0.277 | 0.268 | 0.161 | 0.745 | 0.750 | 0.752 |
| QwenTemporalIR [2] | 0.264 | 0.249 | 0.160 | 0.761 | 0.769 | 0.769 |
| b-bm25-dense-rerank-v2 [air-benders] | 0.088 | 0.260 | 0.155 | 0.200 | 0.752 | 0.754 |
| b-bm25-dense-rerank [air-benders] | 0.089 | 0.264 | 0.154 | 0.200 | 0.752 | 0.755 |
| baseline-bm25 | 0.279 | 0.254 | 0.150 | 0.743 | 0.748 | 0.751 |
| kq-rm3-pl2 [3] | 0.276 | 0.233 | 0.142 | 0.803 | 0.811 | 0.812 |
| kq-rm3-bm25 [3] | 0.273 | 0.219 | 0.139 | 0.814 | 0.823 | 0.823 |
| rm3-conservative [2] | 0.271 | 0.249 | 0.134 | 0.741 | 0.750 | 0.753 |
| bm25-dense-rerank [nlp hashiras] | 0.225 | 0.215 | 0.134 | 0.846 | 0.850 | 0.848 |
| baseline-qwen3-4b | 0.234 | 0.201 | 0.133 | 0.791 | 0.796 | 0.800 |
| put-to-front [4] | 0.261 | 0.217 | 0.132 | **0.914** | **0.926** | **0.925** |
| ce-reranked [air-02] | 0.218 | 0.215 | 0.109 | 0.679 | 0.690 | 0.697 |
| Timely-Boost-V2 [4] | 0.130 | 0.106 | 0.108 | 0.525 | 0.590 | 0.598 |
| kq-rm3-dirichlet [3] | 0.181 | 0.123 | 0.079 | 0.666 | 0.672 | 0.656 |

`kq-rm3-bm25` (0.814). This order remains the same for the other two snapshots. A few systems, such as `b-bm25-dense-rerank` and `b-bm25-dense-rerank-v2` show an exceptional performance increase from the first to the second snapshot. Aside from these systems, the effectiveness across snapshots remains stable. The scores on the DCTR qrels are also much lower than those on the synthetic labels. Figure 1 highlights both findings as well.

### 2.3.2. Robustness

Beyond being effective at individual points in time, a good retrieval system also needs to maintain its effectiveness over time, even when the search setting evolves. We assess how the dynamics influence the system rankings by measuring the correlation between snapshots based on the rank with Kendall's $\tau$ and based on the effectiveness with Pearson's correlation. Comparing Snapshot 2 and Snapshot 1, the Pearson correlation is 0.458, indicating only a medium positive relation. This is similar for Snapshot 3 compared to Snapshot 1, which shows a correlation of 0.479. Likewise, Kendall's $\tau$ is 0.574 and 0.515 for Snapshot 2 and Snapshot 3 comparisons, respectively. On the synthetic relevance judgments with GPT-OSS-120b, the correlation is similar. Pearson's correlations compared to Snapshot 1 are slightly lower, with 0.414 and 0.404 for snapshots 2 and 3. In comparison to the DCTR qrels, Kendall's $\tau$ is slightly higher, with 0.684 and 0.625. All p-values indicate significance at an $\alpha$ level of 0.05.

We also assess the temporal robustness of individual approaches. For this task, González Sáez [33] proposed the Result Delta ($\mathcal{R}\Delta$) framework, by which the effectiveness is measured per snapshot and then compared over time using a comparison strategy. The effectiveness is measured as the Average Retrieval Performance (ARP) across many queries using established measures, such as nDCG@10. Different comparison strategies can be employed to factor or emphasize different aspects of temporal robustness. We primarily focus on two measures to assess the temporal robustness: the Relative Change (RC) and the Delta Relative Improvement (DRI). However, temporal robustness is an active research

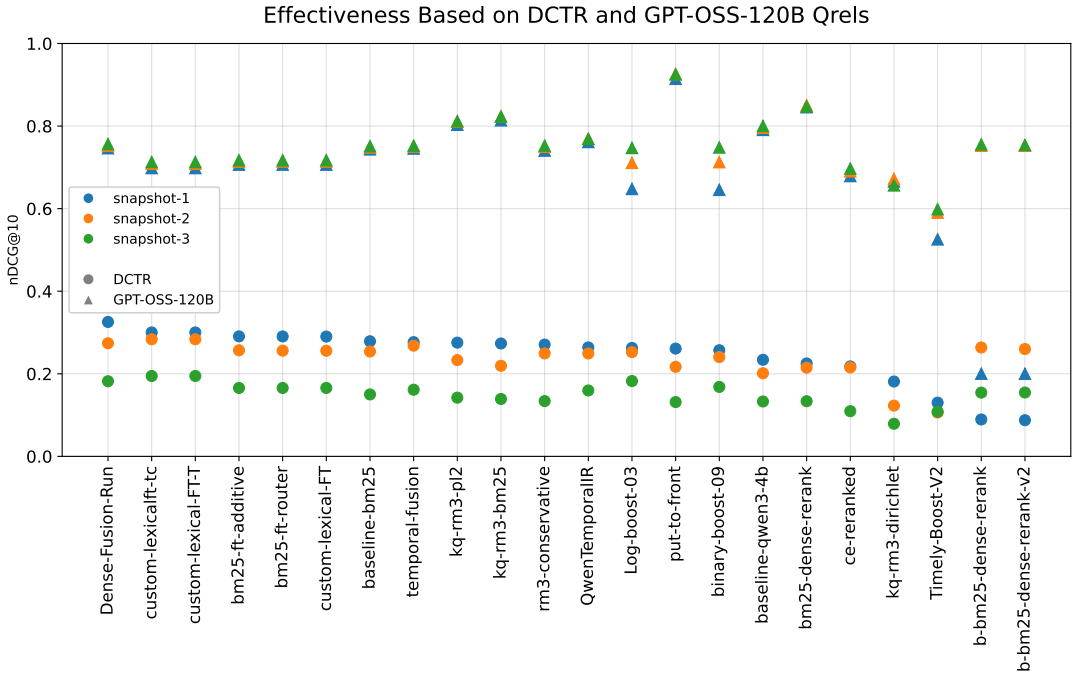


**Figure 1:** Effectiveness measured by nDCG@10 across all approaches and test snapshots using the DCTR qrels (opaque) and the synthetic relevance labels generated with GPT-OSS-120b (faded).

area, and Standards have yet to be established. The RC measures the effectiveness delta between a target snapshot (e.g. $t_2$) and the baseline snapshot ($t_1$), normalized by the baseline effectiveness at $t_1$ . It offers an intuitive view of temporal degradation or improvement and is defined as:

$$\text{RC} = \frac{\text{ARP}_{t_1} - \text{ARP}_{t_2}}{\text{ARP}_{t_1}}. \tag{1}$$

A relative change of 0 indicated that the effectiveness remains stable and the system is robust. A negative RC indicates an improved effectiveness and a positive RC an impaired effectiveness. This measure maintains continuity with methodologies established in previous iterations of the lab [34, 35, 12].

The DRI quantifies the relative change in effectiveness (e.g., nDCG@10) anchored against a designated *Pivot system*. By normalizing the performance of the experimental system at both $t_1$ and $t_2$ against the Pivot system at those exact timestamps, DRI separates system-specific variations from environmental noise (such as collection drift or intent shifts in the testbed) [36, 37]. For this evaluation we rely only on the DCTR qrels, select the first snapshot as the reference point, and the BM25 baseline as Pivot system [38]. The results are displayed in Table 4 together with the nDCG@10 effectiveness (ARP) in the evolved snapshots.

As we noted earlier in the effectiveness evaluation, the effectiveness fluctuates over time. This fluctuation is directly described in the RC. In the second snapshot, compared to the first snapshot, all approaches except `b-bm25-dense-rerank` and `b-bm25-dense-rerank-v2` show reduced effectiveness, indicated by a positive RC. The approaches with the least change (RC close to 0) compared to the second snapshot are `ce-reranked` (0.011), `temporal-fusion` (0.031), and `Log-boost-03` (0.038). Towards the third snapshot, this negative trend is further continued as most approaches impair the effectiveness. The approaches `Timely-Boost-V2` (0.169), `Log-boost-03` (0.305), and `binary-boost-09` (0.345) show the least change. Again, only `b-bm25-dense-rerank` and `b-bm25-dense-rerank-v2` improve their scores.

The DRI dampens the environmental effect and emphasizes the changes in effectiveness due to the system effect. For Snapshot 2, 9 approaches show a low positive DRI, which indicates that the effectiveness compared to BM25 improves in this new snapshot. The remaining 11 systems show decreased relative deltas. `rm3-conservative`, `binary-boost-09`, and `bm25-ft-additive` have

**Table 4**
Robustness of approaches compared to Snapshot 1, based on the DCTR qrels set. Since BM25 is used as the Pivot system, the DRI is 0, indicating no change. The baselines are gray and the best result per measure is **bold**.

| | Snapshot 2 | | | Snapshot 3 | | |
|---|---|---|---|---|---|---|
| | nDCG@10 | RC | DRI | nDCG@10 | RC | DRI |
| kq-rm3-dirichlet [3] | 0.123 | 0.322 | 0.167 | 0.079 | 0.564 | 0.124 |
| rm3-conservative [2] | 0.249 | 0.079 | **-0.010** | 0.134 | 0.506 | 0.079 |
| ce-reranked [air-02] | 0.215 | **0.011** | -0.067 | 0.109 | 0.498 | 0.052 |
| put-to-front [3] | 0.217 | 0.169 | 0.082 | 0.132 | 0.496 | 0.060 |
| kq-rm3-bm25 [3] | 0.219 | 0.198 | 0.117 | 0.139 | 0.492 | 0.054 |
| kq-rm3-pl2 [3] | 0.233 | 0.154 | 0.071 | 0.142 | 0.484 | **0.040** |
| baseline-bm25 | 0.254 | 0.089 | 0.000 | 0.150 | 0.462 | 0.000 |
| Dense-Fusion-Run [4] | 0.274 | 0.158 | 0.089 | 0.182 | 0.441 | -0.045 |
| baseline-qwen3-4b | 0.201 | 0.139 | 0.047 | 0.133 | 0.432 | -0.047 |
| bm25-ft-additive [1] | 0.257 | 0.116 | 0.031 | 0.166 | 0.430 | -0.062 |
| bm25-ft-router [1] | 0.256 | 0.118 | 0.034 | 0.166 | 0.430 | -0.062 |
| custom-lexical-FT [1] | 0.256 | 0.118 | 0.034 | 0.166 | 0.429 | -0.064 |
| temporal-fusion [2] | 0.268 | 0.031 | -0.063 | 0.161 | 0.417 | -0.084 |
| bm25-dense-rerank [nlp hashiras] | 0.215 | 0.045 | -0.039 | 0.134 | 0.406 | -0.084 |
| QwenTemporalIR [2] | 0.249 | 0.056 | -0.034 | 0.160 | 0.395 | -0.118 |
| custom-lexical-FT-T [1] | **0.284** | 0.055 | -0.040 | **0.195** | 0.351 | -0.221 |
| custom-lexicalft-tc [1] | **0.284** | 0.055 | -0.040 | **0.195** | 0.351 | -0.221 |
| binary-boost-09 [4] | 0.241 | 0.064 | -0.025 | 0.168 | 0.345 | -0.200 |
| Log-boost-03 [4] | 0.253 | 0.038 | -0.052 | 0.182 | 0.305 | -0.274 |
| Timely-Boost-V2 [4] | 0.106 | 0.184 | 0.049 | 0.108 | **0.169** | -0.254 |
| b-bm25-dense-rerank [air benders] | 0.264 | -1.951 | -0.718 | 0.154 | -0.728 | -0.709 |
| b-bm25-dense-rerank-v2 [air benders] | 0.260 | -1.969 | -0.709 | 0.155 | -0.766 | -0.717 |

the least change, indicated by a DRI closest to 0. In Snapshot 3, only six approaches have an improved delta to BM25. `kq-rm3-pl2`, `Dense-Fusion-Run`, `bm25-ft-additive`, and `ce-reranked` have the lowest change in Snapshot 3.

