applied sciences

MDPI

Article

A Multi-Stage Hybrid Retrieval Framework for the Scientific Literature with Cross-Encoder Re-Ranking

Walaa Al-Joofi, Alaa Sagheer * and Hala Hamdoun

Department of Computer Science, College of Computer Sciences and Information Technology, King Faisal University, Hufof 36362, Al Ahsa, Saudi Arabia; hhamdoun@kfu.edu.sa (H.H.)
* Correspondence: asagheer@kfu.edu.sa

Abstract

Effective scientific literature retrieval requires moving beyond surface-level term matching toward structured semantic reasoning. This paper presents a controlled empirical study of multi-stage retrieval for scientific literature, integrating lexical matching, dense semantic modeling, hybrid fusion, and cross-encoder re-ranking within a unified evaluation framework. The study is designed to analyze the interactions, trade-offs, and failure modes of these components in claim-based scientific search. Experiments on the SciFact benchmark demonstrate that dense models capture semantic similarity but remain insufficient when used in isolation. Hybrid fusion broadens the candidate pool but does not consistently outperform the best standalone dense retriever, as RRF-based fusion can dilute strong dense rankings when lexical and semantic signals diverge. Cross-encoder re-ranking proves to be the primary driver of final performance gains, with the best configuration, Hybrid (SciNCL + BM25) + Cross-Encoder, reaching NDCG@10 of 0.523, MAP@10 of 0.479, Recall@10 of 0.642, and MRR@10 of 0.497. Ablation analysis shows that lexical pseudo-relevance feedback (RM3) introduces query drift in claim-focused retrieval, and that passage-level max pooling weakens effectiveness by fragmenting document-level evidence. Cross-domain evaluation on SciFact, PubMedQA, and SciDocs demonstrates that the relative ranking of retrieval paradigms remains stable across datasets with varying difficulty levels, while also revealing that the RRF dilution effect intensifies on harder retrieval tasks. These findings suggest that effective scientific retrieval benefits from integrated multi-stage pipelines, and that understanding component-level interactions is essential for designing robust retrieval systems.

Keywords: scientific information retrieval; multi-stage retrieval; hybrid fusion; cross-encoder re-ranking; dense neural retrieval; SciFact; PubMedQA; SciDocs

Check for updates

Academic Editor: George Drosatos

Received: 5 March 2026
Revised: 29 April 2026
Accepted: 7 May 2026
Published: 12 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid growth of digital information has made effective document retrieval a central challenge for modern search systems. In scientific domains, the accelerating volume of publications further strains traditional information retrieval (IR) methods, which rely primarily on exact keyword matching and struggle to scale across specialized and multidisciplinary content [1]. This challenge has motivated the development of retrieval approaches that are more effective, efficient, and sustainable.

Recent advances in natural language processing and machine learning have enabled semantic retrieval systems that move beyond surface-level term matching toward deeper contextual understanding [2]. Progress in generative AI has further strengthened semantic

representations by improving contextual reasoning and language understanding. Despite these advances, academic research libraries continue to face persistent challenges related to domain complexity, implicit scientific claims, and cross-disciplinary terminology, all of which require more robust retrieval strategies [3].

Although semantic retrieval methods have advanced considerably in recent years, effectively retrieving relevant scientific documents for claim-based queries remains a challenging task. Existing approaches often struggle to balance lexical precision with semantic understanding, particularly when queries are short, domain-specific, and require accurate evidence-level matching. This challenge is especially pronounced in datasets such as SciFact [4], where queries are formulated as scientific claims that demand precise semantic alignment with supporting evidence distributed across document abstracts.

While large-scale benchmarking efforts such as BEIR [5] have evaluated retrieval models across multiple datasets, including SciFact, these studies focus primarily on aggregate performance comparisons and do not systematically analyze the interaction between retrieval paradigms within multi-stage pipelines. In particular, the trade-offs introduced by hybrid fusion, the impact of design choices such as query expansion and passage-level aggregation, and the failure modes specific to claim-based scientific retrieval remain underexplored. This limits the ability to design efficient, interpretable, and robust retrieval systems tailored to the scientific literature.

To address these gaps, this paper presents a controlled empirical study of multi-stage retrieval for scientific literature, integrating lexical retrieval, dense semantic matching, hybrid fusion, and cross-encoder re-ranking within a unified evaluation framework. Beyond performance comparison, the study emphasizes component-level ablation, interpretability analysis, and structured error diagnosis to provide actionable insights into retrieval behavior and limitations in claim-based scientific search.

Accordingly, this paper is guided by the following research questions:

1. How does SciBERT-based dense retrieval compare with other dense retrieval models, such as SPECTER and SciNCL, on the SciFact dataset?
2. To what extent does hybrid fusion using RRF improve recall and ranking quality compared to standalone lexical and dense retrieval?
3. How much additional performance gain is achieved through cross-encoder re-ranking beyond dense and hybrid retrieval pipelines?
4. What insights do explainability and error analysis provide regarding the limitations and interpretability of semantic retrieval models in the scientific literature?

To answer these questions, the paper pursues the following objectives:

1. To conduct a controlled empirical comparison of lexical, dense, and hybrid retrieval paradigms, including BM25, SciBERT, SPECTER, SciNCL, and Reciprocal Rank Fusion (RRF), on the SciFact dataset [4].
2. To investigate the impact of key pipeline design choices, including cross-encoder re-ranking, query expansion strategies, and passage-level aggregation, on retrieval effectiveness and ranking precision.
3. To analyze retrieval behavior and failure modes through interpretability techniques, including LIME and attention-based visualizations, and to construct a structured error taxonomy for claim-based scientific retrieval.
4. To provide task-specific insights into claim-based scientific retrieval, with a focus on trade-offs, complementarity, and evidence aggregation challenges that extend beyond the aggregate metrics reported in prior benchmark studies.

The remainder of this paper is organized as follows. Section 2 presents the background concepts, foundational retrieval models, and a review of related work. Section 3 describes

the novelty and contribution of this work. Section 4 describes the proposed methodology, including system architecture, model selection, and evaluation setup. Section 5 reports the experimental results, ablation studies, and error analysis. Section 6 discusses the findings in the context of prior work and their practical implications. Section 7 outlines the limitations of the proposed system and directions for future research. Finally, Section 8 concludes the paper.

2. Background and Related Work

This section provides an overview of the retrieval paradigms relevant to this paper, covering lexical retrieval, dense neural retrieval, hybrid fusion, and cross-encoder re-ranking, along with related work in each area.

2.1. Lexical Retrieval

Information retrieval (IR) systems have traditionally relied on lexical matching, where relevance is estimated from term frequency and keyword overlap between queries and documents [6]. BM25 [7], a probabilistic model from the Okapi family, remains one of the most widely used baselines. It extends TF-IDF weighting with document length normalization and term saturation [8] and continues to perform competitively in large-scale retrieval settings. However, lexical methods struggle when queries and documents express the same concept using different terminology [9]. Query expansion techniques, such as RM3, attempt to address this through pseudo-relevance feedback [10], but these approaches are sensitive to noise and query drift, particularly on small or specialized datasets.

2.2. Dense Neural Retrieval

Dense retrieval models represent queries and documents as continuous vector embeddings, enabling similarity computation based on contextual meaning rather than surface-level term overlap. Transformer-based language models have significantly advanced this paradigm by capturing contextual dependencies through self-attention mechanisms [11]. In scientific domains, several pretrained encoders have been developed. SciBERT [12] is trained on full scientific texts and produces general-purpose contextual embeddings suitable for a range of NLP and retrieval tasks. SPECTER [13] learns document-level representations using citation-based supervision, capturing both topical content and structural relationships between papers. SciNCL [14] builds on this direction by applying contrastive learning with citation signals to improve semantic alignment between related and unrelated documents.

To support efficient similarity search over the resulting high-dimensional embeddings, vector indexing libraries such as FAISS [15] are commonly used. While dense models effectively mitigate vocabulary mismatch, they may still underperform lexical methods in isolation when exact term matching or evidence-level precision is required.

2.3. Hybrid Fusion and Cross-Encoder Re-Ranking

Hybrid retrieval approaches combine lexical precision with dense semantic similarity. Prior work has shown that such combinations often outperform standalone dense retrievers, especially in domain-specific settings where vocabulary mismatch and contextual variation coexist [16]. Reciprocal Rank Fusion (RRF) [17] is a widely adopted fusion technique that aggregates ranked outputs from heterogeneous retrievers. By combining rankings rather than raw scores, RRF improves recall and ranking stability without requiring score calibration across systems.

While hybrid fusion broadens the candidate pool, bi-encoder dense retrievers still approximate relevance through independently encoded representations. Cross-encoder models address this by jointly encoding query-document pairs, enabling fine-grained

semantic interaction and more precise relevance scoring [18]. Due to their computational cost, cross-encoders are typically applied as a second-stage re-ranker over a limited set of top-ranked candidates, a strategy that has been shown to yield substantial gains in early precision while preserving scalability [19]. Zhang et al. [20] further explored this direction with HLATR, a lightweight list-aware transformer that couples retrieval and reranking stage features into a unified re-ranking module, demonstrating that jointly modeling multi-stage signals can improve ranking effectiveness without substantial computational overhead.

Despite progress in each of these components, relatively few studies have evaluated lexical, dense, hybrid, and cross-encoder approaches within a unified retrieval framework for the scientific literature, or systematically analyzed their interactions and failure modes in claim-based retrieval settings.

