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

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 Konstantin 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, LIG, Grenoble, France
sarah.bouaraba@univ-grenoble-alpes.f
[2] TH Köln - University of Applied Sciences, Cologne, Germany
[3] The Open University, Milton Keynes, UK
[4] TU Wien, Vienna, Austria
[5] Friedrich-Schiller-Universität Jena, Jena, Germany
[6] University of Stavanger, Stavanger, Norway

Abstract. The LongEval lab is about the evaluation of information retrieval systems over time. For this purpose, the lab cooperates with the academic search platform CORE that provides access to documents and queries with real-world interactions. The participants get access to different snapshots that are derived from the CORE 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 Robustness · Information Retrieval · Topic Generation · User Simulation · Retrieval Augmented Generation

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 CORE search platform for open access academic

S. Bouaraba, L. Goeuriot, P. Mulhem and D. Schwab—Authors ordered alphabetically.

M. Hagen et al. (Eds.): CLEF 2026, LNCS 17087, pp. 548–573, 2027.
https://doi.org/10.1007/978-3-032-39150-6_32

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 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 be give good results when the candidate set is small.

Overall, LongEval 2026 collected 97 runs 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.

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 [8]. It is a traditional ad-hoc search task in the scientific domain. We evaluate the longitudinal effectiveness and robustness of 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.

2.1 Datasets

We continued to use the COnnecting REpositories (CORE) search engine to construct a dynamic test collection from user queries, SERPs, and click feedback gathered throughout 2025. We created three snapshots spanning three-month periods: Snapshot 1 (March–May), Snapshot 2 (June–August), and Snapshot 3 (September–November). While this collection chronologically extends the 2025 dataset [8], differences in their construction cause that the two are not directly comparable. Namely, the snapshots now span longer time periods, documents are only added but never removed, and qrels are only constructed using certain click types. We introduced these changes to improve the reliability of the test collection and adapt better to the dynamics of scientific search. The dataset is available from the TU Wien Research Data Repository [5] and through the LongEval `IR_datasets` integration [17].[1]

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. For each query, we constructed the document collections by sub-sampling the CORE corpus using the CORE API.[2] We included documents appearing in the logged SERPs, as well as the top 1,000 documents from the January 2026 CORE ranking. To model the growth of scholarly collections, snapshots are cumulative: each snapshot extends the previous one with newly appearing ranked documents and a proportional number of randomly sampled documents. The resulting corpus sizes are approximately 870K, 1.3M, and 1.7M documents for Snapshots 1, 2, and 3, respectively.

We derive two click-based qrel sets from user interactions. The raw qrels treat every clicked document as relevant, while a second set discretizes Document Click Through Rate (DCTR) scores into three relevance grades. For the evaluation of the retrieval task, we have a connection to Task 2 on topic formalization, as the

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

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 this condensed overview, we include qrels generated with GPT-OSS-120b [24] 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 [30].

We provided participants with training and test datasets. The training data consists of the Snapshot 1 documents, 100 queries, 2,275 raw relevance assessments, and 10,406 DCTR relevance assessments. The test datasets consist of documents from Snapshots 1, 2, and 3 and 381 queries.

2.2 Submissions and Approaches

In total, we received 20 runs from eight teams. Compared to last year, two more teams participated, but submitted three fewer runs. Five teams submitted a notebook paper (compared to two in 2025). Submissions were made through the TIRA platform [13], where participants uploaded runs. We also collected metadata, including short run descriptions, which participants provided in the `ir_metadata` specification format [6].

All 20 submitted runs used a lexical ranker at some point in the retrieval pipeline. Three runs used a deep neural approach, and one run used a dense neural component. Five runs are single-stage, and the remaining 15 use a multi-stage approach. BM25 [25] is used most often for the first-stage ranking. Re-ranking varied from classical boosting 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, by incorporating a retrieval signal from previous snapshots, or from the newest citation as a combined proxy for recency and quality. Further approaches included query expansion with RM3 [20] and Reciprocal Rank Fusion [9]. Additionally, two first-stage ranker baselines using the abstract corpora are provided to support re-ranking experiments. The first uses BM25 with default parameters, and the second uses the Qwen3 4B embedding model [36].

2.3 Results

We evaluate the approaches for effectiveness and robustness. We contrast the results with two baselines and compare them to Snapshot 1. Matching the retrieval task, nDCG@10 is selected as the primary effectiveness measure.