## 2.4. Discussion

The LongEval-Sci task continues the research on longitudinal evaluations for the scientific domain, started in the 2025 iteration of the lab. It directly addresses the lab's core goal: to investigate changes in effectiveness, including potential drops, and to develop strategies to maintain it. The new snapshots added this year span longer time frames to capture more dynamics of the slower-paced scientific domain, in comparison to Web search with more diverse dynamics.

For almost all systems, the effectiveness measured on the DCTR qrels decreases over time. Meanwhile, the collection size increases, which might explain this observation. Only six approaches outperform the BM25 baseline consistently on all snapshots. On the synthetic qrels, the Qwen3 baseline performs better and is consistently outperformed by only four approaches over time. Compared to the findings of the SciRetrieval task of 2025, the ranking of systems is more dynamic, indicated by a lower rank correlation. Similarly, we observe again that effectiveness and robustness seem to be independent goals, and improving one often impairs the other. Furthermore, temporal robustness does not seem to depend on any particular approach class.

In this fourth iteration of the LongEval Lab, the submitted approaches increasingly relied on the temporal properties of the dynamic test collection, such as the publication date or document recency, to create effective rankings. The results showed that this metadata can be used to construct valuable relevance signals, but it does not guarantee an effective approach. For example, document recency does not appear to be a driving factor of relevance on this testbed. While temporal signals seem especially valuable when the DCTR qrels are used for evaluation, the synthetic LLM judgments seem much less

influenced by these signals. For the first time, we also provided citation data for the collection. This metadata was used for simple signals by two teams, but had only a marginal positive effect.

For the first time, we used synthetic relevance labels generated by LLMs to measure the effectiveness of systems. While the general viability of LLM judgments is widely discussed [39, 40, 41], they have unique and unexplored implications for longitudinal evaluations. On the positive side, synthetic judgments are highly scalable, cheaper than editorial judgments, and avoid the position bias inherent in click-based pseudo-qrels (provided the judgment pool is diverse). This scalability fixes a critical bottleneck in continuous evaluations, allowing for rapid, high-volume judgments for each new snapshot. However, these benefits are offset by a critical limitation: temporal asymmetry. Because an LLM's world knowledge is permanently frozen at its training cutoff, true in-situ (synchronous) assessments are impossible. Assuming deterministic outputs, a static LLM lacks temporal awareness; it will evaluate a query-document pair identically whether the assessment happens today or a year from now. This limitation is visible in our results (cf. Figure 1), where effectiveness measured by LLMs barely varies across the three snapshots, contrasting sharply with the natural variance seen in DCTR qrels. Ultimately, LLMs cannot capture temporal dependence in relevance, instances where a document's relevance evolves due to real-world, external events, because the model remains blind to the passage of time.

### 2.5. Conclusion

The 2026 iteration of the LongEval lab continued the SciRetrieval task in which we investigate the temporal robustness and effectiveness of ad-hoc retrieval systems. Submitted approaches are systematically evaluated across a cumulatively growing dynamic test collection, of three snapshots each spanning three months. We received submissions from eight teams with a diverse methodological coverage. The majority of approaches rely on multi-stage pipelines, augmenting classic lexical retrieval with cross-encoder re-ranking, query expansion, rank fusion, and explicit temporal features such as publication dates and citation metrics. The results demonstrate that maintaining retrieval effectiveness over time remains a complex challenge. As the collection scaled from 87K to over 1.7M documents, the average system effectiveness fluctuates. Furthermore, the system rankings across time proved highly dynamic. The moderate rank correlations (Kendall's $\tau$ 0.574 and 0.515) across snapshots highlight the variability in effectiveness, indicating that peak performance on an initial dataset does not guarantee sustained performance as the target corpus evolves.

Moving forward, future work in longitudinal IR needs to find more sophisticated integrations of temporal relevance signals and potentially merge with citation metadata, such as dynamic citation networks or intent-dependent recency boosts. Most critically on the evaluation side, reliable testbeds with clear temporal effects and in-time (synchronous) relevance assessments are needed. Ultimately, the challenge remains to design retrieval architectures that not only gracefully degrade effectiveness but also actively adapt to the temporal dynamics of the scholarly landscape.

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

The LongEval-TopEx task, introduced in this edition of LongEval, aims to extract formal TREC-style retrieval topics from query logs. This is to study differences between relevance judgments derived from user interactions and those created through expert annotations in Cranfield-style retrieval evaluations.

Cranfield-style retrieval evaluations should ideally be reliable in the sense that repeating an experiment under similar conditions yields similar observations (e.g., System A is more effective than System B) with a high probability [42]. In the context of LongEval, relevance judgments are derived from user interactions via clicks from the production search engine. Deriving relevance judgments from user interactions has the disadvantage that new retrieval systems can not be reliably evaluated on past clicks, as the new retrieval system can retrieve documents that the production retrieval system has not retrieved and that therefore could not receive clicks. The goal of Task 2 of LongEval is to extract formalized information needs from the query log of Task 1 for reliable Cranfield-style evaluations.

**Table 5**
Example input and output for Task 2. Given an entry from the query log, predict a formalization of the query using a description and narrative that aligns with observed clicks over time and allows for reliable evaluations.

| | Field | Example |
|---|---|---|
| **Input** | Query | `ransomware detection` |
| **Output** | Description | I want to know which algorithms for ransomware detection are effective. |
| | Narrative | Papers that describe algorithms for ransomware detection are relevant when they also have an evaluation. Evaluation papers that just compare multiple ransomware detection algorithms are also relevant. Other aspects, such as describing which ransomware attacks exist, the history of ransomware detection, etc., are not relevant. |

Perfectly formalized topics could explain user behavior of the future (e.g., they could predict clicks) and still allow to evaluate new retrieval systems.

Table 5 shows the input and output of Task 2 in an example. We use the formalized information needs as input to large language model relevance assessors, and the quality of the formalized topics is measured via the relevance judgments created by the large language models. The motivation for the topic extraction task comes from experiments that did show that formalized information needs improve the relevance judgments [43]. Overall, 3 teams submitted 23 formalized topic sets to Task 2. For each of the 23 submitted topic sets, we used four large language model relevance assessors to create relevance judgments, enriching Task 1 with 92 variants of relevance judgments.

### 3.1. Datasets

We use the test dataset of Task 1. Participants have access to the test queries and documents. For the evaluation, we use all retrieval systems that are submitted to Task 1 and pool them via top-10 pooling and use four large language model relevance assessors that use the formalized topics to generate relevance judgments. Altogether, the evaluation setup of Task 2 combines 23 variants of formalized topics for which four large language models generated relevance judgments together with 22 retrieval runs submitted for Task 1, which allows to study the reliability of evaluations in the LongEval scenario.

### 3.2. Submissions and Approaches

In total, three teams submitted a total of 23 runs via TIRA [19].

**Team CIR** had a focus on contrasting human-formalized topics with automatically formalized topics. Information science students manually formalized 30 topics with ChatNoir [44]. For the automatically formalized topics, related concepts, summaries, and rationales of queries were extracted using large language models.

**Team VerbaNexAI** developed an agentic multi-stage approach for automatic topic formalization that combines pseudo relevance feedback in PyTerrier [24], manually developed rules, and multiple iterative prompting techniques for large language models (e.g., Chain-of-Thought, self-critique loops, etc.).

**Team JOINorDIE** developed automatic topic formalization with large language models in a contrastive approach with PyTerrier. In different variants, the top-3 retrieved documents were used as positive documents, whereas documents on position 100 to 102 were used as negative documents to formalize the topics.

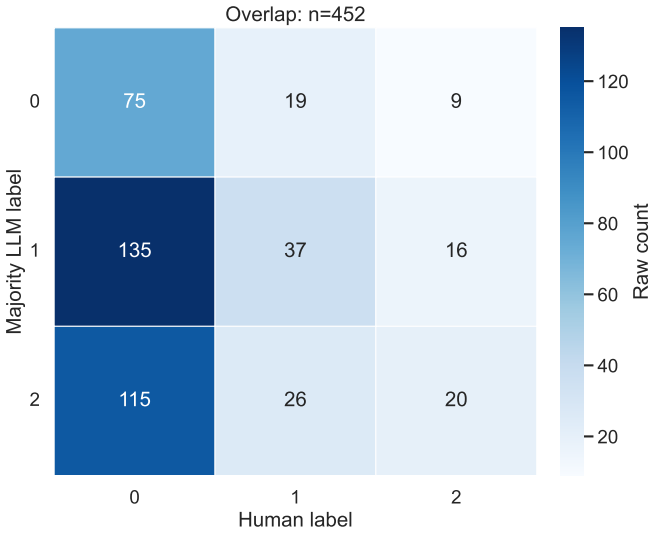


**Figure 2:** The confusion matrix of relevance judgments from large language models (the majority vote of all LLMs with all topic formalizations) and the human labels derived from click logs over the three snapshots.

**Baseline** to compare the submissions of the three teams, we used a template to use the original query as description and narrative.

### 3.3. Results

We use the UMBRELA [45] prompt to make relevance judgments with four large language models for the all formalized topics. We pooled all 24 retrieval systems of Task 1 via depth-10 pooling and use four large language models, namely, qwen-qwen3-32b, openai-gpt-oss-20b, openai-gpt-oss-120b, and llama-3.1-8b. To reduce the costs and ensure that all submissions can be compared with each other, we restricted our evaluation to the 30 topics of the submission by team CIR that manually formulated topics (all automatic submissions are a superset of those 30 topics). We evaluate submissions along three dimenstions:

1. Not too many documents should be judged as relevant for a topic. Multiple studies [46, 47] have shown that large language models judge much more documents as relevant than humans. When too many documents are relevant for a topic, retrieval experiments are not reliable [48].

