SciRet: A Compute-Aware Empirical Study of Retrieval and Reranking for Scientific RAG

Anonymous ACL submission

Abstract

We introduce SciRet, a compute-aware empirical study of retrieval-augmented generation for scientific question answering over CORD-19. Rather than proposing a new model, we evaluate a fixed scientific RAG pipeline across three corpus scales: 1,034 chunks (1K papers), 5,160 chunks (5K papers), and 15,480 chunks (15K papers). The pipeline combines sentence-window chunking, BM25, BGE-M3 dense retrieval, reciprocal rank fusion, optional cross-encoder reranking, and grounded answer generation. Across these settings, hybrid retrieval is more robust than either sparse-only or dense-only retrieval in our setting, reaching Recall@10 of 1.000 at 1K and 15K. In contrast, an MS MARCO-trained cross-encoder reranker reduces precision on the scientific corpus, suggesting that domain mismatch can outweigh the benefits of stronger query-passage interaction. Generation faithfulness measured with RAGAS increases with corpus scale in our setup. Retrieval evaluation uses pseudo-relevance labels derived from the hybrid system, so we treat the results as controlled comparative evidence rather than a benchmark claim. We release code, indexes, and evaluation outputs to support replication and follow-up studies.

1 Introduction

Retrieval-augmented generation (RAG) is now a common architecture for knowledge-intensive question answering, but scientific literature remains a difficult setting. Scientific queries often depend on precise terminology, domain-specific evidence, and citation-grounded answers. A RAG pipeline that works well for web QA may not transfer cleanly to scientific text, especially when retrieval, reranking, and generation are trained or tuned on different domains.

This paper asks a focused empirical question: *how do standard retrieval and reranking components behave when the same scientific RAG pipeline is evaluated across increasing corpus scale?* We study CORD-19, holding preprocessing, chunking, embedding model, retrieval settings, and generation settings fixed across 1K, 5K, and 15K paper samples. This controlled design isolates scale effects and makes failure modes easier to inspect.

Our goal is not to claim a new state-of-the-art architecture. Instead, we report a compact, reproducible study of practical design choices for resource-constrained scientific RAG: BM25 versus dense retrieval, sparse-dense fusion, off-the-shelf cross-encoder reranking, and automated generation evaluation. Figure 1 summarizes the system. This is a focused empirical study with 15 evaluation queries per scale, intended to compare system behavior under controlled conditions rather than establish a benchmark.

Contributions. We make three contributions: (1) a controlled multi-scale evaluation of a fixed scientific RAG pipeline across 1K, 5K, and 15K CORD-19 paper samples; (2) an empirical analysis showing that BM25+BGE-M3 fusion is more robust than either component alone in our setting; and (3) a negative reranking result: an MS MARCO-trained cross-encoder reduces precision on scientific text, highlighting domain mismatch as a practical risk.

2 Related Work

Scientific RAG and QA. RAG combines retrieval with generation for knowledge-intensive NLP (Lewis et al., 2020). Scientific and biomedical QA benchmarks such as BioASQ (Tsatsaronis et al., 2015), SciFact (Wadden et al., 2020), COVID-QA (Möller et al., 2020), and CORD-19 (Wang et al., 2020) emphasize that retrieval quality and evidence grounding are central to scientific question answering. SciRet focuses on the retrieval and reranking behavior of a practical scientific RAG pipeline rather than on training a new generator.

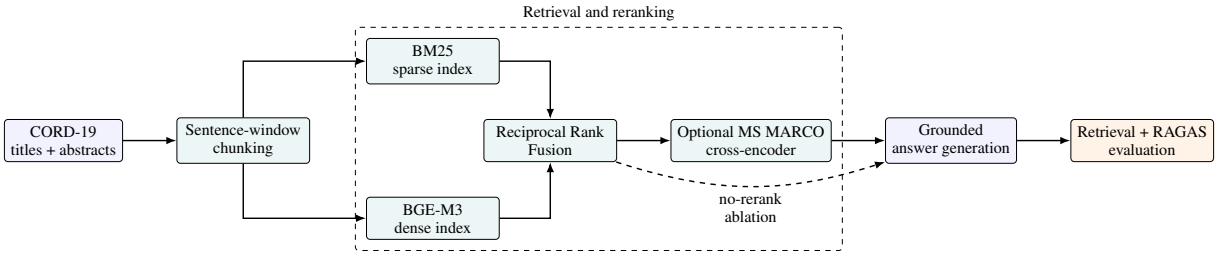

Figure 1: SciRet pipeline evaluated in this paper. Text is chunked once and indexed for both sparse and dense retrieval. Ranked lists are fused with reciprocal rank fusion; an off-the-shelf cross-encoder reranker is evaluated as an ablation before grounded generation and evaluation. The diagram is rendered as vector graphics to avoid raster blurring in the submission PDF.