Effectiveness. 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. Table 1 shows the retrieval effectiveness for each snapshot and approach, and the results are also visualized in Fig. 1. Based on the DCTR labels, investigating the top three systems per snapshot indicates a variability in the

Table 1. Effectiveness of across qrels and snapshots, measured by nDCG@10. The approaches are sorted by the snapshot 3 results on DCTR qrels, the baselines are light and the best approaches are highlighted in **bold**.

Qrels	dctr			gpt-oss-120b		
Run/Snapshot	1	2	3	1	2	3
custom-lexicalft-tc [34]	0.300	**0.284**	**0.195**	0.698	0.709	0.713
custom-lexical-FT-T [34]	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 [34]	0.290	0.256	0.166	0.706	0.714	0.717
bm25-ft-additive [34]	0.291	0.257	0.166	0.706	0.714	0.717
bm25-ft-router [34]	0.290	0.256	0.166	0.707	0.714	0.717
temporal-fusion [10]	0.277	0.268	0.161	0.745	0.750	0.752
QwenTemporalIR [10]	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 [1]	0.276	0.233	0.142	0.803	0.811	0.812
kq-rm3-bm25 [1]	0.273	0.219	0.139	0.814	0.823	0.823
rm3-conservative [10]	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 [1]	0.181	0.123	0.079	0.666	0.672	0.656

system order and effectiveness. On average, the effectiveness drops from the first to the second snapshot and decreases further in the third snapshot.

For the synthetic relevance labels the ranking of systems remains stable across the three snapshots. 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. Besides these systems, the effectiveness across snapshots is stable. The scores on the DCTR qrels are also lower than those on the synthetic labels. Figure 1 highlights both findings as well.

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 τ and based on the effectiveness with Pearson's correlation. Comparing Snapshot 2 and Snapshot 1, the Pearson correlation is 0.458, indicating only

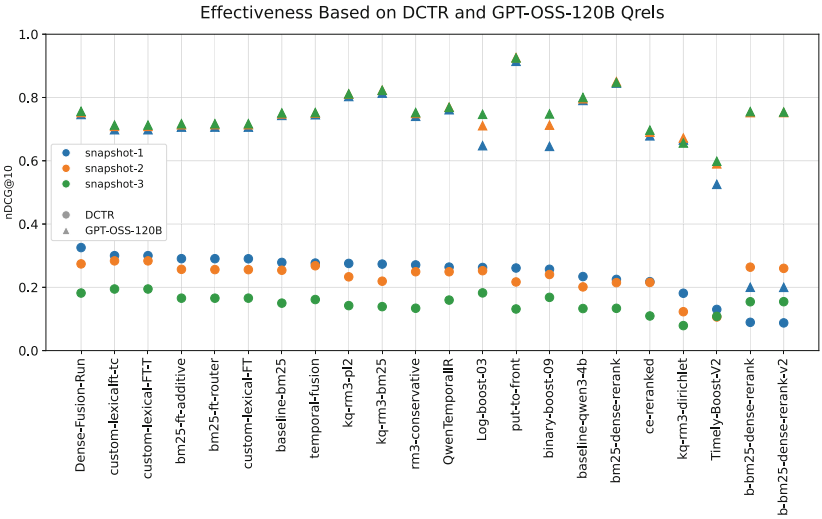

Fig. 1. Effectiveness measured by nDCG@10 across all approaches and test snapshots using the DCTR qrels (circle) and the synthetic relevance labels generated with GPT-OSS-120b (triangle).

a medium positive relation. This is similar for Snapshot 3 compared to Snapshot 1, which shows a correlation of 0.479. Likewise, Kendall's τ 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 τ is slightly higher, with 0.684 and 0.625. All p-values indicate significance at an α level of 0.05.

