Analysis of Retrieval-Augmented Generation (RAG) Pipelines using Gemini for Financial Document Analysis.

Gummadi Sai Dheeraj

October 26, 2024

Contents

Abstract

This research paper explores the efficiency and effectiveness of different Retrieval-Augmented Generation (RAG) pipelines for financial document analysis using Google's Gemini model. Two distinct approaches are compared: Image-Based RAG, which processes images of documents, and Text-and-Image-Based RAG, which extracts text and images separately. The Text-and-Image-Based RAG pipeline proves significantly more efficient, processing data in 30 minutes compared to 2 hours and 30 minutes for the Image-Based pipeline. The study further investigates the impact of four different prompting techniques on Gemini's performance, highlighting the importance of prompt engineering in guiding the model towards accurate and insightful responses. The paper analyzes the effectiveness of various prompting techniques, including chain-of-thought, few-shot learning, and knowledge elicitation, in guiding Gemini's reasoning process. The research concludes by analyzing the strengths and limitations of each RAG pipeline and prompting technique, providing valuable insights for researchers and practitioners seeking to leverage Gemini for financial document analysis.

1 INTRODUCTION

Financial Document analysis is a fundamental pillar of the financial industry (Siddiqui, 2023), playing a crucial role in assessing the creditworthiness of individuals and businesses seeking financial assistance. It involves a meticulous examination of a borrower's financial history, credit history, and ability to repay the debt. This process is essential for lenders to manage risk effectively, make informed credit decisions, and ultimately, ensure the profitability of their loan portfolios.

1.1 Understanding Credit Analysis Using Financial Documents

Credit analysis is the sequential process that covers several aspects of the financial well-being of a borrower to determine the borrower's ability to pay off debt(Segal, 2023). Most often, the process involves the analysis of financial statements, cash flows, and other financial data to generally assess the fiscal health of the borrower and their creditworthiness. Non-financial factors also include the quality of management and industry trends and competitive landscape in pursuit of a balanced view of the borrower's overall credit risk(K, 2024).
Credit analysis refers to the systematic evaluation of an individual or corporation's creditworthiness in the context of a debt. This method assesses a range of financial and non-financial factors to ascertain the borrower's capacity to fulfill loan repayment obligations.
Financial statements are an integral component of the creditworthiness assessment, providing a lender with an overall view of the financial position of a borrower. Such statements include the balance sheet, income statement, and the cash flow statement(Kumaraswamy, 2024). The balance sheet presents a detailed account of a firm's assets and liabilities, whereas the income statement outlines its revenues and expenditures. The cash flow statement delineates the company's cash inflows and outflows, highlighting its capacity to generate cash from operational activities. Through the examination of these financial statements, lenders are able to assess the borrower's debt-to-income ratio, debt service coverage ratio, and various other critical financial indicators. These metrics facilitate lenders in evaluating the borrower's capacity to repay the loan while simultaneously managing associated risks.

1.2 Need of Analyzing Financial Documents of Companies

Corporate financial statements are crucial for examining by a wide group of stakeholders, including creditors, investors, and financial analysts(NSIM, n.d.). It provides a clear and comprehensive view regarding the financial health of an enterprise, including profitability, liquidity, and management of debt. These numbers are essential for making lending, investment, and investment portfolio management decisions. Credit analysis is one of the primary tools used by lenders and investors in determining the creditworthiness of prospective borrowers.

For the lenders, the ability to review the financial statements would offer them an assessment of the borrower's ability to service the loan. They could trace the cash flow ability of a specific borrower and enable them to understand whether the debtor may pay his or her debt obligations in time or not. This information is useful to evaluate their creditworthiness, interest rate, and credit exposures.

To an investor, financial document analysis allows insight into the financial health and growth of a company. Investors can easily determine the profitability, efficiency, and leverage of a company(K, 2024). Thus, this allows them to make well-informed investment decisions and properly allocate capital.

Financial analysts employ financial statement analysis as a method to assess a corporation's financial performance both longitudinally and in comparison to its competitors. Analysts utilize the data derived from financial document analysis to construct financial models that facilitate the forecasting of key metrics such as revenue, expenses, and profitability(Tuovila, 2024). This information plays an essential role in formulating investment recommendations and enabling informed decisions regarding the buying, selling, or retention of a company's stock.

