Analyzing NASDAQ Financial Fundamentals: A RAG Based LLM Financial Analyst System

Mohammed Ali-Khan[1], **Mohammed Usama Bin Hossian**[1], **Uzma Zehra**[1*],

[1]Department of Computer Science and Engineering, University of Texas at Arlington
CSE 6363.004 - Machine Learning - Fall 2025
mohammed.alikhan@mavs.uta.edu, mxh1127@mavs.uta.edu, uxz4241@mavs.uta.edu

Abstract

This work presents a Retrieval Augmented Generation (RAG) based Large Language Model (LLM) system designed to analyze financial information from neatly formatted csv files from NASDAQ sources, as well as specific company reports and text documents from publicly available SEC filings. The analysis is integrated into a vector-based retrieval layer to allow an LLM to return a clear, concise, and insightful response to a user given query.

Introduction

Currently, even as Artificial Intelligence continues to evolve, there are limitations to certain fields that could further be enhanced. Existing Large Language Models face notable limitations, especially when applied to financial analysis. Generic LLMs lack specialization in certain financial fields. One such application is the ability to extract meaningful insights from a given base of financial information. Current LLMs also operate as ”Black Boxes”, making the decision making process difficult to fully understand. This is especially important in regulated industries, where integrity and security are of great concern. Such models also do not possess the specialized knowledge to navigate the particular liability framework of the financial industry. LLMs are also prone to become rapidly outdated, resulting in an analysis being conducted on historical data or obsolete financial methods.

Retrieval Augmented Generation offers substantial benefits in overcoming these limitations, by providing pre-trained LLMs with additional up to date, transparent, and domain specific information, making such a model more effective in speciallized fields. RAG also significantly reduces the possibilities of ”Hallucinations”, when fabriculated information is produced by the LLM, by grounding the responses in a more controlled knowledge base. Using the knowledge base expands the explainability of the LLMs output, allowing a much better analysis to be made, and the source of the information to be traced. Finally, RAG models can effectively utilize descriptive and privitizable data specific to a single institution, allowing for customization, integrity, and security of financial analysis.

Dataset Description

The RAG based LLM architecture implementation is comprised of two main data sources. The first data source is the NASDAQ Fundamentals CSV Dataset from Kaggle. This is the same dataset used from our reference research paper. This database of financial information contains 186,336 rows and 6 columns. These key columns are Period, Company, Tickers, Indicators, Unit, and Amount.

The CSV dataset needs to be preprocessed so that it can be used as part of the RAG implementation. There were three major changes that were done to the CSV dataset, to allow for consistentcy as part of the retrieval process. First, the numerical values within the dataset were modified, so that there were no commas to separate numbers. In this way, both numbers 4.9B (Billion) and 4,960,500,000.00 could be represented as a standardized 4960500000 (4.9 Billion). Such number values could then be converted to a numeric data type, and NULL value rows could be removed. Secondly, Periods were parsed, as fianncial documentation typically represents assessments based on quarters. Each time frame was converted to an appropriate Quarter (i.e. 2015 Q2). Thirdly, For simplification with symbols and currencies, only USD amounts were kept. All records in foreign currencies were removed. The remaining filtered and modified numerical values from the csv were then parsed into an English sentence, that would later be converted to a vector embedding, and saved into a vector database for the retrieval algorithm to take.

The second data source is a collection of SEC Fillings on the top companies listed in the NASDAQ Fundamentals CSV, to match consistency. Various separate pdfs, most of which were company quarter reports, were taken from the publically available governmental site reponsible for holding SEC Fillings and parsed into txt files. These txt files were rpeprocessed, where only relevant lines that contained some type of financial information along with the corresponding quarter date were kept. All other lines, such as page numbers or desctiptive headers, were not needed and removed. These lines could then be converted to vector embeddings and saved in the vector database to be passed to the retreival algoritm.

[*] These authors contributed equally.