We also assess the temporal robustness of individual approaches using the Result Delta ($\mathcal{R}\Delta$) framework proposed by González-Sáez [26], in which the effectiveness is measured per snapshot across many queries and then compared over time. We focus on two measures: the Relative Change (RC) and the Delta Relative Improvement (DRI). 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: $\text{RC} = \frac{\text{ARP}_{t_1} - \text{ARP}_{t_2}}{\text{ARP}_{t_1}}$. A negative RC indicates an improved effectiveness and a positive RC an impaired effectiveness. This measure maintains continuity with previous iterations of the LongEval lab [2,3,8]. The DRI quantifies the relative change in effectiveness anchored against a designated *Pivot system*. By normalizing against the Pivot at both timestamps, it separates system-specific variations from environmental noise, such as collection drift or intent shifts in the testbed [14,27]. 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 [25]. The results are displayed in Table 2. The RC reflects

the effectiveness fluctuation. In the second snapshot, compared to the first, all approaches except `b-bm25-dense-rerank` and `b-bm25-dense-rerank-v2` show a reduced effectiveness, indicated by a positive RC. Towards the third snapshot, this degradation continues: almost all approaches show a larger positive RC, while the same two runs improve their effectiveness. On the DRI, nine approaches show a low positive value in Snapshot 2, indicating an improvement relative to BM25, while the remaining systems have decreased relative deltas. In the third snapshot, fewer approaches retain an improved delta to BM25.

2.4 Discussion

The LongEval-Sci task continues the research on longitudinal evaluations for academic search, started in LongEval 2025. The snapshots added this year are larger to capture more dynamics of the scientific domain. For almost all systems, the effectiveness measured on the DCTR qrels decreases over time, while the

Table 2. 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 shaded and the best results are highlighted in **bold**.

	Snapshot 2			Snapshot 3		
	nDCG@10	RC	DRI	nDCG@10	RC	DRI
kq-rm3-dirichlet [1]	0.123	0.322	0.167	0.079	0.564	0.124
rm3-conservative [10]	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 [1]	0.217	0.169	0.082	0.132	0.496	0.060
kq-rm3-bm25 [1]	0.219	0.198	0.117	0.139	0.492	0.054
kq-rm3-pl2 [1]	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 [34]	0.257	0.116	0.031	0.166	0.430	−0.062
bm25-ft-router [34]	0.256	0.118	0.034	0.166	0.430	−0.062
custom-lexical-FT [34]	0.256	0.118	0.034	0.166	0.429	−0.064
temporal-fusion [10]	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 [10]	0.249	0.056	−0.034	0.160	0.395	−0.118
custom-lexical-FT-T [34]	**0.284**	0.055	−0.040	**0.195**	0.351	−0.221
custom-lexicalft-tc [34]	**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

collection grows. Only six approaches outperform the BM25 baseline consistently on all snapshots, and on the synthetic qrels, just four approaches outperform the Qwen3 baseline. Compared to 2025, the ranking of systems is more dynamic, as indicated by a lower rank correlation, and effectiveness and robustness appear to be competing goals, with improving one often impairing the other.

Submitted approaches increasingly relied on the temporal properties of the dynamic test collection to create effective rankings. 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. Temporal signals seem valuable when the DCTR qrels are used for evaluation, while the synthetic LLM judgments are less influenced by them. This year, we also provided citation data for the collection, which two teams used for simple signals with only a marginal positive effect.

For the first time, we used synthetic relevance labels generated by LLMs to retrospectively measure the effectiveness of systems. The general viability of LLM judgments is debated [11,28,29], but their use has unique implications for longitudinal evaluations. Synthetic judgments are cheaper and more accessible than editorial relevance judgments, which addresses a key obstacle in continuous evaluations, where judgments are required frequently and quickly after each new snapshot. Compared to the click-based pseudo qrels often relied on in the past, they are less biased to the production ranking used to collect the clicks, if the judgment pool is diverse enough, and provide more judgments per query. These positive qualities are offset by a critical limitation: only retrospective assessments are possible. Assuming deterministic outputs, a static LLM lacks temporal awareness and evaluates a query-document pair identically regardless of the assessment timeline. Figure 1 highlights this, contrasting the natural variance over time in DCTR judgments with the static GPT-OSS-120b assessments. Temporal dependence in relevance is therefore not captured by these judgments, which are bound by the LLM's knowledge cutoff.

2.5 Conclusion

The 2026 iteration of the LongEval lab continued the SciRetrieval task, investigating the temporal robustness and effectiveness of ad-hoc retrieval systems across a dynamic test collection. We received submissions from eight teams. Maintaining effectiveness over time remained a challenge: as the collection scaled from 87K to over 1.7M documents, the average system effectiveness declined, and the rankings are highly dynamic, with peak performance on the initial dataset not guaranteeing sustained performance as the corpus evolves. Future work in longitudinal IR needs better integration of temporal and citation signals, and reliable testbeds with timeline-specific relevance assessments, to design retrieval architectures that adapt to the temporal dynamics of the scholarly landscape.