3. Novelty and Contribution

While the individual components employed in this work, BM25, SciBERT, SPECTER, SciNCL, RRF, and cross-encoder re-ranking, are established techniques, the contribution of this paper is integrative and analytical rather than methodological. Existing benchmark studies, most notably BEIR [5], evaluate retrieval models across diverse datasets but primarily report aggregate performance without examining how these components interact within a multi-stage pipeline or why specific configurations succeed or fail on claim-based scientific queries.

This work addresses that gap through three complementary contributions. First, it provides a controlled, component-level comparison of lexical, dense, hybrid, and re-ranked retrieval paradigms within the SciFact benchmark, isolating the effect of each pipeline stage on retrieval effectiveness. Second, it presents ablation analyses that investigate design choices often underexplored in claim-based retrieval settings, including RM3-based query expansion, which is shown to introduce query drift, and passage-level max pooling, which fragments document-level evidence and degrades performance for most dense models. Third, it emphasizes interpretability and structured error analysis, constructing an error taxonomy based on representative failure cases that identifies specific failure modes such as semantic drift, lexical overlap bias, and query ambiguity.

The novelty of this work therefore lies not in proposing new retrieval components, but in offering a task-specific analytical perspective that complements existing benchmark studies. By moving beyond aggregate metrics toward systematic diagnosis of retrieval behavior, design trade-offs, and failure patterns, this paper provides actionable insights for improving retrieval system design in scientific literature search.

4. Methodology

This section presents the design and implementation of the proposed semantic retrieval framework for scientific literature. The methodology is organized as a multi-stage pipeline that integrates lexical, dense, hybrid, and re-ranking components to support effective and interpretable retrieval. It covers the overall system architecture, the selection and integration of retrieval and ranking models, dataset preparation, and the evaluation protocol used to assess retrieval performance.

4.1. System Architecture

The proposed system is designed as a modular, multi-stage retrieval pipeline in which each component performs a distinct role within the retrieval and evaluation process. The architecture integrates lexical and dense retrieval models to balance term-level precision with semantic relevance, and supports hybrid fusion and cross-encoder re-ranking to improve overall retrieval quality. Figure 1 presents the proposed system pipeline,

illustrating the progression from query input through candidate retrieval, hybrid fusion, re-ranking, and explainability analysis.

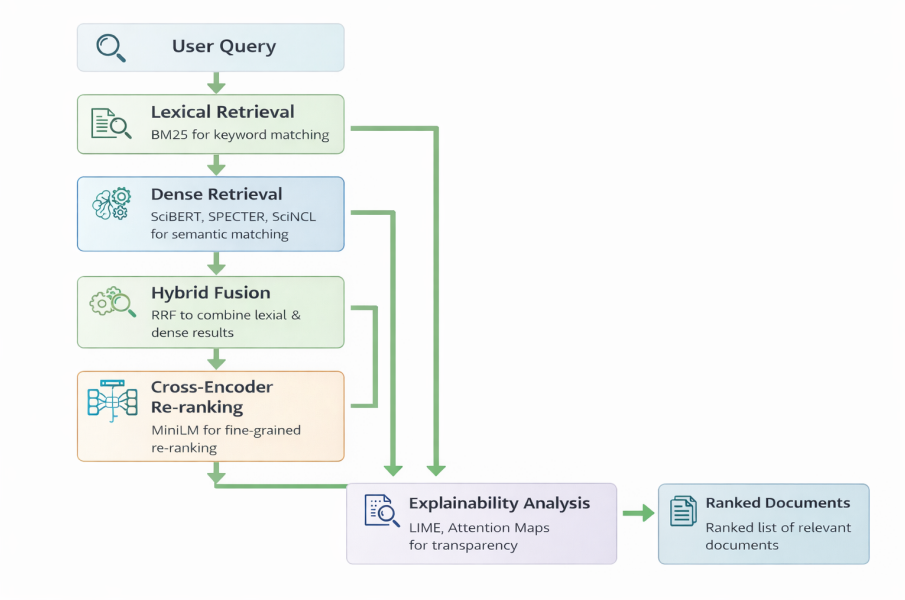


Figure 1. Architecture of the proposed multi-stage semantic retrieval pipeline, combining lexical and dense retrieval, hybrid fusion, cross-encoder re-ranking, and explainability analysis.

4.2. Model Selection and Implementation

This subsection describes the retrieval, ranking, and analysis components integrated into the proposed system.

- Lexical retrieval: BM25 from the Okapi family is used as a classical retrieval baseline, relying on exact term matching and term frequency statistics. The default Pyserini parameters are adopted ($k_1 = 0.9$, $b = 0.4$).
- Query expansion: As an additional lexical variant, we apply RM3-based pseudo-relevance feedback to evaluate query expansion. The top five retrieved documents are used as feedback (`fb_docs` = 5), from which 20 expansion terms are selected (`fb_terms` = 20). This configuration follows standard practice in pseudo-relevance feedback experiments and is evaluated separately to assess whether lexical expansion benefits claim-oriented queries.
- Dense retrieval models: Three domain-specific dense encoders are employed for semantic matching: SciBERT, SPECTER, and SciNCL. We use the pretrained checkpoints `allenai/scibert_scivocab_uncased`, `allenai/specter`, and `malteos/scincl`, respectively, loaded via the SentenceTransformers framework [21]. No additional fine-tuning is applied; all models are used with their original pretrained weights. Embedding extraction follows each model's default pooling configuration as specified in its SentenceTransformers model card: SciBERT uses mean pooling over token embeddings, while SPECTER and SciNCL use `[CLS]` token pooling. Each document is represented by concatenating its title and abstract in the format *title. abstract*, and both queries and documents are encoded into dense vector representations. The resulting embeddings are L2-normalized, and nearest-neighbor search is performed using FAISS [22] with cosine similarity:

$$\mathrm{sim}(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\| \, \|\mathbf{d}\|} \quad (1)$$

- Hybrid fusion: Reciprocal Rank Fusion (RRF) is applied to combine the ranked outputs of lexical and dense retrievers, using complementary signals to improve recall and ranking stability. The fused relevance score for a document d is computed as:

$$\mathrm{RRF}(d) = \sum_{i=1}^{N} \frac{1}{k + r_i(d)}, \tag{2}$$

  where $r_i(d)$ denotes the rank assigned to document d by the i-th retrieval system and k is a smoothing constant set to 60, following the value recommended in the original RRF formulation [17].
- Cross-encoder re-ranking: A cross-encoder based on the MS MARCO MiniLM model [23] is used as a second-stage re-ranker. For each query, the top 100 candidates retrieved by the hybrid fusion stage are re-ranked by jointly encoding query-document pairs. The model processes inputs with a maximum sequence length of 512 tokens, with truncation applied when the concatenated query-document pair exceeds this limit. The cross-encoder relevance scores are used directly for re-ranking without additional score calibration or normalization. Batching during inference is handled internally by the SentenceTransformers `CrossEncoder` implementation [21]
- Explainability analysis:
  - LIME is used to identify influential tokens that contribute to relevance predictions, supporting the analysis of misclassified or incorrectly retrieved documents.
  - Attention heatmaps are used to visualize token-level interactions within the cross-encoder, revealing where the model focuses during query document matching and helping identify attention misalignment and semantic drift.

Figure 2 provides a schematic illustration of the semantic similarity computation and hybrid fusion process, highlighting the use of cosine similarity for ranking dense representations and the RRF [17] for combining retrieval results.

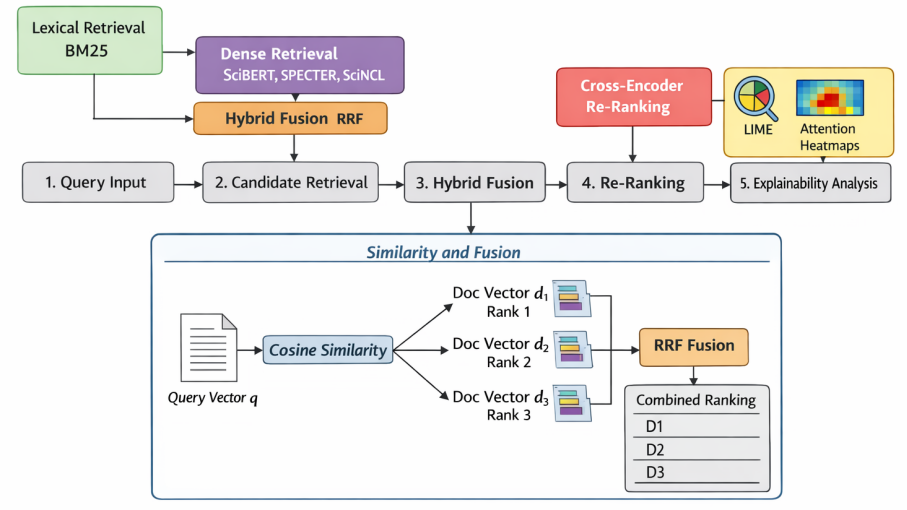


Figure 2. Schematic illustration of semantic similarity computation and hybrid fusion within the proposed retrieval framework, showing cosine similarity-based ranking of dense representations and RRF for combining ranked results.

It is worth noting that the dense retrieval models used in this paper, SciBERT, SPECTER, and SciNCL, were chosen specifically because they are trained on scientific corpora and designed to represent scientific documents. SciBERT is pretrained on full-text scientific papers, SPECTER learns document-level embeddings from citation graphs, and SciNCL refines these representations through contrastive learning with citation-based supervision.

