OPTIMIZING RETRIEVAL STRATEGIES FOR FINANCIAL QUESTION ANSWERING DOCUMENTS IN RETRIEVAL-AUGMENTED GENERATION SYSTEMS

Sejong Kim*, Hyunseo Song*, Hyunwoo Seo*, Hyunjun Kim*†

KAIST
Daejeon, South Korea
{kingsj,ssongzzz,shw4166,hyunjun1121}@kaist.ac.kr

ABSTRACT

Retrieval-Augmented Generation (RAG) has emerged as a promising framework to mitigate hallucinations in Large Language Models (LLMs), yet its overall performance is dependent on the underlying retrieval system. In the finance domain, documents such as 10-K reports pose distinct challenges due to domain-specific vocabulary and multi-hierarchical tabular data. In this work, we introduce an efficient, end-to-end RAG pipeline that enhances retrieval for financial documents through a three-phase approach: pre-retrieval, retrieval, and post-retrieval. In the pre-retrieval phase, various query and corpus preprocessing techniques are employed to enrich input data. During the retrieval phase, we fine-tuned stateof-the-art (SOTA) embedding models with domain-specific knowledge and implemented a hybrid retrieval strategy that combines dense and sparse representations. Finally, the post-retrieval phase leverages Direct Preference Optimization (DPO) training and document selection methods to further refine the results. Evaluations on seven financial question answering datasets—FinDER, FinQABench, FinanceBench, TATQA, FinQA, ConvFinQA, and MultiHiertt—demonstrate substantial improvements in retrieval performance, leading to more accurate and contextually appropriate generation. These findings highlight the critical role of tailored retrieval techniques in advancing the effectiveness of RAG systems for financial applications. A fully replicable pipeline is available on GitHub: https://github.com/seohyunwoo-0407/GAR.

1 Introduction

Ensuring accuracy and reliability is paramount in the financial domain. Even minor errors or misinterpretations within financial statements or regulatory filings can trigger significant economic losses and adversely impact investment decisions and compliance processes. (Rezaee, 2005) Motivated by these high-stakes challenges, this paper presents a specialized framework that integrates Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) tailored to the unique challenges of financial data. While LLMs have demonstrated impressive performance in natural language processing, they are not immune to limitations—particularly in specialized domains where hallucinations and misinterpretations can have critical issues. (Kerner, 2024).

RAG is designed to mitigate these shortcomings by incorporating an additional retrieval step into the generation process. Rather than relying solely on pre-trained knowledge, the system first retrieves pertinent information from external sources and subsequently integrates this data during the generation phase. This two-stage approach ensures that the LLM has access to the most relevant information before generating its response. By leveraging external knowledge sources, RAG significantly enhances performance on domain-specific tasks.

In this paper, we propose a novel RAG pipeline optimized for usage in the finance domain. Our key contributions span robust data preparation, domain adaptation, and effective information retrieval.

^{*}These authors contributed equally

[†]Corresponding author

First, we compared tailored **preprocessing methods** for preparing both user inputs and financial documents, ensuring our system effectively captures and utilizes context from complex and diverse data sources. Next, we utilized **task-specific retrieval methods** that leverages SOTA models to the unique language and structural features of financial information, thus improving retrieval performance. Finally, we implemented a **reranking method**, coupled with a novel method (**document selection**), to guarantee that the final generated responses are grounded in the most accurate and relevant data.

2 RELATED WORK

2.1 Embedder Fine-tuning

Fine-tuning, the process of adapting a pre-trained model to domain-specific tasks using typically smaller datasets, has been widely explored across various applications. While embedding models exhibit strong zero-shot performance on general benchmarks such as MTEB (Muennighoff et al., 2023; Zhang et al., 2017), recent studies have demonstrated that even modestly sized models can benefit substantially from fine-tuning when applied to domain-specific tasks. For instance, fine-tuning embedders on specialized datasets has led to notable improvements in areas such as medical question answering Sengupta et al. (2024) and financial question answering Anderson et al. (2024). In the finance domain, prior research on embedders has underscored several inherent challenges: domain-specific vocabulary and semantic patterns, the complexity of multi-hop queries, and multimodal data (e.g. text, tables, and time-series) Tang & Yang (2024); Kim et al. (2024); Xie et al. (2024). These challenges necessitate tailored fine-tuning strategies that can effectively capture the nuanced information contained in financial documents.