3 Task 2. LongEval-TopEx: Topic Extraction

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 [32]. In the context of LongEval, only clicks from the production search engine are available over time, which 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 3 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 [16]. 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.

Table 3. 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.

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 at scale.

3.2 Submissions and Approaches

In total, three teams submitted a total of 23 runs via TIRA [13].

Team CIR had a focus on contrasting human-formalized topics with automatically formalized topics. Information science students manually formalized 30 topics with ChatNoir [22]. 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 [21], 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.

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 [31] 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 [12,35] 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 [33].
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 4 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

Table 4. 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

Table 5. 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

overall trend is that all LLM relevance judgments are (as expected from the literature [12,35]) 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 5 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 of 0.775), whereas

Table 6. 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

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

Table 6 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 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) [7], 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. 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: (i) The sequence of queries submitted. (2) Timestamps for query submissions. (3) The top-10 SERP entries retrieved for each query. (iv) 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 [19]. 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. A high temperature setting was used to encourage greater diversity among the candidate predictions for each session.

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. 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. [18]. Specifically, we look at two main aspects: (1) the inability to distinguish simulated from real data, evaluated using the Rank-Diversity Score (RDS) [19]; 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) [2,3,8,15]. Similar to the other tasks' definition, robustness is defined as the ability of a simulator to perform consistently well across different snapshots.

Table 7. 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. The table only contains the two best and worst runs for every team according to the average RDS score across snapshots to provide a concise overview.

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_contrastive	**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]
VerbaNexAI_VerbaNexAI_v8b	0.492	**0.433**[5]	**0.252**[5]	**0.119**[4]	**0.487**[4]
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]
Promptly_baseline-cot	0.454	0.148	0.128	0.675	0.718
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-cot	0.22	0.154	0.068	0.301	0.69

Table 7 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, 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 7 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 approach consider diversity as well as robust ranking as part of the query candidate selection, achieve 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.

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.

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 often lead to improved results, likely due to better coverage of the user's 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.

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

5.1 LongEval-RAG Dataset

The Task 4 aimed at studying to which extent does RAG systems cope with evolution of scientific knowledge in time. Do to that, we provided the task participants a dataset composed of 47 questions defined as:

1. A textual query (along with its query id) for which the participating system provided a textual answer;
2. A set of document ids, 10 for each query, mined from the corpus. We expected, for each query, that the answer generation was only based on this set, which

may be seen as a first stage filtering in a RAG system. This set of document **was not** composed only of relevant documents to the query, but it contained some relevant documents. The provided documents were a part of the dataset provided within the Task 1.

Participants were expected to provide their answer in the TREC RAG run format. The answer must contain the generated text, as well as the links to the referenced documents that were used to generate the answer .

The building the LongEval-RAG collection follows a semi-automatic approach that combines vector similarity methods and LLMs. The main challenge was identifying scientific topics exhibiting temporal evolution across multiple research domains.

To address this, we independently processed two test collections derived from the Sci-Retrieval corpus, corresponding to snapshot 2 (June–August 2025) and snapshot 3 (September–November 2025), constructing clusters separately to analyze the evolution of scientific concepts over time.

Paper abstracts were encoded using the SPECTER2[3] classification model to obtain dense scientific document embeddings. We identified semantically similar papers through approximate nearest neighbor (ANN) search, retaining only pairs with similarity scores above 0.95 and at least one shared author to increase the likelihood of capturing genuine scientific progression.

Because the number of candidate pairs remained large, we re-clustered papers based on their titles using the same encoder and generated word clouds to characterize cluster topics. From each snapshot, we manually selected six interpretable clusters containing potentially evolving research themes. We then randomly sampled up to 30 deduplicated paper pairs per cluster, excluding pairs without publication dates. After ordering papers chronologically, we used full-text analysis and LLM-based assessment to retain only pairs in which the later work clearly extended, refined, generalized, or built upon the earlier one.

Finally, using the full text of the selected pairs, we prompted an LLM to generate question-answer pairs and the reasoning hops linking the two papers. Questions were designed to capture how later work addressed research gaps identified in both introductions while minimizing lexical overlap with the source texts. Answers were generated from the full papers and written as concise summaries of the underlying scientific evolution.

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 BertScore, were used for evaluating the generated answer. In addition to this, we

[3] https://huggingface.co/allenai/specter2_classification.

also used the the nugget coverage and an LLM-as-a-judge for evaluating answer generation [23].