2. The agreement between different annotators (large language models in our case) should be high. If the agreement would be low, the formalization of the topic leaves still room for ambiguity. If it is high, annotators come independently to identical results which is good for the reliability (as the experiment could be repeated with different annotators).

3. The alignment between the annotations (large language models in our case) and observations from the production system (clicks from the core search engine in our case) should be high. If the alignment is high, annotations could be used to expand judgments/clicks to new documents.

Table 6 shows the percentage of documents that are judged as relevant for different scenarios of creating relevance judgments. High scores indicate that relevance judgments can not be re-used to evaluate new retrieval systems. The baseline judgments yield in the mean 86.7 % of documents as relevant. This is in a drastic contrast to the relevance judgments derived from the production system via click qrels that only deem 12.7 % of the documents as relevant. The overall trend is that all LLM relevance judgments are (as expected from the literature [46, 47]) more positive than the observations from the production system, but some formalized topics can reduce the percentage of documents that are judged as relevant substantially, down to 33.5 % for the submission cir-l1f.

**Table 6**
The percentage of documents that are judged as relevant. We show the assessor (e.g., one of our four LLMs respectively observations of humans) and the submission method. Too high percentages of relevant documents can indicate that relevance judgments can not be applied to new retrieval systems.

| Team | Method | Relevant | | | | |
|---|---|---|---|---|---|---|
| | | oss-120b | oss-20b | llama-3 | qwen3 | mean |
| CIR | cir-h1-raw | 0.820 | 0.852 | 0.931 | 0.873 | 0.869 |
| Baseline | naive-topics | 0.825 | 0.810 | 0.949 | 0.882 | 0.867 |
| JOINorDIE | o-contrastive-5tfidf | 0.787 | 0.798 | 0.940 | 0.876 | 0.850 |
| VerbaNexAI | v9-query-augmented | 0.766 | 0.751 | 0.926 | 0.840 | 0.821 |
| JOINorDIE | ollama-rel-5tfidf | 0.751 | 0.765 | 0.908 | 0.851 | 0.818 |
| JOINorDIE | ollama-non-rel-5tfid | 0.701 | 0.739 | 0.900 | 0.806 | 0.787 |
| CIR | cir-l1-raw | 0.671 | 0.849 | 0.892 | 0.728 | 0.785 |
| CIR | cir-manual | 0.701 | 0.658 | 0.906 | 0.830 | 0.774 |
| JOINorDIE | openai-cont-tfidf5 | 0.601 | 0.550 | 0.907 | 0.758 | 0.704 |
| JOINorDIE | openai-rel-tfidf5 | 0.597 | 0.545 | 0.881 | 0.754 | 0.694 |
| JOINorDIE | openai-norel-tfidf5 | 0.572 | 0.519 | 0.880 | 0.702 | 0.669 |
| CIR | cir-h1c | 0.498 | 0.433 | 0.858 | 0.687 | 0.619 |
| CIR | cir-h1d | 0.467 | 0.371 | 0.841 | 0.657 | 0.584 |
| CIR | cir-h2f | 0.453 | 0.380 | 0.839 | 0.653 | 0.581 |
| CIR | cir-h1fd | 0.425 | 0.354 | 0.836 | 0.619 | 0.559 |
| CIR | cir-h1 | 0.403 | 0.343 | 0.831 | 0.648 | 0.556 |
| CIR | cir-h2 | 0.417 | 0.349 | 0.813 | 0.636 | 0.554 |
| CIR | cir-h1f | 0.404 | 0.351 | 0.832 | 0.619 | 0.551 |
| CIR | cir-l2 | 0.234 | 0.202 | 0.779 | 0.478 | 0.423 |
| CIR | cir-l1cf | 0.262 | 0.238 | 0.745 | 0.406 | 0.413 |
| CIR | cir-l2f | 0.185 | 0.145 | 0.757 | 0.430 | 0.379 |
| CIR | cir-l1 | 0.192 | 0.162 | 0.741 | 0.389 | 0.371 |
| CIR | cir-l1c | 0.198 | 0.166 | 0.740 | 0.365 | 0.367 |
| CIR | cir-l1f | 0.159 | 0.129 | 0.726 | 0.327 | 0.335 |
| Production | click-qrels | | | | | 0.127 |

To get a better understanding how the relevance judgments of the large language models and the human relevance judgemts from the click logs differ, Figure 2 shows a confusion matrix between those two types of relevance judgments. We combine all LLM relevance judgments that we have into a single set of relevance judgments via majority vote over all LLMs and topic formalizations and show for overlapping judgments (i.e., judgments available in the human relevance judgments and in the LLM judgment pool) how the LLM judgments translate into the human labels. Figure 2 shows that the trend that LLMs are too positive compared to their human judgments is visible for all LLM relevance labels: 71.8 % of the LLM labels of relevance 1 turn out to be 0 in the human labels, and 71.4 % of the LLM labels of relevance 2 turn out to be 0 in the human labels. To better understand the difference, we next investigate the agreement between LLM annotators and human judgments in more detail.

Table 7 shows the agreement of different LLM annotatores for the topics formalized in different ways. We use our four LLMs, and compare the agreement between all pairs of LLMs. We use two notions of agreement. First, we measure Cohens Kappa on the individual relevance decisions (higher values are better). A Cohens Kappa of 1 indicates that two different LLMs would make the same relevance judgments, whereas 0 indicates that all agreement is only random. Second, we measure the agreement on the system ranking via Tau-AP for nDCG@10. For each pair of LLM relevance judgments, we order the retrieval systems of Task 1 by their nDCG@10 score for the given LLM relevance judgments and compare the correlation between the two nDCG@10 system rankings. A Tau-AP of 1 indicates that the system rankings are identical, whereas 0 indicates that there is no correlation. The manually formalized topics in the cir-manual submission yield the highest agreement on the system ranking level (Tau-AP

**Table 7**
The agreement on the observations between all pairs of our four LLM relevance assessors. We show the agreement for individual relevance judgment decisions as the Cohens Kappa (higher is better) and the agreement on the resulting system ranking for nDCG@10 as TauAP correlation (higher is better).

| Team | Method | Cohens Kappa | Tau-AP |
|---|---|---|---|
| CIR | cir-manual | 0.305 | 0.775 |
| Baseline | naive-topics | 0.370 | 0.732 |
| VerbaNexAI | v9-query-augmented | 0.387 | 0.708 |
| CIR | cir-l1-raw | 0.196 | 0.654 |
| JOINorDIE | o-contrastive-5tfidf | 0.318 | 0.644 |
| JOINorDIE | ollama-non-rel-5tfid | 0.293 | 0.642 |
| JOINorDIE | ollama-rel-5tfidf | 0.318 | 0.639 |
| JOINorDIE | openai-norel-tfidf5 | 0.298 | 0.607 |
| JOINorDIE | openai-cont-tfidf5 | 0.311 | 0.605 |
| JOINorDIE | openai-rel-tfidf5 | 0.323 | 0.598 |
| CIR | cir-h1-raw | 0.240 | 0.576 |
| CIR | cir-h1fd | 0.229 | 0.525 |
| CIR | cir-h1d | 0.245 | 0.524 |
| CIR | cir-h1c | 0.244 | 0.519 |
| CIR | cir-h2 | 0.217 | 0.491 |
| CIR | cir-h2f | 0.211 | 0.469 |
| CIR | cir-l2 | 0.127 | 0.440 |
| CIR | cir-h1 | 0.204 | 0.424 |
| CIR | cir-h1f | 0.225 | 0.408 |
| CIR | cir-l1c | 0.154 | 0.296 |
| CIR | cir-l1 | 0.162 | 0.284 |
| CIR | cir-l1f | 0.161 | 0.257 |
| CIR | cir-l1cf | 0.208 | 0.252 |
| CIR | cir-l2f | 0.137 | 0.211 |

of 0.775), whereas the submission of team VerbaNexAI yields the highest agreement on the individual document level (Cohens Kappa of 0.387).

Table 8 shows the alignment of different LLM annotators with the human relevance judgments derived from the click logs. For each of the four LLMs that we use and the different ways to formalize topics. We observe that there is no agreement between the relevance judgments derived from the click logs and the relevance judgments derived from the formalized topics (Cohens Kappa is very close to zero in all cases). Still, the system rankings are moderately correlated, as measured by Tau-AP, reaching up to 0.533 for submission cir-l1-raw, even when the agreement on the individual document level as measured by Cohens Kappa is close to zero. Still, the results highlight that there is substantial disagreement between the click logs and the large language model relevance judgments, highlighting that there is currently no good alignment between the click logs and the formalized topics. Identifying approaches that could increase this alignment would be interesting directions for future work.

## 4. Task 3. LongEval-USim: User Simulation

In this year's edition of the LongEval lab, we included a new task on user simulation, more specifically the simulation of query formulations. For this task, the CORE log file-based dataset in LongEval differs from typical test collections not only in the temporal characteristics described earlier but also in the inclusion of logged user interactions in academic search tasks.

In academic search, we observe interaction patterns that are less common in other (web) search tasks. As described by Li et al. [49], queries in academic search scenarios are special as they have a substantially bigger proportion of entity queries, and a heavier tail in query length distribution and lead to approx. 10% of empty result sets. Some typical session behaviors only occur in academic search

**Table 8**
The alignment for each LLM relevance assessor for each topic generation approach. We show the alignment for individual relevance judgment decisions as the Cohens Kappa (higher values are better) and the alignment on the resulting system ranking for nDCG@10 as TauAP correlation (higher values are better).

