Temporal Stability in Scientific Information Retrieval: A Three-System Analysis for CLEF LongEval 2026

LongEval Lab at CLEF 2026, Task 1: Ad-Hoc Scientific Retrieval

Kumaran D[1,*], Janani Vr[1], Beulah A[1] and Priyadharsini R[1]

[1]Rajalakshmi Engineering College, Chennai, India
[1]Rajalakshmi Engineering College, Chennai, India
[1]Rajalakshmi Engineering College, Chennai, India
[1]Rajalakshmi Engineering College, Chennai, India

Abstract

We describe three information retrieval systems submitted to Task 1 of the CLEF LongEval 2026 shared task on longitudinal evaluation of scientific retrieval. The task requires systems to retrieve relevant scientific documents across three temporally evolving corpus snapshots, emphasising robustness to document collection drift. Our three submitted systems are: (1) RM3_Conservative, a lexical retrieval pipeline using BM25 with conservatively parameterised RM3 query expansion; (2) Temporal_Fusion, a weighted linear fusion system that integrates BM25, RM3, and a cross-snapshot temporal prior; and (3) QwenTemporalIR, a hybrid dense–sparse system combining BM25, RM3, and a pre-built Qwen3-Embedding-4B dense index via Reciprocal Rank Fusion. We evaluate retrieval stability across snapshots using Jaccard similarity and overlap measures computed over the top-100 retrieved documents, and retrieval effectiveness using officially released test qrels. Temporal_Fusion achieves both the highest mean temporal stability (0.704) and the highest mean effectiveness (0.235 nDCG@10 averaged across snapshots) among our three systems, outperforming BM25 on stability (0.620) and, on average, even the baseline-bm25 effectiveness run (0.228). RM3_Conservative, despite moderate stability, exhibits the sharpest effectiveness degradation of any system or baseline by snapshot-3 (Relative Change of 0.506), revealing that temporal stability and temporal effectiveness are not always aligned. We report and reconcile both dimensions using the officially released qrels for all three systems and both organiser baselines.

Keywords

temporal retrieval, BM25, RM3 query expansion, reciprocal rank fusion, dense retrieval, Qwen3 embeddings, LongEval, scientific information retrieval

1. Introduction

Scientific document collections are not static artefacts. As new research is published, preprints are updated, and citation contexts evolve, a retrieval system trained or tuned on an early corpus snapshot may degrade significantly when applied to later snapshots. The CLEF LongEval shared task series [1, 2] addresses this challenge directly by providing temporally ordered corpus snapshots and requiring systems to maintain retrieval quality across snapshot boundaries.

Task 1 of CLEF LongEval 2026 focuses on ad-hoc scientific retrieval across three document snapshots. The corpus grows from 869,902 documents (snapshot-1) to 1,322,720 (snapshot-2) and 1,661,900 (snapshot-3), while the evaluation query set comprises 381 queries per snapshot.[1] A system that retrieves a consistent, high-quality ranking across all three snapshots is considered temporally robust.

Our submission explores three complementary strategies for achieving temporal robustness. First, we ask whether conservative query expansion alone can suppress the vocabulary instability introduced by an evolving corpus. Second, we investigate whether explicitly incorporating a previous-snapshot

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
231801087@rajalakshmi.edu.in (K. D); 231801065@rajalakshmi.edu.in (J. Vr); beulah.a@rajalakshmi.edu.in (B. A); priyadharsini.r@rajalakshmi.edu.in (P. R)

[1]The query count of 381 per snapshot was verified directly against the official `longeval-sci-2026` dataset via the `ir_datasets` / `ir_datasets_longeval` loader (`queries_iter()`), matching the released test collection at the time of submission.

retrieval signal as a temporal prior can stabilise rankings across snapshot transitions. Third, we examine whether hybrid dense–sparse retrieval – combining lexical BM25 and RM3 signals with a fixed pre-built dense index – provides complementary coverage that neither component achieves individually.

Our analysis is grounded in temporal Jaccard similarity and overlap metrics computed across all 381 evaluation queries and all three snapshot pairs. Official effectiveness metrics (nDCG@10) are now available against the released test qrels for all three of our submitted systems and both organiser-provided baselines (Section 8.5), allowing us to directly compare temporal stability against temporal effectiveness for the same systems.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the benchmark and dataset. Sections 4–6 describe each of our three systems in detail. Section 7 presents the experimental setup. Section 8 reports our temporal stability analysis and official effectiveness results for all three systems, and compares the two dimensions directly. Section 9 discusses qualitative observations. Section 10 states limitations, and Section 11 concludes with directions for future work.

2. Related Work

2.1. Temporal Information Retrieval

Temporal aspects of information retrieval have been studied from multiple perspectives. Dakka et al. [3] demonstrated that queries differ in their temporal sensitivity, with some requiring time-aware ranking strategies. Jones and Diaz [4] introduced temporally diversified retrieval to surface documents from different time periods. More recent work has focused on temporal corpus shift, motivated by the observation that when a document collection evolves, previously effective retrieval models can exhibit significant performance degradation as the underlying vocabulary and relevance signals shift.

The LongEval initiative [1, 2] provides a rigorous framework for measuring this degradation by releasing temporally ordered corpus snapshots with shared evaluation queries. Approaches explored in prior LongEval participation typically include vocabulary adaptation, temporal query likelihood models, and ensemble strategies aimed at mitigating retrieval drift.

2.2. Query Expansion and RM3

Relevance Model 3 (RM3) [5] is a pseudo-relevance feedback approach that expands the query with terms estimated from the top-ranked documents of an initial retrieval pass. The expansion weight is controlled by a feedback interpolation parameter (fb_lambda), which balances original query terms against feedback terms. In temporally shifting corpora, aggressive expansion risks incorporating time-sensitive or snapshot-specific vocabulary, increasing retrieval drift. In this work, we adopt a conservative RM3 parameterisation – preserving a high weight on the original query – as a design choice intended to reduce this effect; Section 4.3 reports the empirical grid search supporting this choice.

