Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Measuring and Closing the Retrieval Gap in Financial Question Answering

Amine Kobeissi[†,‡,*], Philippe Langlais[†,‡]
[†] Université de Montréal
[‡] RALI

Abstract

Retrieval-augmented generation (RAG) is increasingly applied to financial question answering over long regulatory documents, yet evaluations typically measure only chunk-level retrieval or end-to-end answer quality, leaving a systematic understanding of where and why pipelines fail out of reach. We introduce an oracle-based evaluation framework that decomposes retrieval performance into document, page, and chunk discovery, providing empirical upper bounds at each granularity and exposing a consistent retrieval gap that persists even when the correct document is found. We systematically evaluate several retrieval strategies on 150 FinanceBench questions, spanning dense, sparse, hybrid, hierarchical, query reformulation, and reranking methods using a shared multi-document index. Our analysis shows that while methods such as Multi-HyDE and cross-encoder reranking improve document recall, page-level retrieval substantially lags behind oracle bounds across all baselines. We further break down performance by question type and document type, revealing that retrieval difficulty varies significantly across these dimensions and that no single strategy closes the gap uniformly. As a targeted intervention, we introduce a domain fine-tuned page scorer that ranks pages before chunk retrieval, achieving strong gains under cross-validation, suggesting that domain-specific and page-level modeling is a promising direction.

Keywords: Retrieval-Augmented Generation, Financial Question Answering, Information Retrieval, Evaluation Framework, Large Language Models

1. Introduction

Financial question answering over financial documents poses a distinct challenge for retrieval-augmented generation (RAG) [1]. Documents are long, repetitive, semi-structured, and numerically sensitive, while answers often hinge on specific tables or statements buried within hundred-page reports. In this setting, retrieval quality is the primary determinant of answer correctness, yet it is rarely measured carefully.

Most prior work on financial RAG evaluates retrieval at a single granularity, typically chunk-level recall, or relies on end-to-end answer quality as a proxy [2, 3]. Neither provides the diagnostic depth needed to understand why and where systems fail. Since generative quality is directly conditioned on what is retrieved, a drop in answer accuracy could reflect poor document-level coverage, failure to locate the right page within the correct document, imprecise chunk selection, or a generation error given otherwise adequate context. Without decomposing retrieval across these levels, it is difficult to attribute failures, compare methods meaningfully, or know which component of the pipeline most needs improvement.

We address this with a systematic evaluation study centered on three contributions. First, an oracle-based evaluation framework that decomposes retrieval into document, page, and chunk discovery, providing empirical upper bounds at each level and a diagnostic framework for any RAG pipeline. Second, a systematic comparison of several retrieval strategies, dense, sparse, hybrid, hierarchical, query rewriting, and reranking, evaluated jointly on document, page, and chunk metrics on a shared index, making results directly comparable across methods. Third, a fine-grained performance breakdown by question type and

[*] amine.kobeissi@umontreal.ca

document type, showing that retrieval difficulty is not uniform and that different strategies have distinct strengths and failure modes. Beyond the evaluation, we introduce a domain fine-tuned page scorer as a proposed method and show that explicit domain and page-level modeling yields further gains. Initial results are promising, though the small dataset and cross-validation setup mean that generalization claims require evaluation on held-out data and larger benchmarks.

2. Task Setup and Evaluation Framework

2.1. Dataset

We use the open-source 150-question subset of FinanceBench [4], which provides gold document annotations, gold page numbers, and reference answers for questions over financial documents. The dataset covers three question types of 50 questions each: domain-relevant, metrics-generated, and novel-generated. It draws from four document types: 10-K (74.7%), 10-Q (10%), earnings call (EC) transcripts (9.3%), and 8-K (6%). The skew toward 10-K document is an important caveat for interpreting aggregate results.

2.2. Oracle Retrieval Conditions

The core of our evaluation framework is a set of three retrieval conditions defined by progressively restricting the candidate search space. Standard retrieval operates over the full corpus. Oracle-Doc restricts candidates to the gold document $d^\star$, simulating perfect document discovery. Oracle-Page further restricts candidates to the gold pages $P^\star$, simulating perfect page localization.

These conditions provide empirical upper bounds for our experimental setup, where Oracle-Doc quantifies headroom from imperfect page and chunk discovery, and Oracle-Page quantifies headroom from imperfect chunk selection given the correct pages.

2.3. Evaluation Metrics

For retrieval we report document recall DocRec@k, page recall PageRec@$k = |P^\star \cap \{p_i \mid d_i = d^\star\}_{i=1}^{k}|/|P^\star|$, and chunk-level maximum BLEU [5] and ROUGE-L [6] against the concatenated gold evidence. For generation, we report ROUGE-L on predicted answers against expected answers, and a numeric match metric on the 50 metrics-generated questions, using $\pm 3\%$ absolute/relative tolerance. These automatic generation metrics are a practical proxy but are limited since they only measure lexical overlap without capturing semantic equivalence or factual correctness. All results use $k = 5$.