1.3 Brief Overview of LLM Technologies

Large Language Models are a particular type of AI algorithm that utilizes deep learning techniques and large datasets to help in the analysis and production of human like text communication (Filippo et al., 2024). LLM's have revolutionized the field of NLP by enabling machines to understand and generate text at levels much higher than previously imagined. Large language models, such as GPT-4, have proven incredibly competent in the question-answering domain, especially through their elaborate structure and immense training data(Kalyan, 2023).

LLMs are based on the Transformer architecture proposed first by (Vaswani, 2017) in 2017. Transformers are highly proficient at capturing context and semantic meaning in sequential data-including but not limited to text. A typical architecture of a transformer is a stack of several transformer blocks or layers. Most layers for such a model consist of self-attention mechanisms, feed-forward neural networks and norm components. This, in principle, helps the model to weigh the importance of different parts of the input through self-attention, whereas the feedforward and normalization layers make computation more efficient and stable, respectively. By layering up these stacks, it is therefore possible to construct progressively deeper and more robust transformer models that can subsequently execute more complicated tasks.

LLMs have developed rapidly on a number of dimensions. Models have shrunk, become smaller in power consumption, and more sensitive to performance on some applications - such as device-bound use cases. Cloud and platform players, meanwhile, drive the future development of AI, expanding access and affordability for LLMs.

1.4 Need of LLMs in Document Analysis and Credit Analysis

The need for LLMs in document analysis and credit analysis is driven by the increasing volume and complexity of data that financial institutions need to process(Kim et al., 2024). Traditional methods for analyzing financial documents, such as manual review and rule-based approaches, are often time-consuming, prone to human error, and lack scalability (Haider, 2024). The LLMs are more efficient and thus more accurate in document analysis, leading to meaningful insights from data, automating tedious repletion tasks, and making better decisions(Lin et al., 2024).

The pace at which LLMs are being integrated into financial services is accelerating(Nucci, 2024). They are no longer a thing of the new world. According to (Fasha et al., 2024), OWASP the Open Source Foundation for Application Security that recently published the Top 10 guidelines on ensuring security aspects of LLM applications as well as older, legacy versions of the application, including web, mobile, desktop, etc.

1.5 RAG vs Finetuning LLMs

This model integrates large language models and external sources of information into a single algorithm to enhance the functionality of LLMs. RAG(Lewis et al., 2020) frameworks retrieve data from sources outside the models - for example, databases or knowledge graphs-and then transmit that data to the LLM for synthesizing and formulating a response. This methodology allows LLMs to access an entire new set of information, making them better placed to provide more complete and accurate responses.

The RAG method involves data processing from the loading phase up to subsequent storage in a database using a vector database optimized for efficient and systematic retrieval of information. At the same time, the process of data retrieval will be integrated into the overarching framework of general LLMs. Unlocking New Horizons of Large Language Models in Financial Services

Finetuning the pre-trained LLM on a specialized dataset to improve its performance on a particular task. This has been applied to meet specific needs in a particular domain or application, for example, financial analysis or customer services.

RAG and finetuning are complementary techniques that may be used in parallel to improve the capability of LLMs. RAG would be very helpful for those tasks that require access to a large amount of external data and get the results with high accuracy, while finetuning would be more expensive task.

1.6 Existing Works Done in Applying LLMs in Finance

The application of LLMs in finance is an area that is rapidly moving forward, whereas tremendous research is in full swing to explore LLM's potential on many tasks. Some of the most important areas in which researchers are using LLMs in finance are as follows:

- Analysis of financial statements(Kim et al., 2024): Researchers are finding an automatic way through which financial statements can be analyzed for several things, such as key trends and ratios, generating financial summary, or predicting future earnings.
- Credit scoring(Feng et al., 2023): LLMs are being educated on historical data to predict the possibility of a borrower defaulting.