Within the framework of Retrieval-Augmented Generation (RAG), embedding models are primarily tasked with Information Retrieval (IR), where the semantic similarity between a query and a corpus is assessed and ranked. A prevalent strategy for enhancing this process is contrastive learning or contrastive fine-tuning—which relies on constructing triplets (query, relevant corpus) to form positive and negative training pairs Karpukhin et al. (2020). Despite the effectiveness of contrastive learning in embedders Lu et al. (2024), there remains a notable gap in the literature regarding the impact of embedder fine-tuning on RAG systems, particularly within the finance domain Setty et al. (2024).

By addressing this gap, our work aims to explore and quantify the benefits of embedder fine-tuning in RAG applications, thereby contributing to the broader understanding of domain-adaptive IR.

2.2 PRE-RETRIEVAL

Effective dataset preprocessing is essential for enhancing the performance of Retrieval-Augmented Generation (RAG) systems, because it directly impacts the clarity and semantic alignment of both queries and documents (Gao et al., 2024). Previous work has shown that short, context-poor queries can lead to significant ambiguity, which in turn hinders retrieval accuracy (Koo et al., 2024). To mitigate this issue, researchers have explored various query enhancement techniques—such as query expansion and rephrasing—to enrich the original input and better capture user intent (Patel, 2024).

In line with these findings, we evaluate **query preprocessing methods**—from raw queries and keyword extraction with linguistic simplification to LLM-based query expansion—and show that adding contextual information significantly improves retrieval performance.

In addition to query enhancement, the heterogeneity of corpus data presents unique challenges that necessitate tailored preprocessing strategies. Unlike approaches that apply a uniform treatment to all documents, recent studies have emphasized the importance of adapting preprocessing to the structural characteristics of the corpus—especially when dealing with diverse formats such as plain text and tabular data.

Building on these insights, we explore various **corpus preprocessing methods**—preserving original formats markdown restructuring, table annotation, and table extraction. Our evaluation shows that markdown restructuring yields the best performance, highlighting the benefits of an optimized, data-driven pipeline. Overall, our framework addresses query ambiguity and document heterogeneity, improving retrieval accuracy in domain-specific applications.

2.3 Hybrid retrieval

Recent advancements in Retrieval-Augmented Generation (RAG) systems have increasingly focused on overcoming the limitations of using a single retrieval modality by fusing the strengths of both dense and sparse retrieval approaches. Dense retrieval methods, which leverage semantic embeddings generated by models such as BERT(Devlin et al., 2019) or SentenceTransformers(Reimers & Gurevych, 2019), excel at capturing deep contextual relationships between queries and documents. However, they may sometimes fail to retrieve documents that contain precise terms, proper nouns, or abbreviations. In contrast, sparse retrieval techniques, employing methods like BM25(Wang et al., 2021), offer excellent keyword matching capabilities and provide high interpretability, although they often lack the ability to grasp nuanced semantic meaning. (Sengupta et al., 2024)

To address these complementary weaknesses, hybrid retrieval methods have been proposed. These methods combine the scores obtained from dense and sparse retrieval, typically through linear weighted fusion or techniques such as Reciprocal Rank Fusion (RRF). For example, the Sawarkar et al. (2024) demonstrates that integrating semantic search techniques with sparse encoder indexes can significantly enhance retrieval performance on benchmarks such as NQ and TREC-COVID, leading to improved overall accuracy in RAG systems(Zhang et al., 2024).

3 METHODOLOGIES TO IMPROVE RETRIEVAL

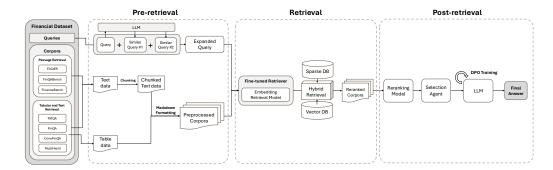


Figure 1: Our enhanced RAG pipeline

3.1 Pre-Retrieval

Data Preprocessing A major shortcoming of conventional RAG approaches is their inability to effectively handle ambiguous queries and the heterogeneous structure of complex documents, leading to issues such as ambiguity and structural fragmentation. Thus, preprocessing must enrich semantic context while reducing noise. Commonly used techniques such as query expansion(Wang et al., 2023), LLMLingua(Jiang et al., 2023) and markdown formatting can significantly enhance retrieval performance. Additionally, tabular data may require specialized modification to be effectively grounded(Singh & Bedathur, 2023). In this phase, various preprocessing methods are considered including novel modifications to *both query and corpus*.

3.2 Retrieval

Model Selection and Fine-Tuning Embedding models often struggle to accurately interpret complex document structures such as tabular and graphical data, and lack sensitivity to domain-specific semantics. Thus, careful selection of a moderately large-scale, high-performing model—and fine-tuning it—are required to ensure base retrieval performance.

