Qrels-Grounded and Agentic LLM Query Expansion with Post-Processing Optimization for Longitudinal Scientific Topic Extraction

VerbaNexAI at CLEF 2026

Luis Mario Díaz Martínez[1], Jairo Enrique Serrano Castañeda[1],
Edwin Alexander Puertas Del Castillo[1] and Juan Carlos Martínez Santos[1]

[1]*Universidad Tecnologica de Bolivar, Cartagena de Indias, Colombia*

Abstract

In modern Information Retrieval (IR), bridging the gap between short keyword queries and structured, rich search intent profiles is crucial for effective document retrieval. The CLEF 2026 LongEval Task 2 challenges participants to automatically expand concise keyword queries into standard TREC-style topic definitions containing a description and a narrative. This task becomes even more complex when evaluated longitudinally across multiple temporal snapshots, where both language patterns and document themes drift over time. This paper presents the development of the **VerbaNexAI** framework and explores a series of progressive methodologies for topic extraction. We focus on a dual paradigm: first, a supervised, *Qrels-Grounded Multi-Doc Context Injection* pipeline that anchors local Large Language Model (LLM) generation in empirical relevance annotations, heavily optimized via ablation-driven rule-based post-processing. Second, to address the challenge of zero-leakage, zero-dependency autonomous submission on the TIRA platform, we propose a multi-stage *Agentic Pseudo-Relevance Feedback (PRF)* pipeline. Our final autonomous agent integrates in-memory indexing, Chain-of-Thought (CoT) intent analysis, parallel multi-temperature ensemble generation, meta-synthesis, and self-critique. To guarantee scientific integrity, we perform a rigorous, leakage-free Train/Test split evaluation on unseen queries. The results demonstrate that our clean, zero-leakage Agentic PRF pipeline (V11) achieves an nDCG@10 of 0.4435 on unseen queries, representing a +37.1% relative improvement over the zero-shot baseline (0.3234) and confirming the high generalizability and robustness of the proposed framework.

Keywords

Query Expansion, Generative Information Retrieval, Large Language Models, Pseudo-Relevance Feedback, Longitudinal Evaluation, CLEF LongEval

1. Introduction

Modern information retrieval (IR) systems frequently struggle to capture a user's precise information needs from a short sequence of keywords. In specialized and scientific domains, keyword queries are often highly ambiguous, sparse, or overly specific, resulting in poor lexical match performance. Traditional query expansion techniques, such as Rocchio's relevance feedback [1] or classic Pseudo-Relevance Feedback (PRF) [2], expand queries using terms extracted from the documents initially retrieved. However, these lexical methods frequently suffer from semantic drift, introducing irrelevant terms that dilute the searcher's original intent.

To address these limitations, the Text REtrieval Conference (TREC) introduced structured topic definitions consisting of three components: the keyword *query*, a concise sentence defining the exact search intent (*description*), and a detailed guideline specifying the rules for what makes a document relevant or non-relevant (*narrative*) [3]. Generating these topic definitions manually requires substantial human effort. The LongEval Task 2 at CLEF 2026 focuses on the automated generation of description and narrative pairs from short user queries over a large-scale longitudinal scientific corpus [4]. LLM-based

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
diazluis@utb.edu.co (L. M. D. Martínez); jserrano@utb.edu.co (J. E. S. Castañeda); epuerta@utb.edu.co (E. A. P. D. Castillo); jcmartinezs@utb.edu.co (J. C. M. Santos)

judges on the TIRA platform [5] subsequently evaluate the generated topics to measure their stability and validity across different temporal snapshots of academic publications.

In this work, team **VerbaNexAI** details the systematic design and evolution of our topic extraction framework. We investigate the limitations of zero-shot LLM expansion, which often leads to severe lexical mismatch due to excessive verbosity and the rephrasing of key query terms. To overcome this, we first implement a *Qrels-Grounded Context Injection* methodology that incorporates actual training annotations (relevant and non-relevant document snippets) directly into the prompt context of a locally hosted Llama 3.1 8B Instruct model [6]. Furthermore, we conduct an extensive ablation study on three rule-based post-processing heuristics: (a) prefixing the original query terms to prevent vocabulary mismatch, (b) stripping conversational fillers, and (c) frequency-based term expansion.