- Risk assessment(Cao et al., 2024): LLMs are being used for the assessment of risk in the decision of extending credit to a borrower. This is keeping into consideration various factors and risks such as the borrower's history of finances, industry trends, and competitive landscape.
- Investment strategy(Basyal & Sanghvi, 2023): LLMs are being generated to produce investment recommendations, analysis of market trends, and identifying potentials where investments can be channeled.
- Fraud detection: LLMs are helping to identify fraud by creating deep fake videos and even voice generating scams.
- Customer service: LLMs are used to power chatbots and thus provide customer support, answering questions, and solving issues.
- Compliance: LLMs are used for automating compliance work that involves reviewing financial documents and making sure they have met the regulatory requirements.
- LLMs in Portfolio Management: LLMs can be applied toward market trends which can themselves be helpful in the detection of investment opportunities along with the optimization of portfolio allocation.
- LLMs in Regulatory Compliance: LLMs can be used to automate compliance-related tasks, such as reviewing financial documents and adherence to regulatory requirements.
- Fraud Detection LLMs: These can be used to detect fraudulent activities, such as deep fake videos and voice generation scams.
- Customer Service LLMs: LLMs can power chatbots that provide customer service, answer questions and solve problems.

The application of LLMs in finance is yet in its infant stages, but the amount of revolution that these models could bring about in this industry is huge. As LLMs continue to be developed and improved, more innovative applications related to models in finance were explored. This is an area where LLMs can revolutionize the financial industry, with a reduction in tasks to be performed, increased efficiency, and new insights that can emerge in the process. These considerations and challenges surrounding the models need to be addressed responsibly and ethically.

2 THE METHOD

After Looking at the multiple approaches stated in the literature which are specific to individual domains, A novel pipeline for generic credit analysis of businesses is proposed in this section.

2.1 Data

This paper explores the application of advanced natural language processing (NLP) techniques, specifically Retrieval-Augmented Generation (RAG), for extracting the key information from the financial documents. To look into this, The dataset comprising financial documents of public-listed companies in Frankfurt Stock market is used. The whole process can be found in this section describing sources, selection criteria, and further processing into an analysis-friendly format. The data has been gathered from the individual company financial websites which have been the primary sources of financial data. Specifically, in the ”Historical Data” section for each company, they provide convenient direct access to annual reports, financial statements, etc., while the stock price data comes from the same source (Investing.com, n.d.) the chosen companies. To cover a wide cross section of industries and domains, companies listed on the Frankfurt Stock Exchange have been randomly selected. Hence, the random selection avoided biases due to any preferred sectors or other factors of company size. The choice criteria were based on companies whose financial documents are easy to obtain and whose historical stock price data are available on the selected websites.

2.2 Vector Databases Benchmarking

A vector database is a database that is specifically designed for the storage and retrieval of high-dimensional vectors, often used to represent data such as in the case of text, images, and audio. They are found to be quite useful for similarity search applications such as RAG, whose objective is to find the closest points which should be similar to a given query vector. Milvus(Wang et al., 2021) is an open source vector database engineered for high-performance in similarity search. Supports multiple types of data, including text, images, and audio files. The indexing algorithms also differ, such as HNSW, IVF_FLAT, and IVF_SQ8. Milvus also provides extensibility in scaling up to large amounts of vector data. In addition to that, batch operations, real-time updates, and distributed deployment are supported. Qdrant is an open-source vector database intended for similarity search, besides vector-based analytics. It is multi-type and works with text, images, and audio, besides offering a number of features for data management and querying - filtering, ranking, and faceting, besides integration capabilities with plenty of tools and frameworks. In addition, it offers a variety of features for data management and querying, including filtering, classification, and face-tracking support. PGVector is an extension for PostgreSQL that sets a specific data type for the fast storing and indexing of high-dimensional vectors. PGVector is designed to interface with PostgreSQL, which is an open source relational database management system widely used in the marketplace and offers several facilities in the form of data management and vector querying. Among these are similarity search, filter operations, and ranking functionality and could perfectly interface with PostgreSQL, allowing its users to benefit from the ecosystem and tools already built around the PostgreSQL server. The finance information being extracted is sourced from financial documents through automated document scraping techniques. This technique basically includes identifying the relevant elements that contain the desired information and then extracting it into structured formats.
The extracted stock prices and financial information are converted to vectors by the model ”textembeddinggecko@003”. It is a very potent text embedding tool, transforming textual data into numerical vectors capturing the semantic relationship between words and phrases. The vectors obtained were stored in an open source vector database. The hardware configurations of the system where the experiment has been carried out are presented in Table 1 and the reference numbers are presented in Table 2