This makes them natural candidates for a study focused on claim-based scientific retrieval, where domain-specific language and reasoning patterns play a central role.

We acknowledge that more recent general-purpose embedding models, such as E5 [24], BGE [25], and GTE [26], have demonstrated strong performance on the BEIR benchmark, including SciFact. However, these models are trained on large-scale heterogeneous web data and are optimized for broad cross-domain generalization rather than domain-specific scientific understanding. Since the goal of this work is to study how different retrieval components interact within a multi-stage pipeline for the scientific literature, not to identify the single highest-scoring model on a given benchmark, we considered domain-specific encoders to be the more appropriate choice for our experimental setup.

4.3. Dataset and Preprocessing

The experiments are conducted on the SciFact dataset [4], which is also included in the BEIR benchmark framework [5]. The dataset focuses on scientific claim verification. It includes a corpus of scientific abstracts, a set of claim-based queries, and ground-truth relevance judgments (qrels) linking claims to supporting documents. The data are downloaded and prepared using the GenericDataLoader utility provided by the BEIR framework. Text normalization and tokenization are applied, followed by concatenation of document titles and abstracts to form document representations. Dense embeddings are then indexed using a FAISS-compatible vector format to support efficient similarity search [22].

4.4. Evaluation Metrics and Protocol

Retrieval performance is evaluated using standard information retrieval metrics that assess ranking quality, early precision, and coverage of relevant documents. Since scientific search is highly sensitive to the ordering of top-ranked results, we report ranking-based metrics at cutoff $k = 10$, reflecting realistic user behavior where only the first page of results is typically examined. This cutoff also aligns with common practice in neural information retrieval benchmarks.

- NDCG@k (Normalized Discounted Cumulative Gain) evaluates ranking quality by assigning higher importance to relevant documents retrieved at higher ranks, using logarithmic discounting to penalize lower-ranked positions. NDCG is particularly suitable for claim-based scientific retrieval because it captures position-sensitive effectiveness and supports graded relevance signals. For a ranked list of documents, NDCG@k is defined as:

$$\text{NDCG@k} = \frac{1}{\text{IDCG@k}} \sum_{i=1}^{k} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)}, \tag{3}$$

- MAP@k (Mean Average Precision) measures ranking effectiveness by averaging precision values at the ranks where relevant documents are retrieved, favoring systems that return relevant results early.
- Recall@k quantifies the proportion of relevant documents successfully retrieved within the top-k results, capturing coverage of relevant evidence.
- MRR@k (Mean Reciprocal Rank) captures how early the first relevant document appears in the ranked list, emphasizing early correct retrieval.

Among these metrics, NDCG@10 is used as the primary comparison measure, as it jointly reflects ranking sensitivity and early precision, making it well suited for evaluating multi-stage retrieval and re-ranking pipelines. For all reported metrics in this paper, higher values indicate better retrieval performance.

4.5. Implementation Details

The system is implemented using a set of established software libraries and executed within a GPU-enabled environment to support efficient retrieval and neural inference.

4.5.1. Software Libraries

The following software libraries are used through the experiments of this paper:

- FAISS is used for dense vector indexing and similarity search [15,22], with document embeddings stored in an `IndexFlatIP` index for exact inner-product search over L2-normalized vectors.
- PyTorch (Version: 2.10.0+cu128) serves as the backend framework for executing transformer-based models.
- HuggingFace Transformers [27] is used to load pretrained model weights for SciBERT, SPECTER, and SciNCL. All three models are accessed through the SentenceTransformers library [21] for consistent encoding and pooling.
- The BEIR toolkit [5] provides standardized data loading and evaluation utilities for benchmark datasets.
- Scikit-learn, NumPy, and Pandas are used for data preprocessing and metric computation.

4.5.2. Computational Environment

Experiments are conducted in a GPU-enabled environment to ensure efficient processing of transformer models. Google Colab is used as the execution platform, with an NVIDIA Tesla T4 GPU V100 (16 GB memory) allocated for model inference and re-ranking. Python 3.x with a CUDA-enabled PyTorch installation is used throughout the experiments.

5. Experimental Results

The results are organized to follow the progressive stages of the retrieval pipeline. We begin with standalone model comparisons (lexical and dense), then examine hybrid fusion, followed by cross-encoder re-ranking. After reporting the main results, we present computational efficiency and statistical significance analyses. We then turn to error analysis and explainability, component-wise ablation, and finally cross-domain evaluation. This structure allows each stage to build on the previous one, making it easier to see how individual components contribute to the overall pipeline performance.

The proposed semantic retrieval framework improves the accuracy and relevance of the scientific literature retrieval across multiple retrieval paradigms. In dense retrieval experiments comparing SciBERT, SPECTER, and SciNCL, the SciNCL achieves the best performance, while SciBERT records the lowest scores among dense models, as summarized in Table 1 and illustrated in Figure 3. The figure shows a clear separation between models, with SciNCL leading across all four metrics and SciBERT falling below the BM25 baseline, particularly in NDCG@10 and MRR@10.

In contrast, the lexical baseline BM25 outperforms SciBERT, highlighting the limitations of using SciBERT as a standalone retriever, as shown in Table 1. To address this limitation, a hybrid fusion combining SciBERT with BM25 is implemented using RRF [17], resulting in a clear performance improvement over dense retrieval alone as summarized in Table 2 and illustrated in Figure 4. The upward trend across all metrics confirms that hybridization consistently benefits weaker dense models like SciBERT, while the gains for SciNCL are more modest.

Table 1. Retrieval effectiveness of BM25 and dense scientific encoders (SciBERT, SPECTER, and SciNCL) on SciFact at cutoff 10, and higher values indicate better effectiveness. Best values are shown in bold.

Model	NDCG@10	MAP@10	Recall@10	MRR@10
BM25	0.348	0.312	0.451	0.322
SciBERT	0.320	0.285	0.423	0.293
SPECTER	0.432	0.390	0.555	0.401
SciNCL	**0.503**	**0.454**	**0.638**	**0.470**

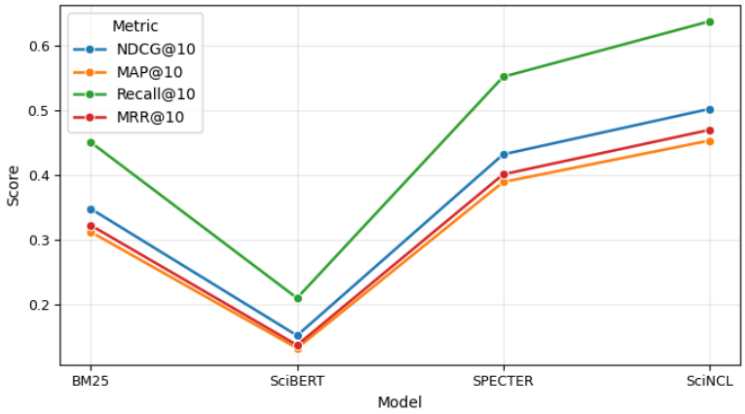


Figure 3. Metric-wise comparison at cutoff 10 for BM25 and dense retrievers on SciFact, illustrating the relative effectiveness of SciBERT, SPECTER, and SciNCL against the lexical baseline.

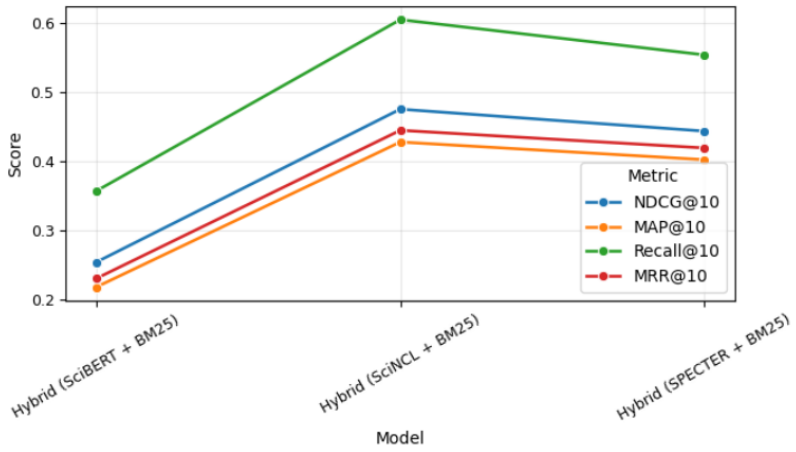


Figure 4. Hybrid retrieval performance at cutoff 10 on SciFact for RRF fusion of BM25 with SciBERT, SPECTER, and SciNCL, visualizing the metric improvements from combining lexical and dense signals.

Table 2. Hybrid fusion results on SciFact at cutoff 10 using RRF to combine BM25 with each dense retriever models, showing the impact of hybridization on ranking effectiveness. Best values are shown in bold.

Model	NDCG@10	MAP@10	Recall@10	MRR@10
Hybrid (SciBERT + BM25)	0.390	0.356	0.488	0.365
Hybrid (SPECTER + BM25)	0.443	0.402	0.554	0.419
Hybrid (SciNCL + BM25)	**0.475**	**0.428**	**0.604**	**0.445**

Additional hybrid configurations using SciNCL and SPECTER are also evaluated. While Hybrid (SciNCL + BM25) outperforms Hybrid (SPECTER + BM25), hybrid fusion does not consistently surpass the best standalone dense retriever. Specifically, SciNCL alone

achieves an NDCG@10 of 0.503, whereas its hybrid combination with BM25 via RRF yields 0.475, a drop of 0.028. This occurs because RRF aggregates rankings from both systems equally, so when BM25 and SciNCL disagree on the relevance of candidate documents, lower-quality BM25 candidates can dilute the stronger SciNCL ranking. In other words, for a dense model that already captures semantic relevance well, the addition of a lexical retriever whose rankings diverge substantially can hurt rather than help.