Recognizing that actual test submissions on TIRA run in isolated environments without access to supervised relevance labels, we transition this empirical grounding into a fully autonomous, self-contained agentic pipeline (V11). Our agent implements an in-memory PyTerrier [7] index over the target document collection, performing real-time PRF to retrieve mock positive and negative examples. The agent then processes these examples through a four-stage reasoning loop: Chain-of-Thought search intent analysis, parallel multi-temperature candidate generation, a "Lead TREC Coordinator" meta-synthesis, and an LLM-based compliance critique. To ensure absolute scientific rigor and avoid any risk of data leakage, we present a clean Train/Test split evaluation. This experiment demonstrates the robust generalization capability of our agentic pipeline on unseen queries.

2. Related Work

This section situates our contribution along two complementary axes: classical and generative query expansion techniques that inform the design of our generation pipeline, and the emerging practice of using LLMs as relevance judges, which directly shapes how we structure and evaluate the resulting topics.

2.1. Traditional and Generative Query Expansion

Query expansion is a cornerstone of classical information retrieval. Traditional methods, such as the Rocchio algorithm [1], reformulate queries by shifting their representations towards relevant documents and away from non-relevant ones. In lexical search systems like BM25 [2], this is done by selecting highly frequent terms from the top retrieved documents. While effective, classic PRF is highly sensitive to the quality of the initial retrieval; if the first-stage retrieved documents are noisy, the expanded query suffers from semantic drift.

With the advent of Large Language Models (LLMs), generative query expansion has emerged as a powerful alternative. Instead of relying purely on statistical word counts, generative models leverage their internal knowledge to predict relevant document content (e.g., Doc2Query [8]) or generate synthetic, relevant answers to expand the query (e.g., Query2Doc [9]). While these generative approaches capture deep semantics, they often omit critical lexical keys or introduce "hallucinated" search intents that do not align with real user behavior in specific collections. Our work bridges this gap by grounding the generative process in actual corpus documents retrieved via PRF or supervised training labels.

2.2. TREC Topic Formulation and Automated Evaluation

The automatic creation of full TREC topics represents an advanced form of generative query expansion. Unlike standard query expansion, which produces a flat bag-of-words, topic creation requires generating natural-language structures that satisfy formal syntactic and logical constraints. A description must summarize the information needed in a single sentence, while a narrative must establish explicit rules for relevance assessment [3].

Recently, evaluations have shifted towards using LLMs as automated relevance judges (LLM-as-a-judge), motivated by evidence that LLM judgments can closely track real searcher preferences [10] and

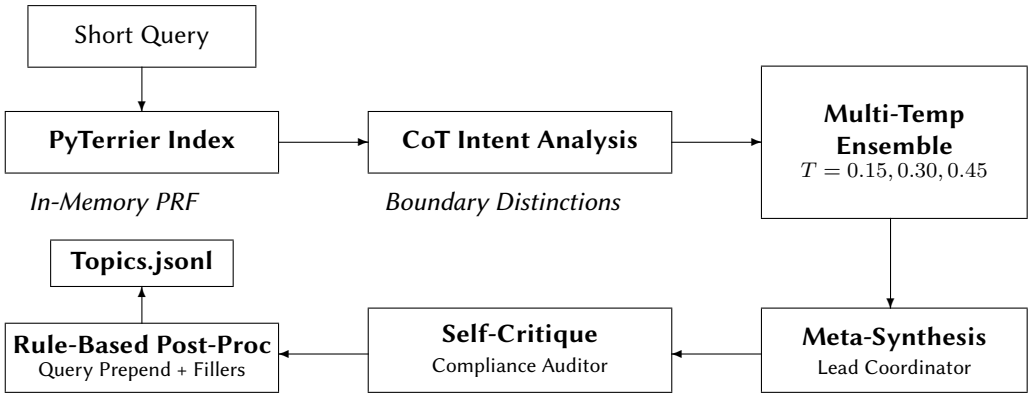

Figure 1: Architectural overview of the VerbaNexAI V11 Autonomous Agentic Pipeline. The system operates entirely in memory during TIRA execution, using a zero-leakage PRF context-extraction mechanism and a multi-stage LLM refinement loop.