2.3 RAG pipelines Analysis

This paper presents the analysis of the effectiveness of two Retrieval-Augmented Generation (RAG) pipelines using ”Gemini 1.5 Flash” (Reid et al., 2024) model for extracting meaningful insights from financial documents. The pipeline leverages the capabilities of large language models (LLMs) and

System OS	Windows
Version	10.0.22631
Machine	AMD64
Processor	Intel64 Family 6 Model 186 Stepping 3
Physical Cores	10
Logical Cores	12
Total Memory	32 gigabytes

Table 1: Hardware Configurations

VectorDB	Indexing Algorithm	Avg Retrieval time
Milvus	HNSW	1.2s
Qdrant	HNSW	1.68s
PGVector	HNSW	0.8s

Table 2: Vector Database benmchmarking results for retrieving top-5 similar results.

vector databases. This paper experimented with two distinct approaches, and the pipelines are given in Fig 1, 2

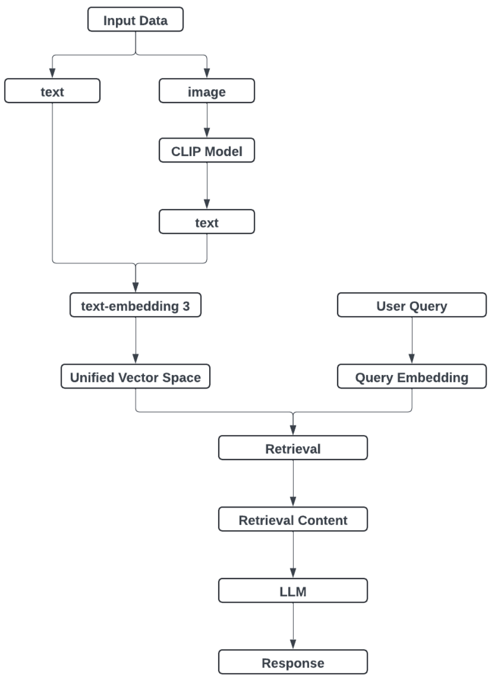

Figure 1: Text based Pipeline. Given the documents text and images are extracted from the documents. Images are converted to text using clip model and the extracted text is uploaded to Vector database (Source: Own Results)

Applying the image-based RAG method, all pages of the PDF were extracted and converted into images. Text extraction and figure summarization also took place, where texts were then conversed into embeddings and stored within a single vector database. During inference, the single collection was queried. This approach was interested in preserving the context of visual content within a document and would not lose the contextual information if extracted on its own as text. This is very helpful for accounting reports, which are replete with graphically represented information such as charts, graphs, and tables that carry critical data. However, rendering pages as images along with text extraction and summarization of figures took over 30 minutes for a 12-page document. Which makes the approach

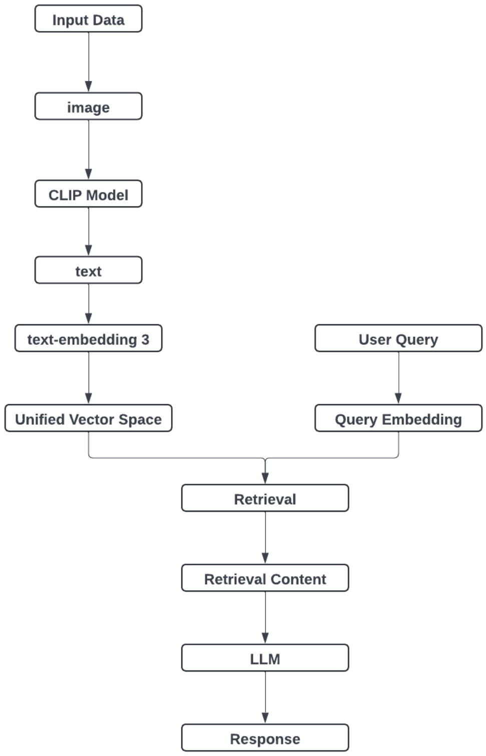

Figure 2: Image Pipeline, Given the documents all the pages are converted to images and the images are converted to text using clip models. This helps in using the information of images mentioned wiht the context. (Source: Own Results)

utterly impractical for real-time use. Although this preserved the context of images, the results produced by this method were less accurate than those anticipated. This might be due to the fact that current techniques in image processing have proved ineffectual in the retrieval and interpretation of extremely complex visual data especially against the background of financial documents.

