CIR at LongEval 2026: Manual Topic Creation and Concept Induction-Based Topic Generation

Timo Breuer[1,*], Jüri Keller[1], Nolwenn Bernard[1] and Philipp Schaer[1]

[1]TH Köln - University of Applied Sciences, Claudiusstr. 1, 50678 Cologne, Germany

Abstract

Information Retrieval (IR) systems are challenged by temporal changes in user behavior, document collections, and language usage. The LongEval Lab addresses these issues by evaluating the long-term generalizability of retrieval systems under evolving information settings. In this work, we participate in the LongEval-TopEx task, which focuses on generating TREC-style topics from search engine query logs and user interaction data. The generated topics consist of queries, descriptions, and narratives that are subsequently used for large language model (LLM)-based relevance assessment. Our contributions are twofold. First, we provide manually-created topic submissions developed by information science students. Second, we investigate concept induction as an automated approach for topic generation, leveraging extracted concepts, summaries, and rationales to construct topic descriptions and narratives or to guide LLM-based generation. In total, we submit 17 topic runs enabling controlled comparisons between manual, template-based, and generative approaches. Our work contributes a human-based testbed with manual topic creation, while demonstrating the potential of concept induction for scalable topic generation.

Keywords

information need, synthetic topic generation, manual topic generation, LLooM, LLM evaluation

1. Introduction

Information Retrieval (IR) systems are commonly evaluated using static datasets and fixed evaluation settings. However, both user behavior and document collections evolve over time, which can significantly affect the performance of retrieval systems. Changes in language usage, societal trends, and user expectations may lead to performance degradation, especially when there is a temporal gap between training and testing data. The LongEval Lab was established to address these challenges by studying the long-term generalizability of IR systems and developing robust methods that remain effective under temporal changes.

The LongEval Lab, currently in its fourth edition, investigates how retrieval systems adapt to evolving documents, queries, and relevance judgments. The lab provides training data from earlier time periods together with more recent, unseen test data, enabling the evaluation of retrieval performance over short- and long-term temporal shifts. For the 2026 edition, the lab expands its scope to four tasks covering ad-hoc retrieval, topic extraction from user behavior, user simulation, and Retrieval-Augmented Generation (RAG) under evolving information settings. For further information, we refer to the official overview papers [1, 2].

This report focuses on the task *LongEval-TopEx: Topic Extraction From Query Logs* (Task 2). The objective of this task is to generate formalized TREC-style topics from search engine query logs and user interaction data. Each extracted topic consists of a query, a description clarifying the users' underlying information need, and a narrative defining which documents should be considered relevant or non-relevant. The generated topics are then used to create relevance judgments with the help of large language models (LLMs).

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

*Corresponding author.

timo.breuer@th-koeln.de (T. Breuer); jueri.keller@th-koeln.de (J. Keller); nolwenn.bernard@th-koeln.de (N. Bernard); philipp.schaer@th-koeln.de (P. Schaer)

0000-0002-1765-2449 (T. Breuer); 0000-0002-9392-8646 (J. Keller); 0009-0007-0565-3210 (N. Bernard); 0000-0002-8817-4632 (P. Schaer)

The task evaluation is based on three main criteria, including *alignment*, *distinguishability*, and *clarity*. First, alignment measures how closely the generated topics correspond to observed user behavior in click logs. Second, distinguishability evaluates whether the generated relevance judgments can effectively differentiate the performance of retrieval systems. Third, clarity assesses the consistency of relevance judgments generated by different LLM assessors for the same topic. Together, these evaluation dimensions aim to ensure that extracted topics provide meaningful and reliable representations of user information needs over time.

Our contributions to the task are twofold. First, we provide manually-created runs written by information science students. Second, we provide automatically-generated topics by leveraging concept induction, which is a powerful technique to extract high level concepts from text input data. The concepts and related features like summaries and rationales, which define relevance and decision criteria, are either used to construct topic submissions directly or to instruct LLMs when generating topics. In total, we contribute 17 topic submissions that facilitate a controlled setting in reference to manually-created topics and allow the comparison of different template-based construction and generative approaches. The source code is provided in a public GitHub repository.[1]

2. Manual Topic Creation

The topic creations were conducted by 15 individual persons (8 female and 7 male, all in the age range of 23-30). All are graduate students in a Computer and Information Science Master's program at a German university. Each student submitted two topics, resulting in a total of 31 submissions for task 2 (due to one duplicate submission). Topic creation was part of a lecture on Web Information Retrieval and included a 90-minute session covering information needs, IR test collections, and topic abstractions. They were instructed for an additional 30 minutes on the specific topic creation task, and we crafted an example topic in class. The students were instructed to pick any two queries from the task 2 query list to avoid intersections and duplicates. After the query selection, they had to fill out an online form that included the following fields and instructions:

- **Title**: The title should be a brief, 2-4 words label for the topic.
- **Description**: The description should clearly summarize the user's information goal in one sentence or a question.
- **Narrative**: The narrative should be a short statement that explains the user's intent and the criteria for relevance.