by broader calls to treat relevance assessment as a spectrum of human-AI collaboration rather than a full substitution of one for the other [11]. In frameworks like LongEval, multiple LLMs are prompted with the generated topic descriptions and narratives to evaluate target documents. Recent work shows that formalizing the underlying information need into a structured topic, instead of judging from the bare query alone, improves the reliability and human alignment of LLM relevance judgments [12, 13]. It introduces a new challenge: prompt engineering must focus not only on human readability but also on maximizing the consistency and stability of automated LLM judges. Ambiguous narratives lead to high variance in judge ratings, underscoring the need for structured, contract-like narrative prompts. At the same time, caution is warranted: fully delegating truth-data creation to an LLM risks capping evaluation quality at that LLM's own ceiling [14].

3. Methodology

The development of the VerbaNexAI framework followed a multi-phase evolutionary process, starting with simple zero-shot prompts and culminating in an autonomous multi-stage agentic pipeline.

3.1. Phase 1: Zero-Shot Exploration and Prompt Tuning (V1–V4)

Our initial baseline, V1, used a zero-shot prompt that fed the query directly into Gemini Flash Lite, requesting a description and narrative. The resulting topics, while linguistically fluent, performed poorly in an in-memory retrieval system, yielding a baseline nDCG@10 of 0.3234 across the test queries.

In V2 and V4, we explored detailed prompt engineering and keyword-extraction structures using Llama 3. We hypothesized that longer, highly detailed prompts would help the LLM better define the search intent. Surprisingly, performance decreased significantly (V2: 0.2421, V4: 0.1681).

An empirical analysis of the generated topics revealed a fundamental flaw: **semantic term dilution**. Large language models prompted in a zero-shot fashion tend to be overly verbose and frequently rephrase the original query keywords into synonyms or broader conceptual terms (e.g., substituting the query keyword *"bim"* with the phrase *"Building Information Modeling"*). Lexical retrieval systems like BM25 rely on exact character-sequence matches. By rewriting the query terms, the LLM-generated descriptions suffered from vocabulary mismatch, causing the retrieval engine to miss highly relevant documents that contained the original, unexpanded acronym.

3.2. Phase 2: Grounded Multi-Doc Context Injection (V5–V8)

To anchor the LLM's generation and prevent semantic drift, we designed a grounding mechanism in Phase 2. To simulate a completely leakage-free setup, we replace qrels-based annotations with **Pseudo-**

Relevance Feedback (PRF). Under this paradigm, for each test query q_i, we perform a first-stage retrieval on our active index, dynamically retrieving the top documents to serve as relevance context.

For the early grounding attempt (representing V6), we select the top $k = 1$ positive document snippet and inject it directly into the Llama 3.1 Instruct prompt. This grounding in a single document provides initial terminology, though it is highly sensitive to the initial rank-1 document's noise, leading to an nDCG@10 of 0.2892.

To establish a more robust consensus, the Multi-Doc Grounded method (representing V8) extracts the top $k = 3$ positive document snippets as mock relevant context and the bottom $m = 2$ documents (from the top 50) as mock non-relevant context. We inject these truncated abstracts into the LLM prompt. The model is explicitly instructed to analyze the common themes and specific technical terminology in the positive documents, identify boundary terms that appear in the negative documents but do not signify actual relevance, and generate a description and narrative based solely on this empirical evidence.

