FinanceRAG with Hybrid Search and Reranking

Jing Wang*
Accrete.AI
Wellesley, MA, USA

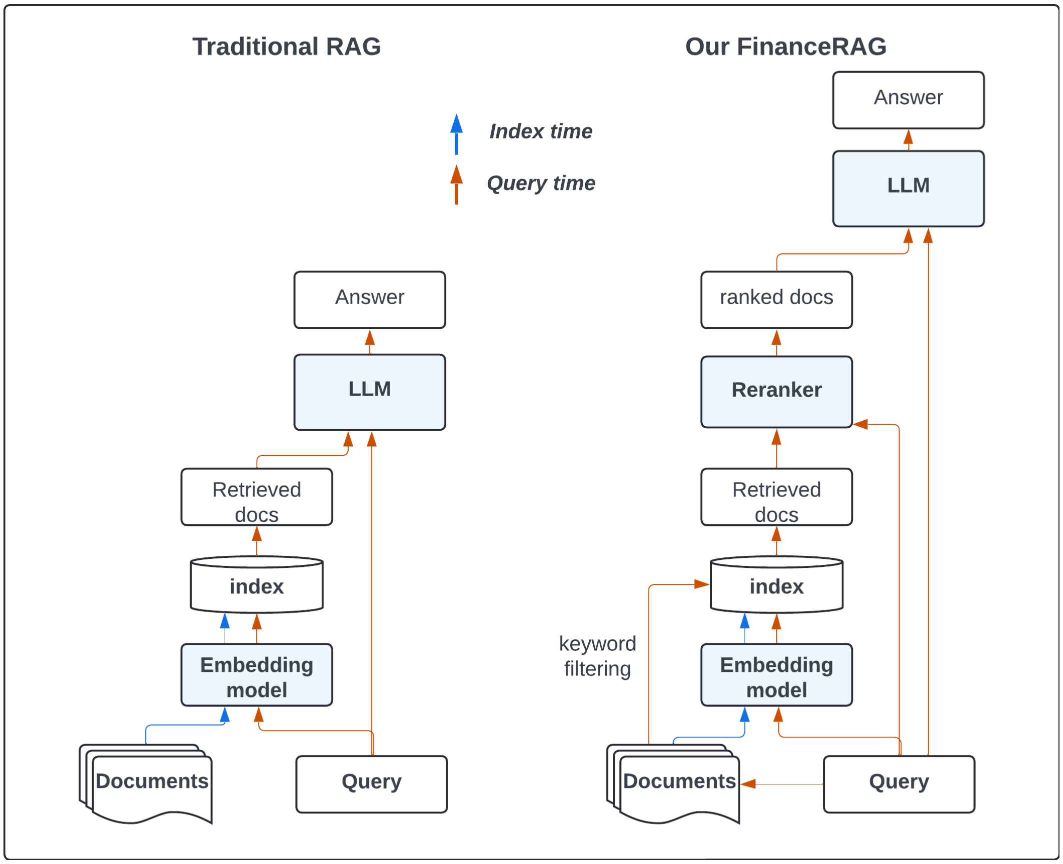

Figure 1: Overview of the RAG workflow

Abstract

In this project [1], we integrate hybrid search with document reranking in the retrieval step in a Retrieval-Augmented Generation (RAG) systems. During the hybrid search phase, we retrieve documents that are not only semantically similar to the query but also contain keywords inferred from the query. Subsequently, a reranking step leverages the cross-attention mechanism to further assess the semantic similarity of the filtered documents. The top-k most relevant documents identified through this combined approach are then used as contexts for a large language model (LLM) to generate the final answer.

CCS Concepts

• **Information systems → Information retrieval**.

Keywords

Retrieval-Augmented Generation(RAG), Information retrieval, Text Embeddings, Approximate Nearest Neighbor (ANN) Search, Generative Language Models

ACM Reference Format:
Jing Wang. 2024. FinanceRAG with Hybrid Search and Reranking. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[*] jing.wang@accrete.ai

ACM-ICAIF'24, Nov 14 -17 2024, Brooklyn, NY, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Retrieval-Augmented Generation (RAG) frameworks leverage external knowledge sources to provide rich and accurate contexts for LLM applications. This study introduces a method that synergistically combines hybrid search techniques with advanced reranking strategies to optimize document retrieval. By employing a combination of semantic similarity and keyword matching in a multi-stage retrieval process, our approach enhances both the relevance and specificity of information retrieved from extensive corpora.