To support the manual topic creation process, the students had access to ChatNoir's longeval-sci-2026-snapshot-3 index[2] [3] to manually check the results based on the original queries. Additionally, we asked them to provide examples of one to three relevant and one to three irrelevant documents. The instructions stated that, for these manually extracted documents, they could also try different search queries or specifically search for documents they had found on another search engine, like Google, but no one provided non-ChatNoir examples. In a few cases, students used the ClueWeb22 index on ChatNoir. Good negative (irrelevant) samples were defined as documents that might seem relevant at first glance but are, in fact, not relevant. Spam was mentioned as a poor example for manual sampling.

Overall, we gathered 2.42 positive examples and 2.1 negative examples per topic. Titles, descriptions, and narratives had average lengths of 3.58, 14.51, and 51.64 words, respectively.

3. Automatic Topic Generation

The following section describes the key methodology of our automatic topic generation approaches. First, we provide a brief introduction to the concept induction algorithm and the LLooM framework.

[1] https://github.com/irgroup/longeval26-concept-topics/
[2] https://www.chatnoir.eu/?index=longeval-sci-2026-snapshot-3

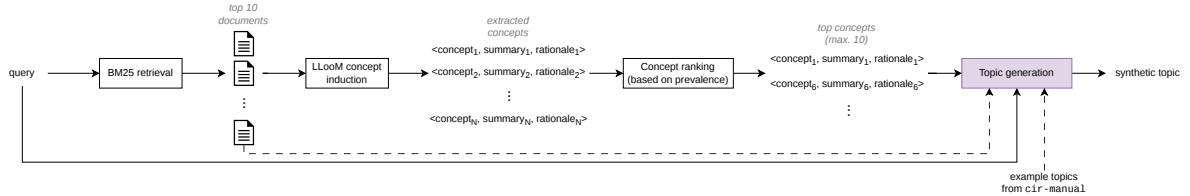

Figure 1: Overview of the automatic topic generation pipeline. Dash lines represent optional input for the topic generation based on variants.

Afterwards, we outline the principles and building blocks of the topic generation methods that are used for the final topic submissions. Figure 1 provides an overview of the automatic topic generation pipeline. For a given query, BM25 retrieves ten documents that are used for the concept induction with LLooM. The concept induction algorithm generates concepts, summaries, and rationales that are ranked by the concept's prevalence score afterwards. At maximum, ten of these ranked concepts are then included in the topic generation step to synthesize the final topic.

3.1. Concept Induction

Concept induction [4] is a powerful topic modeling approach that leverages LLMs to find high-level concepts in text data. Technically, the methodology comprises a multi-stage pipeline, including (1) the distillation of input data, (2) the clustering of related summarized items, and (3) the synthesis to unified concepts. Each stage is executed consecutively, and the pipeline is repeated multiple times, processing the outputs of the previous run to facilitate higher levels of abstraction. Compared to other topic modeling approaches like Latent Dirichlet allocation (LDA) [5] or BERTopic [6], qualitative evaluations suggest that concept induction leads to more meaningful concepts and topics.

Lam et al. [4] introduce concept induction as part of the LLooM workbench that integrates a human-in-the-loop approach to select appropriate concepts for LLM-based scoring of the input text documents. Once concepts are synthesized from the input data, the human actor selects fitting concepts from a list of candidates that are then used to score the input document with regard to how well their content fits the concept. Besides the concept title, the LLooM workbench also synthesizes criteria in the form of question, helping the LLM during scoring, and a more detailed summary of the concept. In addition to the actual scores, LLooM also provides a text-based rationale for the scoring.

The LLooM software package[3] allows customization by integrating different LLMs for the single stages. In our experiments, we employ LLMs from the Qwen3 family, for both text completion as well as text embedding (Qwen3-30B Mixture-of-Expert model [7] for the distillation and synthesis steps and Qwen3-4B embeddings [8]). For each query, we used the 10 documents ranked highest by the BM25 baseline run as the input for the concept induction. In order to fully automate the approach, we include the top 10 concepts for the scoring – or less in case fewer result from the concept synthesis. Eventually, the concepts are ranked by their prevalence in the input data to select the most promising for the generation of the topics.