Sparse, dense, and hybrid retrieval. BM25 remains a strong lexical baseline for scientific text because exact terms, abbreviations, and biomedical entities matter (Robertson and Zaragoza, 2009). Dense retrievers such as DPR (Karpukhin et al., 2020), Sentence-BERT (Reimers and Gurevych, 2019), SPECTER (Cohan et al., 2020), and BGE-M3 (Chen et al., 2024) support semantic matching beyond lexical overlap. Hybrid retrieval is often effective because sparse and dense systems retrieve complementary evidence; reciprocal rank fusion (RRF) is a simple and robust fusion method (Cormack et al., 2009). BEIR further shows that retrieval behavior can vary substantially across domains and datasets (Thakur et al., 2021).

Reranking and evaluation. Cross-encoders can improve ranking by jointly encoding query-passage pairs, but transfer depends on training data. We test an MS MARCO-trained reranker in a scientific setting and find that it hurts precision. For generation, we use RAGAS (Es et al., 2023) as an automated comparative signal, while recognizing that LLM-based evaluation can inherit judge-model biases. Claim-level factuality metrics such as FActScore (Min et al., 2023) motivate future citation-integrity evaluation.

3 Method

3.1 Corpus and Scale Protocol

We evaluate on CORD-19 (Wang et al., 2020). To keep the study compute-aware and reproducible, we index titles and abstracts rather than full text. This reduces storage and embedding cost, but limits evidence depth; we treat this as a limitation rather than a complete scientific-document solution.

We construct three corpus scales while keeping all settings fixed: 1K papers, 5K papers, and 15K papers. The resulting chunk counts are 1,034, 5,160, and 15,480, respectively (Table 1). All experiments use the same sentence-window chunking strategy, embedding model, retrieval cutoffs, RRF parameter, generation prompt style, and evaluation scripts.

Table 1: Dataset statistics for the three evaluation scales.

Statistic	1K	5K	15K
Papers indexed	1,000	5,000	15,000
Text chunks	1,034	5,160	15,480
Mean chunk tokens	215	215	215
Evaluation queries	15	15	15

3.2 Retrieval and Reranking

We compare three retrieval systems: **Dense**: BGE-M3 embeddings with vector search; **BM25**: sparse lexical retrieval; and **Hybrid**: reciprocal rank fusion of the dense and BM25 ranked lists. RRF scores a document d as:

$$\mathrm{RRF}(d) = \sum_i \frac{1}{60 + \mathrm{rank}_i(d)}. \tag{1}$$

Stage 1 retrieves 50 candidates from each retriever before fusion.

We then evaluate an optional cross-encoder reranker, `cross-encoder/ms-marco-MiniLM-L-6-v2`. Because this reranker is trained on web-search data, its performance on scientific abstracts is an empirical question rather than a guaranteed improvement.

3.3 Generation and Metrics

Retrieved passages are assembled into a grounded prompt for GPT-4o-mini, which is instructed to answer only from retrieved context and cite passages by index. We report retrieval Recall@K for $K \in \{1, 3, 5, 10, 20\}$ and Precision@K for reranking ablations. Generation is evaluated with

Table 2: Recall@K across scales. With three pseudo-relevant documents per query, the maximum possible R@1 is $1/3 = 0.333$. Best value per scale and cutoff is bolded.

Scale	System	R@1	R@3	R@5	R@10	R@20
1K	Dense	**0.333**	**0.720**	0.720	0.767	0.793
	BM25	**0.333**	0.513	0.573	0.667	0.693
	Hybrid	0.267	0.627	**0.820**	**1.000**	**1.000**
5K	Dense	**0.333**	**0.733**	0.740	0.807	0.827
	BM25	0.327	0.513	0.573	0.627	0.687
	Hybrid	0.260	0.613	**0.847**	**0.993**	**1.000**
15K	Dense	0.147	0.400	0.560	0.787	0.920
	BM25	0.180	0.387	0.560	0.747	0.847
	Hybrid	**0.333**	**1.000**	**1.000**	**1.000**	**1.000**

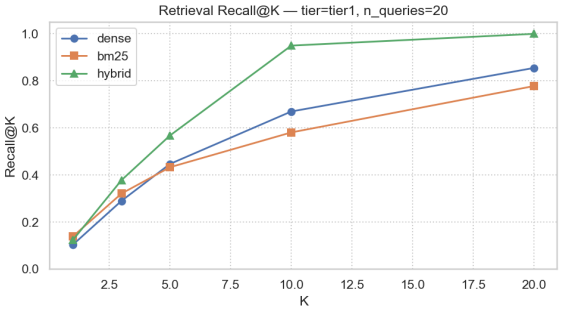

Figure 2: Recall curves at the 1K scale. Hybrid retrieval improves over either component alone at larger cutoffs.