It is worth noting, however, that hybrid fusion still plays a useful role in the overall pipeline. It broadens the candidate pool by surfacing documents that a single retriever might miss, and this becomes important once cross-encoder re-ranking is applied. Indeed, incorporating cross-encoder re-ranking consistently yields the best overall results: Hybrid (SciNCL + BM25) + Cross-Encoder achieves the highest effectiveness across all metrics (NDCG@10 = 0.523, MAP@10 = 0.479, Recall@10 = 0.642, MRR@10 = 0.497). These results indicate that the main performance gains in the full pipeline are driven primarily by the re-ranking stage, which refines the broader candidate set produced by hybrid fusion through fine-grained query–document interaction. Hybrid fusion, therefore, should be understood not as a direct ranking improvement over the best dense retriever, but as a candidate diversification step that enables the cross-encoder to operate over a richer set of documents.

All results are summarized in Table 3 and illustrated in Figure 5. As the figure shows, the three hybrid + CE configurations converge to similar performance levels, with differences of less than 0.011 in NDCG@10, suggesting that cross-encoder re-ranking substantially narrows the gap between underlying dense models. For ease of comparison across all experimental configurations, the complete set of results is consolidated in Table 4.

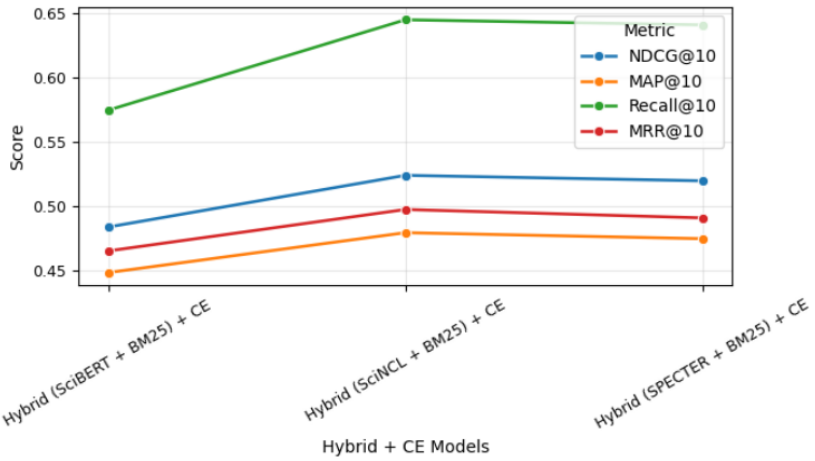


Figure 5. Metric-wise performance at cutoff 10 for hybrid + cross-encoder (CE) configurations on SciFact, showing the effect of second-stage re-ranking.

Table 3. Retrieval performance on SciFact at cutoff 10 after applying cross-encoder re-ranking (CE) to the hybrid candidate sets, where higher values indicate better performance.Best values are shown in bold.

Model	NDCG@10	MAP@10	Recall@10	MRR@10
Hybrid (SciBERT + BM25) + CE	0.512	0.475	0.610	0.490
Hybrid (SPECTER + BM25) + CE	0.519	0.475	0.641	0.491
Hybrid (SciNCL + BM25) + CE	**0.523**	**0.479**	**0.642**	**0.497**

Table 4. Retrieval performance of lexical, dense, hybrid, and hybrid + cross-encoder models on the SciFact dataset. Best values are shown in bold.

Model	NDCG@10	MAP@10	Recall@10	MRR@10
SciBERT	0.320	0.285	0.423	0.293
SPECTER	0.432	0.390	0.553	0.401
SciNCL	0.503	0.454	0.638	0.470
BM25	0.348	0.312	0.451	0.322
BM25_RM3	0.327	0.297	0.407	0.305
Hybrid (SciNCL + BM25)	0.475	0.428	0.604	0.445
Hybrid (SPECTER + BM25)	0.443	0.402	0.554	0.419
Hybrid (SciBERT + BM25)	0.390	0.356	0.488	0.365
Hybrid (SciNCL + BM25) + CE	**0.523**	**0.479**	**0.642**	**0.497**
Hybrid (SPECTER + BM25) + CE	0.519	0.475	0.641	0.491
Hybrid (SciBERT + BM25) + CE	0.512	0.475	0.610	0.490

5.1. Computational Efficiency Analysis

To quantify the computational cost of each retrieval paradigm, we compare setup time, per-query latency, and resource utilization across lexical, dense, hybrid, and cross-encoder configurations. As reported in Table 5, BM25 remains the most computationally efficient baseline, requiring negligible indexing overhead and minimal runtime resources.

Dense retrieval models introduce additional cost due to embedding generation and vector indexing, increasing both preprocessing time and memory consumption. Hybrid pipelines further add latency through score fusion mechanisms. The largest computational overhead is observed with cross-encoder re-ranking, which significantly increases GPU utilization and per-query inference time. These results highlight a clear effectiveness–efficiency trade-off: while multi-stage pipelines yield superior retrieval quality, they require substantially greater computational resources compared to standalone lexical or dense approaches.

Table 5. Computational cost comparison of lexical, dense, hybrid, and cross-encoder retrieval pipelines.

Evaluating	Setup Time (s)	Avg Retrieval Latency (ms/Query)	CPU Usage (%)	GPU Usage (%)	n_queries
BM25	0.000	16.321	7.50	0.04	100
SciBERT	6.071	14.414	4.50	7.98	100
Hybrid (SciBERT + BM25)	6.071	29.829	5.92	5.68	100
Hybrid (SciBERT + BM25) + CE	6.071	50.061	3.57	17.63	100
SPECTER	5.191	14.175	5.00	7.94	100
Hybrid (SPECTER + BM25)	5.191	29.009	5.34	5.40	100
Hybrid (SPECTER + BM25) + CE	5.191	49.001	3.16	17.25	100
SciNCL	5.546	13.377	7.03	7.40	100
Hybrid (SciNCL + BM25)	5.546	29.673	1.75	5.46	100
Hybrid (SciNCL + BM25) + CE	5.546	48.434	6.13	17.57	100

5.2. Statistical Significance and Effect Size Analysis

To determine whether the observed differences among dense retrieval models are statistically meaningful, we conducted paired t-tests on per-query NDCG@10 scores. Each query in the SciFact test set (n = 100) produces an NDCG@10 value under each model, and the paired design accounts for query-level variability by comparing score differences within the same query across models. Table 6 reports the mean performance difference (Δ), standard deviation, t-statistics, p-values, and corresponding effect sizes (Cohen's d) for each pairwise comparison.

We note that paired t-tests assume approximate normality of the score differences. To assess this assumption, we applied the Shapiro-Wilk test to the per-query NDCG@10

difference distributions for each model pair. As reported in Table 6, the Shapiro–Wilk test rejects normality in all three cases ($p < 0.001$), indicating that the score differences are not normally distributed. To ensure that our significance conclusions are not affected by this violation, we additionally conducted the Wilcoxon signed-rank test, a non-parametric alternative that does not require normality. The Wilcoxon test confirms statistically significant differences across all three comparisons ($p < 0.001$ for SciNCL vs. SciBERT and SPECTER vs. SciBERT; $p = 0.0011$ for SciNCL vs. SPECTER), consistent with the paired t-test results. We therefore report both tests in Table 6 and consider the observed performance differences to be statistically robust.

Table 6. Pairwise statistical comparison of dense retrieval models on SciFact using NDCG@10.

Comparison	Mean Δ	Std Δ	Shapiro (p-Value)	t-Value	(p-Value)	Wilcoxon W-Value	Wilcoxon (p-Value)	Cohen's d
SciNCL vs. SciBERT	0.351	0.447	<0.001	13.61	<0.001	6547.0	<0.001	0.786
SPECTER vs. SciBERT	0.281	0.438	<0.001	11.10	<0.001	8893.0	<0.001	0.641
SciNCL vs. SPECTER	0.070	0.347	<0.001	3.51	<0.001	17,781.0	0.0011	0.203

All model comparisons yield statistically significant differences ($p < 0.001$). SciNCL demonstrates a substantial improvement over SciBERT (mean $\Delta = 0.351$, $d = 0.786$) and a moderate improvement over SPECTER (mean $\Delta = 0.070$, $d = 0.203$). Similarly, SPECTER significantly outperforms SciBERT ($d = 0.641$). The effect size estimates indicate that the performance gap between SciNCL and SciBERT is large, while differences involving SPECTER are moderate to small. These results confirm that the ranking differences among dense models are not due to random variation but reflect consistent and practically meaningful performance differences.

5.3. Error Analysis and Explainability

This subsection examines representative failure case to better understand retrieval limitations and model behavior through interpretability analysis.

Case study: *A deficiency of vitamin B12 increases blood levels of homocysteine.*

In this example, the retrieval pipeline assigns a high relevance score (0.65) to a document discussing associations between elevated plasma homocysteine and vascular diseases, despite the absence of explicit evidence linking vitamin B12 deficiency to the stated claim. Although the retrieved document is topically related, it does not directly address the causal relationship expressed in the query.