Hybrid Retrieval Strategy Conventional RAG systems typically suffer from limitations stemming from their exclusive reliance on either dense or sparse retrieval techniques. Dense retrieval, which leverages continuous vector representations, excels at capturing deep semantic relationships but

often struggles with precise keyword matching—especially when dealing with domain-specific terminology. In contrast, sparse retrieval methods, such as BM25 or SPLADE, are effective for exact keyword matching but tend to overlook the richer contextual and semantic nuances inherent in complex documents. These shortcomings can lead to suboptimal retrieval performance in specialized domains like finance, where both exact term matching and deep semantic understanding are essential. To address these issues, our method introduces a hybrid retrieval strategy that fuses the strengths of both dense and sparse retrieval.

$$S_{\text{total}} = \alpha \cdot S_{\text{dense}} + (1 - \alpha) \cdot S_{\text{sparse}} \tag{1}$$

where S_{dense} represents the dense retrieval score (e.g., embedding similarity), S_{sparse} represents the sparse retrieval score (e.g., BM25), and α is a hyperparameter ranging from 0 to 1 that determines the contribution of each retrieval method.

By appropriately tuning α on a small subset, our hybrid approach can adapt to different query types, giving weight to either exact matching or deeper semantic interpretation when necessary. This strategy is backed up by ensemble learning principles, where combining complementary methods leads to a more robust and effective retrieval system.

3.3 Post-Retrieval

Reranking Reranking aims to further capture relevancy by leveraging larger models on top-K retrieved documents. Refined relevancy scores are calculated (usually) pairwise by Cross-Encoders and LLMs(Déjean et al., 2024). Despite requiring exponentially larger compute compared to Bi-Encoders, rerankers could be utilized on a small set of filtered documents to effectively emphasize underrepresented semantics with respect to the query.

Document Selection While reranking effectively provides the top-K relevant documents, whether LLMs can correctly process the long context remains a challenge, e.g., Lost in the Middle(Liu et al., 2024). To enable optimal utilization of retrieved documents, we proposed document selection—selecting *only the documents required* to answer the query.

4 EXPERIMENTS

This section details the experimental procedures, settings, and methodologies implemented to validate the theoretical contributions presented in Section 3.

4.1 EXPERIMENTAL SETUP

DATASET

Datasets To evaluate our proposed pipeline, we utilize a comprehensive set of 7 distinct financial document datasets, each representing various query types and document structures (refer to Table 1). These datasets are provided in the FinanceRAG Challenge(Choi et al., 2024), selected to cover

Table 1: Dataset Descriptions

DESCRIPTION

FinDER*	Real-world questions written by financial experts
FinQABench	Generated queries by LLM
FinanceBench(Islam et al., 2023)	Real-world questions written by non-experts
TATQA(Zhu et al., 2021)	Basic arithmetic questions
FinQA(Chen et al., 2021)	Complex arithmetic questions
ConvFinQA(Chen et al., 2022)	Questions asking for specific value from table values
MultiHiertt(Zhao et al., 2022)	Questions requiring multi-hop reasoning

*To be announced

comprehensive and real-world financial question answering scenarios. Also, each text is chunked at

512 tokens, following the results of Yepes et al. (2024).

Experiment Settings Our experiments were conducted in a notebook-based development environment using Google Colab, with access to 40GB NVIDIA A100 GPUs.

4.2 EVALUATION METRIC: NDCG@10

To assess the ranking quality of the retrieved documents, we employed the Normalized Discounted Cumulative Gain(Wang et al., 2013) at 10 (NDCG@10) metric. NDCG is a widely used measure in information retrieval that evaluates how well the predicted ranking of documents aligns with an ideal ranking based on ground-truth relevance. In our experiments, the provided labels of all 7 benchmarks were used to measure the total weighted NDCG@10 score.

In this metric, the **DCG** (**Discounted Cumulative Gain**) is computed by summing the relevance scores of the retrieved documents, with each score discounted by the logarithm of its rank position. The **IDCG** (**Ideal Discounted Cumulative Gain**) represents the maximum possible DCG achievable when the documents are ideally ranked in descending order of relevance. Normalizing the DCG by the IDCG yields the **NDCG**, a score between 0 and 1, where a higher value indicates a ranking that more closely approximates the ideal ordering.

$$NDCG = \frac{DCG}{IDCG} = \frac{1}{m} \sum_{u=1}^{m} \sum_{j \in I_u, v_j \le L} \frac{g_{uj}}{\log_2(v_j + 1)}$$
(2)

This metric is particularly useful because it simultaneously accounts for the relevance of each document and its position in the ranking. Improved NDCG@10 scores in our experiments indicate that our system retrieves a more complete and relevant set of documents for each query.