Text-based RAG approach extracts both text and images separately from the PDF document. The extracted text is encoded as embeddings, and images are summarized using image captioning, and combined and stored the summaries of both text and images together in one vector database, then targeted inference queries at that overall collection. This enabled a single collection to be searched for information both about the text and the images related to it, thereby simplifying retrieval while at the same time enabling a holistic understanding of the document as well. In this approach, in comparison with the RAG pipeline based on images, significantly better accuracy was displayed by retrieving the pertinent information. This is because LLMs process and understand textual data much better than usual, using the summaries of images to contextualise. The approach loses the detailed visual context that images provide, however, which may be a limitation in cases where complex financial documents have visual elements that play significant roles in passing on information.

Our experiments show that the text-based RAG pipeline outperforms its image-based counterpart in terms of accuracy and efficiency. This can be primarily explained by the strengths of LLMs in processing and understanding textual data, whereas the image-based approach has the advantage of preserving context from visual features that is a great feature in certain scenarios. In inference, latency of a text-based RAG pipeline is significantly lower compared to the image-based approach. This latency difference is critical for real-time applications in which immediate responses are of the essence.

The shortcomings of each approach therefore reveal an area that needs further research and development in financial document understanding. Some possible areas of future research may include improvement in image processing techniques, enhanced summarization for images, and hybrid approaches of RAG that combine the strengths of image-based as well as text-based methods, along with domain-specific knowledge infused into the pipeline of RAG to further enhance the accuracy and relevance of information retrieved. And combining domain-specific knowledge and advancing the capability of LLMs as well as image processing techniques can make more robust and effective pipelines for RAG to be helpful in extracting insights from financial documents.

Approach	Text Extraction + Embedding Creation time
Image based Approach	2 hour 34 minutes
Text based approach	38 minutes

Table 3: Text extraction and Embedding creation time for Image based RAG and Text based RAG pipelines.

3 RESULTS

This section explains our experiments on how various prompts perform in extracting the information from the financial documents using text based RAG approach. The focus has been placed on the extraction of critical information like net sales, gross profit, and EBITDA, which are considered critical in terms of financial analysis. Four different strategies are being compared:

3.1 RAG with Question

This baseline approach involved simply providing the LLM with the question. The sample prompt is shown in Table 4

Prompt	Looking at the financial documents given information of adidas: $information$. Answer the following question: $question$
Question	Predict the net sales, gross profit and EBITDA in 2023
Answer	The following projections for adidas in 2023 are based on the data provided:- **Net Sales:** The company's net sales grew by 16% in 2021 and by 1% in 2022. The global economic recovery is expected to continue in 2023, albeit at a slower rate. sporting goods industry is expected to benefit from major events in 2023. Based on these factors, we project that adidas's net sales will increase by 5% in 2023, reaching €23.25 billion. - **Gross Profit:** - adidas's gross profit margin decreased to 47.5% in 2022 from 52.2% in 2021. - This decrease was mainly driven by higher supply chain costs, as well as increased discounting and lower average selling prices. - While we expect the supply chain challenges to persist in 2023, we also anticipate some improvements in the latter half of the year. - Therefore, we project that adidas's gross profit margin will stabilize at around 47% in 2023.- **EBITDA:** - EBITDA decreased by 28% in 2022, primarily due to the sale of the Reebok business and higher operating expenses. Assuming a continued focus on cost optimization and efficiency, we expect EBITDA to increase by 10% in 2023, reaching €1.7 billion. These projections are based on a number of assumptions, and the actual results may vary. However, they provide a reasonable estimate of what adidas's financial performance could be in 2023.

