Domain-Specific Retrieval-Augmented Generation in Finance : FinanceRAG Challenge

Joohyun Lee
Financial Security Institute
Republic of Korea
dlee110600@gmail.com

Minji Roh
Financial Security Institute
Republic of Korea
nohmin0710@gmail.com

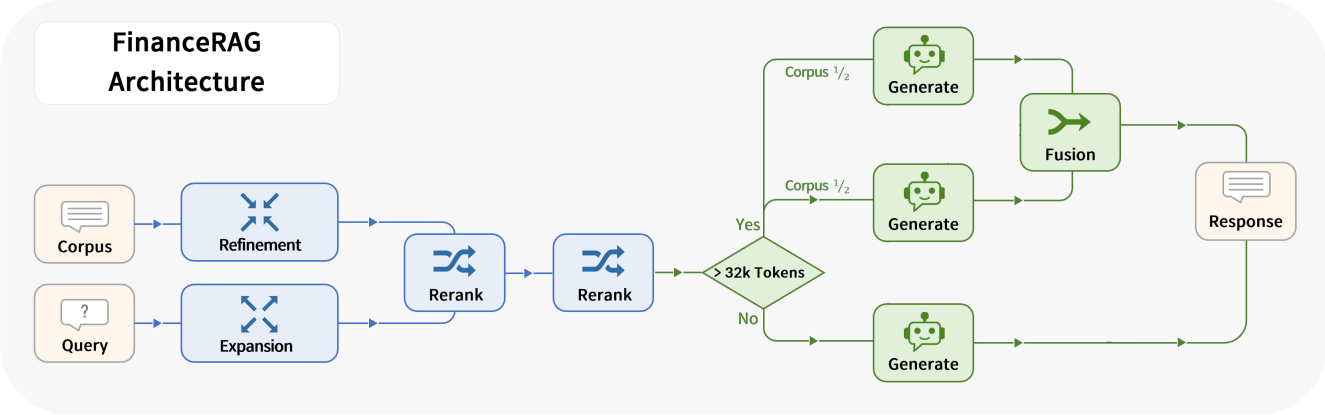

Figure 1: Overview of Retrieval Augmented Generation System

Abstract

As Large Language Models (LLMs) increasingly address domain-specific problems, their application in the financial sector has expanded rapidly. Tasks that are both highly valuable and time-consuming, such as analyzing financial statements, disclosures, and related documents, are now being effectively tackled using LLMs. This paper details the development of a high-performance, finance-specific Retrieval-Augmented Generation (RAG) system for the ACM-ICAIF '24 FinanceRAG competition. We optimized performance through ablation studies on query expansion and corpus refinement during the pre-retrieval phase. To enhance retrieval accuracy, we employed multiple reranker models. Notably, we introduced an efficient method for managing long context sizes during the generation phase, significantly improving response quality without sacrificing performance. Our key contributions include: (1) pre-retrieval ablation analysis, (2) an enhanced retrieval algorithm, and (3) a novel approach for long-context management. This work demonstrates the potential of LLMs in effectively processing and analyzing complex financial data to generate accurate and valuable insights. The source code and further details are available at https://github.com/cv-lee/FinanceRAG.

AMC-ICAIF '24, Nov 14–17, 2024, Brooklyn, NY

CCS Concepts

• **Computing methodologies** → **Natural language generation**.

Keywords

Retrieval Augmented Generation, Large Language Model, Finance Analysis, Retrieval, Rerank

ACM Reference Format:
Joohyun Lee and Minji Roh. 2024. Domain-Specific Retrieval-Augmented Generation in Finance : FinanceRAG Challenge. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (AMC-ICAIF '24).* ACM, New York, NY, USA, 5 pages.