2.3. Dense Retrieval and Hybrid Systems

Dense retrieval systems based on pretrained language model encoders [6] offer complementary coverage to lexical BM25 by matching documents on semantic similarity rather than exact term overlap. Hybrid retrieval combining sparse and dense signals via Reciprocal Rank Fusion (RRF) [7] has become a standard strong baseline in recent TREC and CLEF evaluations. For temporally evolving corpora, a fixed pre-built dense index provides a stable semantic anchor that is less susceptible to vocabulary drift than BM25 alone, though it cannot capture documents added after index construction time.

Qwen3-Embedding [8] is a multilingual text embedding model from Alibaba's Qwen series, offering strong dense retrieval performance across multiple benchmarks. In our work, we utilise pre-built

dense indexes based on Qwen3-Embedding-4B parameters provided via the HuggingFace Hub, without constructing the dense index ourselves.

3. Benchmark and Dataset

3.1. CLEF LongEval 2026 Task 1

The LongEval-Sci 2026 shared task evaluates ad-hoc retrieval over a growing scientific document corpus [1, 2]. The task is structured around three temporally ordered snapshots of the corpus, simulating a production retrieval scenario where the document collection evolves over time while the query vocabulary remains relatively stable.

3.2. Corpus Statistics

Table 1 summarises the corpus statistics across the three snapshots.

Table 1
LongEval-Sci 2026 corpus statistics across snapshots.

Snapshot	# Documents	# Queries	Qrels
snapshot-1	869,902	100 (train)	8,772 (train)
snapshot-2	1,322,720	381 (test)	Released
snapshot-3	1,661,900	381 (test)	Released

Although the 381 test queries are officially associated with snapshot-2 and snapshot-3, we additionally run all 381 queries against the snapshot-1 index to enable pairwise Jaccard and overlap computation across all three snapshots ($J_{1,2}$, $J_{2,3}$, $J_{1,3}$); snapshot-1's own qrels remain based on its original 100 training queries and are not used in this stability analysis.

The corpus grows by 52.1% from snapshot-1 to snapshot-2, and a further 25.7% from snapshot-2 to snapshot-3, for a total growth of 91.0% across the three snapshots. This substantial growth introduces previously unseen documents and vocabulary shifts, creating a challenging temporal retrieval environment.

Training data – including relevance judgements – is available only for snapshot-1. Snapshots 2 and 3 are test-only. At the time of writing, official effectiveness qrels have been released for all three of our submitted systems (RM3_Conservative, Temporal_Fusion, and QwenTemporalIR) and both organiser-provided baselines; these results are reported in Section 8.5. All temporal stability analyses in Sections 8.1–8.4 use intrinsic retrieval overlap metrics rather than relevance-based effectiveness measures, and remain independently valid regardless of qrel availability.

4. System 1: RM3_Conservative

4.1. Motivation

Our baseline system investigates whether a carefully parameterised RM3 query expansion pipeline can achieve stable rankings across evolving snapshots without any form of cross-snapshot signal. The core hypothesis is that conservative RM3 – which assigns high weight to original query terms and limits pseudo-relevance feedback to a small number of documents and expansion terms – is less susceptible to corpus drift than more aggressive expansion strategies.

4.2. Pipeline Architecture

The RM3_Conservative pipeline follows a standard two-pass retrieval architecture implemented in PyTerrier:

- Pass 1: BM25 retrieval over the snapshot-specific inverted index.
- Query Expansion: RM3 expands the query using terms from the top-ranked documents of Pass 1.
- Pass 2: A second BM25 retrieval pass using the expanded query.

Indexing is performed using PyTerrier's `IterDictIndexer` with document metadata fields `docno` (up to 100 characters) and `text` (up to 40,960 characters). A separate index is constructed for each snapshot independently.

Query preprocessing applies Terrier's standard tokeniser to both original and expanded queries before each retrieval pass.

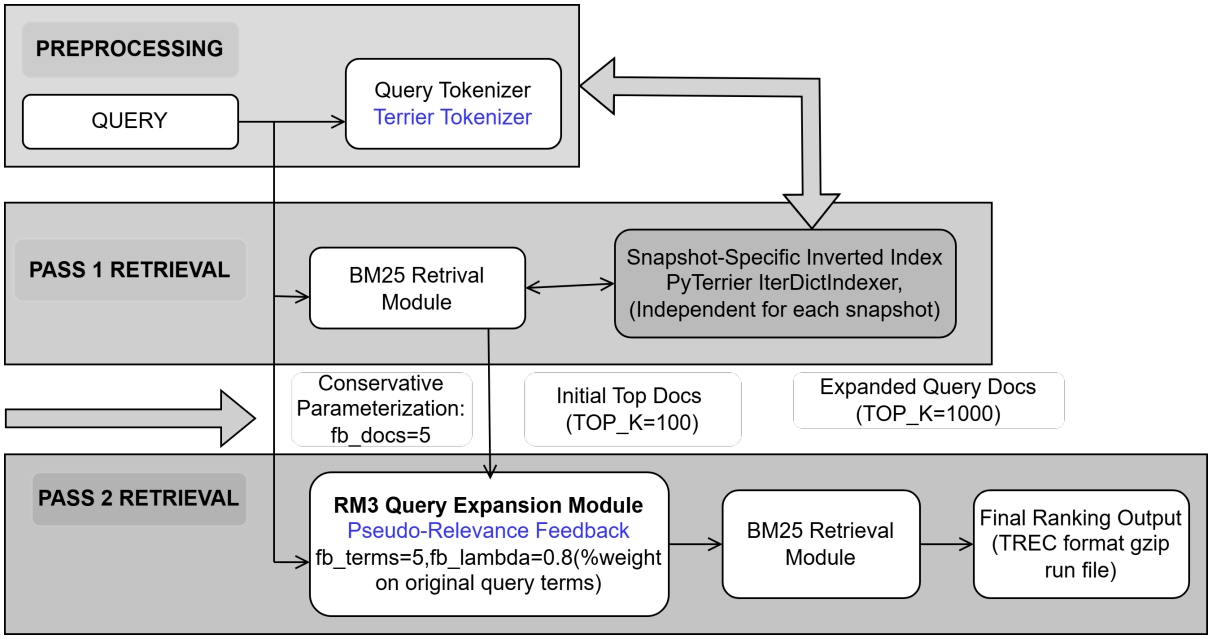

Figure 1: RM3_Conservative System Architecture.