Project Description

Description

The process of our RAG based LLM can be demonstrated in the diagram below.

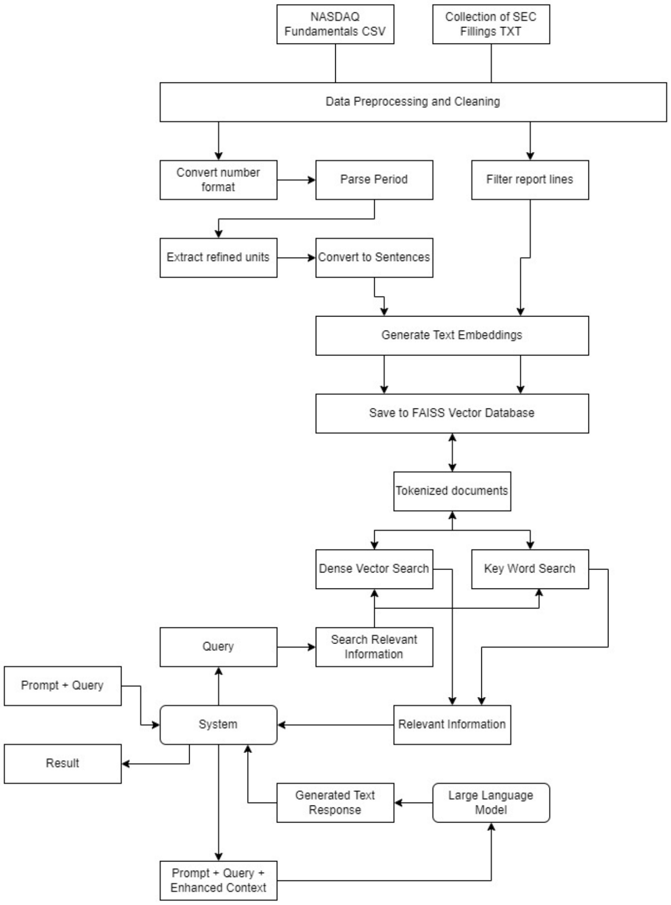
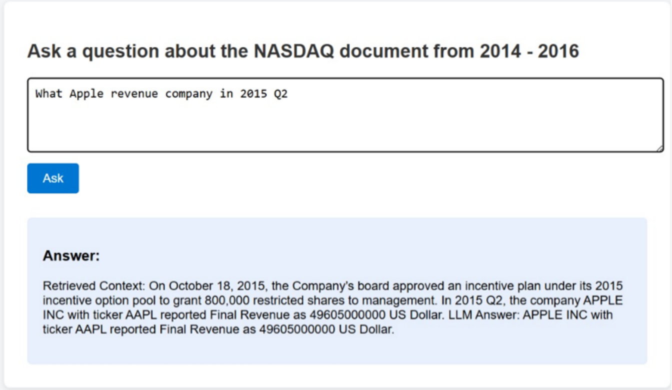


Figure 1: Modified RAG-based LLM Implementation

A user can interact with the system by Prompting a question. This question would contain some form of query, to which the answer will be based off and delivered in accordance to the relevant information already provided in the financial database. The Query is then sent to the Retrieval Algorithm of the system, which searches the relevant information within the Knowledge Sources, or financial database, and searches for the relevant material on which the LLM can base its answer on. There are two methods by which the Retrieval method works on. Dense Vector Search with FAISS, and Sparse Vector Search with BM25. Both methods will be discussed in a later section. The top-k, (for most of our purposes k=5) relevant documents are then returned by both search methods. Duplicate references are removed and are passed on to the LLM, along with the users original prompt and query. The LLM (both the gemeni model and google's flan-t5 model have been used in separate implementations) is able to give the user a response based on the selected query.

A sample of the implementation can be shown below. Where the user types a query, and the LLM returns an answer. Relevant context is also displayed for visualization purposes.