4.3 EXPERIMENTAL PROCEDURES

Retrieval Model Selection Selecting an optimal retrieval model is crucial for maximizing performance in RAG. We evaluated six candidate models based on the Information Retrieval performance on the MTEB Leaderboard (Muennighoff et al., 2023): e5-large-v2 (intfloat), GritLM-7B, Fin-BERT (ProsusAI), TAPAS (Google), stella_en_400M_v5 (NovaSearch), and stella_en_1.5B_v5 (NovaSearch).

Embedder Fine-tuning We fine-tuned our selected retrieval models to better align with financial texts and improve overall performance in our RAG pipeline. We focused on two models—"stella_en_1.5B_v5 (NovaSearch)" and "stella_en_400M_v5 (NovaSearch)"

For fine-tuning, we prepared relevant query-document pairs, splitting the dataset with a ratio of 8 (train): 2 (eval). Positive pairs were given a similarity score of 1.0, while negative pairs were scored 0.0, with negatives sampled randomly to ensure diversity. We utilized contrastive learning for fine-tuning, leveraging Multiple Negatives Ranking Loss (MNRLoss)(Henderson et al., 2017). More detailed information regarding the fine-tuned model and hyperparameter settings can be found in the Appendix 10, 11.

Query data preprocessing Query data tends to be brief and lack sufficient contextual cues, which can hinder the retrieval model's ability to fully interpret user intent. To address this, we experimented with three distinct preprocessing methods. First, **Default (FT_stella_400M)** used raw queries without any modifications. Second, **Keyword Extraction + LLMLingua** involved extracting key terms from each query while removing redundant words. Lastly, **Query Expansion with LLM** employed a large language model to enrich the queries with additional contextual information.

Corpus data preprocessing Corpus data utilized in our study comprised a variety of formats, from plain text to tabular data. Recognizing that a simple preprocessing strategy would be insufficient for such a diverse dataset features, we implemented task-specific methods. First, the **Default** dataset is the raw corpus without any modifications. For the **Corpus Markdown Restructuring**, we restructured documents using markdown formatting to enhance clarity and preserve inherent structural elements. Additionally, we implemented two specialized methods for the MultiHiertt dataset, where tabular data is emphasized. **Corpus Table Augmentation** refers to

augmenting table cells with textual annotations of rows and columns, as in *Investment Return*, 2016 = \$192 (in millions). This better captures the implications within the table by attaching distant row and column data. **Corpus Table Extraction** focuses on isolating and emphasizing the intrinsic structure of tabular data by removing non-tabular text within the chunk (Lee & Roh, 2024).

Hybrid Retrieval To determine the optimal balance between sparse search and dense search that fully reflects the characteristics of the task, we sought to identify the optimal alpha value. We incremented alpha from 0 to 1 in steps of 0.025, computed the total score corresponding to each ratio, and then evaluated the resulting matches using the NDCG@10 metric to observe the trend in score changes. The alpha value that yielded the highest NDCG@10 score was designated as the optimal alpha.

Reranking In the reranking stage, the selected models were "bge-ranker-v2-m3 (BAAI)" and "voyage-rerank-2 (Voyage AI)", based on MTEB. We utilized the top-20 retrieval results from the fine-tuned stella_en_1.5B_v5 model as the reranking target. The reranked results were evaluated using NDCG@10.

Selection Agent The selection agent processes the top-10 retrieved documents. Acting as a financial expert, the agent selects only the documents actually useful in answering the query, based on factual accuracy, relevance, and clarity. This reduces token overhead and improves response quality. The prompt we used can be found in the Appendix A.2

Generation We evaluated the generated responses, referencing two metrics from RAGAS (Es et al., 2023): Answer Relevance and Context Precision without reference. Answer Relevance quantifies how directly the generated answer addresses the query by reverse-engineering questions from the answer and averaging their cosine similarities with the query. Context Precision without reference employs GPT-40 mini to compare context chunks with the generated response, ensuring that relevant information is prioritized.

A key component is our DPO-trained(Rafailov et al., 2023) GPT-40 mini. We generated answers using the gpt-40-2024-08-06 at two temperature settings (0.1 and 1.0) for identical queries, then evaluated responses on financial terminology and clarity to divide preferred and non-preferred responses. These pairs were used to fine-tune gpt-40-mini-2024-07-18 via OpenAI's API (beta = 0.1).