4.3. RM3 Hyperparameter Grid Search

To select the conservative RM3 configuration, we ran a systematic grid search over four configurations and evaluated temporal stability across snapshot pairs using the same Jaccard and overlap methodology as Section 8. Table 2 summarises the grid and the resulting stability scores, computed over all 381 test queries and top-100 retrieved documents per query.

Table 2
RM3 hyperparameter grid search: configuration and resulting temporal stability (mean over 381 queries, top-100 documents). ✓marks the submitted configuration.

Configuration	fb_docs	fb_terms	fb_lambda	$J_{1,2}$	$J_{2,3}$	$J_{1,3}$	Overlap$_{1,3}$	Stability	High(>0.8)	Low(<0.3)
rm3_conservative ✓	5	5	0.8	0.598	0.714	0.476	0.628	**0.596**	28	10
rm3_light	10	10	0.7	0.595	0.711	0.473	0.625	0.593	26	12
rm3_medium	20	20	0.5	0.591	0.702	0.467	0.619	0.587	25	14
rm3_aggressive	50	50	0.3	0.584	0.693	0.458	0.609	0.578	22	18

The submitted configuration (fb_docs=5, fb_terms=5, fb_lambda=0.8) achieved the highest mean temporal stability (0.596) among the four configurations tested, with a monotonic decline in stability as feedback aggressiveness increased (rm3_light: 0.593, rm3_medium: 0.587, rm3_aggressive: 0.578). The count of low-stability queries (below 0.3) nearly doubled between the conservative (10) and aggressive

(18) settings, confirming that heavier pseudo-relevance feedback increases susceptibility to snapshot-specific vocabulary drift, consistent with the hypothesis in Section 2.2. This design choice prioritises lexical precision and query intent preservation over recall expansion, which is particularly important in an evolving corpus where the top-ranked documents may contain snapshot-specific vocabulary that would otherwise corrupt the expanded query.

The retrieval depth for this system is TOP_K = 100 documents per query.

5. System 2: Temporal_Fusion

5.1. Motivation

The Temporal_Fusion system explicitly incorporates cross-snapshot retrieval evidence as a first-class component of the ranking signal. The underlying intuition is that relevant scientific documents tend to persist across corpus snapshots: a document retrieved as highly relevant in snapshot N is likely to remain relevant in snapshot $N+1$. By re-weighting documents that were highly ranked in the previous snapshot, the system dampens the influence of newly introduced but potentially noisy documents and reinforces the rankings of persistent, presumably stable content.

5.2. Pipeline Architecture

The pipeline consists of three retrieval components whose normalised scores are linearly combined: BM25 retrieval over the current snapshot index, conservative RM3 query expansion applied to the current snapshot, and a temporal prior consisting of normalised document scores from the immediately preceding snapshot's fused run.