3. Retrieval Methods

Documents are chunked into overlapping spans of 1024 tokens with 128-token overlap [7]. Each chunk (c, d, p) carries its source document d and page number p. For a query q, a retriever returns the top-k chunks $\mathcal{R}_k(q) = \{(c_i, d_i, p_i)\}_{i=1}^{k}$, and the generator produces $\hat{y} = G_\theta(q, \mathcal{R}_k(q))$ using Qwen-2.5-7B-Instruct with fixed prompting and decoding. All methods operate over this shared index.

We evaluate several retrieval strategies. Dense retrieval uses BGE-M3 [8]. Sparse methods include BM25 [9] with finance-oriented tokenization and SPLADE [10]. Hybrid fusion combines BM25 and BGE-M3 via Reciprocal Rank Fusion [11]. A Parent-Child hierarchy indexes fine-grained child chunks and maps retrievals to larger parent spans [12]. HyDE and Multi-HyDE [3, 13] generate hypothetical passages to bridge lexical mismatch, using Qwen-2.5-7B-Instruct at temperature 0.7. Cross-encoder reranking uses `BAAI/bge-reranker-v2-m3`

over 20 first-stage candidates. Finally, we include a domain fine-tuned page scorer as described below.

To directly target the within-document retrieval gap, we fine-tune a bi-encoder E_{θ_p} initialized from BGE-M3 to score page-level relevance. Pages are scored by cosine similarity $s_{\text{page}}(q, d, p) = \cos(E_{\theta_p}(q), E_{\theta_p}(\mathcal{N}(d, p)))$, where $\mathcal{N}$ normalizes page text. The top $P = 20$ pages are selected and chunk retrieval is restricted to those pages. Training uses Multiple Negatives Ranking loss with in-batch negatives [14], pairing each question with its gold pages, and document-level 5-fold cross-validation to prevent leakage.

4. Results and Analysis

4.1. Results Across All Methods

Table 1 presents the main comparison. We observe consistent patterns across all methods. First, the retrieval gap is present, across all methods, page recall substantially lags document recall, confirming that retrieving the correct document does not guarantee finding the right page or chunk. Dense retrieval (BGE-M3) outperforms sparse methods, consistent with prior work [15], owing to lexical mismatch between conversational queries and formal document language. BM25 in particular achieves only 0.32 document recall, showing that keyword matching alone is insufficient for document retrieval in this domain.

Second, query reformulation and reranking help but do not close the gap. Multi-HyDE and cross-encoder reranking progressively improve both page recall and generative performance, with the best combined pipeline reaching a page recall of 0.46 and numeric match of 0.38, yet page recall remains 0.14 below the Oracle-Doc bound of 0.60. This gap persists despite the pipeline achieving near-perfect document recall (0.93), underscoring that within-document retrieval is the primary bottleneck for current methods. The page scorer closes this gap further to a page recall of 0.55 and numeric match of 0.50, surpassing even Oracle-Doc (0.44). We surpass Oracle-Doc because it does not optimize for ranking. The oracle bounds confirm substantial remaining headroom, Oracle-Page reaches a numeric match of 0.70, suggesting that perfect page retrieval would imrpove performance of the best baseline.

	Retrieval				Generative	
Method	**DocRec**	**PageRec**	**BLEU**	**ROUGE-L**	**ROUGE-L**	**Num. Match**
BM25	0.32	0.07	0.04	0.12	0.05	0.04
SPLADE	0.81	0.31	0.24	0.36	0.06	0.22
Dense (BGE-M3)	0.88	0.34	0.26	0.35	0.06	0.24
Hybrid (BM25 + BGE-M3)	0.61	0.23	0.13	0.27	0.09	0.20
Parent-Child	0.91	0.32	0.23	0.34	0.09	0.25
Dense + Multi-HyDE	0.85	0.42	0.27	0.39	0.10	0.32
Dense + ReRanker	0.87	0.41	0.19	0.34	0.11	0.32
Dense + Multi-HyDE + ReRanker	0.93	0.46	0.28	0.40	0.12	0.38
Learned Page Scorer (Ours)	0.95	0.55	0.33	0.46	0.15	0.50
Oracle-Doc	1.00	0.60	0.25	0.42	0.13	0.44
Oracle-Page	1.00	1.00	0.40	0.59	0.16	0.70

Table 1. Retrieval and Generative Results at k=5.

4.2. Breakdown by Question and Document Type

Figure 1 breaks down document and page recall by question type, revealing that the retrieval gap is not uniform. On metrics-generated questions, most methods achieve near-perfect document recall and strong page recall, reflecting that these questions target financial tables in structurally predictable locations. On domain-relevant and novel questions, the gap between document recall and page recall is wider, as these require locating narrative context

that is harder to distinguish from surrounding boilerplate. This suggests that different question types call for different retrieval strategies, and that aggregate numbers can mask significant variation.

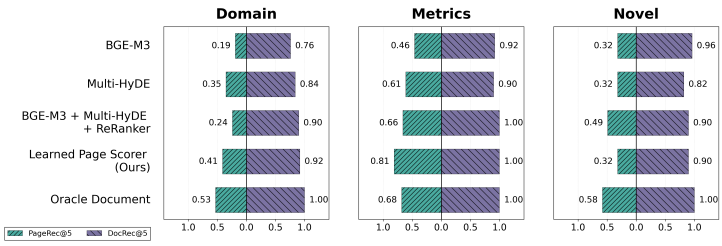

Figure 1. Document and page recall at $k = 5$ by question type (50 questions per type).