5 Result

Our experiment evaluation demonstrates that the enhancements introduced to the RAG framework yield substantial improvements in retrieving and processing domain-specific financial data.

5.1 RETRIEVAL MODEL SELECTION

Table 2: Performance comparison of retrievers

MODELS	NDCG@10
e5-large-v2	0.29746
GritLM	0.21262
FinBERT	0.25595
TAPAS	0.10124
stella_400M	0.32006
stella_1.5B	0.32178

We evaluated several candidate retrieval models on the FinanceRAG dataset using NDCG@10 (Table 2). Among the models tested, stella_1.5B achieved the highest NDCG score. It outperformed alternatives that either provided stable but suboptimal results or showed limitations in processing general textual content. For example, FinBERT, despite being fine-tuned on financial text, performed

poorly compared to stella_1.5B. TAPAS excelled at handling tabular data but was less effective in overall text retrieval.

5.2 Embedder Fine-tuning

Table 3: Comparison of NDCG@10 Performance for the Fine-tuned 400M & 1.5B Embedder

MODELS	NDCG@10	
FT_stella_400M*	0.40186	
FT_stella_1.5B*	0.50864	

*Fine-Tuned

Subsequent fine-tuning of the selected retrieval model further enhanced performance, as shown in Table 3. As a result, the fine-tuned version of "stella_en_1.5B_v5 (NovaSearch)"—referred to as FT stella_1.5B_achieved an NDCG@10 score of 0.50864. This is a significant improvement over the baseline performance of the "stella_en_400M_v5 (NovaSearch)" model, which achieved an NDCG@10 score of 0.40186. These results confirm that the fine-tuning process, effectively improves retrieval precision and overall performance in the finance domain within the RAG pipeline.

5.3 Dataset Preprocessing

To further improve retrieval performance, we implemented tailored preprocessing strategies for both queries and corpus documents. For query preprocessing, we tested three approaches (Table 4). Among these methods, **Query Expansion with LLM** reached the highest NDCG@10 score.

Table 4: NDCG@10 results of query preprocessing methods

METHODS	NDCG@10
Default (FT_stella_400M) Keyword Extraction + LLMLingua	0.40186 0.43613
Query Expansion with LLM	0.43613 0.48601

For corpus preprocessing, we compared four methods and **Corpus Markdown Restructuring** demonstrated the best NDCG@10 score (Table 5).

Table 5: NDCG@10 results of corpus preprocessing methods

METHODS	NDCG@10
Default (Query Expansion)	0.48601
Corpus Markdown Restructuring	0.48645
Corpus Table Augmentation	0.45411
Corpus Table Extraction	0.43604

The notable point is that only **Corpus Markdown Restructuring** showed the highest performance. This is attributable to the better readability of Markdown, which led to further contextual comprehension. On the other hand, focusing on tabular data—emphasizing it by extraction or text-aided augmentation—resulted in worse performance than before. Despite enriching contextual information in cells, Table Augmentation led to excessive noise within the corpus. Similarly, Table Extraction assumed significance only in tabular data, discarding all other text. This suggests that fine-tuned retrievers can sufficiently understand knowledge in tabular data without the need for modification.

These results indicate that, in the context of financial documents, using Query Expansion for queries and Markdown Restructuring for the corpus yields the best retrieval performance.

5.4 Hybrid retrieval

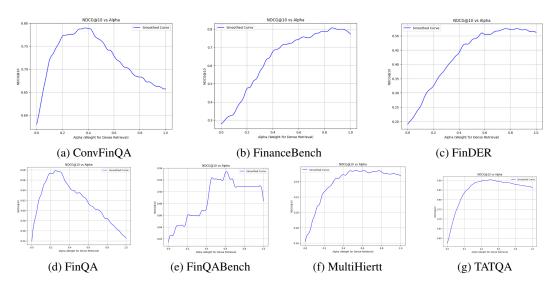


Figure 2: α -NDCG@10 Graphs by each task

Table 6: Comparison of optimal α and its following NDCG scores across the different datasets.

DATASET	OPTIMAL α	NDCG@10
FinDER	0.85	0.52572
FinQABench	0.6	0.93214
FinanceBench	0.85	0.80464
FinQA	0.25	0.67849
TATQA	0.5	0.90288
ConvFinQA	0.375	0.78959
MultiHiertt	0.525	0.24736

We conducted experiments by varying α at intervals of 0.025 and evaluated performance using NDCG@10 across 7 datasets. Table 6 illustrates the optimal α value and performance for each dataset. In Figure 2, subfigures from (a) to (g) illustrate the relationship between the α value (x-axis) and the corresponding NDCG score (y-axis). These graphs reveal that tasks requiring precise, fact-based retrieval generally achieved optimal performance at lower α values (emphasizing sparse retrieval), while those demanding deeper semantic interpretation performed best at higher α values (emphasizing dense retrieval).