LIME explanations, Figures 6 and 7, reveal that the model's relevance estimation is largely driven by overlapping biomedical terms, particularly "homocysteine" and "increased plasma." These tokens receive the highest contribution weights, indicating that lexical co-occurrence plays a dominant role in the decision process. This pattern reflects a topical similarity error, where shared terminology is interpreted as semantic relevance without verifying claim-specific alignment. We categorize this failure as lexical overlap bias.

Attention analysis of the cross-encoder shown in Figure 8 provides further insight. The attention maps show concentrated weights on frequent domain-specific terms rather than on causal indicators related to "vitamin B12 deficiency." This suggests that, even in the re-ranking stage, the model emphasizes topic-level similarity over precise claim evidence matching.

This case illustrates a broader limitation observed across multiple failure examples: while hybrid retrieval and cross-encoder re-ranking improve overall ranking metrics,

semantic alignment at the level of explicit scientific claims remains challenging. These observations motivate the structured error taxonomy presented in the following subsection.

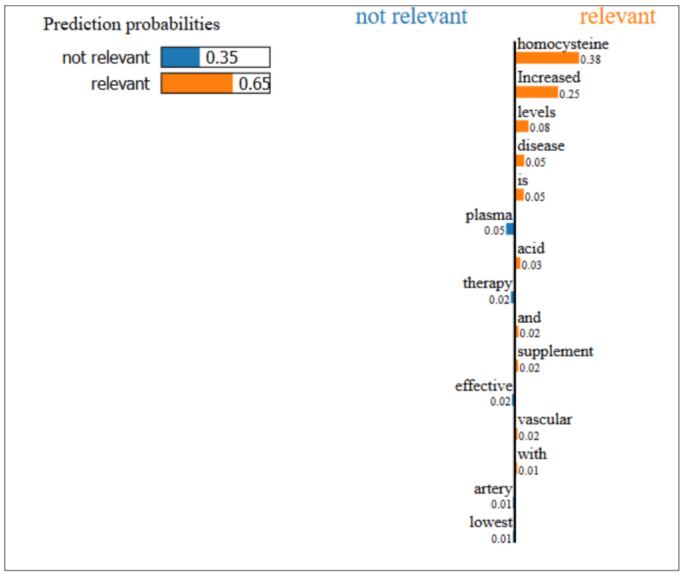


Figure 6. LIME explanation highlighting the most influential tokens contributing to the incorrect relevance prediction. Lexical overlap terms receive the highest weights.

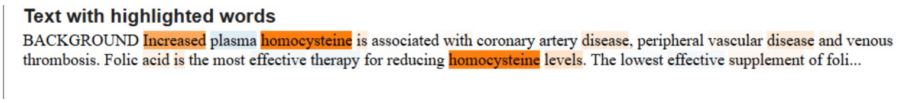


Figure 7. Highlighted text showing LIME-identified tokens in the incorrectly retrieved document for the vitamin B12 query. Orange-highlighted terms ("homocysteine", "increased", and "plasma") indicate tokens with the highest contribution to the relevance prediction, confirming that the model relies on shared biomedical vocabulary rather than evidence directly supporting the claim.

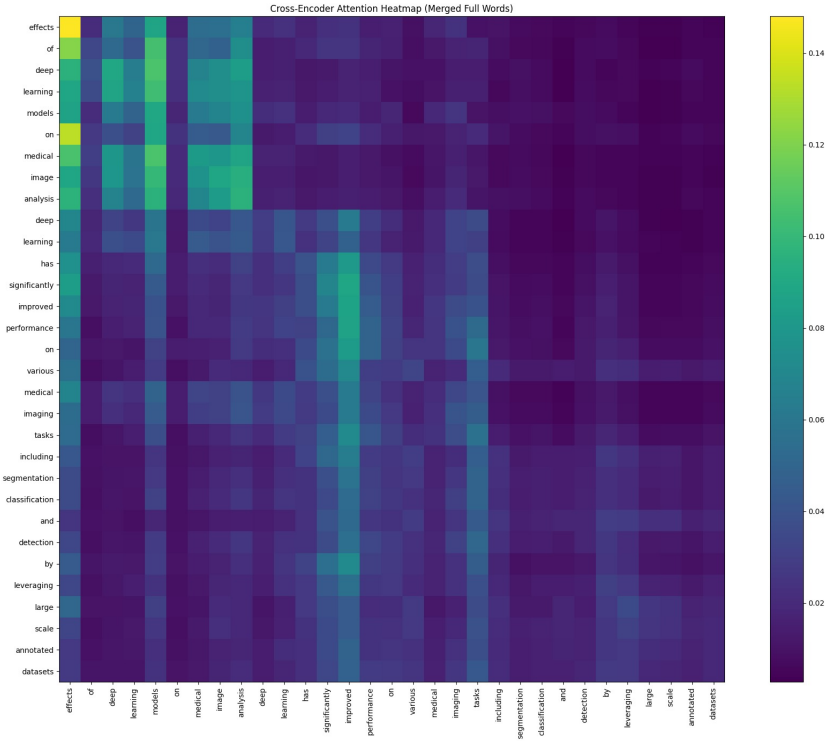


Figure 8. Attention distribution of the cross-encoder for the incorrectly retrieved document. Attention weights concentrate on high-frequency domain terms instead of causal indicators related to vitamin B12 deficiency, illustrating a topical similarity bias in relevance estimation.

5.4. Error Taxonomy and Failure Case Analysis

To systematically characterize retrieval failures, we construct an error taxonomy based on ten representative cases in which a correctly relevant document is ranked below an incorrectly top-ranked one. Each case compares the gold-standard relevant document with the highest-ranked incorrect result. The full set of examples is provided in Appendix A.

Qualitative analysis reveals that *semantic drift* is the most prevalent failure mode. In such cases, retrieved documents are topically related to the query but do not contain explicit evidence supporting the claim. The model captures domain similarity but fails to verify claim-specific alignment, resulting in relevance overestimation.

A second recurrent pattern is *lexical overlap bias*, where shared terminology between query and document dominates the relevance signal despite underlying semantic inconsistency. We also observe frequent *query ambiguity and intent mismatch*, particularly for short or underspecified claims. In these cases, multiple plausible interpretations exist, and the retrieval system selects documents aligned with an alternative interpretation rather than the intended claim. Less frequent errors include *vocabulary mismatch* (different lexical forms for the same concept), *domain shift* (related but cross-subfield retrieval), *topic confusion* (broad yet misaligned content), and *entity mismatch* (conflation of closely related biomedical entities).

Overall, this taxonomy indicates that while hybrid retrieval and cross-encoder re-ranking substantially improve ranking effectiveness, residual errors are primarily driven by limitations in fine-grained semantic alignment and claim-level reasoning. These findings suggest that improvements in scientific retrieval require models that move beyond topical similarity toward structured evidence modeling. These recurring error types suggest that future retrieval systems should incorporate more explicit claim–evidence modeling and robust disambiguation mechanisms to reduce semantic drift and topical bias.

5.5. Component-Wise Ablation Analysis

An ablation analysis is conducted to evaluate the individual contribution of key components in the retrieval pipeline, including query expansion, hybrid fusion, and cross-encoder re-ranking.

5.5.1. Lexical Versus Semantic Query Expansion

This experiment evaluates the impact of lexical and semantic query expansion strategies on retrieval effectiveness. The lexical baseline BM25 is compared against BM25 enhanced with RM3 pseudo-relevance feedback and a semantic expansion approach based on SPECTER embeddings.

Results indicate that BM25 outperforms BM25 + RM3 across all evaluation metrics. Although RM3 is designed to improve recall by incorporating top-ranked expansion terms, it introduces semantic noise and query drift on the SciFact dataset [4], particularly for short, claim-oriented queries. This leads to reduced ranking quality, as reflected in lower NDCG and MAP scores (Table 7, Figure 9). The figure clearly illustrates the divergent behavior: while RM3 pulls all metrics downward relative to BM25, SPECTER-based expansion pushes them sharply upward, with the largest gain visible in Recall@10.

In contrast, the semantic expansion strategy based on SPECTER yields consistent improvements over both BM25 and BM25 + RM3. By embedding the query in a dense semantic space and retrieving expansion candidates through vector similarity, SPECTER produces contextually aligned and more stable expansions. These results highlight the advantage of embedding-based expansion over purely lexical pseudo-relevance feedback in scientific retrieval settings.

Table 7. Retrieval performance of BM25, BM25 + RM3 (lexical expansion), and BM25 + SPECTER (semantic expansion) on SciFact. Higher values indicate better effectiveness. Best values are shown in bold.

Model	NDCG@10	MAP@10	Recall@10	MRR@10
BM25	0.348	0.312	0.451	0.322
BM25_RM3	0.327	0.297	0.407	0.305
BM25_SPECTER Expansion	**0.523**	**0.457**	**0.716**	**0.468**

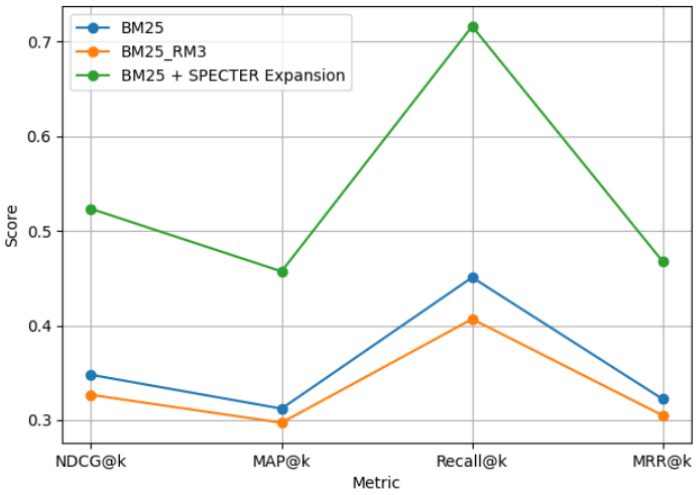


Figure 9. Visual comparison of query expansion strategies on SciFact, contrasting lexical feedback (RM3) with embedding-based semantic expansion (SPECTER).