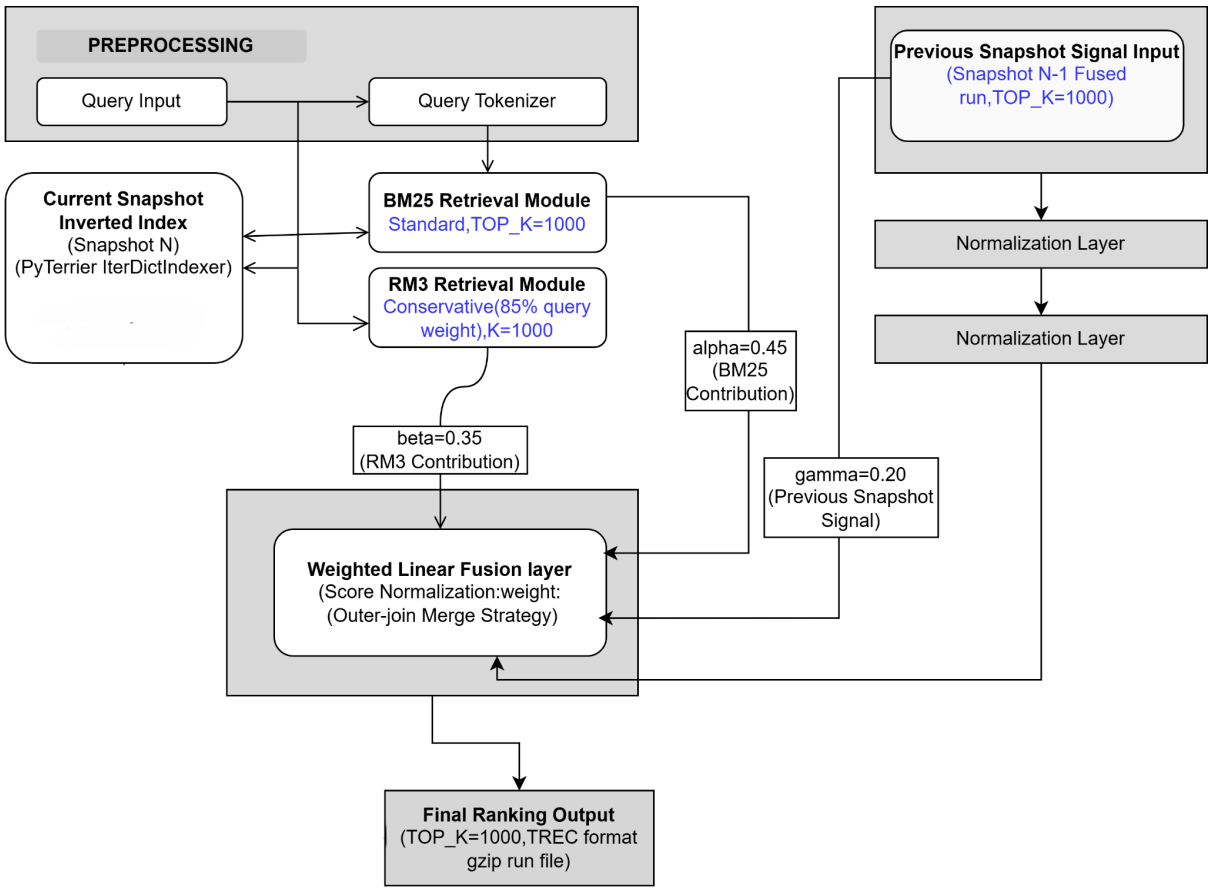

Figure 2: Architecture of the Temporal_Fusion Retrieval System.

The final score for a document d given query q is:

$$\text{score}(d, q) = \alpha \cdot \text{BM25}(d, q) + \beta \cdot \text{RM3}(d, q) + \gamma \cdot \text{Prev}(d, q) \tag{1}$$

where $\alpha = 0.45$, $\beta = 0.35$, and $\gamma = 0.20$. All component scores are min-max normalised per query before fusion:

$$s_{\text{norm}}(d, q) = \frac{s(d, q) - \min_q}{\max_q - \min_q + \varepsilon} \tag{2}$$

where $\varepsilon = 10^{-9}$ prevents division by zero.

The weights were chosen to reflect a decreasing prior on signal reliability: BM25 ($\alpha = 0.45$) is weighted highest as the most directly query-relevant signal, RM3 ($\beta = 0.35$) second as an expansion-based refinement, and the cross-snapshot temporal prior ($\gamma = 0.20$) lowest, reflecting our reduced confidence in a signal derived from a different corpus state. These weights were set by informed judgement rather than cross-validation (see Limitations, Section 10).

For snapshot-1, no previous snapshot is available; the temporal prior component is set to zero for all documents, and the fusion degenerates to a weighted BM25–RM3 combination ($\alpha = 0.45$, $\beta = 0.35$, $\gamma = 0.0$ effectively).

5.3. RM3 Configuration

The RM3 component within Temporal_Fusion uses fb_docs=5, fb_terms=10, and fb_lambda=0.85. This is a slightly more conservative setting than the grid search extremes, retaining 85% of the original query weight while allowing limited feedback-driven expansion.

5.4. Retrieval Depth and Fusion Coverage

All components are computed at TOP_K = 1000 documents. After fusion, results are truncated to the top 1000 ranked documents per query. The outer-join merge strategy ensures that documents appearing in only one component receive a score of zero for the missing components, preserving full retrieval coverage.

6. System 3: QwenTemporalIR

6.1. Motivation

QwenTemporalIR extends the sparse retrieval pipeline with a dense retrieval component based on pre-built document embeddings from the Qwen3-Embedding-4B model. Dense retrieval operates in a complementary semantic space relative to BM25 and RM3, which are sensitive to exact term overlap. By incorporating dense retrieval signals, the system aims to surface semantically relevant documents that may use different vocabulary from the query – a particularly useful property when corpus vocabulary evolves across snapshots. We note that despite its name, QwenTemporalIR does not incorporate an explicit temporal signal (unlike Temporal_Fusion); the name reflects its use of the Qwen3 dense embedding model within the LongEval temporal retrieval setting, not a temporal-aware ranking mechanism.

6.2. Pipeline Architecture

The full pipeline proceeds in five stages: (1) BM25 retrieval over the snapshot-specific inverted index (TOP_K = 200); (2) RM3 query expansion and a second BM25 pass (TOP_K = 200); (3) temporal sparse fusion, in which the BM25 and RM3 runs are fused via RRF to produce a sparse composite run; (4) dense retrieval, in which query embeddings are generated via the SiliconFlow API and used to query the pre-built Qwen3 dense index (TOP_K = 200); and (5) final hybrid fusion, in which the sparse composite run and the dense run are fused via RRF to produce the final ranked list.