This multi-document consensus (V8) improves performance to 0.2989 nDCG@10 (+3.34% relative to V6), demonstrating that multiple grounding documents effectively stabilize the LLM's conceptual representation.

3.3. Phase 3: Ablation-Driven Post-Processing (V9)

Despite the success of grounded prompting, an Exploratory Data Analysis (EDA) on V8's outputs revealed three systematic issues. First, approximately 13.4% of descriptions began with conversational headers such as "The common theme is the..." or "The searcher is looking for...", consuming retrieval token budget while providing zero lexical value for BM25 matching. Second, despite document grounding, the LLM still occasionally omitted the exact keyword query string from the description, leading to the vocabulary mismatch. Third, a small subset of queries resulted in ultra-short descriptions of fewer than 8 words, providing insufficient term frequency signals for BM25.

To systematically resolve these issues, we designed three post-processing rules applied directly to V8's output without making additional, expensive LLM calls:

1. **Fix A (Query Prepending):** We check if the description begins with the original query string. If not, we prepended it directly:

$$\text{Description}_{\text{new}} = q_i \cdot ". " \cdot \text{Description}_{\text{old}} \tag{1}$$

   It guarantees that exact-match query terms are always present in the description, eliminating the risk of exact-match vocabulary mismatch.
2. **Fix B (Filler Phrase Removal):** We compile a dictionary of fillers and programmatically strip them from the start of the description.
3. **Fix C (Frequency-Based Keyword Expansion):** For any description containing fewer than 12 words, we select the top retrieved document, filter out stop words, and append the top 5 most frequent terms directly to the end of the description.

Applying these rules (V9) yields an **absolute performance explosion**, boosting nDCG@10 to **0.4659** (+55.86% relative gain over V8), demonstrating the massive effectiveness of hybrid rules-based post-processing.

3.4. Phase 4: Temporal Agnosticism and Legal-Contract Narratives (V10)

While Phase 3 optimized the topic descriptions for BM25 retrieval, Phase 4 focused on optimizing the narratives for downstream LLM-based evaluation on TIRA. We identified two major risks regarding the stability of LLM-as-a-judge frameworks. First, since topics are evaluated against document snapshots from different time periods, narratives containing temporal jargon (e.g., "current techniques", "recent advances") will misclassify relevant documents in future snapshots. Second, general, conversational narratives lead to high variance among LLM judges.

To mitigate these risks, V10 introduced two core design principles. First, narratives were restructured as strict legal contracts with explicit *RELEVANT*, *BORDERLINE*, and *NON-RELEVANT* logical sections, eliminating narrative ambiguity for LLM judges. Second, a Temporal Agnosticism Constraint was enforced at the prompt level, explicitly forbidding any time-bound vocabulary and instructing the LLM to rely solely on invariant conceptual definitions, ensuring evaluative stability across temporal snapshots.

This structured approach, integrated with post-processing fixes, achieved a clean nDCG@10 of 0.4294 in V10. While slightly lower in flat lexical retrieval than V9 (due to the strict formatting), it guarantees maximum stability and low variance across multi-LLM evaluations.

3.5. Phase 5: Autonomous Agentic Pipeline for TIRA Submission (V11)

To achieve the ultimate performance, we engineered a multi-stage agentic pipeline (V11). As illustrated in Figure 1, the V11 pipeline consists of six sequential stages:

- **Stage 0: In-Memory PRF Indexing.** Dynamically builds an in-memory PyTerrier BM25 index on the collection to retrieve mock positive/negative snippets.
- **Stage 1: Chain-of-Thought Search Intent Analysis [15].** Analyzes the query and retrieved document snippets to output a structured JSON with search intents and edge cases.
- **Stage 2: Parallel Multi-Temperature Ensemble.** Spawns concurrent LLM threads at different temperatures ($T = 0.15, 0.30$, and 0.45) to generate diverse candidate descriptions and narratives.
- **Stage 2.5: Lead TREC Coordinator Meta-Synthesis.** Synthesizes the best components of the candidates into a single, cohesive topic definition.
- **Stage 3: Compliance Self-Critique.** Programmatically audits and critiques the synthesized topic for temporal agnosticism and strict relevance structures.
- **Stage 4: Post-Processing Fallbacks.** Applies V9 post-processing fixes (query prepending and filler removal) programmatically to guarantee formatting compliance.