5.5 RERANKING

Table 7 demonstrates the positive impact of reranking for each of the 7 tasks. Reranking the top-20 retrieved documents yielded a positive impact across all seven tasks. In most tasks, reranking led to noticeably improved NDCG@10 scores, confirming the effectiveness of Cross-Encoders over BERT-based retrievers. However, while the total score was improved (Table 8), a slight performance drop was observed in certain tasks such as FinQABench, hinting at potential task-specific adjustments in reranking.

Table 7: NDCG@10 for the initial retrieval stage versus the subsequent reranking stage (voyage-rerank-2) across various datasets.

DATASET	NDCG@10	
	Retrieval	Reranking
FinDER	0.52572	0.58001
FinQABench	0.93214	0.89974
FinanceBench	0.80464	0.91191
FinQA	0.67849	0.80054
TATQA	0.90288	0.90781
ConvFinQA	0.78959	0.86574
MultiHiertt	0.24736	0.45837

Table 8: NDCG@10 for the final results with reranking

MODELS	NDCG@10
1.5B FT + reranking with bge-reranker-v2-m3	0.51508
1.5B FT + reranking with voyage-rerank-2	0.59898

Table 9: Generation Score by RAGAS

MODELS	Answer Relevance	Context Precision
Selection Agent + DPO-trained GPT-40 mini GPT-40	0.8924 0.8663	0.3962 0.3418

5.6 GENERATION

The experimental results in Table 9 indicate that the Direct Preference Optimization (DPO) agent and selection agent, implemented with the lightweight GPT-40 mini, outperformed GPT-40 on both evaluated metrics. Despite its smaller architecture, the DPO agent generated responses with enhanced alignment to the input query and stronger support from the most pertinent contextual information, as shown in the Appendix A.3. The enhanced performance is also attributable to the targeted use of the selection agent to carefully utilize the top-10 documents. This integrated approach significantly improves both the quality and efficiency of answer generation in financial tasks.

6 Conclusion

In this study, we presented an enhanced Retrieval-Augmented Generation (RAG) pipeline designed specifically for financial question answering documents. Our work integrates query and corpus preprocessing, embedder fine-tuning, hybrid retrieval, and reranking. These components work in accordance with each other to process complex financial documents effectively and enable Large Language Models (LLMs) to generate responses with improved accuracy and relevancy.

Specifically, we employed multiple preprocessing techniques such as query expansion and corpus markdown restructuring to preserve the heterogeneous structure and subtle contextual information present in financial question answering datasets. In parallel, we fine-tuned embedding models to align them with the nuances of financial documents. By fusing sparse retrieval—which excels at precise keyword matching—with dense retrieval—which captures deep semantic relationships—our hybrid retrieval approach ensures that a comprehensive set of relevant documents is retrieved for each query. Moreover, the subsequent reranking stage refines the ordering of these documents to maximize their contextual relevance.

Experimental results demonstrate significant improvements in retrieval as well as generation, most noted by the NDCG@10 scores. The observed increase in NDCG@10 indicates that our system re-

trieves a more complete and relevant set of documents for each query, thereby resolving challenges related to information omission and hallucination.

Overall, the results validate the robustness and effectiveness of our RAG pipeline in the finance domain. The substantial performance gains and enhanced retrieval quality provide a reliable foundation for applications in financial contexts, where precise information retrieval and specialized expertise are critical. This work lays the groundwork for future research aimed at optimizing RAG in finance and outlines possible applicability to other domain-specific natural language processing tasks.

7 Future Work

Our proposed method overcomes the limitations of the existing RAG pipeline through improvements on retrieval tailored to the finance domain, thereby opening new possibilities for natural language processing in specialized contexts. However, several remaining challenges warrant further investigation.

First, our proposed method faces difficulties in incorporating rapidly changing financial data. Because financial markets continuously generate time-sensitive information such as corporate disclosures, news, and fluctuations in stock prices, in-depth research on techniques for efficiently retrieving and indexing streaming data is essential. Moreover, the need for a multilingual extension of the RAG framework becomes evident when considering the global nature of financial environments (Zhao et al., 2024).

Second, security problems must be addressed. Derner et al. (2024) has already highlighted the susceptibility of large language model-based systems to various threats, including malicious prompt attacks. To mitigate these risks and ensure AI safety, it is crucial to adopt the layered security strategies presented in Shamsujjoha et al. (2025), as exemplified by LLaMA Guard (Inan et al., 2023).