5.5.2. Impact of Retrieval Paradigms and Multi-Stage Ranking

This experiment compares lexical, dense, and hybrid retrieval paradigms to assess their relative contribution to ranking effectiveness. Dense retrievers consistently outperform the lexical baseline, reflecting their ability to capture semantic similarity beyond exact term matching. However, hybrid retrieval achieves the strongest overall performance by integrating dense representations with lexical signals. This combination mitigates vocabulary mismatch while preserving term-level precision, resulting in more stable and robust retrieval outcomes, as illustrated in Figure 10. Across all four metric panels, the hybrid bars consistently exceed both the dense and lexical bars, visually confirming the complementarity of combining retrieval signals.

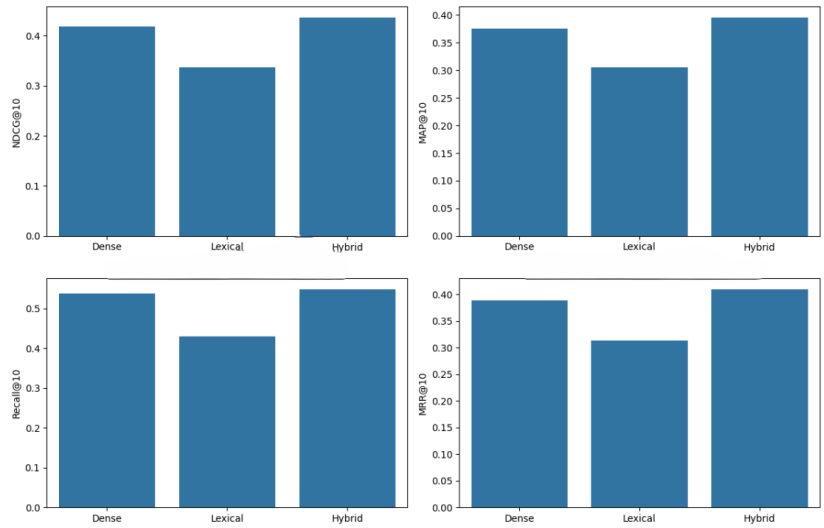


Figure 10. Bar chart comparing lexical, dense, and hybrid retrieval paradigms on the SciFact dataset at cutoff 10.

5.5.3. Effect of Cross-Encoder Re-Ranking on Hybrid Retrieval

This experiment evaluates the impact of adding a cross-encoder re-ranking stage to the hybrid retrieval pipeline. While hybrid retrieval improves candidate selection by integrating lexical and dense signals, relevance estimation is still based on independently encoded representations.

Introducing cross-encoder re-ranking enables joint query–document encoding, allowing fine-grained semantic interaction and more precise relevance scoring. As shown in Table 3 and Figure 5, this second-stage refinement consistently improves NDCG, MAP, Recall, and MRR. These results highlight the complementary roles of multi-stage retrieval: hybrid models enhance recall and candidate quality, whereas cross-encoders refine early precision through deeper semantic alignment. estimation.

5.5.4. Passage-Level Max Pooling Versus Document-Level Dense Encoding

This experiment examines the effect of applying passage-level max pooling (MaxP) within dense retrieval models. In the MaxP setting, each document is segmented into multiple passages, embeddings are computed independently, and the final document score is determined solely by the highest-scoring passage relative to the query [28]. The underlying assumption is that relevance is concentrated within a single highly matching segment, allowing the model to focus on the most informative textual unit.

However, empirical results indicate that this assumption is problematic for claim-based scientific retrieval. As shown in Table 8, replacing full-document dense representations with MaxP leads to substantial performance degradation for both SciBERT and SPECTER. The decline is particularly pronounced for SciBERT, where NDCG@10 drops from 0.320 to 0.153, and for SPECTER, where performance decreases from 0.432 to 0.251. These results suggest that isolating a single top-scoring passage discards complementary contextual evidence distributed across the document.

In scientific abstracts, relevant evidence is often distributed across multiple sentences rather than concentrated in one passage. By selecting only the maximum-scoring segment, MaxP introduces evidence fragmentation, ignoring supporting information that may collectively establish claim relevance. This effect is especially detrimental for models that rely on global semantic coherence within the abstract.

Table 8. Performance comparison between document-level dense encoding and passage-level MaxP on SciFact, illustrating the degradation introduced by max pooling for most dense models.

Model	Variant	NDCG@10	MAP@10	Recall@10	MRR@10
SPECTER	Dense	0.432	0.390	0.553	0.401
SPECTER (MaxP)	MaxP	0.251	0.214	0.356	0.226
SciBERT	Dense	0.320	0.285	0.423	0.293
SciBERT (MaxP)	MaxP	0.153	0.128	0.218	0.144
SciNCL	Dense	0.503	0.454	0.638	0.470
SciNCL (MaxP)	MaxP	0.497	0.443	0.644	0.464

Interestingly, SciNCL exhibits comparatively stable behavior under MaxP, with only marginal performance changes (e.g., NDCG@10 from 0.503 to 0.497). This robustness may be attributed to its contrastive training objective, which encourages stronger document-level semantic alignment and potentially yields more discriminative passage embeddings. Nevertheless, even in this case, MaxP does not produce measurable gains over the dense document-level representation.

Overall, these findings indicate that for claim-oriented scientific retrieval, document-level dense encoding is more reliable than passage-level max pooling. While MaxP may be beneficial

in long-document retrieval scenarios where relevance is localized, it appears less suitable for short scientific abstracts in which evidence is contextually integrated rather than isolated.

5.6. Cross-Domain Robustness and Generalization

To evaluate robustness beyond a single benchmark, we assess the proposed retrieval pipelines on three datasets with distinct structural properties: SciFact [4], a claim verification benchmark; PubMedQA [29], a biomedical question-answering dataset; and SciDocs [13], a scientific document similarity benchmark based on citation prediction. This three-way comparison allows us to examine whether the relative effectiveness of lexical, dense, and hybrid paradigms holds across different query formulations, relevance distributions, and difficulty levels.

As reported in Table 9, the three datasets produce markedly different absolute score ranges. On PubMedQA, nearly all models achieve high scores, with BM25 alone reaching an NDCG@10 of 0.945 and the best hybrid + cross-encoder configuration reaching 0.986. In contrast, SciDocs yields much lower absolute scores across all models, with the best NDCG@10 being 0.178 for standalone SciNCL. SciFact falls in between, with scores ranging from 0.320 (SciBERT) to 0.523 (Hybrid SciNCL + BM25 + CE). These differences are consistent with the structural characteristics summarized in Table 10: PubMedQA queries exhibit strong lexical alignment with relevant documents and high relevance density, making it a comparatively easy retrieval setting. SciDocs, on the other hand, uses short paper titles as queries against a large corpus of abstracts, with relevance defined by citation relationships rather than explicit textual overlap, a much harder retrieval task.

Several patterns emerge from this cross-domain comparison. First, the relative ranking among dense models is stable: SciNCL consistently outperforms SPECTER, which in turn outperforms SciBERT, regardless of the dataset. This ordering holds across the full range of difficulty levels, from the near-ceiling PubMedQA scores to the challenging SciDocs setting.

Table 9. Retrieval performance comparison at $k = 10$ across SciFact, PubMedQA, and SciDocs. Best values are shown in bold.

Model	SciFact				PubMedQA				SciDocs			
	NDCG	MAP	Recall	MRR	NDCG	MAP	Recall	MRR	NDCG	MAP	Recall	MRR
Dense												
SciBERT	0.320	0.285	0.423	0.293	0.371	0.319	0.536	0.319	0.025	0.012	0.025	0.057
SPECTER	0.432	0.390	0.553	0.401	0.839	0.807	0.941	0.807	0.132	0.074	0.138	0.238
SciNCL	0.503	0.454	0.638	0.470	0.972	0.965	0.995	0.966	0.178	0.102	0.190	0.304
Lexical												
BM25	0.348	0.312	0.451	0.322	0.945	0.936	0.975	0.936	0.099	0.056	0.101	0.186
Hybrid												
Hybrid (SciNCL + BM25)	0.475	0.428	0.604	0.446	0.969	0.964	0.983	0.965	0.135	0.073	0.150	0.241
Hybrid (SPECTER + BM25)	0.443	0.402	0.556	0.419	0.936	0.923	0.975	0.923	0.107	0.056	0.119	0.195
Hybrid (SciBERT + BM25)	0.390	0.356	0.488	0.364	0.687	0.645	0.821	0.645	0.031	0.015	0.032	0.064
Hybrid + CE												
Hybrid (SciNCL + BM25) + CE	**0.523**	**0.479**	**0.642**	**0.497**	**0.986**	**0.983**	**0.995**	**0.983**	**0.164**	**0.095**	**0.173**	**0.286**
Hybrid (SPECTER + BM25) + CE	0.520	0.475	0.641	0.491	0.984	0.982	0.992	0.982	0.153	0.088	0.159	0.276
Hybrid (SciBERT + BM25) + CE	0.512	0.475	0.610	0.490	0.981	0.978	0.988	0.979	0.070	0.040	0.061	0.152

Second, the interaction between hybrid fusion and the best dense retriever varies by dataset. On both SciFact and SciDocs, Hybrid (SciNCL + BM25) scores lower than standalone SciNCL, dropping from 0.503 to 0.475 on SciFact and from 0.178 to 0.135 on SciDocs. This pattern reinforces the observation discussed in the begining of Section 6 that RRF-based fusion

can dilute the rankings of a strong dense retriever when BM25 candidates diverge substantially from the dense model's top results. Notably, this dilution effect is even more pronounced on SciDocs, where the gap between SciNCL alone and the hybrid is 0.043 points compared to 0.028 on SciFact. On PubMedQA, where both BM25 and SciNCL already perform near the ceiling, the hybrid combination shows negligible change (0.969 vs. 0.972).