Table 2 reports performance by document type and reinforces this point. On 10-K documents, which make up 75% of the data, the page scorer achieves page recall of 0.62, exceeding both the best baseline (0.45) and Oracle-Doc (0.56), suggesting that domain-tuned page embeddings generalize well within the structured format of annual reports. On earnings call transcripts, however, the page scorer drops sharply to 0.10 against an Oracle-Doc bound of 0.64, while the best baseline reaches 0.36. The conversational and unstructured nature of earnings calls is poorly matched to a model trained predominantly on 10-K documents, and this effect may be further amplified by class imbalance. Performance on 10-Q and 8-K documents similarly underperforms relative to oracle bounds, pointing to both distribution shift and data imbalance as key limitations. These results highlight that document type is a critical axis of evaluation that aggregate benchmarks tend to obscure.

	10K (n=112)		**10Q (n=15)**		**8K (n=9)**		**EC (n=14)**	
Method	**D@5**	**P@5**	**D@5**	**P@5**	**D@5**	**P@5**	**D@5**	**P@5**
Dense + Multi-HyDE + ReRanker	0.96	0.45	0.87	0.47	0.89	0.78	0.86	0.36
Learned Page Scorer (Ours)	0.97	0.62	0.87	0.40	0.89	0.56	0.86	0.10
Oracle-Doc	1.00	0.56	1.00	0.60	1.00	0.89	1.00	0.64

Table 2. Document (D@5) and page (P@5) recall at $k = 5$ by document type.

5. Conclusion

We presented a systematic evaluation of retrieval strategies for financial question answering organized around an oracle-based framework that decomposes retrieval failures into document discovery, page localization, and chunk-level retrieval. Across several methods on FinanceBench, page recall consistently lags document recall across all baselines, and oracle analysis confirms that closing this within-document gap would yield meaningful generative gains. Performance varies by question type and document type, suggesting that aggregate benchmarks can obscure important failure patterns and that no single retrieval strategy consistently performs best across all settings. These findings suggest that improving within-document retrieval is the key bottleneck for reliable financial RAG systems.

The dataset is small and skewed toward 10-K documents, limiting generalization, and the page scorer requires a more rigorous evaluation on held-out external data. Generative evaluation relies on automatic metrics that capture lexical overlap and numerical precision but lack the context needed to assess factual correctness. Future work includes retraining the page scorer on external financial corpora, a systematic study of chunking strategies, and applying the oracle diagnostic framework to larger and more diverse financial benchmarks.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474.

[2] X. Wang, J. Chi, Z. Tai, T. S. T. Kwok, H. He, Z. Li, Y. Hua, M. Li, P. Lu, S. Wang, et al. "Finsage: A multi-aspect rag system for financial filings question answering". In: *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 2025, pp. 6144–6152.

[3] R. George, A. G. Srinivasan, J. K. Joe, H. MR, H. Kant, R. Vimalkanth, S. Suresh, et al. "Enhancing Financial RAG with Agentic AI and Multi-HyDE: A Novel Approach to Knowledge Retrieval and Hallucination Reduction". In: *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*. 2025, pp. 19–32.

[4] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen. *FinanceBench: A New Benchmark for Financial Question Answering*. 2023. arXiv: 2311.11944 [cs.CL].

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[6] C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[7] A. J. Yepes, Y. You, J. Milczek, S. Laverde, and R. Li. *Financial Report Chunking for Effective Retrieval Augmented Generation*. 2024. arXiv: 2402.05131 [cs.CL].

[8] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. *M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. 2025. arXiv: 2402.03216 [cs.CL].

[9] S. E. Robertson and H. Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Found. Trends Inf. Retr.* 3 (2009), pp. 333–389.

[10] T. Formal, B. Piwowarski, and S. Clinchant. "SPLADE: Sparse lexical and expansion model for first stage ranking". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 2288–2292.

[11] G. V. Cormack, C. L. Clarke, and S. Buettcher. "Reciprocal rank fusion outperforms condorcet and individual rank learning methods". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, pp. 758–759.

[12] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, et al. "Retrieval-augmented generation for large language models: A survey". In: *arXiv preprint arXiv:2312.10997* 2.1 (2023), p. 32.

[13] L. Gao, X. Ma, J. Lin, and J. Callan. "Precise zero-shot dense retrieval without relevance labels". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 1762–1777.

[14] M. Henderson, R. Al-Rfou, B. Strope, Y. hsuan Sung, L. Lukacs, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil. *Efficient Natural Language Response Suggestion for Smart Reply*. 2017. arXiv: 1705.00652 [cs.CL].

[15] S. Kim, H. Song, H. Seo, and H. Kim. *Optimizing Retrieval Strategies for Financial Question Answering Documents in Retrieval-Augmented Generation Systems*. 2025. arXiv: 2503.15191 [cs.IR].