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: 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: This team provided 9 runs that cover many different ideas.
- Lab Individual: 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: 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 okham submit a notebook paper describing their approach.

5.3 LongEval-RAG Results

The results of the evaluations for LongEval RAG are presented in Table 8. From this table, here are the most interesting facts:

- The best generated runs, according to both Rouge (0.206) and BertScore (0.263), is *pairjudge-qwen3* from Verba Nex AI. It is a automatic run with selection of pairs followed by a reranker step containing temporal elements, the generation is supported by Qwen3-14B-4bit. We notice though that this run is far from being the best one regarding the retrieval evaluation using precision (0.681). This shows that a good generation may compensate a good (but not high-quality) retrieval.
- The second best run, *GPT DS@GT (MANUAL)* from the team DS-GT, obtains quite lower rouge (0.188) and BertScore (0.227). It uses a manual post-processing to ensure the expected submission format. This system has a high retrieval precision (0.904). In this case, a high retrieval capability does

Table 8. 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 [23].

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

not lead to a better generation. Such comment also applies to the top runs of the team GLHF that are based on a retrieval and a generation step 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 behave similarly regarding Rouge, BertScore and Precision. Here again a higher precision does not lead to a better generation. If we consider that the runs *multi-agent-rag* and *multi-agent-rag-v2* differ almost exclusively by the LLM used (Qwen for *multi-agent-rag* and deepseek-v4-pro for *multi-agent-rag-v2*), there is an advantage of using Qwen over DeepSeek according to Rouge and BertScore.
- The *official-naive-baseline* behaves quite badly relatively to the participant runs with a Rouge score lower than 0.1 and a negative BerstScore. This

run only beats the ockham team runs 3–8 (based on information-theoretic pruning of chunks and BART as LLM), that achieves both low retrieval (precision=0.438) values as well as low generation scores.
- The three runs that were based only on generative models, 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 at LongEval-RAG show that the highest retrieval capabilities of a RAG does not lead to the best generation results. This is easily explained by the fact that the LLM used integrates in the generation elements from its training set.

Generation-only systems were used by two participants of the task. Such approaches are possible as we did provide relatively small number of documents with the queries. Such systems do behave well compared to retrieval and generate approaches, as they correspond to the second, fifth and fourteeth best run (on 26 runs). Such results show that retrieving a relatively large number of documents is tractable by LLMs.

Next year, we plan to release more documents in a way to force participants to use a retrieval step. Such methodology will then be more realistic when facing large documents corpora.

6 Conclusion

We have presented an overview of the LongEval 2026 lab, describing the data provided to participants, the submissions, and the evaluation. In this edition, we evaluated 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 SciRetrieval dataset, built from CORE search logs as three cumulative snapshots spanning 2025. In total, we received 97 submissions. In Task 1 and Task 3, we evaluated the runs for effectiveness and temporal robustness and observed the same pattern: effectiveness declines as the collection scales, and system rankings remain dynamic between snapshots. Effectiveness on the click-based qrels decreases from snapshot to snapshot, effectiveness and robustness behave as competing goals, and simulator performance degrades over time despite comparable session lengths. Strong performance on an initial snapshot does not guarantee sustained performance as the corpus evolves, so temporal robustness emerges as a key challenge.

We also investigated the limits of using LLMs to support evaluation. Synthetic LLM relevance labels (Task 1) are cheaper and less biased to the production ranking, but they are static and capture only retrospective relevance, missing the temporal dependence bound by the model's knowledge cutoff. In

Task 2, the LLM assessors over-label relevance and disagree substantially 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 cut 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 Task 4 document sets in future editions to necessitate retrieval under more realistic conditions.