RAGAS faithfulness, answer relevancy, context precision, and context recall.

3.4 Retrieval Evaluation Note

Retrieval evaluation uses pseudo-relevance labels: the top-3 hybrid results per query are treated as relevant. This makes our results useful for controlled comparison and debugging, but it introduces circularity that may favor the hybrid system, and a 15-query set limits statistical power. All retrieval figures should therefore be read as controlled comparative evidence rather than benchmark scores. We return to this limitation in Section 6.

4 Results

4.1 Hybrid Retrieval Is Most Robust

Table 2 reports Recall@K across scales. Hybrid retrieval dominates at $K \geq 5$ and reaches Recall@10 of 1.000 at 1K and 15K. Dense retrieval is competitive at a small scale, while BM25 remains competitive at R@1, especially at 15K. Figure 2 visualizes the recall curves at the 1K scale.

Table 3: Precision@K before and after cross-encoder reranking.

Scale	System	P@1	P@3	P@5	P@10
1K	No rerank	1.000	1.000	0.600	0.300
	Reranked	0.680	0.527	0.404	0.234
5K	No rerank	1.000	1.000	0.600	0.300
	Reranked	0.620	0.467	0.368	0.250

Table 4: RAGAS generation quality across scales.

Scale	Faith.	Ans. Rel.	Ctx. Prec.	Ctx. Rec.
1K	0.917	0.680	0.095	0.260
5K	0.940	0.703	0.108	0.240
15K	0.960	0.870	0.122	0.100

4.2 Generic Reranking Hurts Precision

Table 3 shows that the MS MARCO-trained cross-encoder reduces precision at all reported cutoffs. At 1K, P@5 drops from 0.600 to 0.404. At 5K, it drops from 0.600 to 0.368. Since the no-rerank baseline is identical at both scales, the degradation is attributable to reranking rather than to weaker Stage 1 retrieval.

4.3 Generation Scores Improve With Scale

Table 4 reports RAGAS generation metrics. Faithfulness increases from 0.917 at 1K to 0.960 at 15K, and answer relevancy increases from 0.680 to 0.870. Context precision remains low, indicating that retrieved contexts include topic-related but not always tightly targeted passages.

5 Discussion

The results support two practical lessons. First, sparse-dense fusion is a strong default for scientific RAG under limited compute: BM25 and BGE-M3 retrieve complementary evidence, and RRF is simple enough to run across scales. Second, reranking is not automatically beneficial. The MS MARCO cross-encoder harms ranking on our scientific corpus, suggesting that domain-adapted reranking should be tested before deployment.

The study also shows why compute-aware evaluation matters. The 1K scale is useful for debugging, but some retrieval behavior changes at 15K. Small-scale experiments should therefore be treated as development checks rather than final evidence.

6 Limitations

This paper has important limitations. First, we index only titles and abstracts, not full papers, figures,

or tables. Second, retrieval evaluation uses pseudo-relevance labels derived from hybrid top-3 results, which can favor hybrid retrieval and cannot replace independent relevance annotation. Third, the evaluation set contains 15 queries, limiting statistical power. Fourth, RAGAS provides an automated signal but not a substitute for human or claim-level citation verification. We frame SciRet as a reproducible empirical study and a basis for stronger follow-up work, not as a complete scientific QA benchmark.

7 Conclusion

We presented SciRet, a compute-aware empirical study of retrieval and reranking for scientific RAG over CORD-19. Across 1K, 5K, and 15K paper samples, BM25+BGE-M3 fusion is more robust than either component alone in our setting, while an MS MARCO-trained cross-encoder reranker consistently reduces precision. These results suggest that standard RAG components should be tested carefully in scientific QA rather than assumed to transfer from web search settings. Hybrid retrieval appears to be a strong default under limited compute, but the retrieval evaluation here is based on pseudo-relevance labels and a small query set, so the findings should be read as controlled empirical evidence rather than final benchmark results.Future work should prioritize independent relevance annotation to remove the pseudo-label circularity, full-document evidence beyond titles and abstracts, domain-adapted rerankers trained on scientific text, and claim-level citation verification to ground generated answers in verifiable evidence.