| Method | Cohens Kappa | | | | | Tau-AP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | oss-120b | oss-20b | llama-3 | qwen3 | mean | oss-120b | oss-20b | llama-3 | qwen3 | mean |
| cir-l1-raw | -0.018 | 0.004 | 0.011 | -0.011 | -0.003 | 0.527 | 0.527 | 0.490 | 0.589 | 0.533 |
| cir-h1c | -0.082 | -0.033 | -0.034 | 0.005 | -0.036 | 0.487 | 0.404 | 0.541 | 0.632 | 0.516 |
| cir-h1-raw | -0.017 | 0.018 | 0.013 | 0.007 | 0.005 | 0.530 | 0.514 | 0.483 | 0.517 | 0.511 |
| rel-5tfid | -0.019 | -0.026 | -0.007 | 0.019 | -0.008 | 0.393 | 0.497 | 0.502 | 0.605 | 0.499 |
| cir-manual | -0.031 | -0.029 | 0.014 | 0.010 | -0.009 | 0.458 | 0.459 | 0.526 | 0.497 | 0.485 |
| norel-tfidf5 | -0.011 | -0.014 | 0.011 | -0.009 | -0.006 | 0.413 | 0.398 | 0.474 | 0.614 | 0.475 |
| Baseline | -0.008 | 0.010 | -0.018 | -0.007 | -0.006 | 0.432 | 0.494 | 0.503 | 0.458 | 0.472 |
| v9 | 0.013 | 0.023 | 0.016 | 0.012 | 0.016 | 0.463 | 0.519 | 0.452 | 0.438 | 0.468 |
| rel-5tfidf | -0.008 | -0.021 | -0.008 | -0.014 | -0.012 | 0.446 | 0.403 | 0.452 | 0.518 | 0.455 |
| cir-h2f | 0.007 | 0.012 | 0.006 | -0.046 | -0.005 | 0.352 | 0.402 | 0.525 | 0.504 | 0.446 |
| contr-5tfidf | 0.008 | -0.027 | 0.037 | -0.002 | 0.004 | 0.408 | 0.445 | 0.509 | 0.412 | 0.443 |
| rel-tfidf5 | 0.031 | -0.011 | -0.017 | 0.026 | 0.007 | 0.365 | 0.373 | 0.551 | 0.419 | 0.427 |
| cont-tfidf5 | -0.003 | 0.044 | -0.014 | -0.002 | 0.006 | 0.393 | 0.445 | 0.419 | 0.424 | 0.420 |
| cir-h1fd | -0.016 | 0.003 | -0.006 | -0.023 | -0.011 | 0.315 | 0.377 | 0.457 | 0.523 | 0.418 |
| cir-h1 | 0.002 | 0.012 | 0.002 | -0.052 | -0.009 | 0.342 | 0.288 | 0.444 | 0.594 | 0.417 |
| cir-h1d | 0.002 | 0.002 | -0.048 | -0.017 | -0.015 | 0.393 | 0.245 | 0.458 | 0.517 | 0.403 |
| cir-h1f | 0.008 | 0.007 | 0.016 | -0.023 | 0.002 | 0.246 | 0.382 | 0.449 | 0.519 | 0.399 |
| cir-h2 | -0.015 | 0.012 | -0.059 | -0.018 | -0.020 | 0.420 | 0.111 | 0.460 | 0.560 | 0.388 |
| cir-l2f | -0.003 | 0.015 | -0.023 | -0.023 | -0.009 | 0.157 | 0.455 | 0.568 | 0.313 | 0.373 |
| cir-l2 | -0.014 | -0.015 | 0.026 | 0.024 | 0.005 | 0.313 | 0.256 | 0.389 | 0.491 | 0.362 |
| cir-l1 | 0.016 | 0.013 | 0.010 | -0.004 | 0.009 | 0.324 | 0.289 | 0.463 | 0.256 | 0.333 |
| cir-l1c | 0.002 | 0.011 | -0.000 | -0.014 | -0.000 | 0.225 | 0.358 | 0.450 | 0.289 | 0.330 |
| cir-l1f | 0.005 | 0.010 | -0.032 | -0.040 | -0.014 | 0.196 | 0.238 | 0.536 | 0.336 | 0.326 |
| cir-l1cf | 0.007 | 0.021 | -0.006 | 0.030 | 0.013 | 0.196 | 0.129 | 0.345 | 0.283 | 0.238 |

**Table 9**
Session statistics for the three snapshots for Task 3

| Snapshot | # Sessions | Avg. Len. | Med. Len. | Std. Len. |
|---|---|---|---|---|
| Snapshot-1 (Train, Mar-May) | 180 | 3.19 | 3.0 | 0.58 |
| Snapshot-2 (Test, Jun-Aug) | 119 | 3.33 | 3.0 | 1.27 |
| Snapshot-3 (Test, Sep-Nov) | 182 | 3.13 | 3.0 | 0.39 |

settings, such as journal runs, footnote chasing [50]. These and other characteristics like a high level of boolean queries may lead to interaction behaviors that are not in the focus of other user modeling and simulation tasks, enabling a wide range of interesting research scenarios.

Therefore, in this new LongEval task, we invite participants to model different user behaviors and types of user simulations, with a particular focus on query formulations. This approach was explored in the SIGIR 2025 Workshop on Simulations in Information Access (Sim4IA) [51], in the form of a micro-shared task, where the evaluation of query (re-)formulation simulations was first tested out.

## 4.1. Dataset

From the original CORE search log files, we extract a set of interaction sessions, similar to the ones in Sim4IA (see Table 9). With the `search_id`, a fingerprint `uid`, and different session detection 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. More specifically, we provided access to:

(1) The sequence of queries submitted.

(2) Timestamps for query submissions.

(3) The top-10 SERP entries retrieved for each query.

(4) Clicked documents (and click type).

Participants in this task get access to pre-filtered session log excerpts. From here, they had to predict the next query in the sequence of interactions. The participants had a wide range of options to detect and train specific user models on historical interaction logs and to test those models on the test data in the final LongEval round. To simplify this task, we provide a pre-configured SimIIR environment that is part of the Sim4IA Bench [52]. It allows 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 them can provide more information about user intentions than queries alone.

## 4.2. Submissions

In total, we received 29 runs from four teams. The submissions were done through the TIRA platform, where participants could upload their runs. We give a brief summary of the submissions based on the information provided by the participants.

**Team JOINorDIE** submitted a total of 18 runs using an in-context prompting approach. For prediction, the LLM was provided with different in-context signals, including selections of previous queries, a LLM-generated TREC-style topic derived from the query history, and prompt scenario variations for Prompt A. For Prompt B, the input consisted of previous queries and pseudo-relevant and non-relevant documents (title and abstract), where interacted documents were treated as relevant, while the remaining documents from the top-10 retrieval list were considered non-relevant.

For the document context in Prompt B, three prompting strategies—namely relevant, non-relevant, and contrastive—were evaluated. These strategies determined which documents were included as context in the prompt, with up to three documents for the relevant and non-relevant strategies and up to six documents for the contrastive strategy. Based on these varying context configurations, the LLM generated candidate queries.

**Team Split 'n Simulate** submitted one run in which they first performed a qualitative analysis of the provided sessions to identify eight distinct user groups. These personas were assigned by an LLM based on additional descriptive observations of each user group. Using the assigned persona together with the complete previous session history as context, the LLM was then used to predict the next query. For the prediction the LLM was additionally provided with characteristic formulation patterns for the assigned persona. A high temperature setting was used to encourage greater diversity among the candidate predictions for each session. After generation, the five candidate queries were re-ranked using their cosine similarity to the second-to-last query in the session as a heuristic.

**Team Promptly** performed six experimental runs to evaluate different reasoning strategies, including Chain-of-Thought (CoT) and Self-Refinement. Furthermore, prompts were personalized through the integration of personas and tags representing characteristics of the respective session, such as the verbosity of previous queries or the depth of the session. Across all runs, the LLM was provided with the complete history of query reformulations as additional context.

**Team VerbaNexAI** submitted four runs based on the idea of treating Candidate Generation and Candidate Selection as two distinct but complementary problems. Candidate Generation combines four complementary strategies: heuristic reformulation of the last visible query, transition memory to retrieve query candidates from similar sessions from the training dataset, a controlled LLM-based reformulation using the USimAgent framework, and role-guided refinement of strong seed queries.

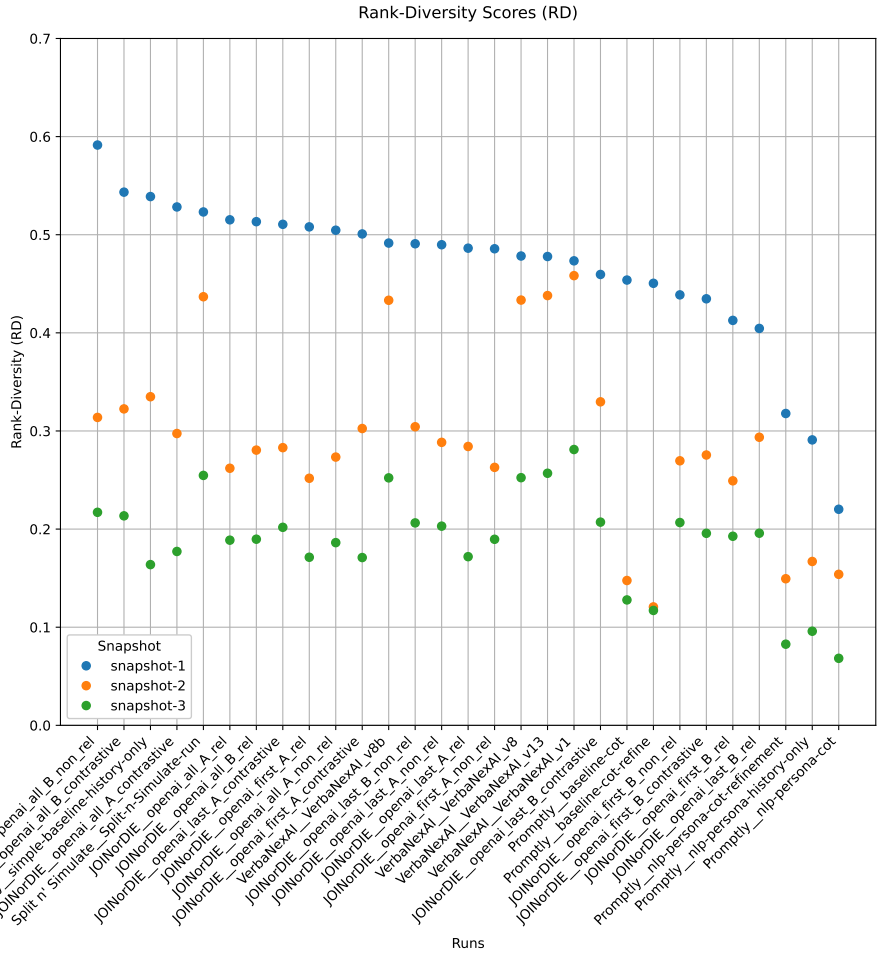


**Figure 3:** Similarity to the Ground-Truth Next Query evaluated using Rank-Diversity Score across sessions.

For Candidate Selection, two approaches were used. The first is a selector that balances relevance to the session context with diversity among selected queries, incorporating additional quality filters. The second is a portfolio-based selector that aggregates candidates across generation strategies using source reliability weights and cross-source agreement, followed by a diversity filter to remove near-duplicates.

## 4.3. Analysis

For the evaluation, we compare the simulated queries to the queries from the session logs, for which we rely on query comparison measures evaluated in earlier work by Kruff et al. [53]. Specifically, we look at two main aspects: (1) the inability to distinguish simulated from real data, evaluated using the Rank-Diversity Score (RDS) [52]; and, in addition, a separate evaluation focusing on indistinguishability, where the top-1 candidate (based on the submitted ranking) is compared to the ground truth using the average cosine similarity across sessions, and (2) the robustness of the approach, measured via the Relative Change (RC) [34, 35, 12, 54]. Similar to the other tasks' definition, robustness is defined as the ability of a simulator to perform consistently well across different snapshots.