Acknowledgments. This work was supported by Deutsche Forschungsgemeinschaft (STELLA 407518790 and RESIRE 509543643), NFDIxCS grant number 501930651, and the German Federal Ministry of Education and Research (BMBF) and Joint Science Conference (GWK) in the PLan_CV project (03FHP109).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alexander, D., et al.: Team openwebsearch at longeval: Using large language models for relevance feedback for scientific search. In: Salido, E.S., no, A.B.C., de Herrera, A.G.S., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2026)
2. Alkhalifa, R., et al.: Overview of the CLEF-2023 longeval lab on longitudinal evaluation of model performance. In: CLEF. Lecture Notes in Computer Science, vol. 14163, pp. 440–458. Springer (2023)
3. Alkhalifa, R., et al.: Overview of the CLEF 2024 longeval lab on longitudinal evaluation of model performance. In: CLEF (2). Lecture Notes in Computer Science, vol. 14959, pp. 208–230. Springer (2024)
4. Bakirci, C., et al.: Cir at longeval 2026: ad-hoc scientific retrieval, topic extraction from query logs, and user simulation. In: Salido, E.S., no, A.B.C., de Herrera, A.G.S., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2026)
5. Bouaraba, S., et al.: Longeval 2026 dataset (2026). https://doi.org/10.48436/6avmz-esd52
6. Breuer, T., Keller, J., Schaer, P.: ir_metadata: an extensible metadata schema for IR experiments. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11 - 15 July 2022, pp. 3078–3089. ACM (2022). https://doi.org/10.1145/3477495.3531738
7. Breuer, T., et al.: Report on the workshop on simulations for information access (sim4ia 2024) at SIGIR 2024. CoRR abs/2409.18024 (2024). https://doi.org/10.48550/arXiv.2409.18024

8. Cancellieri, M., et al.: Longeval at CLEF 2025: longitudinal evaluation of IR systems on web and scientific data. In: Carrillo-de-Albornoz, J., et al., (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, 9-12 September 2025, Proceedings, pp. 363–387. Lecture Notes in Computer Science, Springer (2025). https://doi.org/10.1007/978-3-032-04354-2_20
9. Cormack, G.V., Clarke, C.L.A., Büttcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J. (eds.) Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, 19-23 July 2009, pp. 758–759. ACM (2009). https://doi.org/10.1145/1571941.1572114
10. D, K., Vr, J., A, B., R, P.: Temporal stability in scientific information retrieval: a three-system analysis for clef longeval 2026. In: Salido, E.S., no, A.B.C., de Herrera, A.G.S., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2026)
11. Faggioli, G., et al.: Perspectives on large language models for relevance judgment. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 39–50. ACM, Taipei Taiwan (2023). https://doi.org/10.1145/3578337.3605136
12. Fröbe, M., et al.: Large language model relevance assessors agree with one another more than with human assessors. In: Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S. (eds.) Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, 13-18 July 2025, pp. 2858–2863. ACM (2025). https://doi.org/10.1145/3726302.3730218
13. Fröbe, M., et al.: Continuous integration for reproducible shared tasks with TIRA.io. In: Kamps, J., et al., (eds.) Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), pp. 236–241. Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (2023).https://doi.org/10.1007/978-3-031-28241-6_20
14. Keller, J., Breuer, T., Schaer, P.: Evaluation of temporal change in IR test collections. In: Oosterhuis, H., Bast, H., Xiong, C. (eds.) Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024, pp. 3–13. ACM (2024). https://doi.org/10.1145/3664190.3672530
15. Keller, J., Breuer, T., Schaer, P.: Replicability measures for longitudinal information retrieval evaluation. In: Goeuriot, L., et al., (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, 9-12 September 2024, Proceedings, Part I. Lecture Notes in Computer Science, vol. 14958, pp. 215–226. Springer (2024). https://doi.org/10.1007/978-3-031-71736-9_16
16. Keller, J., et al.: Formalized information needs improve large-language-model relevance judgments. In: Moffat, A., Scholer, F., Bast, H., Najork, M., Zhang, M. (eds.) Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, Melbourne VIC Australia, 20-24 July 2026, pp. 778–789. ACM (2026). https://doi.org/10.1145/3805712.3809561
17. Keller, J., Fröbe, M., Hendriksen, G., Alexander, D., Potthast, M., Schaer, P.: Simplified longitudinal retrieval experiments: a case study on query expansion and document boosting. In: Experimental IR Meets Multilinguality, Multimodality,