Finally, adherence to ethical regulations in the finance domain and broader AI governance guidelines, represented by EU's Ethics guidelines for trustworthy AI (Commission et al., 2019), calls for robust monitoring and oversight when deploying this pipeline in real-world settings. Future research could focus on establishing systems that continually evaluate inference processes and generated outputs to prevent the spread of unnecessary or inaccurate information. Such systems must also detect and block any potential exposure of sensitive data or violations of regulatory standards, thereby ensuring responsible and compliant use of the proposed pipeline.

REFERENCES

- Peter Anderson, Mano Vikash Janardhanan, Jason He, Wei Cheng, and Charlie Flanagan. Greenback bears and fiscal hawks: Finance is a jungle and text embeddings must adapt. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 362–370, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.26.
- Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. arXiv preprint arXiv:2210.03849, 2022.
- Chanyeol Choi, Jy-Yong Sohn, Yongjae Lee, Subeen Pang, Jaeseon Ha, Hoyeon Ryoo, Yongjin Kim, Hojun Choi, and Jihoon Kwon. Acm-icaif '24 financerag challenge. https://kaggle.com/competitions/icaif-24-finance-rag-challenge, 2024. Kaggle.
- European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics guidelines for trustworthy AI*. Publications Office, 2019. doi: doi:10.2759/346720.
- Erik Derner, Kristina Batistič, Jan Zahálka, and Robert Babuška. A security risk taxonomy for prompt-based interaction with large language models. *IEEE Access*, 12:126176–126187, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3450388.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.
- Hervé Déjean, Stéphane Clinchant, and Thibault Formal. A thorough comparison of cross-encoders and llms for reranking splade, 2024.
- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023.
- Yunfan Gao, Yun Xiong, Meng Wang, and Haofen Wang. Modular rag: Transforming rag systems into lego-like reconfigurable frameworks. *arXiv preprint arXiv:2407.21059*, 2024.
- Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply, 2017.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llmbased input-output safeguard for human-ai conversations, 2023.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv* preprint *arXiv*:2311.11944, 2023.
- Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.825.

- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550.
- Tobias Kerner. Domain-specific pretraining of language models: A comparative study in the medical field. *arXiv preprint arXiv:2407.14076*, 2024.
- Seunghee Kim, Changhyeon Kim, and Taeuk Kim. Fcmr: Robust evaluation of financial cross-modal multi-hop reasoning. *arXiv preprint arXiv:2412.12567*, 2024.
- Hamin Koo, Minseon Kim, and Sung Ju Hwang. Optimizing query generation for enhanced document retrieval in rag. *arXiv* preprint arXiv:2407.12325, 2024.
- Joohyun Lee and Minji Roh. Multi-reranker: Maximizing performance of retrieval-augmented generation in the financerag challenge, 2024.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.
- Jun Lu, David Li, Bill Ding, and Yu Kang. Improving embedding with contrastive fine-tuning on small datasets with expert-augmented scores, 2024.
- Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148.
- Chaitanya Patel. Hypothetical retrieval-augmented generation (hypothetical rag): Advancing ai for enhanced contextual understanding and creative problem-solving. *Scientific Research Journal of Science, Engineering and Technology*, 2(1), 2024. ISSN 2584-0584.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023.
- Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.
- Zabihollah Rezaee. Causes, consequences, and deterence of financial statement fraud. *Critical Perspectives on Accounting*, 16(3):277–298, 2005. ISSN 1045-2354. doi: https://doi.org/10.1016/S1045-2354(03)00072-8.
- Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 155–161, 2024.
- Saptarshi Sengupta, Connor Heaton, Suhan Cui, Soumalya Sarkar, and Prasenjit Mitra. Towards Efficient Methods in Medical Question Answering using Knowledge Graph Embeddings . In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 5089–5096, Los Alamitos, CA, USA, December 2024. IEEE Computer Society. doi: 10.1109/BIBM62325.2024. 10821824.
- Spurthi Setty, Harsh Thakkar, Alyssa Lee, Eden Chung, and Natan Vidra. Improving retrieval for rag based question answering models on financial documents, 2024.