1 Introduction

The advent of Large Language Models (LLMs) has revolutionized the approach to solving domain-specific problems, leading to a surge in their application within the financial sector [11–13]. Tasks that are both high in market value and traditionally time-consuming, such as analyzing financial statements, disclosures, press releases, and related documents, are now being actively addressed using LLMs [11]. Historically, the process of sifting through hundreds of pages to extract necessary information for investment decision-making was not only labor-intensive but also time-consuming, often causing delays that could result in financial losses. LLMs are effectively mitigating these issues by efficiently extracting relevant paragraphs from extensive documents and isolating the required information. [6, 12]

In this paper, we present the development of a high-performance, finance-specific Retrieval-Augmented Generation (RAG) system, detailing our solution approach for the ACM-ICAIF '24 FinanceRAG

competition task [3]. We optimized our system through ablation studies of various techniques, including query expansion and corpus optimization during the pre-retrieval phase. To push the retrieval performance to its limits, we employed a combination of multiple reranker models. Moreover, we designed an innovative method for efficiently managing long context sizes during the generation phase, which significantly elevated the quality and accuracy of the responses. This approach allowed us to harness the full potential of LLMs even when dealing with exceedingly lengthy input contexts, thereby maximizing answer performance. Our contributions in this paper are:

- **Pre-Retrieval Ablation Study** : We conducted comprehensive ablation studies to optimize pre-retrieval techniques, enhancing the overall retrieval effectiveness.
- **Accurate Retrieval Algorithm** : We developed an accurate retrieval algorithm by leveraging multiple reranker models, improving the relevance of the retrieved corpora.
- **Efficient Context Size Management** : We designed a method to efficiently manage context sizes, enabling the processing of very long inputs without sacrificing performance.

2 Task and Dataset

To build a high-performance RAG system in the financial domain, we addressed two primary tasks as defined by the FinanceRAG competition at ACM-ICAIF '24 [3]. Task 1 involves retrieving the top 10 most relevant corpora based on a given query from a large corpus. The key challenge is to accurately identify the necessary corpora using a RAG system that incorporates techniques such as embedding and reranking. Task 2 requires generating precise answers to the query based on the corpora retrieved in Task 1. Notably, the corpora may include not only textual data but also numerical tables, and the context sizes of the reference corpora are frequently very large. The critical aspect is to locate the necessary numerical information within the vast corpus and generate accurate responses using LLMs.

The dataset used in the competition is finance-specific and consists of queries and corpora. Each dataset involves extracting the correct top 10 corpora based on the given query and generating the correct answer. The datasets used are as follows:

- **FinDER**: 10-K Reports and financial disclosures, to evaluate understanding domain-specific professionals, jargon, and abbreviations.
- **FinQABench**: 10-K Reports, focusing on detecting hallucinations and ensuring factual correctness in the generated responses.
- **FinanceBench**: 10-K Reports, to evaluate how well systems handle straightforward, real-world financial questions.
- **TATQA**: Financial Reports, involving numerical reasoning over hybrid data (tabular and text) to evaluate basic arithmetic, comparisons, and logical reasoning.
- **FinQA**: Earnings Reports, involving tabular and text data to evaluate multi-step numerical reasoning
- **ConvFinQA**: Earnings Reports, involving tabular and text data to evaluate handling conversational queries.
- **MultiHiertt**: Annual Reports, involving hierarchical tables and unstructured text to evaluate complex reasoning tasks involving multiple steps across various document sections.

3 Method and Results

Our proposed RAG system comprises three main stages. The first two stages focus on the algorithm for Task 1, while the third stage addresses the algorithm for Task 2. Initially, a pre-retrieval process prepares the data for the retrieval stage. In the retrieval stage, we extract a preliminary set of relevant corpora using reranker models, followed by a second reranking to extract more precise corpora. Finally, based on the extracted corpora, the response generation stage produces the final answer.

3.1 Pre-Retrieval

The queries in the dataset are typically composed of simple sentences; however, they often include financial-specific abbreviations, ambiguous meanings, or require multi-step reasoning to resolve. To address these challenges, we opted not to use the raw queries directly. Instead, we enhanced them through query expansion techniques [6, 8, 10] aimed at clarifying sentence meanings, interpreting abbreviations, and decomposing complex queries into simpler, more manageable steps. These enhancements were conducted using the *OpenAI/GPT-4o-mini*. We employed several query expansion methods, including paraphrasing, keyword extraction, and the creation of hypothetical documents. By combining the results of these methods with the original queries, we aimed to improve the effectiveness of the retrieval process.

Additionally, to handle the extensive corpus efficiently, we utilized summarization and table extraction. For summarization, we used the *OpenAI/GPT-4o* to create condensed summaries of the corpus, which were subsequently used for retrieval. Table extraction was specifically applied to the MultiHiertt dataset, as this dataset contained corpora with extremely large token counts, and much of the critical information was embedded within tables. We conducted an ablation study using various combinations of query expansion techniques and corpus optimizations, with the results presented in Table 1.