Table 10 illustrates the performance of the various runs based on the RDS, differentiated by the underlying snapshot and sorted by the score of snapshot 1. While most runs achieve strong performance on snapshot 1, a clear performance gap can be observed for snapshots 2 and 3. Overall, performance on snapshot 1 is consistently higher than on snapshot 2, which in turn outperforms snapshot 3 across all submitted runs, suggesting potential temporal effects influencing the users' query behavior that is to be predicted.

While most runs perform substantially better on snapshot 1 compared to the other two snapshots,

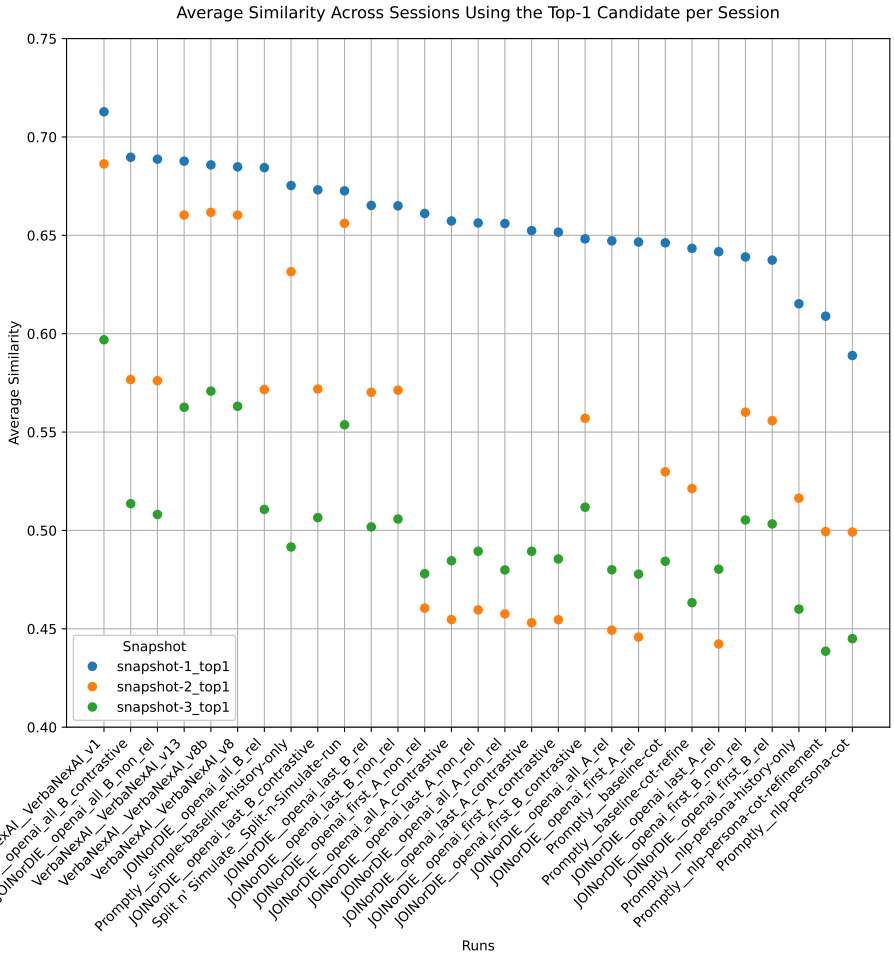


**Figure 4:** Cosine Similarity of the Top-1 Candidate per Session.

some runs, in particular those by Team Split n' Simulate and Team VerbaNexAI, remain relatively strong on snapshot 2. Moreover, snapshot 3 performance for these runs is largely on par with the snapshot 2 results of other teams. To further assess robustness across snapshots, Table 10 also reports the RC, rewarding stability in performance over time.

Overall, the systems by Split n' Simulate and VerbaNexAI show the most stable performance across snapshots. Both approaches consider diversity as well as robust ranking as part of the query candidate selection, achieving consistently high RDS scores. This indicates that their systems generate high-quality, non-redundant query sets while also reliably selecting strong top-1 candidates with high semantic relevance. Overall, this suggests that query candidate selection is as critical as query candidate generation for achieving strong performance. In an additional analysis comparing the average performance of the Top-1 query selected by participants with that of the best-performing query among the five candidates, across both cosine similarity and SERP overlap, it was observed that the heuristic re-ranking approach used by Team Split n' Simulate yielded only marginal benefits, if any. In contrast, Team VerbaNexAI achieved notably stronger results, with only very small differences between Top 1 and Best, indicating that the highest-quality queries were consistently ranked first.

## 4.4. Discussion

The following discussion focuses on the effectiveness of the submitted approaches in terms of their ability to produce queries that are difficult to distinguish from the original data, as measured by the Rank-Diversity Score (RDS) and the Relative Change (RC). Across the submitted runs, several consistent patterns emerge.

**Table 10**
Effectiveness of the simulators for the three snapshots, measured by RDS (Rank-Diversity Score) and RC (Relative Change). The systems are sorted by the RDS based on the first snapshot in decreasing order.

| **Measures** | | **RDS** | | **RC** | |
|---|---|---|---|---|---|
| Run / Snapshot | 1 | 2 | 3 | 1-2 | 1-3 |
| JOINorDIE_openai_all_B_non_rel | **0.591**[1] | 0.314 | 0.217 | 0.469 | 0.633 |
| JOINorDIE_openai_all_B_contrastive | **0.543**[2] | 0.322 | 0.214 | 0.407 | 0.607 |
| Promptly_simple-baseline-history-only | **0.539**[3] | 0.335 | 0.164 | 0.379 | 0.696 |
| JOINorDIE_openai_all_A_scenario_3 | **0.528**[4] | 0.297 | 0.177 | 0.437 | 0.665 |
| Split n' Simulate_Split-n-Simulate-run | **0.523**[5] | **0.437**[3] | **0.255**[3] | **0.165**[5] | **0.513**[5] |
| JOINorDIE_openai_all_A_scenario_1 | 0.515 | 0.262 | 0.189 | 0.491 | 0.634 |
| JOINorDIE_openai_all_B_rel | 0.513 | 0.28 | 0.19 | 0.454 | 0.63 |
| JOINorDIE_openai_last_A_scenario_3 | 0.511 | 0.283 | 0.202 | 0.446 | 0.605 |
| JOINorDIE_openai_first_A_scenario_1 | 0.508 | 0.252 | 0.171 | 0.504 | 0.663 |
| JOINorDIE_openai_all_A_scenario_2 | 0.505 | 0.273 | 0.186 | 0.458 | 0.631 |
| JOINorDIE_openai_first_A_scenario_3 | 0.501 | 0.302 | 0.171 | 0.396 | 0.659 |
| VerbaNexAI_VerbaNexAI_v8b | 0.492 | **0.433**[5] | **0.252**[5] | **0.119**[4] | **0.487**[4] |
| JOINorDIE_openai_last_B_non_rel | 0.491 | 0.304 | 0.206 | 0.38 | 0.58 |
| JOINorDIE_openai_last_A_scenario_2 | 0.49 | 0.288 | 0.203 | 0.411 | 0.586 |
| JOINorDIE_openai_last_A_scenario_1 | 0.486 | 0.284 | 0.172 | 0.416 | 0.647 |
| JOINorDIE_openai_first_A_scenario_2 | 0.486 | 0.263 | 0.19 | 0.459 | 0.61 |
| VerbaNexAI_VerbaNexAI_v8 | 0.478 | **0.433**[4] | **0.252**[4] | **0.094**[3] | **0.472**[3] |
| VerbaNexAI_VerbaNexAI_v13 | 0.478 | **0.438**[2] | **0.257**[2] | **0.083**[2] | **0.462**[2] |
| VerbaNexAI_VerbaNexAI_v1 | 0.473 | **0.458**[1] | **0.281**[1] | **0.032**[1] | **0.406**[1] |
| JOINorDIE_openai_last_B_contrastive | 0.46 | 0.33 | 0.207 | 0.282 | 0.549 |
| Promptly_baseline-cot | 0.454 | 0.148 | 0.128 | 0.675 | 0.718 |
| Promptly_baseline-cot-refine | 0.45 | 0.121 | 0.117 | 0.732 | 0.74 |
| JOINorDIE_openai_first_B_non_rel | 0.439 | 0.27 | 0.207 | 0.385 | 0.529 |
| JOINorDIE_openai_first_B_contrastive | 0.435 | 0.276 | 0.196 | 0.366 | 0.55 |
| JOINorDIE_openai_first_B_rel | 0.413 | 0.249 | 0.193 | 0.396 | 0.533 |
| JOINorDIE_openai_last_B_rel | 0.404 | 0.294 | 0.196 | 0.274 | 0.516 |
| Promptly_nlp-persona-cot-refinement | 0.318 | 0.149 | 0.083 | 0.53 | 0.74 |
| Promptly_nlp-persona-history-only | 0.291 | 0.167 | 0.096 | 0.426 | 0.67 |
| Promptly_nlp-persona-cot | 0.22 | 0.154 | 0.068 | 0.301 | 0.69 |

Incorporating more context generally improves performance. In particular, using the last visible query consistently outperforms using the first query in a session, confirming the incremental nature of query reformulation. Similarly, contrastive prompting strategies that expand or enrich the prompt with additional terms yield better results, likely due to improved representation of the information need.

In contrast, more complex reasoning strategies such as chain-of-thought prompting and self-refinement do not outperform the other approaches in most cases. These methods may introduce unnecessary complexity or overthinking, especially in combination with smaller language models. These strategies were mostly implemented with rather small models and should be reassessed using stronger, larger models in future work.

The use of personas shows mixed effects. While two teams incorporated persona-based approaches, results differ substantially: for Team Promptly, NLP-based personas consistently degraded performance across runs, whereas Team Split n' Simulate achieved strong and stable performance with a single persona-based run across snapshots. This suggests that the effectiveness of persona modeling is highly sensitive to implementation details and prompt engineering.

In contrast to other teams, JOINorDIE used a proprietary LLM that is rather large in size compared to the other team's LLMs. However, the experimental outcomes suggest that the LLM alone is not the primary determinant of effectiveness, as strong results were also achieved with smaller models when combined with effective query construction strategies.

The Relative Change (RC) captures performance consistency across different snapshots, which is

particularly relevant, considering the observed variability across runs, raising questions about temporal stability of the simulators beyond single-snapshot effectiveness.

Regarding candidate selection, the VerbaNexAI systems' results suggest that the choice between a hybrid direct system and a more conservative ensemble had limited impact on final performance. However, both VerbaNexAI and Split n' Simulate explicitly optimized for diversity and effective ranking of candidate lists, which appears to have contributed positively to their overall results.

Furthermore, most teams relied primarily on LLM-based generation strategies. Additionally, Team VerbaNexAI integrated rule-based heuristics and k-nearest-neighbor retrieval for query generation. This hybrid design likely contributed to more robust performance. Possibly, non-neural components reduce variability and enforce more stable query generation, which may explain the consistent performance across snapshots.