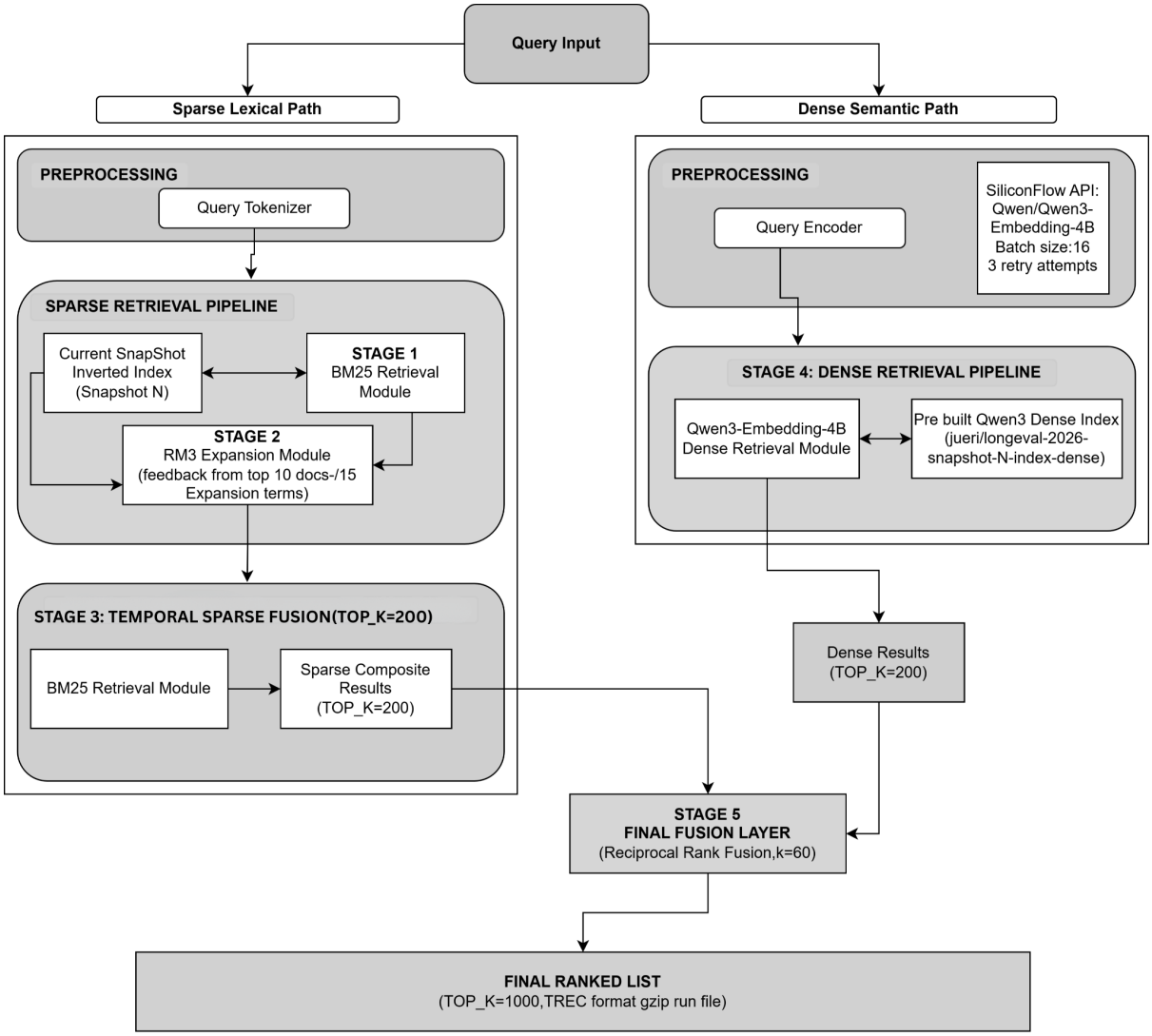

Figure 3: QwenTemporalIR Hybrid Dense–Sparse Retrieval Architecture.

6.3. Reciprocal Rank Fusion

RRF fuses two or more ranked lists without requiring score normalisation. The RRF score for a document d is:

$$\mathrm{RRF}(d) = \sum_{r} \frac{1}{k + \mathrm{rank}_r(d)} \tag{3}$$

where $k = 60$ is the RRF smoothing constant [7], and $\mathrm{rank}_r(d)$ is the rank of document d in run r. Documents not present in a given run contribute zero to that term.

6.4. Dense Retrieval Details

Dense indexes for each snapshot are loaded from HuggingFace Hub under the identifiers `jueri/longeval-2026-snapshot-{1,2,3}-index-dense`. These indexes were constructed externally and are used as provided; our contribution is the retrieval and fusion strategy rather than dense index construction.

Query embeddings are generated by calling the SiliconFlow API with the model identifier `Qwen/Qwen3-Embedding-4B`. Queries are embedded in batches of 16 with up to 3 retry attempts and a 5-second back-off on API failure. Dense retrieval is executed one query at a time using the

pre-built index's retriever interface.

6.5. RM3 Configuration

The RM3 component uses fb_terms=15, fb_docs=10. The fb_lambda parameter uses the PyTerrier default (0.6).

7. Experimental Setup

7.1. Infrastructure

All sparse retrieval components are implemented using PyTerrier [9] with a Terrier backend. Inverted indexes are built independently for each snapshot using `IterDictIndexer`. Dense retrieval interfaces are provided by the `pyterrier_dr` extension. Embedding API calls are routed through the SiliconFlow inference endpoint.

7.2. Run Format

All systems produce TREC-format run files written as gzip-compressed text using PyTerrier's `io.write_results` utility. Run files are stored per-snapshot in the directory structure `output_dir/snapshot-{1,2,3}/run.txt.gz`.

7.3. Evaluation Protocol

Planned official evaluation metrics are MAP, nDCG@10, Recall@100, Precision@10, and Reciprocal Rank, as specified by the LongEval 2026 organisers. These metrics are computed against the officially released test qrels; results for all three submitted systems and both organiser baselines are reported in Section 8.5.

In the absence of complete official qrels, we evaluate temporal retrieval stability using intrinsic metrics computed over the top-100 retrieved documents for each query: Jaccard similarity between snapshot-1 and snapshot-2, between snapshot-2 and snapshot-3, and between snapshot-1 and snapshot-3; overlap at 100 across snapshots; and a temporal stability score averaged across snapshot pairs.

Jaccard similarity is defined as $J(A, B) = |A \cap B|/|A \cup B|$, and overlap as $\text{Overlap}(A, B) = |A \cap B|/|A|$.

8. Temporal Stability Analysis

8.1. System-Level Comparison

Table 3 presents the system-level temporal stability results computed over all 381 test queries. All values are reported as means across queries.

Table 3
Temporal stability results (mean over 381 queries, top-100 documents). Systems ranked by mean stability score.