Ethics Statement

SciRet is evaluated on scientific literature and does not introduce new human-subject data. Because scientific QA systems can influence user interpretation of biomedical evidence, generated answers should not be used for medical decision-making without expert review. The current system uses title and abstract evidence only and may omit important full-text context.

Generative AI Disclosure. The authors utilized Claude and ChatGPT to assist in refining prose, improving readability, and copyediting select portions of this manuscript. The core ideas, technical methodology, and experimental evaluations were fully developed by the authors. The authors assume full responsibility for the final content.

Acknowledgments

Omitted for anonymous review.

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. 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*, pages 758–759.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference*

on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7534–7550.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Baber, Kathryn Eide, Brendan Ros, Nascence Kim, Spencer Wilhelm, and 1 others. 2020. CORD-19: The COVID-19 open research dataset. *arXiv preprint arXiv:2004.10706*.

A Appendix Overview

The appendix preserves supporting evidence that is useful for review but not essential to the short-paper narrative: chunking analysis, retrieval complementarity, embedding sanity checks, question-set details, compute budget, and supplementary visualizations.

B Chunking Analysis

We compared fixed-character, fixed-token, sentence-window, and paragraph-based chunking at the 1K scale. Sentence-window chunking achieved the strongest boundary cleanliness and was used across all scales.

Table 5: Chunking strategy comparison on the 1K corpus.

Strategy	Chunks	Tokens	Clean %
Fixed character	629	182.4	69.2
Fixed token	535	210.3	81.9
Sentence window	519	215.8	85.4
Paragraph	500	222.0	84.8

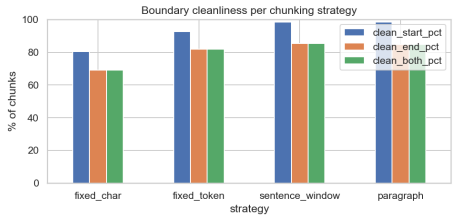

Figure 3: Boundary cleanliness by chunking strategy.

C Retrieval Complementarity

At the 1K scale, the mean Jaccard similarity between dense and BM25 top-10 result sets is 0.213, meaning roughly 79% of retrieved documents differ between the two methods. This supports the use of rank fusion.

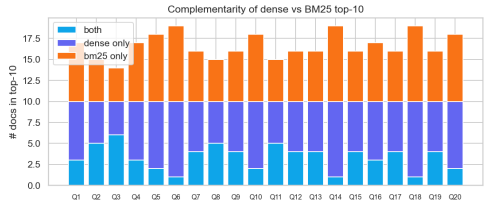

Figure 4: Overlap between dense and BM25 top-10 retrieval results across queries.

D Embedding Sanity Check

BGE-M3 related-passage similarities exceed unrelated-passage similarities by 0.16 to 0.41 cosine units across biomedical term, symptom, and treatment queries.

Table 6: Cosine similarity sanity check for BGE-M3 embeddings.

Query type	Related	Unrelated	Margin
Biomedical term	0.81	0.65	+0.16
Symptom-based	0.79	0.44	+0.35
Treatment	0.83	0.42	+0.41

E Reranker Rank Movement

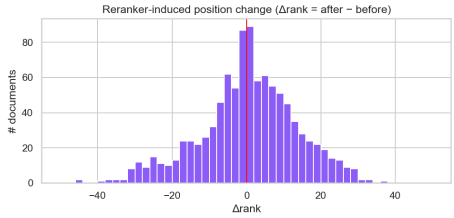

Figure 5: Distribution of reranker-induced rank shifts at the 1K scale.

F Evaluation Questions

The 15 evaluation questions cover clinical, epidemiological, and treatment themes in CORD-19. Examples include: "What are the primary transmission routes of SARS-CoV-2?", "Which comorbidities increase COVID-19 mortality risk?", and "What is the efficacy of remdesivir in treating COVID-19?" We release the question set, pseudo-relevance labels, and raw result files with the project.

G Compute Budget

Table 7: Approximate wall-clock time by pipeline stage.

Stage	1K	5K	15K
Chunking	<1 min	3 min	8 min
Embedding	18 min	25 min	73 min
BM25 index	<1 min	<1 min	2 min
Retrieval	<1 min	<1 min	<1 min
Reranking	4 min	12 min	N/A
Generation	8 min	8 min	8 min