Based on these findings, we opted to combine the original query with keyword extraction for all datasets and applied corpus table extraction specifically to the MultiHiertt dataset to optimize retrieval performance.

3.2 Retrieval

For the retrieval process, we utilized the preprocessed queries and corpora from the pre-retrieval stage. To maximize performance, we directly employed reranker models instead of relying on embedding similarity comparisons. Reranker models offer superior performance as they perform binary classification by processing both the query and the corpus simultaneously.

We initially extracted the top 200 relevant corpora for each query-corpus pair using a relatively lightweight reranker model (*jina-reranker-v2-base-multilingual*) [1]. Subsequently, these top 200 corpora were reranked using more precise reranker models, ultimately selecting the top 10 most relevant corpora.

Table 1: Ablation study of Pre-Retrieval

Query				Corpus			NDCG@10
Original	Paraphrased	Keywords extraction	Hypothetical documents	Original	Summary	Table only*	
✓	-	-	-	✓	-	-	0.48949
✓	✓	-	-	✓	-	-	0.51228
✓	-	✓	-	✓	-	-	0.54090
✓	-	-	✓	✓	-	-	0.43707
✓	-	-	-	-	✓	-	0.45589
✓	✓	-	-	-	✓	-	0.48495
✓	-	✓	-	-	✓	-	0.48949
✓	-	-	✓	-	✓	-	0.43707
✓	-	-	-	-	-	✓	0.51228
✓	✓	-	-	-	-	✓	0.55602
✓	-	✓	-	-	-	✓	**0.58102**
✓	-	-	✓	-	-	✓	0.45589

† Reranker: *jina-reranker-v2-base-multilingual* * Applied only to the MultiHiertt dataset

To identify the more refined reranker models, we conducted extensive experiments using various top-ranking models from the MTEB leaderboard as well as open-source models. Specifically, we utilized the labels provided by the competition organizers to determine the best-performing reranker model for each dataset. Based on these evaluations, we selected *jina-reranker-v2-base-multilingual*, *gte-multilingual-reranker-base* [9], and *bge-reranker-v2-m3* [2], as summarized in Table 2. Through this process, we achieved a final NDCG@10 score of 0.60161 for Task 1.

3.3 Geneartion

The context size of LLMs has increased exponentially in recent times, with some models capable of handling contexts exceeding 2 million tokens [4]. However, studies [5, 7] have shown that LLM performance degrades as the context size increases. Therefore, selecting an appropriate context size is crucial for optimal performance.

According to recent research [5], models like *OpenAI/o1-preview* exhibit high performance up to a context size of 64k tokens, beyond which performance declines. In our manual analysis of the competition data, we observed significant performance degradation beyond 32k tokens. Consequently, we utilized up to the top 20 corpora retrieved to limit the context size.

If the input token size (query plus top 1-20 corpora) was under 32k tokens, we processed it directly. If it exceeded 32k tokens, we split the corpora into halves, processed them separately through the LLM, and then fused the resulting answers.

Another critical consideration during generation was the answer format. Financial experts often require specific and concise numerical values rather than lengthy, complex explanations. Therefore, we performed prompt engineering to ensure that the LLM focused on providing the key information and values requested in the query. The overall process of the proposed method is presented in Algorithm 1.

Algorithm 1 Proposed FinanceRAG System

Input: Query Q, Corpus C
Output: Response R
$Q' \leftarrow Q$ + Extracted Keywords from$Q$
$C' \leftarrow$ Extract tables from MultiHiertt, retain other corpus as is
$C_{top200} \leftarrow$ 1st Reranker(Q', C')
$C_{top20} \leftarrow$ 2nd Reranker(Q', C_{top200})
if Token count of $(Q' + C_{top20}) \leq 32k$ **then**
  $R \leftarrow \text{LLM}(Q', C_{top20})$
else
  $R_1 \leftarrow \text{LLM}(Q', C_{top1-10})$
  $R_2 \leftarrow \text{LLM}(Q', C_{top11-20})$
  $R \leftarrow \text{Fusion}(R_1, R_2)$

4 Discussion

While the competition rules prohibited training models using the source data, future work could involve leveraging the source data for model fine-tuning to achieve better reranking performance. Specifically, parameter-efficient fine-tuning (PEFT) techniques could be applied to train only the final layer (e.g., the scoring layer) of the reranker models, thereby enhancing performance on the subtasks.