Ask a question about the NASDAQ document from 2014 - 2016

```
What Apple revenue company in 2015 Q2
```

Answer:

Retrieved Context: On October 18, 2015, the Company's board approved an incentive plan under its 2015 incentive option pool to grant 800,000 restricted shares to management. In 2015 Q2, the company APPLE INC with ticker AAPL reported Final Revenue as 49605000000 US Dollar. LLM Answer: APPLE INC with ticker AAPL reported Final Revenue as 49605000000 US Dollar.

Figure 2: Sample Output of RAG-LLM System

Above, the user has asked a question on ”What the Revenue of Apple was in 2015 Quarter 2”. The RAG retrieval sub-system pulls forth the context. In this example, two pieces of content is pulled from two separate sources. The statement on Apple's plan to grant shares to management is taken from Apple's Financial Document from the SEC Filling Reports. AAPL stock ticker reporting a 4.9 Billion is taken from the NASDAQ csv data source. Both of these contents, along with the user's original prompt and query is passed to the LLM, and generates the answer. The LLM (flan-t5) correctly chooses to ignore the company share statement, and finds the answer to the user's question in the revenue reported. That content is then written in an answer format and returned to the user.

Main References

The main reference used for this project is a research paper titled ”Financial Analysis: Intelligent Financial Data Analysis System Based on LLM-RAG by Wang, J., Ding, W., and Zhu, X. ”. The research paper discusses an implementation of RAG based LLM for financial analysis using the NASDAQ Financial Fundamentals CSV data source. The researchers suggest the use of a Large Language Model to be used with a Retrieval Augmented Generation system to retrieve relevant financial information in order to do an analysis on financial trends. An implementation of a vector based storage system is implemented in order to hold all of the context necessary to be retrieved by RAG query module. The researchers demonstrate their findings, where the LLM was able to significantly improve the accuracy of its responses, as well as reduce response times. This is demonstrated to enhance efficiency in the LLM's ability to respond to queries, but seems to show an increase in memory utillization. All in all, the researchers showcase the value in such a LLM using RAG as an improvement to previous methods of financial analysis.

Other references include ”Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks” as part of Facebook AI Research. This group of researchers also highlight the many advantages that RAG based LLM models have over traditional baseline and even pre trained LLM models. Different implementations of RAG based models are compared to baseline LLM models, to showcase RAG based models are most effective in generating more specific and fact based responses, as opposed to the LLM outputs of vast and generalized responses.

Another reference is ”Enhancing Medical AI with Retrieval-Augmented Generation: A Mini Narrative Review” by Omid Kohandel Gargari and Gholamreza Habibi. This research paper demonstrates the usage of a RAG based LLM system in order to understand and analyze medical data. This is insightful, as our project focuses on financial analysis, which is simply another field where numerical data can be contextualized into quatified insights and applications, demonstrates RAG's ability to be used for a multitude of subjects and focuses based on the systems needs.

Difference in Method between our project and implementation shown in the reference papers

There were two main differences between our implementation and the implementation of the reference paper. The first being the inclusion of an additional data source. Given that RAG implementations are mainly used by test data sources, it seemed fitting that a text data source be used, as opposed to the research paper's sole reliance on formatted tabular data. Hence, the SEC Filling Quarterly Reports for various companies were taken and put alongside the CSV data set. This not only justifies the use of a RAG based LLM, given that RAG based systems are designed to be effective at textual data, but also allowed us to provide more insight to the LLM in order to further showcase the RAG ability in making an analysis.

The second difference in our system, compared to the reference paper is the implementation of a hybrid semantic search by combining Dense Vector Search, and Sparse Vector Search.

For the Dense Vector Search, FAISS (Facebook AI Similiraty Search) is used to index the documents. The documents, or content of the retrieval databse, is tokenized, and compared to the semantic similarity of the user query. FAISS indexes based on the closest documents, and the top-k documents are sent back to the system.