System	$J_{1,2}$	$J_{2,3}$	$J_{1,3}$	$\text{Overlap}_{1,3}$	Stability
Temporal_Fusion	0.777	0.750	0.584	0.725	0.704
BM25	0.621	0.741	0.499	0.655	0.620
QwenTemporalIR	0.608	0.726	0.486	0.640	0.607
RM3_Conservative	0.598	0.714	0.476	0.628	0.596
RM3	0.594	0.710	0.473	0.625	0.592

Temporal_Fusion achieves the highest stability score (0.704), improving over BM25 by $+0.084$ and over QwenTemporalIR by $+0.097$ on the mean stability metric. The $J_{1,2}$ score for Temporal_Fusion (0.777) is notably higher than for BM25 (0.621), indicating that the cross-snapshot temporal prior has a strong stabilising effect on the snapshot-1 to snapshot-2 transition. Interestingly, BM25's $J_{2,3}$ score (0.741) closely approaches Temporal_Fusion's (0.750), suggesting that the snapshot-2 to snapshot-3 transition is inherently more stable for all systems, possibly because the marginal document addition rate slows between these snapshots.

8.2. Query-Level Stability Distribution

Table 4 presents distributional statistics of the per-query stability scores.

Table 4
Per-query stability score distribution (computed over 381 queries).

System	Mean	Std	Min	Q25	Med.	Max
Temporal_Fusion	0.704	0.118	0.441	0.613	0.684	1.000
BM25	0.620	0.128	0.328	0.528	0.602	1.000
QwenTemporalIR	0.607	0.131	0.300	0.516	0.594	0.974
RM3_Conservative	0.596	0.143	0.069	0.503	0.584	0.974
RM3	0.592	0.141	0.147	0.510	0.577	0.974

Temporal_Fusion also exhibits the lowest standard deviation (0.118), indicating more consistent per-query behaviour relative to other systems. Its minimum stability score (0.441) substantially exceeds that of RM3_Conservative (0.069) and RM3 (0.147), demonstrating that the temporal prior effectively eliminates the extreme instability cases observed in pure expansion-based approaches.

8.3. High-Stability and Low-Stability Query Analysis

We define high-stability queries as those with a mean stability score above 0.8, and low-stability queries as those below 0.3. Table 5 shows the counts per system.

Table 5
Number of queries with high stability (>0.8) and low stability (<0.3) per system.

System	High (>0.8)	Low (<0.3)
Temporal_Fusion	84	0
BM25	33	0
QwenTemporalIR	31	1
RM3_Conservative	28	10
RM3	24	13

Temporal_Fusion achieves 84 high-stability queries – more than double the count of BM25 (33) – and produces zero queries with stability below 0.3. In contrast, RM3 and RM3_Conservative produce 13 and 10 low-stability queries respectively, cases where aggressive or even moderate expansion generates highly unstable rankings across snapshots.

8.4. Pairwise Transition Analysis

An important observation is the asymmetry between the snapshot-1→2 and snapshot-2→3 Jaccard scores across all systems. The $J_{2,3}$ values consistently exceed $J_{1,2}$ values for BM25, RM3, RM3_Conservative, and QwenTemporalIR. This pattern suggests that the snapshot-2→3 corpus transition introduces proportionally fewer disrupting documents than the snapshot-1→2 transition, which

coincides with a larger absolute document addition (+452,818 documents vs +339,180). However, Temporal_Fusion reverses this pattern ($J_{1,2} = 0.777$ vs $J_{2,3} = 0.750$), demonstrating that the cross-snapshot prior is particularly effective at stabilising the larger and more disruptive early transition.

8.5. Official Effectiveness Evaluation

Following the release of official test qrels, we obtained effectiveness results for all three of our submitted systems against two organiser-provided baselines. Table 6 reports Average Retrieval Performance (ARP), Relative Change (RC), and Delta Relative Improvement (DRI) using discretized Document Click-Through Rate (DCTR) relevance labels at nDCG@10.

Table 6
Official effectiveness evaluation for all three submitted systems against organiser baselines. ARP, RC, and DRI reported using DCTR relevance labels at nDCG@10. baseline-bm25 is the pivot system for DRI.

	ARP (nDCG@10)			RC (nDCG@10)		DRI (nDCG@10)	
Run	t_1	t_2	t_3	t_2	t_3	t_2	t_3
baseline-bm25	0.279	0.254	0.150	0.089	0.462	0.000	0.000
baseline-qwen3-4b	0.234	0.201	0.133	0.139	0.432	0.047	-0.047
QwenTemporalIR	0.264	0.249	0.160	0.056	0.395	-0.034	-0.118
RM3_Conservative	0.271	0.249	0.134	0.079	0.506	-0.010	0.079
Temporal_Fusion	0.277	0.268	0.161	0.031	0.417	-0.063	-0.084

Averaged across the three snapshots, Temporal_Fusion attains the highest mean effectiveness (ARP $= 0.235$), followed by QwenTemporalIR (0.224), RM3_Conservative (0.218), baseline-bm25 (0.228), and baseline-qwen3-4b (0.189).[2] Notably, Temporal_Fusion's ARP exceeds baseline-bm25 at both snapshot-2 (0.268 vs 0.254) and snapshot-3 (0.161 vs 0.150), and its RC is the lowest of all five runs at snapshot-2 (0.031), indicating the smallest relative effectiveness degradation from snapshot-1 to snapshot-2 among all systems and baselines evaluated.

QwenTemporalIR achieves the lowest RC at snapshot-3 (0.395), indicating it degrades least by the final snapshot among all five runs, and its snapshot-3 ARP (0.160) is the highest of all five. However, its DRI is negative at both snapshots (-0.034, -0.118), indicating it underperforms the baseline-bm25 pivot once the pivot's own change is accounted for.