3.2. Topic Generation Approaches

In general, all of our automatic topic generation approaches build up on the results of the concept induction. In the following, we provide an overview of the features considered for the topic generation, in which we transfer the mined concepts into retrieval topics that simulate information needs.

- **Document:** Given the logged query, we retrieve 10 documents with BM25. All documents are used for the concept induction. Two topic generation approaches include documents of this ranking in the prompt as relevant examples, namely `cir-h1d` and `cir-h1fd` (cf. Section 4, denoted as **Documents** in Table 1).

[3]https://stanfordhci.github.io/lloom/

Description Prompt

Given a user query and a concept described in the following blocks, write the description field of a TREC-style topic that simulates a nuanced user information need. The description should clearly summarize the user's information goal in a single sentence or a question.

Query:
A person has typed this query into a search engine:
<query>

Relevant Concept:
Title: *<concept>*
Summary: *<summary>*

Contrastive
Not relevant Concept:
Title: *<concept>*
Summary: *<summary>*

Documents
Relevant document:
<document>

Few-shot
Below are some examples:
<examples>

Narrative Prompt

Given a user query, a description, and a concept described in the following blocks, write the narrative field of a TREC-style topic. The narrative should be a short statement that explains the user's intent and the criteria for relevance.

Query:
A person has typed this query into a search engine:
<query>

Description of the information need:
<description>

Relevant Concept:
Title: *<concept>*
Rationale: *<rationale>*

Contrastive
Not relevant Concept:
Title: *<concept>*
Rationale: *<rationale>*

Documents
Relevant document:
<document>

Few-shot
Below are some examples:
<examples>

Figure 2: General form of the prompts used to synthesize the topics. The left prompt is used for the description generation, and the right prompt is used for the narrative. The narrative prompt contains the output based on the description prompt in the *description* field. *<Italicised>* words are placeholders, filled with appropriate contexts. Shaded text is optional, included in some prompt variants indicated by the color coding.

- **Concept:** Based on provided text input data, the concept induction algorithm produces high-level concepts with the help of synthesized inclusion criteria. A concept is comparable to a topic resulting from conventional topic modeling approaches. However, in direct comparisons, human evaluation suggests that concepts have a higher level of meaningfulness [4]. All of the automatic topic generation approaches make use of concepts with different numbers and with different levels of prevalence (cf. Section 4, denoted as **High** and **Low** in Table 1). Compared to the other features the concept string is rather short, comparable to a topic title of ad hoc IR test collections.
- **Summary:** In addition to the concept, the induction algorithm also produces a summary for each identified concept. A summary is a more detailed explanation of the concept, describing the inclusion criteria for the concept scoring. It is somewhat comparable to the topic description of ad hoc IR test collections. All automatic topic generation approaches make use of the summaries corresponding to the concepts included in the generation process.
- **Rationale:** The scoring of single documents with regard to how well they fit a concept is based on a rationale. The rationale is comparable to the description of relevant documents in the topic narrative of ad hoc IR test collections. All automatic topic generation approaches make use of the rationales corresponding to the concepts included in the generation process.

In addition to the actual text features, we build upon the following strategies and methodologies for the topic generation with different combinations:

- **Prevalence-based ranking of concepts:** The prevalence scores account for the number of occurrences of the identified concept in the text input data. Concepts that occur more often in

the documents have a higher level of prevalence. Based on the prevalence scores, it is possible to define a ranking of concepts, e.g., to better identify high- and low-ranked concepts for the topic generation.