Both the simple runs of Team Promptly and the more complex JOINorDIE runs primarily rely on minimal or direct contextual information from previous queries, achieving strong performance overall. However, neither approach maintains consistently strong results across snapshots. Despite competitive effectiveness in individual settings, these methods lack robustness over time and struggle to generalize reliably across different snapshot conditions.

For almost all approaches and snapshots, we observe a general degradation in simulator performance over time, indicating the presence of temporal effects. Notably, the snapshots were constructed to have comparable average session lengths, suggesting that performance differences are unlikely to be driven by variations in available context. Since the snapshots were otherwise used with minimal filtering, the observed decline remains unexpected. As part of future work, a more in-depth understanding of of dataset characteristics are required to ensure fair and stable longitudinal evaluation of user simulators.

### 4.5. Conclusion

The introduction of the LongEval-USim task marks the first time that user simulation has been explicitly integrated into the LongEval lab, broadening its perspective on temporal information retrieval beyond document collections and retrieval effectiveness. The task resulted in a novel benchmark for evaluating next-query prediction in session-based academic search, combining realistic user interaction logs, query sequences, retrieval results, and strong baseline systems that provide a solid foundation for future research on user simulation and session modeling.

The task also demonstrated the potential of exploiting temporal dynamics at the session level. While previous LongEval tasks primarily investigated temporal changes across document collection snapshots spanning months or years, user interactions introduce a complementary temporal dimension operating on much shorter time scales. Future editions of the lab could further explore this idea by considering multiple temporal resolutions, ranging from individual search sessions and daily interactions to long-term temporal evolution over weeks, months, and years.

The submitted approaches further highlighted that next-query prediction remains a highly challenging problem. Even seemingly simple reformulations, such as predicting a single-term query expansion, admit a large number of plausible alternatives, making exact prediction difficult. Future task designs should therefore consider a broader spectrum of query reformulation behaviors, moving beyond minor edits such as single-term additions or removals toward more diverse and realistic reformulation strategies.

Beyond query reformulations, future work should extend user simulation to richer interaction signals, including clicks, scrolling behavior, and result page interactions. Depending on the capabilities of future collaborating search platforms, additional interaction modalities such as spoken queries or mobile search behavior could further increase the realism and applicability of the benchmark.

From an organizational perspective, the task was successfully integrated into the existing LongEval infrastructure, allowing participants to reuse the same datasets, software, and submission platform across multiple tasks. This facilitated participation in several tasks and encouraged methodological cross-fertilization between research problems. Future iterations should strengthen this integration at the conceptual level as well. Techniques developed for session-based user simulation could inspire user-aware methodologies for the other LongEval tasks, including ad hoc retrieval, topic extraction,

**Table 11**
Statistics of the LongEval-RAG dataset.

| Dataset property | Value | Description |
|---|---|---|
| Number of queries | 47 | Total number of user queries in the dataset. |
| Candidate document IDs per query | 10 | Number of candidate document identifiers provided for each query. |
| Average query length (tokens) | 12.49 | Mean query length measured using whitespace tokenization. |
| Minimum query length (tokens) | 9 | Length of the shortest query in the dataset. |
| Maximum query length (tokens) | 16 | Length of the longest query in the dataset. |

**Table 12**
Example entry from the LongEval-RAG dataset.

| Field | Example |
|---|---|
| Query ID | b7c19f3e4a8d52c0e6f1ab9472d35c8e |
| Query | How can single-person motion capture insights enable robust tracking in crowded real-time scenes? |
| Candidate document IDs | 19531678, 129391837, 39516304, 122690614, 85598700, 144926638, 150709332, 156069094, 162758998, 64848716 |

and retrieval-augmented generation, ultimately promoting a more unified view of temporally evolving information retrieval systems that jointly model documents, users, and their interactions.

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

## 5.1. LongEval-RAG Dataset

Task 4 aimed to study the extent to which RAG systems can cope with the evolution of scientific knowledge over time. To this end, we provided participants with a dataset composed of 47 questions. Table 11 summarizes the main statistics of the dataset.

Each question is defined by:

1. A textual query, along with its query ID, for which the participating system has to provide a textual answer;
2. A set of candidate document IDs, with 10 documents provide for each query. For each query, the answer generation is expected to be based only on this set, which can be seen as a first-stage filtering step in a RAG system.

The provided candidate document set is **not** composed only of documents relevant to the query; rather, it contained a mixture of relevant and non-relevant documents. These documents are part of the dataset provided within Task 1. Table 12 shows an example of a query and its associated candidate document IDs.

Participants are expected to provide their answers in the TREC RAG run format[4]. The answer has to contain the generated text, as well as the links to the referenced documents that are used to generate the answer. Table 13 summarizes the expected submission fields.

The construction of the LongEval-RAG collection followed a semi-automatic approach combining dense vector similarity methods, manual inspection, and LLM-based analysis. The main objective was to identify scientific topics that exhibited temporal evolution across multiple research domains, and to transform these evolving topics into question-answer instances suitable for evaluating RAG systems over time.

To this end, we independently processed two test collections derived from the Sci-Retrieval corpus, corresponding to snapshot 2, covering June–August 2025, and snapshot 3, covering September–November

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

**Table 13**
Expected submission format for Task 4 runs.

| Field | Description |
|---|---|
| `metadata` | Contains run-level information, including the team identifier, run identifier, run type, query ID, and the original query. |
| `references` | List of candidate document IDs considered by the system for generating the answer. |
| `answer` | List containing the generated textual answer and its supporting citations. |
| `text` | Generated answer produced by the participating system. |
| `citations` | Indices pointing to the documents in the `references` list that support the generated answer. |

2025. The two snapshots were processed separately in order to identify clusters of scientific work within each time period and to analyze how research concepts evolved across time.

First, paper abstracts were encoded using the SPECTER2[5] classification model to obtain dense scientific document embeddings. These embeddings were then used to identify semantically similar papers through approximate nearest neighbor (ANN) search. We retained only candidate paper pairs with similarity scores above 0.95 and with at least one shared author. The shared-author constraint was introduced to increase the likelihood that the paired papers reflected genuine scientific progression, such as follow-up work, methodological refinement, or extensions of previous studies, rather than merely topical similarity.

Since the number of candidate pairs remained large after this filtering step, we further organized the papers into topical groups. Specifically, papers were re-clustered based on their titles using the same encoder, and word clouds were generated to support the interpretation of the resulting clusters. From each snapshot, we manually selected six interpretable clusters that contained potentially evolving research themes. For each selected cluster, we randomly sampled up to 30 deduplicated paper pairs, excluding pairs for which publication dates were not available.

The sampled paper pairs were then ordered chronologically using the published date indicated in the Sci-Retrieval collection metadata, allowing us to distinguish the earlier work from the later one. We performed full-text analysis and LLM-based assessment using `gpt-4o-mini`[6], following recent work that has used this model for related tasks such as relevance assessment, question generation, and RAG-based scientific text generation [55, 56]. This step retained only those pairs in which the later paper clearly extended, refined, generalized, or built upon the earlier paper. The filtering step was intended to ensure that the final collection captured meaningful scientific evolution rather than simple topical relatedness.

Finally, using the full text of the selected paper pairs, we prompted an LLM to generate question-answer pairs and reasoning hops linking the two papers. The questions were designed to capture how later work addressed research gaps identified in the introductions of both papers, while minimizing lexical overlap with the source texts. The answers were generated from the full papers and written as concise summaries of the underlying scientific evolution. This process resulted in a collection of temporally grounded questions where answering required understanding how scientific knowledge changed or progressed across related publications.

For each generated query, participants received a set of 10 candidate documents from the Sci-Retrieval collection. Two documents corresponded to the paper pair used to generate the query and answer, while the remaining eight are randomly sampled from the same topic cluster to introduce relevant distractors.

The evaluation of retrieval of the relevant documents (out of 10 provided documents) was done using calculating precision of the set of documents marked by the participants as those used in the generation and the ground truth relevant documents. Classical text generation evaluation measures, ROUGE and

[5] https://huggingface.co/allenai/specter2_classification
[6] https://openai.com/es-419/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

BertScore, were used for evaluating the generated answer. In addition to this, we also used the the nugget coverage and an LLM-as-a-judge for evaluating answer generation [57].

## 5.2. LongEval-RAG Submissions

We did receive 26 submissions, from 5 teams, for the Longeval-RAG task. The submissions are evaluated according to retrieval focus (i.e., the capacity of the systems to select the relevant documents during the retrieval step), and to the generation focus (i.e., the capacity of the system to generate answers similar to the ground truth answers).

The teams that participated to the RAG-LongEVal task, ranked by alphabetical order, are:

- DS-GT [10]: This team provided 4 runs. 2 out of these 4 runs use a retrieval stage then generation stage (Gemma-4-31B Instruct or Claude Opus 4.7 Thinking). Other two are generation-only, using GPT-5 with the documents provided.
- glhf [9]: This team provided 9 runs that cover several ideas. It explores several chunking of documents: sentences-based, semantic-based. The retrieval are all lexical-based (BM25, query expansion, pseudo-relevance feedback. Temporal elements are taken into account.
- Lab Individual [11]: This team provided 4 runs. One of these runs does not use retrieval, LLMs deepseek-v4-pro and Qwen3.6-35B-A3B are used.
- ockham: The 6 runs submitted by this team are identical, which explains the equality of the evaluated scores. Their proposal is dedicated to evaluate the efficiency and quality of Shannon Surprisal context pruning for RAG. The generation model, if used, is BART large.
- Verba Nex AI [8]: This team submitted 2 runs, each of them using Qwen3-14B-4bit. These submissions use temporal exponential decay to weight the documents retrieved. One run used pairs of document for ranking, then reranking using one cross-encoder.

All teams except the ockham team submitted a notebook paper describing their approach.

## 5.3. LongEval-RAG results

The results of the evaluation are presented in Table 14. Submissions are assessed along two complementary axes. For retrieval, we measure the precision of the documents cited by the participant against the ground-truth relevant documents, assessing the system's ability to identify the two relevant papers from the ten provided candidates. For generation, the produced answer is compared to the gold answer using ROUGE [58] (ROUGE-1, ROUGE-2, and ROUGE-L F1), which captures n-gram overlap, and BERTScore [59] (Precision, Recall, and F1), which measures semantic similarity using contextual embeddings from RoBERTa-large [60] with baseline rescaling to improve score interpretability. Both metrics are computed per query and averaged over all 47 queries; if a submission does not provide an answer for a given query, a score of 0.0 is assigned for that query. We additionally report two axiomatic LLM-as-a-judge measures: Nugget Coverage and Term Frequency Constraint (TFC1) [57]. TFC1 axiom uses a heuristics which ranks higher documents with more occurrences of query terms [61], while Nugget Coverage uses atomic units of information (nuggets) and measures their coverage in the generated answer. Table 14 reports ROUGE-L F1 and BERTScore F1 as primary measures, alongside retrieval precision, nugget coverage, and TFC1.