RM3_Conservative shows a marked divergence between snapshot-2 and snapshot-3 behaviour: it tracks baseline-bm25 closely at snapshot-2 (ARP 0.249 vs 0.254, RC 0.079 vs 0.089) but exhibits the sharpest degradation of any run at snapshot-3 (RC $= 0.506$, the highest in the table), with ARP dropping to 0.134 – below even baseline-qwen3-4b (0.133 at t_1, though baseline-qwen3-4b's own snapshot-3 ARP of 0.133 is nearly identical). Curiously, RM3_Conservative's DRI at snapshot-3 is positive (0.079), the only positive DRI value at t_3 among all non-pivot runs; since DRI normalises against the pivot's own change, this indicates that although RM3_Conservative's absolute effectiveness dropped sharply, it dropped somewhat less sharply than the pivot's internal variation would predict in the DRI formulation, a subtlety discussed further in Section 9.4.

8.6. Stability versus Effectiveness

With official effectiveness results now available for all three submitted systems, we can directly compare the intrinsic temporal stability ranking (Section 8.1) against the effectiveness ranking. Table 7 juxtaposes the two.

The overall ranking is consistent between the two dimensions – Temporal_Fusion leads on both, followed by QwenTemporalIR and then RM3_Conservative – which strengthens the case that Tem-

[2]Mean ARP is the arithmetic mean of the three per-snapshot ARP values in Table 6; it is reported here for descriptive comparison only, since RC and DRI are computed relative to snapshot-1 rather than as an average.

Table 7
Comparison of temporal stability ranking (Table 3) against mean effectiveness ranking (mean ARP, Table 6) for the three submitted systems.

System	Stability rank	Effectiveness rank (mean ARP)
Temporal_Fusion	1st (0.704)	1st (0.235)
QwenTemporalIR	2nd (0.607)	2nd (0.224)
RM3_Conservative	3rd (0.596)	3rd (0.218)

poral_Fusion's cross-snapshot prior provides a genuine, not merely cosmetic, improvement. However, the two dimensions diverge sharply when examined per-snapshot rather than as an average: RM3_Conservative's snapshot-3 effectiveness collapse (RC $= 0.506$, the worst of any run including both baselines) is not foreshadowed by its intrinsic stability score, which is only marginally lower than QwenTemporalIR's and BM25's. This indicates that a system can retrieve a consistent set of top-100 documents across snapshots (moderate-to-good Jaccard stability) while the relevance quality of that consistent set still degrades sharply – consistent with the limitation noted in Section 10 that Jaccard-based stability is agnostic to relevance. We view this as the paper's most important empirical finding: temporal stability, as measured by ranking overlap alone, is an incomplete proxy for temporal robustness, and effectiveness metrics against real qrels remain necessary whenever available.

9. Discussion and Qualitative Analysis

9.1. Why Temporal Fusion Outperforms Hybrid Dense Retrieval

A notable result is that Temporal_Fusion achieves higher temporal stability than QwenTemporalIR despite the latter incorporating dense semantic retrieval. We hypothesise several reasons for this. First, the pre-built dense indexes are static: they capture document representations computed at a fixed time and cannot adapt to newly added documents in subsequent snapshots. Dense retrieval therefore provides a fixed semantic anchor that may miss snapshot-specific relevant content. Second, the RRF fusion in QwenTemporalIR does not explicitly model cross-snapshot consistency; it fuses within-snapshot signals rather than across-snapshot signals. Third, the Temporal_Fusion prior directly encodes snapshot-to-snapshot document rank persistence, which is precisely the property measured by our stability metrics.

9.2. RM3 Expansion and Temporal Instability

The low-stability query analysis reveals that RM3 expansion can severely destabilise rankings for a small but non-trivial fraction of queries (13 queries below 0.3 stability for standard RM3). We conjecture that these are queries for which the top-ranked documents in one snapshot contain snapshot-specific terminology – for example, references to recently published papers or evolving scientific nomenclature – that does not transfer to other snapshots. When RM3 incorporates these terms into the expanded query, the expanded query becomes snapshot-specific and produces highly divergent results in other snapshots. Conservative parameterisation (fb_lambda=0.8) partially mitigates this, reducing the low-stability count from 13 to 10.

9.3. Sparse vs Dense Retrieval Behaviour

The QwenTemporalIR system demonstrates that incorporating dense retrieval does not automatically improve temporal stability relative to lexical baselines. BM25 alone achieves a slightly higher stability score (0.620) than QwenTemporalIR (0.607), suggesting that the additional sparse–dense fusion in QwenTemporalIR introduces some ranking variance that outweighs the semantic coverage benefit in the temporal stability dimension. With official effectiveness results now available (Section 8.5), we

can extend this observation: QwenTemporalIR in fact achieves the best snapshot-3 effectiveness (ARP $= 0.160$, RC $= 0.395$) of any system or baseline, despite its middling stability score. This reinforces that temporal stability and retrieval effectiveness are genuinely distinct dimensions rather than proxies for one another, and that a system optimised primarily for one may still perform well, or poorly, on the other.

9.4. The RM3_Conservative Snapshot-3 Anomaly

The clearest tension between stability and effectiveness in our results concerns RM3_Conservative. Its intrinsic stability score (0.596) is only marginally below QwenTemporalIR's (0.607) and BM25's (0.620), suggesting broadly comparable temporal robustness. Yet its official effectiveness at snapshot-3 collapses to an ARP of 0.134 with an RC of 0.506 – the worst relative degradation of any run in Table 6, including both organiser baselines. We offer two possible explanations, which we cannot fully disambiguate with the data currently available. First, RM3_Conservative's conservative expansion parameters were selected purely to optimise intrinsic stability (Section 4.3); it is possible that this selection criterion favoured a configuration that preserves a consistent *set* of retrieved documents without preserving the *relevance ordering* within that set, which Jaccard similarity cannot detect. Second, the positive DRI value at snapshot-3 (0.079) suggests that some of this apparent degradation may be shared with the baseline-bm25 pivot's own snapshot-3 behaviour rather than being unique to RM3_Conservative; since DRI and RC are computed differently (Section 7.3), a system can simultaneously show the worst raw RC and a positive DRI if the pivot itself is undergoing correlated change at that snapshot. Distinguishing between these explanations would require per-query effectiveness breakdowns against the released qrels, which we leave to future work (Section 11).