For the Sparse Vector Search BM25 is used to index the documents. The documents, or contents of the retrieval database, are tokenized. The term frequency, based on the query being sent by the user, is used to index the documents. The top-k documents that are the most similar in exact words to the query are sent back to the system.

Taking both the semantics of the query as well as the keyword relevance, the LLM is able to utilize a much better understanding of relevant information in order to respond to the user. Numeric facts and domain specific terminologies can then be gathered much more accurately. This gives a better recall and higher accuracy of the LLM overall.

Difference in Accuracy between our project and the implementation of our reference

- Metric Difference: Our project uses F1/BLEU/Cosine (quality metrics), while the paper uses Accuracy/Recall (success rate metrics).
- Performance Result: The project's average RAG F1 (94.0) is +15.4 percentage points higher than the paper's Accuracy (78.6).
- Methodological Explanation: The project uses a Hybrid Search (FAISS + BM25) and Dual Data Sources, which likely accounts for the high F1 score by improving grounding.

Analysis

What was done well?

We successfully implemented a **Retrieval Augmented Generation (RAG) based Large Language Model (LLM)** system for financial analysis, which effectively addresses several limitations of generic LLMs. Our key accomplishments include:

- **Hybrid Data Sourcing:** We utilized a dual data source approach, combining **structured CSV data** (NASDAQ Fundamentals) with **unstructured text data** (SEC Filling Quarterly Reports). This was a deliberate difference from the main reference paper's sole reliance on formatted tabular data, which better justifies the use of a RAG system and allowed us to provide more insight to the LLM.
- **Hybrid Search Implementation:** We implemented a robust **hybrid semantic search** by combining **Dense Vector Search (FAISS)** and **Sparse Vector Search (BM25)**. By taking both the semantics of the query as well as the keyword relevance, the LLM is able to utilize a much better understanding of relevant information. This is expected to give a better recall and higher overall accuracy.
- **Comprehensive Data Preprocessing:** We performed significant preprocessing on the NASDAQ CSV dataset, including standardizing numerical values (e.g., converting 4.9B to 4960500000), parsing periods into Quarters (e.g., 2015 Q2), and filtering records to keep only USD amounts. For the SEC filings, we preprocessed them to keep only relevant lines containing financial information and the corresponding quarter date.
- **Results:** The comparative data demonstrates the efficacy of Retrieval Augmented Generation (RAG) for grounded factual queries from private knowledge sources [Table 1].
  - Factual Accuracy (F1 0.92-0.96): Near-perfect scores confirm successful retrieval and accurate extraction of facts without hallucination.
  - Semantic Consistency (Cosine Similarity 0.97-0.99): Scores near 1.0 confirm the RAG response is semantically equivalent to the Ground Truth.
  - Lexical Quality (BLEU 0.82-0.88): High scores confirm fluent, structurally similar, and readable answers.

- The LLM-Only model scores were consistently near zero, indicating critical failure due to generating a refusal (e.g., ”I cannot find the specific Q1 2014 total assets...”). The near-zero F1/BLEU scores and low Cosine Similarity confirm a complete failure to provide the correct factual answer.

What could we have done better?

While the implementation was successful, we recognize that the project lacks critical validation and comparison data, which is essential to fully verify the system's superiority. Areas where we could have done better include:

- **Insufficient Performance Testing:** Much testing and comparisons between models and implementations were not fully done. Specifically, the hybrid search RAG-LLM system has not been compared to the performance of an LLM alone or to the baseline (non-hybrid) RAG-LLM. We acknowledge that testing results for the hybrid search seem erroneous for the time being and require more and better test cases.
- **Lack of Hallucination Testing:** Testing was not done on hallucinations. It is unknown if the RAG-based LLM will begin to give incorrect or false information, by either referencing irrelevant information to the query, or making up information not found at all in the financial vector database.

What is left for future work?