- **Template-based topic construction:** Provided with the results of the concept induction, it is feasible to construct topic in a principled way with the help of templates. The actual topic construction does not include generative components based on LLMs at this stage. All generated output from LLMs was part of the concept induction. Approaches following this strategy, simply include the concept, summary, and rationale in a template. The two topic submissions following this approach are `cir-h1-raw` and `cir-l1-raw` (cf. Section 4, denoted as **Template** in Table 1).
- **Generative topic construction:** Topics based on this strategy imply a consecutive generation step after the concept induction. Instead of simply including the concept induction results in a final template, prompt templates are used to hand over the features explained earlier to the LLM with the instructions to generate topics for ad hoc IR test collections (cf. Section 4, denoted as **Generative** in Table 1). In the first step, the concept name and the summary are used, among the other features, to generate a topic description. Subsequently, the concept title and rationale, and also the newly generated topic description, are used to generate the narrative. Figure 2 shows both prompts. All of the generative approaches use at least one concept, and the corresponding summaries and rationales. They differ by what kinds of concepts and how these are integrated and combined in the prompt. Note that the description and the narrative are generated with separate prompts. The majority of our submissions (14 out of 17) follow a generative approach.
- **Few-shot topic examples:** Several topic generation approaches follow a few-shot prompting strategy that includes selected examples from the manual creation process described in Section 2. In general, all of these topic submissions used three examples of manual topics (cf. Section 4, denoted as **Few-shot** in Table 1).
- **Contrastive concept examples:** Several topic generation approaches combine high- and low-ranked concepts in a contrastive manner. (cf. Section 4, denoted as **Contrastive** in Table 1). Overall, four submissions imply a contrastive prompting strategy in the topic generation. These approaches mainly differ by the number of concepts included, and the way high- and low-ranked concepts are treated as positive and negative examples. For instance, a strategy that generates a topic based on well-represented concepts in the input data, i.e., concepts with a high prevalence score, would use high-ranked concepts as positive examples and low-ranked as negative ones. The two runs following this strategy are `cir-h1c` and `cir-h1cf`. Another way to combine concepts in a contrastive manner is prioritizing concepts with a low level of prevalence in the input data, e.g., in order to emphasize concepts that are rather not well-represented to enforce fairness and diversity in the topic generation process. The two runs following this strategy are `cir-l1c` and `cir-l1cf`.

For the topic generation, we used GPT-OSS-120B,[4] an open-weight model from OpenAI. The model implements a sparse mixture of experts architecture with 128 experts and a total of 117B parameters. Each expert uses 5.1B parameters actively for efficient inference [9]. The dataset, including queries and document texts, is interfaced through the `ir_dataset_longeval` extension [10] and the generation was facilitated by the TopicGen toolkit [11].

4. Submitted Runs

Table 1 provides an overview of all submitted topic runs that were either manually created or automatically generated. In addition to the structured overview, we provide a brief description of each run in the following.

[4]Hugging Face: https://huggingface.co/openai/gpt-oss-120b

Table 1
Overview of topic generation approaches and their underlying strategies.

Approach	Manual	High	Low	Template	Generative	Contrastive	Few-shot	Documents
cir-manual	✓	-	-	-	-	-	-	-
cir-h1-raw	-	1	-	✓	-	-	-	-
cir-l1-raw	-	-	1	✓	-	-	-	-
cir-h1	-	1	-	-	✓	-	-	-
cir-h1c	-	1	1	-	✓	✓	-	-
cir-h1cf	-	1	1	-	✓	✓	✓	-
cir-h1d	-	1	-	-	✓	-	-	✓
cir-h1f	-	1	-	-	✓	-	✓	-
cir-h1fd	-	1	-	-	✓	-	-	✓
cir-h2	-	2	-	-	✓	-	-	-
cir-h2f	-	2	-	-	✓	-	✓	-
cir-l1	-	-	1	-	✓	-	-	-
cir-l1c	-	1	1	-	✓	✓	-	-
cir-l1cf	-	1	1	-	✓	✓	✓	-
cir-l1f	-	-	1	-	✓	-	✓	-
cir-l2	-	-	2	-	✓	-	-	-
cir-l2f	-	-	2	-	✓	-	✓	-

cir-manual This manual topic submission was created by students of the Master study program "Data and Information Science" at TH Köln. Students freely selected their queries from the longeval-sci dataset and could investigate documents using the ChatNoir retrieval system. More details about the topic creation are provided above (cf. Section 2).

cir-h1-raw The concept with the highest prevalence and the corresponding summary and rationale are used as the topic for the respective query, wrapped by a minimal static template.

cir-l1-raw Contrary to `cir-h1-raw`, the concept with the lowest prevalence and the corresponding summary and rationale are used as the topic for the respective query, wrapped by a minimal static template.

cir-h1 The concept with the highest prevalence and the corresponding summary and rationale are included in a prompt that is given to the LLMs that generates the topic for the respective query.

cir-h1c The concept with the highest prevalence is used as a positive example, and the concept with the lowest prevalence is used as a negative example, as contrastive context for the LLM to generate the topic. Besides the concept strings, the summaries and rationales are also included.