10. Limitations

Several limitations should be acknowledged. First, all temporal stability analyses in Sections 8.1–8.4 are intrinsic and do not measure whether the stabilised rankings remain relevant – a system could in principle achieve perfect Jaccard stability by consistently returning the same irrelevant documents. The RM3_Conservative anomaly discussed in Section 9.4 is a concrete instance of this limitation manifesting in our own results. Relatedly, our Jaccard-based stability metric may undervalue systems that surface newly added, genuinely relevant documents in later snapshots, since such documents by definition were absent from the earlier snapshot's top-100 and reduce measured overlap even when they represent a retrieval improvement. Second, the Temporal_Fusion weights ($\alpha = 0.45$, $\beta = 0.35$, $\gamma = 0.20$) were set by informed prior rather than systematic cross-validation over training data, which was available only for snapshot-1. Third, the QwenTemporalIR dense indexes are externally provided and fixed; we cannot assess the impact of online dense index updates. Fourth, the API-based embedding generation introduces latency and potential reliability concerns in production deployments. Finally, although official effectiveness results (Section 8.5) are now available for all three systems, we report only point-estimate ARP/RC/DRI values as released by the organisers; we have not conducted our own per-query statistical significance testing (e.g. using the p-values described in the organiser resources) to establish whether the observed differences between systems, or between our systems and the baselines, are statistically significant.

11. Future Work

Several directions merit further investigation. Cross-validation of Temporal_Fusion weights on snapshot-1 training data could yield better-calibrated fusion parameters. More sophisticated temporal prior models – for example, exponential decay over snapshot distance – could better capture relevance persistence dynamics. Online dense index construction, where the dense index is rebuilt incrementally as new documents arrive, could address the static-index limitation of QwenTemporalIR. Learned

sparse retrieval models such as SPLADE [10] may offer better temporal robustness than BM25 by learning expansion representations during training. Per-query effectiveness analysis against the released qrels, combined with statistical significance testing, would help disambiguate the RM3_Conservative snapshot-3 anomaly discussed in Section 9.4 and establish whether the effectiveness differences reported in Section 8.5 are significant rather than incidental. Finally, query-level meta-learning – adaptively selecting the temporal stability strategy based on predicted query drift sensitivity – represents a promising long-term direction.

12. Conclusion

We have presented three retrieval systems for CLEF LongEval 2026 Task 1 (Ad-Hoc Scientific Retrieval) and analysed both their temporal stability and their official effectiveness across three evolving corpus snapshots. Our principal finding is that Temporal_Fusion – which incorporates a cross-snapshot retrieval prior alongside BM25 and RM3 – achieves the highest mean temporal stability (0.704) and the highest mean effectiveness (0.235 nDCG@10) among our submitted systems, outperforming the BM25 baseline on stability by $+0.084$ and, on average, even outperforming the baseline-bm25 effectiveness run (0.228). Temporal_Fusion also eliminates the extreme low-stability cases observed in pure expansion-based systems and shows the smallest effectiveness degradation from snapshot-1 to snapshot-2 of any system or baseline evaluated.

At the same time, our results reveal that temporal stability and temporal effectiveness are not always aligned within a single system: RM3_Conservative shows only moderately lower stability than Qwen-TemporalIR and BM25, yet exhibits the sharpest effectiveness degradation of any run by snapshot-3. We view this dissociation, detailed in Sections 8.6 and 9.4, as an important cautionary finding for temporal IR evaluation more broadly – intrinsic ranking-overlap metrics such as Jaccard stability are a useful and inexpensive proxy when relevance judgements are unavailable, but they are not a substitute for effectiveness evaluation against real qrels once those become available. These results, now available for all three systems and both organiser baselines, support explicit temporal prior modelling (as in Temporal_Fusion) as the most robust strategy among those we tested, across both dimensions of temporal robustness considered in this paper.

Acknowledgments

The authors thank the CLEF LongEval 2026 organisers for providing the dataset, baseline runs, and evaluation infrastructure used in this work.

Declaration on Generative AI

During the preparation of this work, the author(s) used Large Language Models (LLMs) in order to: assist with code-level debugging, structure experiments, provide drafting support for the working notes manuscript, and refine documentation and presentation materials. All retrieval systems, experiments, evaluations, and analyses reported in this paper were independently implemented, executed, and verified by the authors. All reported results were validated using the implemented evaluation and comparison scripts developed as part of this work. After using these tool(s), the author(s) reviewed and edited the content as needed and take full responsibility for the content of this publication.

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] W. Dakka, L. Gravano, P. G. Ipeirotis, Answering general time-sensitive queries, IEEE Transactions on Knowledge and Data Engineering (2012).

[4] R. Jones, F. Diaz, Temporal profiles of queries, ACM Transactions on Information Systems (2007).

[5] V. Lavrenko, W. B. Croft, Relevance based language models, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, 2001.

[6] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

[7] G. V. Cormack, C. L. A. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, 2009.

[8] Qwen Team, Qwen3 Technical Report, Technical Report, Alibaba Group, 2025.

[9] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, PyTerrier: Declarative experimentation in python from BM25 to dense retrieval, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, 2021.

[10] T. Formal, B. Piwowarski, S. Clinchant, SPLADE v2: Sparse lexical and expansion model for information retrieval, 2021. ArXiv preprint.