Table 10. Structural differences between SciFact, PubMedQA, and SciDocs affecting retrieval difficulty and generalization behavior.

Indicator	SciFact	PubMedQA	SciDocs
Domain	Scientific fact verification	Biomedical QA	Scientific document similarity
Query type	Claims (often long)	Questions (natural language)	Paper titles (short)
Lexical overlap	Lower/paraphrase-heavy	Higher (medical terminology)	Very low (title-to-abstract)
Relevance density	Often sparse	Typically denser	Sparse and distributed
Reasoning required	High (evidence retrieval)	Moderate (supporting passages)	High (citation-level relatedness)
Expected difficulty	Higher	Lower	Highest

Third, cross-encoder re-ranking recovers much of the lost performance on SciFact (from 0.475 to 0.523, exceeding the standalone SciNCL score of 0.503), but on SciDocs the recovery is partial: Hybrid (SciNCL + BM25) + CE reaches 0.164, which remains below the standalone SciNCL score of 0.178. This suggests that when the underlying retrieval task is very difficult and the candidate pool produced by hybrid fusion contains many weakly relevant documents, even cross-encoder re-ranking cannot fully compensate for the dilution introduced at the fusion stage.

Fourth, BM25 shows interesting variation across datasets. On PubMedQA, it is remarkably strong (0.945), nearly matching the best dense model. On SciFact, it falls between SciBERT and SPECTER. On SciDocs, it scores 0.099, below SPECTER (0.132) and SciNCL (0.178) but above SciBERT (0.025). This confirms that lexical retrieval effectiveness is highly dependent on the degree of vocabulary overlap between queries and relevant documents, and that dense models offer the greatest advantage precisely in settings where such overlap is low.

Overall, these results demonstrate that while absolute performance varies substantially with dataset characteristics, the key findings of this paper generalize across all three benchmarks: domain-specific dense models outperform the lexical baseline in semantically challenging settings, hybrid fusion benefits weaker dense retrievers more than strong ones, and cross-encoder re-ranking provides the largest incremental gains when applied to a sufficiently diverse candidate pool. The addition of SciDocs as a harder evaluation benchmark strengthens our confidence that these patterns are not artifacts of a particular dataset's properties but reflect genuine properties of the retrieval pipeline.

6. Results Discussion

The experiments presented in this paper yield several findings that are worth situating in the context of prior work and their practical implications.

Our results confirm the well-established observation that dense retrieval models outperform lexical baselines in semantically challenging settings [11,16]. However, the finding that hybrid fusion via RRF can actually hurt the best dense retriever is less commonly reported. Most prior studies, including BEIR [5], present hybrid fusion as generally beneficial. Our results show this is not always the case: when the dense model is already strong (as with SciNCL), introducing a weaker lexical signal through equal-weight rank fusion can degrade performance. This finding was consistent across SciFact and

SciDocs, suggesting it is not an artifact of a single dataset. From a practical standpoint, this implies that retrieval system designers should consider weighted or adaptive fusion strategies rather than applying RRF uniformly.

A related finding is that cross-encoder re-ranking is the single most impactful component in the pipeline. This is broadly consistent with recent work on multi-stage retrieval [18,19], but our ablation results add nuance: the cross-encoder's ability to recover from fusion-induced dilution depends on the difficulty of the retrieval task. On SciFact, it fully compensates and exceeds standalone SciNCL performance. On SciDocs, recovery is only partial. This suggests that for very hard retrieval tasks, the quality of the candidate pool passed to the re-ranker matters more than is often assumed.

The error taxonomy reveals that the dominant failure modes, semantic drift and lexical overlap bias, persist even after cross-encoder re-ranking. This aligns with observations in related work on claim verification [4] and indicates that current retrieval models, including cross-encoders, still operate primarily at the level of topical similarity rather than structured claim-evidence reasoning. Addressing these failures likely requires architectural changes, such as incorporating explicit claim decomposition or entailment-aware scoring, rather than further tuning of existing retrieval components.

Finally, the RM3 query drift finding reinforces earlier concerns about pseudo-relevance feedback on specialized datasets [10]. In claim-based retrieval, where queries are short and precise, feedback-based expansion tends to introduce off-topic terms that dilute the original query intent. The superior performance of SPECTER-based semantic expansion suggests that embedding-based approaches are a more promising direction for query enrichment in scientific search.

7. Limitations and Future Work

Despite the consistent gains achieved by the multi-stage retrieval framework, several limitations should be noted. First, the best performance is obtained when hybrid retrieval is combined with cross-encoder re-ranking. While this configuration substantially improves ranking quality, it increases computational cost and per-query latency. The current pipeline is therefore better suited for high-accuracy retrieval scenarios rather than strict real-time settings. Future work should explore lightweight or adaptive re-ranking strategies to better balance effectiveness and efficiency.

Second, although cross-domain evaluation on SciFact and PubMedQA demonstrates stable relative model behavior, the paper remains limited to two biomedical benchmarks. Since dataset characteristics, such as lexical overlap and relevance density, significantly influence absolute performance, broader evaluation on more diverse scientific corpora is necessary to further validate generalization.

Third, the dense retrieval models evaluated in this paper, SciBERT [12], SPECTER [13], and SciNCL [14], were selected for their domain specificity rather than their recency. Since the time these models were released, several general-purpose embedding models have emerged, including E5 [24], BGE [25], and GTE [26], which achieve considerably higher scores on BEIR benchmarks owing to large-scale contrastive pretraining on diverse web data. Our study does not compare against these models, as the focus here is on understanding pipeline-level interactions (fusion, re-ranking, query expansion) rather than on maximizing absolute retrieval scores. That said, it remains an open question whether the patterns we observe, such as the detrimental effect of RRF fusion on strong dense retrievers, or the dominance of cross-encoder re-ranking in driving final performance, would hold when newer, higher-capacity encoders are used as the dense retrieval backbone. Investigating this is a natural and important direction for our future work.

Fourth, the explainability analysis presented in this paper is diagnostic in nature and relies on qualitative interpretation of LIME token contributions and cross-encoder attention patterns. While this approach is sufficient for identifying recurring failure modes and constructing the error taxonomy in Section 5.3, we do not quantitatively evaluate the fidelity or reliability of the generated explanations, for instance, through perturbation-based faithfulness tests or systematic agreement measures between LIME and attention-based attributions. Such evaluation would require a dedicated experimental protocol that falls outside the retrieval-focused scope of this paper. Future research should address this gap by incorporating quantitative interpretability metrics, which would strengthen confidence in the diagnostic conclusions drawn from the explainability analysis.

8. Conclusions

This paper presented a controlled empirical study of multi-stage retrieval for the scientific literature, evaluating lexical, dense, hybrid, and cross-encoder retrieval paradigms within a unified framework. Experimental results on SciFact demonstrate that dense models capture semantic similarity but remain insufficient when used in isolation for claim-based retrieval. Hybrid fusion via RRF improves performance for weaker dense retrievers such as SciBERT and SPECTER, but does not consistently outperform the strongest standalone dense model, SciNCL, whose rankings can be diluted when combined with a divergent lexical signal. Cross-encoder re-ranking emerges as the primary driver of final performance gains, refining the broader candidate pool produced by hybrid fusion through fine-grained query-document interaction. The ablation analysis provides additional insight. Lexical pseudo-relevance feedback (BM25 + RM3) degrades performance due to query drift in claim-focused retrieval settings, while passage-level max pooling weakens effectiveness for most dense models by fragmenting document-level evidence. In contrast, document-level dense encoding offers a more stable representation of scientific abstracts, and embedding-based query expansion outperforms its lexical counterpart. Cross-domain evaluation on SciFact, PubMedQA, and SciDocs demonstrates that, although absolute performance varies substantially with dataset characteristics, the relative ranking of retrieval paradigms remains stable across all three benchmarks. The RRF dilution effect intensifies on harder tasks such as SciDocs, and cross-encoder re-ranking only partially recovers the lost performance in that setting, underscoring the importance of candidate pool quality. Finally, structured error analysis reveals that remaining failures are primarily driven by semantic drift, lexical overlap bias, and query ambiguity, suggesting that current models still operate at the level of topical similarity rather than structured claim-evidence reasoning. Addressing these limitations will likely require incorporating explicit claim decomposition or entailment-aware scoring mechanisms in future retrieval architectures.