cir-h1cf This submission follows the same methodology as `cir-h1c` but adds three topics from the `cir-manual` submission as few-shot examples.

cir-h1d This submission follows the same methodology as `cir-h1`, i.e., the concept with the highest prevalence and the corresponding summary and rationale are included in a prompt that is given to the LLMs that generates the topic. Additionally, a single document that fits the concept the best (with the highest concept score) is used as additional context for the prompt.

cir-h1f This submission follows the same methodology as `cir-h1`, i.e., the concept with the highest prevalence and the corresponding summary and rationale are included in a prompt that is given to the LLMs that generates the topic. Additionally, three topics from the `cir-manual` submission are included in the prompt as few-shot examples.

cir-h1fd This submission follows the same methodology as `cir-h1f`. Additionally, a single document that fits the concept the best (with the highest concept score) is used as additional context for the prompt.

cir-h2 This submission follows the same methodology as `cir-h1` but includes two instead of one concept (and the corresponding summaries and rationales) in the prompt given to the LLM that generates the topic.

cir-h2f This submission follows the same methodology as `cir-h2`, and additionally includes three topics from the `cir-manual` submission as few-shot examples.

cir-l1 This submission is complementary to `cir-h1`. It replaces the high-ranked and the concept with the lowest prevalence and the corresponding summary and rationale are included in a prompt that is given to the LLMs that generates the topic for the respective query.

cir-l1c This submission is complementary to `cir-h1c`. It basically swaps the role of the high- and low-ranked concepts. The concept with the lowest prevalence is used as a positive example, and the concept with the lowest prevalence is used as a negative examples, as contrastive context for the LLM to generate the topic. Besides the concept strings, the summaries and rationales are also included.

cir-l1cf This submission follows the same methodology as `cir-l1c` but adds three topics from the `cir-manual` submission as few-shot examples. It is complementary to `cir-h1cf`.

cir-l1f This submission follows the same methodology as `cir-l1`, i.e., the concept with the lowest prevalence and the corresponding summary and rationale are included in a prompt that is given to the LLMs that generates the topic. Additionally, three topics from the `cir-manual` submission are included in the prompt as few-shot examples.

cir-l2 This submission follows the same methodology as `cir-l1` but includes two instead of one concept (and the corresponding summaries and rationales) in the prompt given to the LLM that generates the topic. It is complementary to `cir-h2`.

cir-l2f This submission follows the same methodology as `cir-h2`, and additionally includes three topics from the `cir-manual` submission as few-shot examples. It is complementary to `cir-h2f`.

5. Results

This section reports the experimental evaluations of our topics by comparing them to the baseline of the lab organizers. We refer the reader to the overview reports [1, 2] for a more detailed experimental evaluation, contextualizing them with submissions from other participants. The organizer's baseline approach lets the LLM generate relevance labels based on the query without including a description or narrative. The experimental setup makes use of four different LLMs of different families and sizes, including qwen3-32b, llama-3.1-8b, gpt-oss-120b, and gpt-oss-20b, which are used to generate relevance labels for the provided topics.

Table 2 shows the agreement on the observations between all pairs of our four LLM relevance assessors. Specifically, Cohen's Kappa quantifies the agreement for individual relevance judgment decisions and TauAP quantifies the agreement on the resulting system ranking for nDCG@10. In comparison, for both measures, the scores of our manual topic submissions are slightly lower than the baseline. This indicates that the LLMs agree less on the human topics compared to the baseline. The scores of the automated submissions are all lower, often substantially, than those of the manual

Table 2
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 values are better) and the agreement on the resulting system ranking for nDCG@10 as TauAP correlation (higher values are better).

Team	Method	Cohens Kappa	Tau-AP
Baseline	no-description-no-narrative	0.334	0.792
CIR	cir-manual	0.305	0.775
CIR	cir-l1-raw	0.196	0.654
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

submission and the baseline. According to Cohen's Kappa, topics generated from concepts with low prevalence scores show lower agreement than those from high-prevalence concepts. The only exception is the cir-l1cf run that uses the contrastive and few-shot method. The Tau-AP scores, indicating the correlation of system rankings, show a slightly different result. The raw submissions that directly rely on the concepts without further context or generation steps outperform the generated topics, including the ones that use high-prevalence concepts. Regarding the generated topics, using high-prevalence concepts most often also leads to a higher Tau-AP agreement.