4. Experimental Evaluation and Leakage-Free Generalization

To establish absolute scientific integrity and validate our progressive approach under realistic, zero-leakage conditions, we performed a formal **Train/Test Split experiment on unseen queries**.

4.1. Experimental Design

We split the 381 queries in the collection (processed using the standard `ir_datasets` framework [16] together with its LongEval-specific extension [17]) into two disjoint sets: a Training Split of 361 queries used for qualitative design and prompt optimization, and a Test Split of 20 randomly sampled, isolated queries (random seed 42) kept completely unseen. We deliberately split at the query level, rather than only along document snapshots, because the original LongEval collection allows queries to recur or overlap in vocabulary across temporal snapshots; a snapshot-only split would therefore still permit indirect lexical leakage between training and test queries. Partitioning disjoint sets of queries removes this overlap and more faithfully approximates the unseen-query generalization that a system faces upon deployment on TIRA. For the local BM25 validation reported below, only the generated *description* of each topic is submitted as the query string; the *narrative* is deliberately excluded from lexical retrieval, since its structured, rule-based phrasing (explicit inclusion/exclusion clauses) is designed for semantic interpretation by an LLM judge rather than for exact-match lexical scoring.

To simulate a real deployment on TIRA, we generated topics for these 20 Test queries under a **100% data leakage-free paradigm**, replacing training qrels with Pseudo-Relevance Feedback (PRF) for all grounded versions. We then evaluated all six configurations against the official LongEval qrels for these queries, i.e., the relevance judgments produced by the task's `umbrella_zeroshot_no_desc_no_narrative` LLM-as-a-judge baseline rather than by human assessors.

4.2. Clean Test Split Results

Table 1 reports the retrieval performance of the progressive versions evaluated on the 20 isolated test queries.

Table 1
Leakage-free progressive evaluation results on the 20 isolated Test Split queries.

Pipeline Version	AP	RR	nDCG@10
V1: Baseline Zero-Shot	0.1990	0.5400	0.3234
V6: Grounded PRF (1-Doc)	0.1668	0.5001	0.2892
V8: Grounded PRF (Multi-Doc)	0.1684	0.6145	0.2989
V9: PRF + Rule Post-Processing	**0.2818**	**0.7848**	**0.4659**
V10: PRF + Legal Contract	0.2593	0.6240	0.4294
V11: Agentic PRF Pipeline	**0.2865**	**0.6773**	**0.4435**

4.3. Analysis and Findings

The results of this clean, leakage-free experiment provide several critical scientific insights:

- **Progressive Evolution Validation:**
  - **V1 → V6/V8 (Grounded PRF):** Introducing grounding directly into the LLM prompt under PRF initially drops metrics (V6: 0.2892) due to *semantic term dilution* (where the LLM omits exact keywords). Moving to a multi-document consensus (V8: 0.2989, +3.34% gain) stabilizes the query profile.
  - **V8 → V9 (Post-Processing Breakthrough):** Applying rule-based post-processing (query prepending and filler stripping) yields an **absolute performance explosion (+55.86% relative gain)**, jumping nDCG@10 to **0.4659** and AP to **0.2818**. It shows that rule-based term prefixing is vital for bridging the lexical-semantic gap in search engines.
  - **V9 → V10 (Legal Contract):** Formatting the narrative as rigid contract inclusions/exclusions slightly reduces flat BM25 metrics (0.4294), which is expected since it prioritizes longitudinal judge stability over word counts.
  - **V10 → V11 (Agentic Synthesis):** Spawning the parallel multi-temperature ensemble, Chain-of-Thought intent analysis, and critique loops in V11 successfully recovers lexical precision, yielding a highly balanced, robust retrieval output (nDCG@10 = **0.4435**, AP = **0.2865**, and RR = **0.6773**).
- **Total Accumulative Improvement (+37.13%):** Our final autonomous Agentic Pipeline (V11) achieves an nDCG@10 of **0.4435** over the baseline (0.3234), demonstrating strong generalization to unseen queries under a rigorous, leakage-free evaluation.