2 Method

The objective of Task 1 is to retrieve the most relevant document chunks (contexts) from a substantial corpus in response to a given query. This process aims to maximize retrieval relevancy. The evaluation framework assesses the system's ability to rank retrieved contexts compared to established ground truth.

2.1 Retrieval

1. Data Preprocessing. Each document consists of an article title and chunked text. We remove unnecessary markdowns and punctuation and correct invalid bytes. Texts that are excessively long and exceed the token limit of the embedding model are truncated.

2. Query Embedding Generation. Each input query is transformed into a high-dimensional vector representation (embedding) using OpenAI's `text-embedding-3-small` model. This embedding captures the semantic nuances of the query, enabling the similarity comparisons with document embeddings.

3. Initial Retrieval. The generated query embedding is utilized to perform a nearest neighbor search within a pre-indexed vector database containing all document chunks in the corpus. This database is optimized for rapid similarity searches, facilitating the retrieval of an initial set of candidate documents that are semantically aligned with the query. To enhance retrieval precision, a hybrid search approach is employed that combines keyword matching. This ensures that the retrieved documents are not only contextually relevant but also contain specific keywords inferred from the query. If the full-text keyword search returns no results, the system defaults to using semantic search only.

4. Reranking Mechanism. These candidate documents undergo a reranking process using ColBERT, a novel ranking model that employs contextualized late interaction over BERT [3]. We have tested other SOTA open-source models, as well as the Cohere Reranker `rerank-english-v3.0`, and found that our reranking system is on par with commercially available alternatives. This step assesses the contextual token similarity between the query and each document chunk, allowing the system to prioritize documents that exhibit higher relevance. Each document chunk is assigned a SumMax similarity score based on late cross-attention evaluation. These scores are used to reorder the documents, ensuring that the most pertinent contexts are ranked higher.

5. LLM Answer Generation. The top-k documents, as filtered by the reranking scores, are aggregated to serve as contexts for subsequent processing, such as answer generation by an LLM.

2.2 Retrieval Evaluation

The retrieval system is primarily evaluated using nDCG@10, MAP@10 and Recall@10, which assess the ranking quality of the top 10 retrieved document chunks against the ground truth on a subset of data. This metrics consider both the position and relevance of each retrieved document, providing a comprehensive assessment of the system's retrieval accuracy. Table 1 provides the full metrics for the retrieval step across seven datasets.

Table 1: Evaluation metrics across all datasets

Dataset	NDCG@10	MAP@10	Recall@10
ConvFinQA	0.522	0.498	0.595
FinDER	0.425	0.361	0.583
FinQA	0.331	0.311	0.395
FinQABench	0.871	0.850	0.933
FinanceBench	0.699	0.609	0.922
MultiHiertt	0.322	0.291	0.312
TATQA	0.801	0.762	0.922

2.3 Generation

In the answer generation step, we deployed OpenAI's `O1` model and Anthropic `Claude-3.5-Sonnet`. The quality of the generated responses varies based on the specific LLM and its application context. We evaluate the generated answers using the Retrieval-Augmented Generation Assessment Strategy (RAGAS) framework[2], which will be detailed elsewhere.

3 Conclusion

In this project, we developed and evaluated a robust retrieval and reranking system designed to identify the most relevant document contexts from large-scale corpora in response to user queries. By integrating hybrid search techniques with a reranking mechanism, our approach effectively balances semantic similarity and keyword specificity, ensuring that the retrieved documents are both keyword-anchored and contextually pertinent.

References

[1] 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

[2] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217 [cs.CL] https://arxiv.org/abs/2309.15217

[3] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. arXiv:2004.12832 [cs.IR] https://arxiv.org/abs/2004.12832