Author Contributions: Conceptualization, W.A.-J. and A.S.; methodology, W.A.-J. and A.S.; software, W.A.-J. and A.S.; validation, A.S. and H.H.; formal analysis, W.A.-J., A.S. and H.H.; investigation, W.A.-J., A.S. and H.H.; resources, W.A.-J.; data curation, W.A.-J.; writing—original draft preparation, W.A.-J.; writing—review and editing, A.S. and H.H.; visualization, W.A.-J., A.S. and H.H.; supervision, A.S.; project administration, A.S.; funding acquisition, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU260632].

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in SciFACT at https://huggingface.co/datasets/allenai/scifact (accessed on 13 January 2026).

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BEIR	Benchmarking Information Retrieval
BM25	Best Matching 25
BoW	Bag of Words
CE	Cross-Encoder
FAISS	Facebook AI Similarity Search
IR	Information Retrieval
LIME	Local Interpretable Model-agnostic Explanations
MAP	Mean Average Precision
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
RM3	Relevance Model 3
RRF	Reciprocal Rank Fusion
SciBERT	Scientific BERT
SciFact	Scientific Fact Verification Dataset
SciNCL	Scientific Contrastive Learning
SPECTER	Scientific Paper Embeddings using Citation-informed Transformers
TF-IDF	Term Frequency-Inverse Document Frequency

Appendix A

This appendix presents the complete error taxonomy covering all ten analyzed failure cases. As summarized in Table A1, each case reports the query, the correctly relevant document, the incorrectly top-ranked document, and the associated error type. This detailed analysis complements the results discussed in the main paper and supports the explainability findings.

Table A1. Illustrative examples of correct and incorrect retrieval outcomes categorized according to the proposed error taxonomy.

Query ID	Query	Correct Retrieved Document	Incorrect Retrieved Document	Error Type
1	0-dimensional biomaterials for tissue engineering applications	Nanotechnologies are emerging as promising platforms for the design of zero-dimensional biomaterials used in tissue regeneration and biomedical applications.	It has been proposed that epithelial–mesenchymal transition plays a critical role in cancer metastasis and tumor progression.	Semantic Drift
3	The 1000 Genomes Project enables mapping of human genetic variation	Genome-wide association studies enabled by the 1000 Genomes Project provide a comprehensive map of human genetic variation across populations.	Genome-wide studies have greatly expanded our understanding of common genetic variants without explicitly focusing on the 1000 Genomes reference framework.	Semantic Drift

Table A1. *Cont.*

Query ID	Query	Correct Retrieved Document	Incorrect Retrieved Document	Error Type
5	Abnormal prion protein positivity in UK population surveys	Objectives to carry out a further population-based survey to estimate abnormal prion protein prevalence in the United Kingdom.	Hemolysates of erythrocytes from more than 1000 donors were analyzed to study protein aggregation unrelated to prion diseases.	Lexical Overlap Bias
13	Perinatal mortality due to low birth weight in developing countries	One key target of the United Nations health agenda is reducing perinatal mortality through maternal and neonatal health interventions.	Objectives to explore the use of local civil registration data for demographic monitoring without specific focus on perinatal mortality.	Intent Mismatch
36	Vitamin B12 deficiency increases blood homocysteine levels	Lowering serum homocysteine levels through vitamin B12 supplementation has been shown to reduce cardiovascular risk.	Increased plasma homocysteine is associated with vascular disease independent of vitamin B12 deficiency mechanisms.	Lexical Overlap Bias
42	Microerythrocyte counts in hereditary blood disorders	Heritable haemoglobinopathies are characterized by abnormal erythrocyte morphology and reduced cell volume.	Frequency evaluation of genetic mutations affecting erythrocyte production in unrelated metabolic disorders.	Query Ambiguity
48	Asymptomatic individuals in population-based health screening	Population-based screening programs aim to identify asymptomatic individuals at early disease stages.	Body mass index is a major risk factor associated with metabolic syndrome and cardiovascular disease.	Topic Confusion
49	ADAR1 binding to Dicer during RNA processing	ADAR proteins interact with Dicer to regulate RNA editing and post-transcriptional gene regulation.	RNA methylation mechanisms regulate gene expression independently of ADAR-mediated RNA editing.	Vocabulary Mismatch
50	AIRE expression in tumor immune environments	Autoimmune regulator expression has been observed in tumor immune microenvironments affecting immune tolerance.	Intermediate filament proteins play a structural role in epithelial cell organization unrelated to immune regulation.	Domain Shift
51	ALDH1 expression as a cancer stem cell marker	ALDH1 has been identified as a functional marker of cancer stem cells in breast cancer progression.	PD-L1 is an immunoinhibitory molecule involved in immune checkpoint regulation in cancer therapy.	Entity Mismatch

References

1. Tang, M.; Chen, J.; Chen, H.; Xu, Z.; Wang, Y.; Xie, M.; Lin, J. An ontology-improved vector space model for semantic retrieval. *Electron. Libr.* **2020**, *38*, 919–942. [CrossRef]
2. Salsabilla, N.; Wiharja, K. Implementation of Semantic Search Based on Vector Database for Personal Documents. In Proceedings of the 2025 International Conference on Advancement in Data Science, E-Learning and Information System (ICADEIS), Bandung, Indonesia, 3–4 February 2025; pp. 1–6. [CrossRef]

3. Kumar Bevara, R.V.; Lund, B.D.; Mannuru, N.R.; Karedla, S.P.; Mohammed, Y.; Kolapudi, S.T.; Mannuru, A. Prospects of Retrieval-Augmented Generation (RAG) for Academic Library Search and Retrieval. *Inf. Technol. Libr.* **2025**, *44*, 1–15. [CrossRef]
4. Wadden, D.; Lin, S.; Lo, K.; Wang, L.L.; van Zuylen, M.; Cohan, A.; Hajishirzi, H. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 7534–7550.
5. Thakur, N.; Garg, N.; Wallace, E.; Roberts, A.; Raffel, C.; Dehghani, M. BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021.
6. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
7. Robertson, S.; Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
8. Kim, M.Y.; Rabelo, J.; Okeke, K.; Goebel, R. Legal Information Retrieval and Entailment Based on BM25, Transformer and Semantic Thesaurus Methods. *Rev. Socionetw. Strateg.* **2022**, *16*, 157–174. [CrossRef] [PubMed]
9. Lashkari, F.; Bagheri, E.; Ghorbani, A.A. Neural embedding-based indices for semantic search. *Inf. Process. Manag.* **2019**, *56*, 733–755. [CrossRef]
10. Yang, W.; Lu, K.; Yang, P.; Lin, J. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In Proceedings of the 42nd International ACM SIGIR Conference, Paris, France, 21–25 July 2019; pp. 1129–1132.
11. Kamil, M.; Çakır, D. Advances in Transformer-Based Semantic Search: Techniques, Benchmarks, and Future Directions. *Turk. J. Math. Comput. Sci.* **2025**, *17*, 145–166. [CrossRef]
12. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv* **2019**, arXiv:1903.10676. [CrossRef]
13. Cohan, A.; Feldman, S.; Beltagy, I.; Downey, D.; Weld, D. SPECTER: Document-Level Representation Learning Using Citation-Informed Transformers. *arXiv* **2020**, arXiv:2004.07180.
14. Ostendorff, M.; Rethmeier, N.; Augenstein, I.; Gipp, B.; Rehm, G. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. *arXiv* **2022**, arXiv:2202.06671. [CrossRef]
15. Johnson, J.; Douze, M.; Jégou, H. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* **2019**, *7*, 535–547. [CrossRef]
16. Zhang, K.; Tao, C.; Shen, T.; Xu, C.; Geng, X.; Jiao, B.; Jiang, D. LED: Lexicon-Enlightened Dense Retriever for Large-Scale Retrieval. *arXiv* **2022**, arXiv:2208.13661.
17. Cormack, G.V.; Clarke, C.L.A.; Buettcher, S. 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, Boston, MA, USA, 19–23 July 2009; pp. 758–759.
18. Pezzuti, F.; MacAvaney, S.; Tonellotto, N. Exploring the Effectiveness of Multi-stage Fine-tuning for Cross-encoder Re-rankers. *arXiv* **2025**, arXiv:2503.22672.
19. Sager, P.J.; Kamaraj, A.; Grewe, B.F.; Stadelmann, T. Deep Retrieval at CheckThat! 2025: Identifying Scientific Papers from Implicit Social Media Mentions via Hybrid Retrieval and Re-Ranking. *arXiv* **2025**, arXiv:2505.23250. [CrossRef]
20. Zhang, Y.; Long, D.; Xu, G.; Xie, P. HLATR: Enhance Multi-stage Text Retrieval with Hybrid List Aware Transformer Reranking. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 22–27 May 2022.
21. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019.
22. Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.E.; Lomeli, M.; Hosseini, L.; Jégou, H. The Faiss Library. *IEEE Trans. Big Data* **2026**, *12*, 346–361. [CrossRef]
23. Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; Zhou, M. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5776–5788.
24. Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; Wei, F. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv* **2022**, arXiv:2212.03533.
25. Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N. C-Pack: Packaged Resources to Advance General Chinese Embedding. *arXiv* **2023**, arXiv:2309.07597.
26. Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; Zhang, M. Towards General Text Embeddings with Multi-stage Contrastive Learning. *arXiv* **2023**, arXiv:2308.03281. [CrossRef]
27. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020.

28. Dai, Z.; Callan, J. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 985–988.
29. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 2567–2577.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.