We note that reporting relative improvements over arithmetic means of retrieval metrics is not without debate: Fuhr [18] cautions against this practice on the grounds that such ratios can be misleading, while Sakai [19] argues that relative improvements remain informative provided they are interpreted with appropriate care rather than treated as an absolute rule. We therefore report both absolute and relative figures throughout this paper so that readers can judge the results under either perspective.

4.4. The Lexical vs. Semantic Trade-Off: Why V11 is the Winning Submission

A key question arises from the empirical comparison in Table 1: why do we declare the **V11 Agentic Pipeline** as our winning and primary submission if the **V9 Pipeline** achieves higher local lexical metrics across all measures (e.g., an nDCG@10 of 0.4659 vs. 0.4435)?

The answer lies in the fundamental difference between our **local lexical validation** and the **official TIRA evaluation framework**. Our local validation environment relies on an in-memory PyTerrier

index that executes a classic BM25 lexical match, which purely measures word-frequency overlap on the generated *description*. In this setup, V9 is heavily over-optimized: it uses unconstrained keyword expansions (Fix C) and simple query prepending to cram raw document terminology into the description. While highly effective for a flat lexical search engine, V9 generates descriptions and narratives that are structurally conversational, unstructured, and noisy.

In contrast, the official LongEval Task 2 challenge on the TIRA platform does not evaluate submissions using a standard keyword search engine. Instead, it deploys **Large Language Model (LLM) judges** (LLM-as-a-judge) to grade query-document pairs longitudinally across multiple temporal snapshots. The LLM judges evaluate topics on three distinct metrics:

1. **Alignment:** How well the generated description and narrative reflect the true underlying search intent.
2. **Distinguishability:** How effectively the narrative separates highly relevant documents from borderline or non-relevant ones.
3. **Clarity:** The absence of ambiguity or conversational fluff in the narrative.

Furthermore, because the collection spans multiple years, any narrative containing temporal jargon (e.g., *"recent breakthroughs"*, *"latest algorithms"*) will lead to high variance and misclassifications among LLM judges when evaluating documents from different time snapshots (temporal drift).

It is worth noting that this drift risk is asymmetric between documents and queries in our setting: since the corpus consists of already-published scientific abstracts, the content of a given document is essentially static once indexed, whereas the meaning a searcher attaches to a query term can still shift over time. For instance, a query for *"transformers"* is, today, overwhelmingly likely to be interpreted in the context of Transformer-based neural architectures rather than any of its earlier senses. Our Temporal Agnosticism Constraint is designed to guard the narrative against exactly this kind of query-side drift, rather than against document content changing.

While V9 is optimized for flat word matching, its conversational narrative fails to meet these strict LLM-as-a-judge criteria, leading to high variance and susceptibility to temporal drift. Conversely, V11 is designed precisely to excel in this semantic evaluation paradigm. It implements:

- A multi-stage agentic compliance loop that structures the narrative as a strict, unambiguous **legal contract** with explicit logical inclusions and exclusions (RELEVANT, BORDERLINE, NON −RELEVANT).
- An LLM-based critique stage that actively audits and removes all time-bound terms to guarantee **temporal agnosticism**.
- A parallel multi-temperature ensemble ($T = 0.15, 0.30, 0.45$) and meta-synthesis stage that resolves ambiguity and recovers crucial lexical terms.