- Md Shamsujjoha, Qinghua Lu, Dehai Zhao, and Liming Zhu. Swiss cheese model for ai safety: A taxonomy and reference architecture for multi-layered guardrails of foundation model based agents, 2025.
- Rajat Singh and Srikanta Bedathur. Embeddings for tabular data: A survey, 2023.
- Yixuan Tang and Yi Yang. Do we need domain-specific embedding models? an empirical investigation, 2024.
- Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9414–9423, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.585.
- Shuai Wang, Shengyao Zhuang, and Guido Zuccon. Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, pp. 317–324, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386111. doi: 10.1145/3471158. 3472233.
- Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In *Annual Conference Computational Learning Theory*, 2013.
- Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, Shunian Chen, Yifei Zhang, Lihang Shen, Daniel Kim, Zhiwei Liu, Zheheng Luo, Yangyang Yu, Yupeng Cao, Zhiyang Deng, Zhiyuan Yao, Haohang Li, Duanyu Feng, Yongfu Dai, VijayaSai Somasundaram, Peng Lu, Yilun Zhao, Yitao Long, Guojun Xiong, Kaleb Smith, Honghai Yu, Yanzhao Lai, Min Peng, Jianyun Nie, Jordan W. Suchow, Xiao-Yang Liu, Benyou Wang, Alejandro Lopez-Lira, Jimin Huang, and Sophia Ananiadou. Openfinllms: Open multimodal large language models for financial applications, 2024.
- Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. Financial report chunking for effective retrieval augmented generation, 2024.
- Haoyu Zhang, Jun Liu, Zhenhua Zhu, Shulin Zeng, Maojia Sheng, Tao Yang, Guohao Dai, and Yu Wang. Efficient and effective retrieval of dense-sparse hybrid vectors using graph-based approximate nearest neighbor search, 2024.
- Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3010–3019, July 2017. doi: 10.1109/CVPR.2017.321.
- Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv preprint arXiv:2206.01347*, 2022.
- Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 15296–15319. Curran Associates, Inc., 2024.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021.

A APPENDIX

A.1 FINE-TUNED MODEL INFO

Table 10: Evaluated Performance of stella_en_400M_v5-FinanceRAG

METRIC	COSINE	DOT
Accuracy@10	0.8519	0.8422
Precision@10	0.1024	0.0998
Recall@10	0.8398	0.8224
NDCG@10	0.6409	0.6195
MRR@10	0.5902	0.5712

Table 11: Fine-tuning Details

Model Name stella_en_400M_v5-FinanceRAG Train: Eval Ratio 8:2**Epochs** 2 2×10^{-5} LR Warmup 0.1Batch Loss MNRLoss (Multiple Negatives Ranking Loss) Weight Evaluation Information Retrieval Eval

A.2 SELECTION AGENT PROMPT

You are an expert financial advisor and evaluator focused on improving responses. Your task is to enhance answers based on detailed evaluation scores while:

- Maintaining factual accuracy with the provided documents
- Ensuring responses are clear and well-structured for financial contexts
- Providing comprehensive answers that address all aspects of the query
- Using professional financial terminology appropriately

You are given the pair of Query, Corpus (same query) Out of the 10 documents, only provide the list of indices of those that are RELEVANT (e.g. the content is somehow needed to answer the question), from 0 to 9. Example: [0, 2, 8, 9]

Query : **query** Corpus #0 : **corpus 1** Corpus #1 : **corpus 2**

••

A.3 DPO TRAINING EXAMPLES

OUERY:

Did Pfizer grow its PPNE between FY20 and FY21?

PREFERRED:

To determine whether Pfizer grew its Property, Plant, and Equipment (PPNE) between FY20 and FY21, we need to examine the relevant figures from the provided financial data.

According to the data from the Pfizer 2021 10-K report:

- In 2020, the value of Property, Plant, and Equipment was \$13,745 million. - In 2021, this value increased to \$14,882 million.

This indicates that Pfizer's Property, Plant, and Equipment (PPNE) grew from FY20 to FY21. The increase in PPNE from \$13,745 million to \$14,882 million represents a growth of \$1,137 million, which is approximately an 8.27% increase year-over-year. This growth could be attributed to investments in new facilities, upgrades, or expansions of existing infrastructure, reflecting Pfizer's strategic initiatives to enhance its operational capabilities.

NON-PREFERRED:

To determine whether Pfizer grew its Property, Plant, and Equipment (PPNE) between FY20 and FY21, we need to examine the relevant figures from the provided financial data.

According to the data from Pfizer's 2021 10-K report:

- Property, Plant, and Equipment (PPNE) for FY2021 was \$14,882 million. - Property, Plant, and Equipment (PPNE) for FY2020 was \$13,745 million.

Comparing these values, Pfizer's PPNE increased from \$13,745 million in 2020 to \$14,882 million in 2021. This indicates a growth in PPNE between FY20 and FY21.