Moreover, incorporating simple client inputs, such as specifying the search scope (e.g., "APPLE Stock"), could significantly reduce the corpus size, enabling the construction of an accurate RAG system at a lower computational cost. Proper planning and system design can make the solution more efficient and practical for real-world applications.

A notable challenge we encountered was the substantial computational cost and time required. Although financial investment experts often deal with high-value queries where such costs may be justifiable, optimization remains essential. The retrieval and reranking components can be optimized through quantization techniques.

Table 2: Reranker Models Used for Each Dataset

Dataset	Reranker		
	jina-reranker-v2-base-multilingual	gte-multilingual-reranker-base	bge-reranker-v2-m3
FinDER	-	-	✓
FinQABench	✓	-	-
FinanceBench	✓	-	-
TATQA	-	-	✓
FinQA	-	✓	-
ConvFinQA	-	-	✓
MultiHiertt	✓	-	-

However, the LLM input, consisting of the query and corpora, still poses challenges due to its size.

We experimented with using pre-summarized corpora to reduce input size, but this approach led to a significant drop in performance for queries requiring specific numerical responses. Addressing this limitation remains an open area for future research.

5 Conclusion

The analysis of financial statements, disclosures, press releases, and related materials is a task of high market value and one that LLMs and RAG systems are well-suited to perform. Developing a finance-specific RAG system entails handling large-scale corpora and requires domain-specific knowledge and numerical data analysis capabilities. By leveraging hybrid embedding similarity functions, long context management techniques, and comprehensive ablation studies on query expansions, we successfully developed a high-performance finance-specific RAG system.

Our work demonstrates that with careful system design and optimization, LLMs can effectively process and analyze complex financial data to generate accurate and valuable insights. We hope that our contributions will further the adoption of AI technologies in the financial industry and inspire future research in this area.

Acknowledgments

We would like to thank the Finance Security Institute (FSI) for the support.

References

[1] Jina AI. 2024. Jina Reranker v2: Base Multilingual. https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual. Accessed: November 8, 2024.

[2] BAAI. 2024. BGE Reranker v2 M3. https://huggingface.co/BAAI/bge-reranker-v2-m3. Accessed: November 8, 2024.

[3] Chanyeol Choi, Jy-Yong Sohn, Yongjae Lee, Subeen Pang, Jaeseon Ha, Hoyeon Ryoo, Yongjin Kim, Hojun Choi, and Jihoon Kwon. 2024. ACM-ICAIF '24 FinanceRAG Challenge. https://kaggle.com/competitions/icaif-24-finance-rag-challenge. Kaggle.

[4] Google Cloud. 2024. *Long Context Capabilities of Vertex AI Generative AI.* Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/long-context?hl=ko

[5] Databricks. 2024. *The Long Context RAG Capabilities of OpenAI o1 and Google Gemini.* Databricks. https://www.databricks.com/blog/long-context-rag-capabilities-openai-o1-and-google-gemini

[6] Yunfan Gao, Yun Xiong, Meng Wang, and Haofen Wang. 2024. Modular rag: Transforming rag systems into lego-like reconfigurable frameworks. *arXiv preprint arXiv:2407.21059* (2024).

[7] Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2024. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. *arXiv preprint arXiv:2410.05983* (2024).

[8] Hamin Koo, Minseon Kim, and Sung Ju Hwang. 2024. Optimizing Query Generation for Enhanced Document Retrieval in RAG. *arXiv preprint arXiv:2407.12325* (2024).

[9] Alibaba NLP. 2024. GTE Multilingual Reranker Base. https://huggingface.co/Alibaba-NLP/gte-multilingual-reranker-base. Accessed: November 8, 2024.

[10] Chaitanya Patel. 2024. 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), 1–4.

[11] Spurthi Setty, Harsh Thakkar, Alyssa Lee, Eden Chung, and Natan Vidra. 2024. Improving retrieval for rag based question answering models on financial documents. *arXiv preprint arXiv:2404.07221* (2024).

[12] Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial report chunking for effective retrieval augmented generation. *arXiv preprint arXiv:2402.05131* (2024).

[13] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance.* 349–356.