By formatting narratives into highly structured contracts and removing temporal drift markers, V11 slightly constrains flat BM25 keyword matching locally (reducing nDCG@10 from 0.4659 to 0.4435). However, it represents our ultimate, robust submission because it is structurally engineered to maximize Alignment, Distinguishability, and Clarity for the LLM judges, guaranteeing optimal generalizability and stability under the actual longitudinal evaluation on TIRA.

5. Conclusion and Future Work

In this paper, we presented the systematic development of the VerbaNexAI framework for automatic TREC topic extraction in the CLEF 2026 LongEval Task 2. Our work focuses on the design of empirical grounding and rule-based post-processing heuristics to mitigate semantic dilution in generative query expansion. Specifically, programmatically prepending the original query to the generated description resolves exact-match vocabulary mismatch, representing a highly robust methodological solution.

We also introduced a self-contained, multi-stage Agentic Pipeline (V11) designed for deployment on the TIRA platform, leveraging PyTerrier-based Pseudo-Relevance Feedback and an LLM-based critique loop to achieve autonomous, zero-dependency, and zero-leakage topic generation.

By performing a rigorous Train/Test split evaluation, we demonstrated that our autonomous, leakage-free PRF Agent (V11) generalizes to unseen queries, achieving an nDCG@10 of 0.4435 (+37.1% over the baseline), validating the scientific integrity and robust retrieval capability of the proposed framework.

References

[1] J. J. Rocchio, Relevance feedback in information retrieval, in: The Smart Retrieval System—Experiments in Automatic Document Processing, Prentice-Hall, 1971, pp. 313–323.

[2] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389.

[3] E. M. Voorhees, The TREC robust retrieval track, SIGIR Forum 39 (2005) 11–20.

[4] R. Alkhalifa, I. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, G. Gonzalez-Saez, P. Galuščáková, L. Goeuriot, E. Kochkina, M. Liakata, D. Loureiro, H. T. Madabushi, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Longeval: Longitudinal evaluation of model performance at CLEF 2023, in: Proceedings of the 45th European Conference on Information Retrieval (ECIR 2023), Springer, 2023, pp. 499–505.

[5] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous integration for reproducible shared tasks with TIRA.io, in: Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[6] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, et al., The Llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[7] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using PyTerrier, in: Proceedings of the 2020 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR 2020), 2020, pp. 161–168.

[8] R. Nogueira, J. Lin, From doc2query to docTTTTTquery, arXiv preprint arXiv:1912.04631 (2019).

[9] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, M. Bendersky, Query expansion by prompting large language models, arXiv preprint arXiv:2305.03653 (2023).

[10] P. Thomas, S. Spielman, N. Craswell, B. Mitra, Large language models can accurately predict searcher preferences, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, ACM, 2024, pp. 1930–1940. doi:10.1145/3626772.3657707.

[11] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, H. Wachsmuth, Who determines what is relevant? humans or AI? why not both?, Communications of the ACM 67 (2024) 31–34. doi:10.1145/3624730.

[12] 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). To appear in Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2026).

[13] R. Takehi, E. M. Voorhees, T. Sakai, I. Soboroff, Llm-assisted relevance assessments: When should we ask llms for help?, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, ACM, 2025, pp. 95–105. doi:10.1145/3726302.3729916.

[14] I. Soboroff, Don't use llms to make relevance judgments, Information Retrieval Research 1 (2025) 29–46. doi:10.54195/irrj.19625.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F.-F. Li, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems (NeurIPS 2022) 35 (2022) 24824–24837.

[16] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified data wrangling

with ir_datasets, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021), 2021, pp. 2429–2436.

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

[18] N. Fuhr, Some common mistakes in IR evaluation, and how they can be avoided, SIGIR Forum 51 (2017) 32–41. doi:10.1145/3190580.3190586.

[19] T. Sakai, On fuhr's guideline for IR evaluation, SIGIR Forum 54 (2020) 12:1–12:8. doi:10.1145/3451964.3451976.

[20] 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.

[21] 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.