Table 3 provides more insights by reporting the percentage of documents in the pool that are judged as relevant. It can be seen that the baseline approach has a comparatively high number of documents judged as relevant for all models that are used as assessors. On average, the generated topics, especially on low-prevalence concepts, reduce the percentage of relevant documents. In comparison to the baseline, the manually-created topics have a wider spread of scores across the different models. For instance, llama-3.1-8b has the highest score (0.906) and gpt-oss-20b has the lowest score (0.658). This disagreement is even stronger for the generated topics. Especially the GPT-OSS LLMs, the same model family that we used to synthesize the topics, judge fewer documents relevant compared to the other LLMs. The outcomes suggest that the formalized topics lead to fewer relevant documents but also less agreement among the models than the baseline topics.

6. Conclusion

This paper presents our participation in the LongEval-TopEx 2026 task on topic extraction from query logs. We contributed both manually-created and automatically-generated topic submissions, with a particular focus on concept induction as a mechanism for deriving structured representations of user information needs. Furthermore, we introduced a sequential generation process, first generating the topic description and then using it to generate the narrative. Our submissions facilitate comparisons between manual topic construction, template-based approaches, and LLM-assisted generative methods in a controlled evaluation setting that should be analyzed in more detail as part of follow-up studies.

Table 3
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 method (i.e., the topic extraction approach or how humans have been observed in the production system). Too high percentages of relevant documents can indicate that relevance judgments can not be applied to new retrieval systems that retrieve many unjudged documents.

Team	Method	Relevant				
		oss-120b	oss-20b	llama-3	qwen3	mean
Baseline	no-description-no-narrative	0.858	0.859	0.943	0.908	0.892
CIR	cir-h1-raw	0.820	0.852	0.931	0.873	0.869
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
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
Observations	click-qrels					0.127

Acknowledgments

Thanks to the group of students who generated the manual topics: David Hoffmann, Divashini Shasitharan, Janina Sophie Janßen, Muhammad Fakhar, Muhammed Sirac Coban, Rafael Rodrigues Barros, Salman Tariq, Shih-Han Pan, Temiloluwa Sangobiyi, Teresa Lacerda, and Valentin Casas Stöldt.

This work was supported by Deutsche Forschungsgemeinschaft (407518790) 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 spelling and grammar check and Gemini 3.1 pro for formatting assistance. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

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

[2] S. Bouaraba, T. Breuer, M. Cancellieri, A. El-Ebshihy, M. Fröbe, P. Galuscáková, L. Goeuriot, G. Iturra-Bocaz, J. Keller, P. Knoth, A. K. Kruff, P. Mulhem, F. Piroi, D. Pride, P. Schaer, D. Schwab,

Extended Overview of the CLEF 2026 LongEval Lab on Longitudinal Evaluation of Model Performance, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[3] M. Potthast, M. Hagen, B. Stein, J. Graßegger, M. Michel, M. Tippmann, C. Welsch, ChatNoir: a search engine for the ClueWeb09 corpus, in: W. R. Hersh, J. Callan, Y. Maarek, M. Sanderson (Eds.), The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012, ACM, 2012, p. 1004. URL: https://doi.org/10.1145/2348283.2348429. doi:10.1145/2348283.2348429.

[4] M. S. Lam, J. Teoh, J. A. Landay, J. Heer, M. S. Bernstein, Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM, in: F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. O. T. Dugas, I. Shklovski (Eds.), Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024, ACM, 2024, pp. 766:1–766:28. URL: https://doi.org/10.1145/3613904.3642830. doi:10.1145/3613904.3642830.

[5] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada], MIT Press, 2001, pp. 601–608. URL: https://proceedings.neurips.cc/paper/2001/hash/296472c9542ad4d4788d543508116cbc-Abstract.html.

[6] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, CoRR abs/2203.05794 (2022). URL: https://doi.org/10.48550/arXiv.2203.05794. doi:10.48550/ARXIV.2203.05794. arXiv:2203.05794.

[7] Q. Team, Qwen3 Technical Report, CoRR abs/2505.09388 (2025). URL: https://doi.org/10.48550/arXiv.2505.09388. doi:10.48550/ARXIV.2505.09388. arXiv:2505.09388.

[8] 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. doi:10.48550/ARXIV.2506.05176. arXiv:2506.05176.

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

[10] 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: 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. 117–127. URL: https://doi.org/10.1007/978-3-032-04354-2_8. doi:10.1007/978-3-032-04354-2_8.

[11] 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. doi:10.48550/ARXIV.2604.04140. arXiv:2604.04140.