Much testing and comparisons between models and implementations were not fully done. For instance, a hybrid search RAG based LLM was implemented (and is shown in the sample output), but the testing results do seem erroneous for the time being as a more and better test cases need to be written to compare them to. Additionally, the hybrid search model has not been compared to the performance of an LLM alone, or compared to the baseline (non hybrid) RAG-LLM. We hope to see that by doing so there is an increased improvement as the LLM is able to gather much more relevant sources.

Testing was also not done on hallucinations. The LLM model alone provides no references to a specific query, but it is unknown if the RAG based LLM will do the same. Or if the RAG-based LLM will begin to give incorrect or false information, by either referencing irrelevant information to the query, or making up information not found at all in the financial vector database.

These are implementations and testing validations that can be further improved and implemented in order to fully verify the superiority of RAG based LLMs.

Conclusion

In conclusion, the RAG based LLM system was succeful, as the system was able to take in the formatted NASDAQ Fundamentals CSV dataset, along with select individual SEC Quarterly Report Fillings, and parse them into a vector database using Sentence Transformers. A user is able to ask a query about the information held inside the financial database. The hybrid implementation enables the system to index the database based on the queries using both FAISS indexing and BM25 indexing and returns the top relevant information. The LLM gemeni-model and flan-t5 model are able to take the retrieved context, and develop an answer back to the user.

It has also been shown that a RAG-based LLM model performs significantly better than baseline LLM models, given the needed context is provided, allowing for relavent information to be retreived at much higher speeds. Augmentation happens with more insight thanks to the hybrid search process, and the LLM is able to generate a better answer grounded in a factual base of information.

Our implementation of a RAG based LLM for Financial Analysis is able to take context from both structured and formatted sources, such as a clean csv file containing financial information, as well as unstructured, raw text file, such as the ones taken from SEC Fillings of various companies, and still generate a proper response.

All in all, this implementation, and others similar to it, demonstrates the superiority of RAG assited LLMs as being more accurate at factual data, and more useful for privatized and niche based institutional usage. Such an implementation allows for further development in expanding the world of Generative AI, and brings the power of Large Language Models to even the most specialized and distant disciplines.

References

- Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11418.
- Gargari OK, Habibi G. Enhancing medical AI with retrieval-augmented generation: a mini narrative review. Digit Health. 2025; 11:20552076251337177.
- Oche et al. (2025) A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions. arXiv:2507.18910
- Gupta, S. (2023). The Rise of Retrieval-Augmented Generation (RAG) in Enterprise AI. Forbes.
- Beaumont, R. (2020, December 1). Semantic search with embeddings: Index anything. Medium.
- Wang et al., 2025, arxiv.org - “Financial Analysis: Intelligent Financial Data Analysis System Based on LLM-RAG” https://doi.org/10.48550/arXiv.2504.06279

Table 1: Comparative Analysis of RAG vs. LLM-Only Performance with Dummy Metrics

Query	RAG System (Augmented Output)				LLM-Only (Zero-Shot Output)			
	Response	**F1**	**BLEU**	**Cosine**	**Response**	**F1**	**BLEU**	**Cosine**
1. What were the total assets for PIH in 2014 Q1?	The assets for PIH in 2014 Q1 were **42,854,000 US Dollar**.	0.96	0.88	0.99	I cannot find the specific Q1 2014 total assets for PIH.	0.05	0.00	0.15
2. Tell me the income from continuing operations before taxes for the company.	The Income from Continuing Operations before Taxes was **2,307,000 US Dollar** in 2014 Q1.	0.92	0.85	0.97	I don't have enough current or historical financial data for PIH.	0.07	0.00	0.18
3. How much did the company's cash equivalents increase or decrease during that period?	The Cash and Cash Equivalents, Period Increase was **3,323,000 US Dollar** (an increase).	0.94	0.82	0.98	I cannot access the specific change in cash equivalents for PIH.	0.04	0.00	0.11