and Interaction - 16th International Conference of the CLEF Association, CLEF 2024, Madrid, Spain, 9-12 September 2025, Proceedings, Part I. Lecture Notes in Computer Science, Springer (2025)
18. Kruff, A.K., Bernard, N., Schaer, P.: Validating search query simulations: A taxonomy of measures. In: Campos, R., et al., (eds.) Advances in Information Retrieval - 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29 - April 2, 2026, Proceedings, Part I, pp. 243–256. Lecture Notes in Computer Science, Springer (2026). https://doi.org/10.1007/978-3-032-21289-4_16
19. Kruff, A.K., Kreutz, C.K., Breuer, T., Schaer, P., Balog, K.: Sim4ia-bench: a user simulation benchmark suite for next query and utterance prediction. In: Campos, R., et al., (eds.) Advances in Information Retrieval - 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29 - April 2, 2026, Proceedings, Part IV, pp. 594–609. Lecture Notes in Computer Science, Springer (2026). https://doi.org/10.1007/978-3-032-21321-1_61
20. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Croft, W.B., Harper, D.J., Kraft, D.H., Zobel, J. (eds.) SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 9-13 September 2001, New Orleans, Louisiana, USA, pp. 120–127. ACM (2001). https://doi.org/10.1145/383952.383972
21. Macdonald, C., Tonellotto, N.: Declarative experimentation in information retrieval using pyterrier. In: Balog, K., Setty, V., Lioma, C., Liu, Y., Zhang, M., Berberich, K. (eds.) ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, 14-17 September 2020, pp. 161–168. ACM (2020). https://doi.org/10.1145/3409256.3409829
22. Merker, J.H., et al.: Web-scale retrieval experimentation with chatnoir-pyterrier. In: Hauff, C., et al., (eds.) Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, 6-10 April 2025, Proceedings, Part V, pp. 96–104. Lecture Notes in Computer Science, Springer (2025). https://doi.org/10.1007/978-3-031-88720-8_17
23. Merker, J.H., Fröbe, M., Stein, B., Potthast, M., Hagen, M.: Axioms for retrieval-augmented generation. In: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR). ICTIR 2025, pp. 67–77. Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3731120.3744601
24. OpenAI: gpt-oss-120b & gpt-oss-20b model card. CoRR abs/2508.10925 (2025). https://doi.org/10.48550/ARXIV.2508.10925
25. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Harman, D.K. (ed.) Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, 2-4 November 1994, pp. 109–126. NIST Special Publication, National Institute of Standards and Technology (NIST) (1994). http://trec.nist.gov/pubs/trec3/papers/city.ps.gz
26. Sáez, G.N.G.: 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). https://tel.archives-ouvertes.fr/tel-04547265
27. González-Sáez, G.N., Mulhem, P., Goeuriot, L.: Towards the evaluation of information retrieval systems on evolving datasets with pivot systems. In: Candan, K.S., et al., (eds.) CLEF 2021. LNCS, vol. 12880, pp. 91–102. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_8

28. Soboroff, I.: Don't use LLMs to make relevance judgments. Inf. Retr. Res. J. **1**(1), 29–46 (2025). https://doi.org/10.54195/IRRJ.19625
29. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, 14-18 July 2024, pp. 1930–1940. ACM (2024). https://doi.org/10.1145/3626772.3657707
30. Upadhyay, S., et al.: A large-scale study of relevance assessments with large language models using UMBRELA. In: Zamani, H., Dietz, L., Piwowarski, B., Bruch, S. (eds.) Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, Padua, Italy, 18 July 2025, pp. 358–368. ACM (2025). https://doi.org/10.1145/3731120.3744605
31. Upadhyay, S., Pradeep, R., Thakur, N., Craswell, N., Lin, J.: UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor (2024). https://doi.org/10.48550/arXiv.2406.06519
32. Voorhees, E.M.: The evolution of cranfield. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, pp. 45–69. The Information Retrieval Series, Springer (2019). https://doi.org/10.1007/978-3-030-22948-1_2
33. Voorhees, E.M., Craswell, N., Lin, J.: Too many relevants: whither cranfield test collections? In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) SIGIR 2022: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11 - 15 July 2022, pp. 2970–2980. ACM (2022). https://doi.org/10.1145/3477495.3531728
34. Wu, H., Yang, Y.: Official and diagnostic analysis of full-text temporal retrieval for longeval-sci. In: Salido, E.S., no, A.B.C., de Herrera, A.G.S., MacAvaney, S., Struß, J.M. (eds.) CLEF 2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2026)
35. Yu, C., Li, H., Zuccon, G., Mackenzie, J., Leelanupab, T.: When LLM judges inflate scores: exploring overrating in relevance assessment. CoRR abs/2602.17170 (2026). https://doi.org/10.48550/arXiv.2602.17170
36. Zhang, Y., et al.: Qwen3 embedding: advancing text embedding and reranking through foundation models. CoRR abs/2506.05176 (2025). https://doi.org/10.48550/arXiv.2506.05176