Based on these evaluation measures, we observe several interesting facts:

- The best-performing run, as measured by both ROUGE (0.206) and BERTScore (0.263) is *pairjudge-qwen3* from Verba Nex AI. This is an automated run that involves selecting pairs followed by a reranking step incorporating temporal elements, with generation based on Qwen3-14B-4bit. However, this run ranks far lower in retrieval evaluation, achieving a precision of only 0.681 (when the best precision achieved in 1.0). This suggests that strong generation is able to compensate for a good, but not of the highest quality, retrieval.

**Table 14**
Evaluation for Task 4 of LongEval sorted by BertScores. We report the $F_1$ score of Rouge-L, BertScore as primary measures. Additionally, we report the precision of retrieval, the nugget coverage, and TFC1.

| Team | Approach | Rouge | BertScore | Precision | Nugget Cov. | TFC1 |
|---|---|---|---|---|---|---|
| Verba Nex AI | pairjudge-qwen3 | **0.206** | **0.263** | 0.681 | 0.251 | **0.809** |
| DS-GT | GPT DS@GT (MANUAL) | 0.188 | 0.227 | 0.904 | 0.308 | 0.434 |
| glhf | rule-minilm | 0.157 | 0.168 | 0.975 | 0.367 | 0.305 |
| glhf | topic-shift-minilm | 0.157 | 0.163 | 0.940 | 0.288 | 0.373 |
| DS-GT | Opus DS@GT (MANUAL) | 0.169 | 0.157 | **1.000** | 0.443 | 0.017 |
| glhf | semantic-minilm | 0.160 | 0.157 | 0.922 | 0.291 | 0.349 |
| Lab Individual | multi-agent-rag | 0.158 | 0.157 | 0.950 | 0.505 | -0.051 |
| glhf | topic-shift-current | 0.167 | 0.156 | 0.929 | 0.292 | 0.215 |
| glhf | semantic-curren | 0.156 | 0.150 | 0.933 | 0.282 | 0.262 |
| glhf | single-query-bm25 | 0.156 | 0.152 | 0.940 | 0.308 | 0.120 |
| glhf | rrf-no-rerank | 0.155 | 0.151 | 0.929 | 0.336 | 0.087 |
| glhf | default | 0.147 | 0.146 | 0.931 | 0.334 | 0.025 |
| glhf | caes-rag-rrf | 0.155 | 0.144 | 0.942 | 0.338 | 0.059 |
| Lab Individual | single-agent-rag-v1 | 0.155 | 0.138 | 0.769 | 0.473 | -0.082 |
| Verba Nex AI | qwen3-minmax-temp | 0.139 | 0.126 | 0.975 | 0.493 | -0.197 |
| DS-GT | Baseline DS@GT | 0.132 | 0.118 | 0.766 | 0.210 | -0.322 |
| Lab Individual | multi-agent-rag-v2 | 0.136 | 0.113 | 0.940 | **0.537** | -0.119 |
| Lab Individual | multi-agent-rag-v3 | 0.132 | 0.109 | 0.943 | 0.504 | -0.128 |
| DS-GT | CiteFix DS@GT | 0.126 | 0.076 | 0.692 | 0.152 | -0.505 |
| – | official-naive-baseline | 0.082 | -0.118 | 0.200 | 0.124 | -0.758 |
| ockham | run 3 | 0.093 | -0.039 | 0.438 | 0.402 | -0.894 |
| ockham | run 4 | 0.093 | -0.039 | 0.438 | 0.402 | -0.894 |
| ockham | run 5 | 0.093 | -0.039 | 0.438 | 0.402 | -0.894 |
| ockham | run 6 | 0.093 | -0.039 | 0.438 | 0.402 | -0.894 |
| ockham | run 7 | 0.093 | -0.039 | 0.438 | 0.402 | -0.894 |
| ockham | run 8 | 0.093 | -0.039 | 0.438 | 0.402 | -0.894 |

- The second-best run, *GPT DS@GT (MANUAL)* from team DS-GT, achieves lower ROUGE (0.188) and BERTScore (0.227) scores. It relies on manual post-processing to ensure the expected submission format. This system demonstrates high retrieval precision (0.904), showing that strong retrieval does not necessarily translate to better generation. A similar observation applies to the top runs from team GLHF, which combine retrieval and generation using OpenAI GPT-5.4. The other manual run of GLHF, *Opus DS@GT (MANUAL)*, that uses Claude Opus 4.7 Thinking, achieves an ideal precision of 1.0, but has lower Rouge and Bertscore evaluations scores.
- The runs from Lab Individual follow a similar pattern in terms of ROUGE, BERTScore, and precision. Again, higher precision does not result in better generation performances. If we compare the runs multi-agent-rag and multi-agent-rag-v2 (which differ almost exclusively in the LLM used: Qwen for multi-agent-rag and DeepSeek-v4-Pro for multi-agent-rag-v2), Qwen appears to have an advantage over DeepSeek according to ROUGE and BERTScore metrics.
- The *official-naive-baseline* performs significantly worse compared to participant runs, with a ROUGE score below 0.1 and a negative BERTScore. It only outperforms the Ockham team runs 3–8, which rely on information-theoretic pruning of chunks and use BART as the LLM, achieving both low retrieval precision (0.438) and poor generation scores.
- The three runs based that do not rely on retrieval but only on generation, i.e., the manual runs of DS-GT and the run *single-agent-rag-v1*, do behave very well compared to classical RAG approaches.

## 5.4. Discussion

The results obtained from the participants at LongEval-RAG demonstrate that the highest retrieval capabilities in a RAG system do not necessarily lead to the best overall generation. This is explained by the fact that the LLM used in these systems incorporates elements from its training data into the generation process.

Two participants employed generation-only systems, as the relatively small number of documents provided with the queries allowed such methods. These approaches performed well compared to retrieve and generate methods, ranking as the second, fifth, and fourteenth best runs (out of 26 total). These results suggest that LLMs are able of effectively processing a relatively large number of documents without explicit retrieval.

For next year, our plan for the RAG task is to increase the number of documents to encourage participants to rely on a retrieval step. This will make the task more realistic, especially when dealing with large document corpora. However, such element will complicate even more the selection of these documents, as we need to provide *hard negative* documents in a away to mimic a realistic retrieval.

### 5.5. Conclusion

The first edition of the LongEval-RAG task establishes a novel benchmark for evaluating RAG systems under temporal evolution of scientific knowledge. The results reveal that retrieval quality and generation quality are not tightly coupled: the best-performing run in terms of ROUGE and BERTScore achieved only moderate retrieval precision, while systems with near-perfect precision did not produce the best answers, suggesting that generation quality is not solely determined by retrieval quality. The surprisingly strong performance of generation-only approaches further supports this observation. At the other end of the spectrum, systems relying on information-theoretic chunk pruning with BART as the generation model consistently underperformed, indicating that both the retrieval strategy and the choice of language model have a significant impact on generation quality. For future editions, we plan to increase the number of candidate documents per query to make the retrieval step more necessary and challenging, and to introduce harder negative documents that better mimic realistic retrieval conditions.

## 6. Conclusion

We have presented an overview of the LongEval 2026 Lab, describing the data provided to participants, submissions, and evaluation. In this edition, we evaluate IR systems over time across four tasks: (1) ad-hoc scientific retrieval, (2) topic extraction from query logs, (3) user simulation via query (re-)formulation, and (4) retrieval-augmented generation over an evolving scientific corpus. All tasks share the 2026 LongEval SciRetrieval dataset built from CORE search logs as three cumulative snapshots spanning 2025. In total, we received 97 submissions from 18 teams.

In Task 1 and Task 3, we evaluated the runs for effectiveness and temporal robustness, where we observe the same pattern: effectiveness declines as the collection evolves and scales, and system rankings are not stable between snapshots. In Task 1, effectiveness on the click-based qrels decreases from snapshot to snapshot, and effectiveness and robustness behave as competing goals. In Task 3, simulator performance degrades over time, though the snapshots have comparable session lengths. Strong performance on an initial snapshot does not guarantee sustained performance as the corpus evolves. Temporal robustness emerges as a key challenge for sustained performance.

We also investigated the limits of using LLMs to support evaluation. In Task 1, synthetic LLM relevance labels are cheaper and less biased to the production ranking, but they are static and capture only asynchronous relevance, missing the temporal dependence bound by the model's knowledge cutoff. In Task 2, the LLM assessors over-label relevance and show substantial disagreement with the click-based judgments at the document level, so there is currently no good alignment between the click logs and the formalized topics. In Task 4, higher retrieval precision does not lead to better generation, and generation-only systems remain competitive when the candidate set is small.

Two lessons are learned across the tasks: (1) temporal signals help but do not guarantee an effective approach, and (2) the design of the evaluation, from topic formalization to candidate set size, shapes the conclusions as much as the systems. For future work, we plan to: (1) develop temporally aware features and timeline-specific relevance assessments, (2) find approaches that increase the alignment between click logs and formalized topics, and (3) scale the document sets for Task 4 in future editions to necessitate retrieval under more realistic conditions.

## Acknowledgments

This work was supported by Deutsche Forschungsgemeinschaft (STELLA, project number 407518790 and NFDIxCS, project number 501930651) and the German Federal Ministry of Education and Research (BMBF) and Joint Science Conference (GWK) in the PLan_CV project (03FHP109).

## Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for Grammar and spelling checks and Gemini 3.1 pro for formatting assistance. Further, the authors used GitHub Copilot to draft the analysis code to create the confusion matrix in Figure 2. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

[1] H. Wu, Y. Yang, Official and diagnostic analysis of full-text temporal retrieval for longeval-sci, in: E. S. Salido, A. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[2] K. D, J. Vr, B. A, P. R, Temporal stability in scientific information retrieval: A three-system analysis for clef longeval 2026, in: E. S. Salido, A. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[3] D. Alexander, M. Fröbe, G. Hendriksen, M. Hagen, D. Hiemstra, M. Potthast, A. de Vries, Team openwebsearch at longeval: Using large language models for relevance feedback for scientific search, in: E. S. Salido, A. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[4] C. Bakirci, A. Brückner, J. M. Christian, N. K. Debrezion, C. H. A. D. Pino, A. Imam, K. Kartal, A. B. Klatt, R. Klinkhammer, Y. Mählmann, L. Michel, E. Sahin, L. D. Thies, V. Timmer, A. Kruff, J. Keller, Cir at longeval 2026: Ad-hoc scientific retrieval, topic extraction from query logs, and user simulation, in: E. S. Salido, A. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[5] T. Breuer, J. Keller, N. Bernard, P. Schaer, Cir at longeval 2026: Manual topic creation and concept induction-based topic generation, in: E. S. Salido, A. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[6] L. M. D. Martínez, J. E. S. Castañeda, E. A. P. D. Castillo, J. C. M. Santos, Qrels-grounded and agentic llm query expansion with post-processing optimization for longitudinal scientific topic extraction, in: E. S. Salido, A. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[7] S. A. O. Cueter, J. E. S. Castañeda, J. C. Martinez-Santos, E. Puertas, Hybrid evidence-aware user simulation for scientific search sessions at longeval 2026 task 3, in: E. S. Salido, A. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[8] E. P. Isaac David Sánchez Sánchez, Jairo Enrique Serrano Castañeda, J. C. Martínez-Santos, VerbaNexAI at LongEval 2026 Task 4: Fully Local Temporal and Pairwise Evidence Selection for Scientific RAG, in: E. S. Salido, A. Barrón-Cedeno, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[9] Y. Yang, H. Wu, Candidate-Constrained Retrieval-Augmented Generation for LongEval-RAG: System Design and Empirical Analysis, in: E. S. Salido, A. Barrón-Cedeno, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[10] B. Michaels, B. Johnson, DS@GT ARC at LongEval: Citation Integrity and Factual Grounding in

Scientific QA, in: E. S. Salido, A. Barrón-Cedeno, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[11] H. Q. Zhenyang Cheng, Y. Yi, A Multi-Agent Approach with Similarity-Guided Judgment for Scientific Knowledge Retrieval and Answer Generation, in: E. S. Salido, A. Barrón-Cedeno, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[12] 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, Longeval at CLEF 2025: Longitudinal evaluation of IR systems on web and scientific data, in: J. Carrillo-de-Albornoz, A. G. S. de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Lecture Notes in Computer Science, Springer, 2025, pp. 363–387. URL: https://doi.org/10.1007/978-3-032-04354-2_20.

[13] M. Cancellieri, T. Breuer, A. M. Elebshihy, M. Fröbe, P. Galuščáková, L. Goeuriot, G. Iturra-Bocaz, J. Keller, P. Knoth, A. K. Kruff, F. Piroi, D. Pride, P. Schaer, P. Mulhem, D. Schwab, Longeval 2026 train dataset, 2026. doi:10.48436/6avmz-esd52.

[14] M. Cancellieri, T. Breuer, A. M. Elebshihy, M. Fröbe, P. Galuščáková, L. Goeuriot, G. Iturra-Bocaz, J. Keller, P. Knoth, A. K. Kruff, F. Piroi, D. Pride, P. Schaer, P. Mulhem, D. Schwab, Longeval 2026 test dataset, 2026. doi:10.48436/4rg9e-47328.

[15] 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.

[16] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified data wrangling with ir_datasets, in: SIGIR, ACM, 2021, pp. 2429–2436.

[17] OpenAI, gpt-oss-120b & gpt-oss-20b model card, CoRR abs/2508.10925 (2025). doi:10.48550/ARXIV.2508.10925. arXiv:2508.10925.

[18] S. Upadhyay, R. Pradeep, N. Thakur, D. Campos, N. Craswell, I. Soboroff, J. Lin, A large-scale study of relevance assessments with large language models using UMBRELA, in: H. Zamani, L. Dietz, B. Piwowarski, S. Bruch (Eds.), Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, Padua, Italy, 18 July 2025, ACM, 2025, pp. 358–368. URL: https://doi.org/10.1145/3731120.3744605. doi:10.1145/3731120.3744605.

[19] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[20] T. Breuer, J. Keller, P. Schaer, ir_metadata: An extensible metadata schema for IR experiments, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 3078–3089. URL: https://doi.org/10.1145/3477495.3531738.

[21] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, NIST Special Publication, National Institute of Standards and Technology (NIST), 1994, pp. 109–126. URL: http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[22] V. Lavrenko, W. B. Croft, Relevance-based language models, in: W. B. Croft, D. J. Harper, D. H. Kraft, J. Zobel (Eds.), SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New

Orleans, Louisiana, USA, ACM, 2001, pp. 120–127. URL: https://doi.org/10.1145/383952.383972.

[23] G. V. Cormack, C. L. A. Clarke, S. Büttcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, J. Zobel (Eds.), Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, ACM, 2009, pp. 758–759. URL: https://doi.org/10.1145/1571941.1572114.

[24] 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.

[25] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy, Nature 585 (2020) 357–362. URL: https://doi.org/10.1038/s41586-020-2649-2. doi:10.1038/s41586-020-2649-2.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 8024–8035. URL: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

[28] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[29] T. Hagen, M. Fröbe, J. H. Merker, H. Scells, M. Hagen, M. Potthast, Tirex tracker: The information retrieval experiment tracker, in: N. Ferro, M. Maistro, G. Pasi, O. Alonso, A. Trotman, S. Verberne (Eds.), Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, ACM, 2025, pp. 3764–3771. URL: https://doi.org/10.1145/3726302.3730297. doi:10.1145/3726302.3730297.

[30] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou, Qwen3 embedding: Advancing text embedding and reranking through foundation models, CoRR abs/2506.05176 (2025). URL: https://doi.org/10.48550/arXiv.2506.05176. arXiv:2506.05176.

[31] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with pagedattention, in: J. Flinn, M. I. Seltzer, P. Druschel, A. Kaufmann, J. Mace (Eds.), Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023, ACM, 2023, pp. 611–626. URL: https://doi.org/10.1145/3600006.3613165. doi:10.1145/3600006.3613165.

[32] 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.

[33] G. N. G. Sáez, Continuous Evaluation Framework for Information Retrieval Systems. (Cadre d'évaluation continue pour les systèmes de recherche d'information), Ph.D. thesis, Grenoble Alpes University, France, 2023. URL: https://tel.archives-ouvertes.fr/tel-04547265.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3., 1994.

[39] P. Thomas, S. Spielman, N. Craswell, B. Mitra, Large language models can accurately predict searcher preferences, in: G. H. Yang, H. Wang, S. Han, C. Hauff, G. Zuccon, Y. Zhang (Eds.), Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, ACM, 2024, pp. 1930–1940. doi:10.1145/3626772.3657707.

[40] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, H. Wachsmuth, Perspectives on Large Language Models for Relevance Judgment, in: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ACM, Taipei Taiwan, 2023, pp. 39–50. doi:10.1145/3578337.3605136.

[41] I. Soboroff, Don't use llms to make relevance judgments, Inf. Retr. Res. J. 1 (2025) 29–46. doi:10.54195/IRRJ.19625.

[42] E. M. Voorhees, The evolution of cranfield, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, The Information Retrieval Series, Springer, 2019, pp. 45–69. doi:10.1007/978-3-030-22948-1_2.

[43] J. Keller, M. Fröbe, B. Engelmann, F. Haak, T. Breuer, B. Larsen, P. Schaer, Formalized information needs improve large-language-model relevance judgments, CoRR abs/2604.04140 (2026). URL: https://doi.org/10.48550/arXiv.2604.04140. arXiv:2604.04140.

[44] J. H. Merker, J. Bevendorff, M. Fröbe, T. Hagen, H. Scells, M. Wiegmann, B. Stein, M. Hagen, M. Potthast, Web-scale retrieval experimentation with chatnoir-pyterrier, 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 V, Lecture Notes in Computer Science, Springer, 2025, pp. 96–104. URL: https://doi.org/10.1007/978-3-031-88720-8_17.

[45] S. Upadhyay, R. Pradeep, N. Thakur, N. Craswell, J. Lin, UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor, 2024. doi:10.48550/arXiv.2406.06519. arXiv:2406.06519.

[46] C. Yu, H. Li, G. Zuccon, J. Mackenzie, T. Leelanupab, When LLM judges inflate scores: Exploring overrating in relevance assessment, CoRR abs/2602.17170 (2026). URL: https://doi.org/10.48550/arXiv.2602.17170. arXiv:2602.17170.

[47] M. Fröbe, A. Parry, F. Schlatt, S. MacAvaney, B. Stein, M. Potthast, M. Hagen, Large language model relevance assessors agree with one another more than with human assessors, in: N. Ferro,

M. Maistro, G. Pasi, O. Alonso, A. Trotman, S. Verberne (Eds.), Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, ACM, 2025, pp. 2858–2863. URL: https://doi.org/10.1145/3726302.3730218.

[48] E. M. Voorhees, N. Craswell, J. Lin, Too many relevants: Whither cranfield test collections?, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 2970–2980. URL: https://doi.org/10.1145/3477495.3531728.

[49] X. Li, B. J. A. Schijvenaars, M. de Rijke, Investigating queries and search failures in academic search, Inf. Process. Manag. 53 (2017) 666–683. URL: https://doi.org/10.1016/j.ipm.2017.01.005.

[50] A. Kacem, P. Mayr, Analysis of search stratagem utilisation, Scientometrics 116 (2018) 1383–1400. URL: https://doi.org/10.1007/s11192-018-2821-8.

[51] T. Breuer, C. K. Kreutz, N. Fuhr, K. Balog, P. Schaer, N. Bernard, I. Frommholz, M. Gohsen, K. Ji, G. J. F. Jones, J. Keller, J. Liu, M. Mladenov, G. Pasi, J. R. Trippas, X. Wang, S. Zerhoudi, C. Zhai, Report on the workshop on simulations for information access (sim4ia 2024) at SIGIR 2024, CoRR abs/2409.18024 (2024). URL: https://doi.org/10.48550/arXiv.2409.18024. arXiv:2409.18024.

[52] A. K. Kruff, C. K. Kreutz, T. Breuer, P. Schaer, K. Balog, Sim4ia-bench: A user simulation benchmark suite for next query and utterance prediction, 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, Lecture Notes in Computer Science, Springer, 2026, pp. 594–609. URL: https://doi.org/10.1007/978-3-032-21321-1_61.

[53] A. K. Kruff, N. Bernard, P. Schaer, Validating search query simulations: A taxonomy of measures, 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 I, Lecture Notes in Computer Science, Springer, 2026, pp. 243–256. URL: https://doi.org/10.1007/978-3-032-21289-4_16.

[54] 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.

[55] J. A. Schnabel, J. R. Trippas, F. Scholer, D. Hettiachchi, Multi-stage large language model pipelines can outperform gpt-4o in relevance assessment, in: Companion Proceedings of the ACM on Web Conference 2025, 2025, pp. 1288–1292.

[56] D. Di Nuzzo, E. Vakaj, H. Saadany, E. Grishti, N. Mihindukulasooriya, Automated generation of competency questions using large language models and knowledge graphs., in: NLP4KGC@SEMANTiCS, 2024, pp. 128–153.

[57] J. H. Merker, M. Fröbe, B. Stein, M. Potthast, M. Hagen, Axioms for retrieval-augmented generation, in: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), ICTIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 67–77. URL: https://doi.org/10.1145/3731120.3744601.

[58] C. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain, July 2004, Association for Computational Linguistics, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[59] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[60] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov,

Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[61] H. Fang, T. Tao, C. Zhai, A formal study of information retrieval heuristics, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 49–56. URL: https://doi.org/10.1145/1008992.1009004. doi:10.1145/1008992.1009004.