Table 4: Sample Input and Output of RAG

3.2 RAG with Few shot prompting

Few shot prompting (Brown, 2020) approach provides the LLM with a few example question-answer pairs along with the target question.The sample prompt is shown in Table 5

Prompt	Question: Looking at the financial documents information of adidas. Try to predict the net sales and gross profit and EBITDA in 2022. Response: $'netsales' : 22511million,' grossprofit' : 10,644million,' EBITDA' : 1874million$ Question: Based on the financial documents information, estimate the operating profit and total assets for Adidas in 2022. Response: $'operating_profit' : 669million,' totalassets' : 20,296million$ Looking at the financial documents given information of adidas: $information$. Answer the following question: $question$
Question	Predict the net sales, gross profit and EBITDA in 2023
Answer	$'netsales' : 21,457million,' grossprofit' : 10,444million,' EBITDA' : 1,358million$

Table 5: Sample Input and Output of RAG with Few Shot prompt

3.3 RAG with Chain of Thought

Chain of Thought (Wei et al., 2022) approach involved prompting the LLM to explicitly articulate its reasoning steps, breaking down the complex question into a series of simpler sub-questions. The sample prompt is shown in Table 6

Prompt	Question: Looking at the financial documents information of adidas. Try to Project the net sales and gross profit and EBITDA in 2022. Response: net sales in 2019 is 23,640 and in 2020 decreased by 16 percent to 19,844 and increased back again by 7 percent to 21,234. Gross profit in 2019 is 12293 got reduced by 19.83 percent to 9855 and increased by 9 percent to 10,765 in 2021. EBITDA is 3845 in 2019 which got reduced by 45 percent to 2079 and again increased back again by 327 percent to 8,892. Based on the above trends the values of net sales, gross profit and EBITDA of 2023 are projected as follows: $'netsales' : 22511million,' grossprofit' : 10,644million,' EBITDA' : 1874million$ Question: Based on the financial documents, estimate the operating profit and total assets for Adidas in 2022. Response: operating profit in the 2019 is 2,660 and in 2020 decreased by 71.5 percent to 756 and increased back again by 162.7 percent to 1,986 in 2021. total assets in the 2019 is 20,680 and in 2020 increase by 1.8 percent to 21,053 and increased again by 5.1 percent to 22,137 in 2021. Based on the above trends the values of operating_profit, and total assets of 2023 are projected as follows: $'operating\_profit' : 669million,' totalassets' : 20,296million$ Question: Looking at the financial documents given information of adidas: $information$. Answer the following question: $question$
Question	Predict the net sales, gross profit and EBITDA in 2023
Answer	net sales in 2021 is 21,234 and in 2022 increased by 1 percent to 21,457. Gross profit in 2021 is 10,765 got reduced by 1.12 percent to 10,644 and decreased by 1.6 percent to 10,466 in 2023. EBITDA is 1874 in 2021 which got reduced by 27.5 percent to 1,358 and again increased back again by 27.5 percent to 1,874 in 2023. Based on the above trends the values of net sales, gross profit and EBITDA of 2023 are projected as follows: $'netsales' : 21,457million,' grossprofit' : 10,466million,' EBITDA' : 1874million$

Table 6: Sample Input and Output of RAG with Chain of Thoughts prompt

3.4 RAG with Tree of Thoughts

Tree of Thought prompting(Yao et al., 2023) helps LLM's to explore multiple lines of thought before arriving at a final answer.The sample prompt is shown in Table 7

Prompt	Imagine three different financial experts are Looking at the financial document information of adidas question: $question$ and Consider possible scenarios for moderate growth, low growth, and high growth.
Question	Predict the net sales, gross profit and EBITDA in 2023
Answer	Here are the projections of the net revenue, gross profit, and EBITDA for 2023, considering three possible scenarios:— Scenario — Net Revenue (in millions) — Gross Profit (in millions) — EBITDA (in millions) Moderate Growth €22,800 €10,600 €3,000 Low Growth €21,600 €9,900 €2,700 High Growth €23,400 €11,200 €3,300 **Assumptions:** **Moderate Growth:** This scenario assumes a continuation of the recent trend of moderate revenue growth. The growth rate would be around 3%, aligning with the company's long-term guidance.* **Low Growth:** This scenario considers a slower growth rate due to potential headwinds like economic uncertainty, inflationary pressure, and the ongoing war in Ukraine. The growth rate could be in the range of 1-2%.* **High Growth:** This scenario reflects a more optimistic outlook, assuming a strong recovery in consumer spending and a favorable global economic environment. The growth rate could exceed 5%.**Methodology:**The projections are based on historical data and current trends, considering the company's performance in recent years, market conditions, and the expected impact of the war in Ukraine. **Important Notes:*** These are just projections, and actual results may vary.* The company's financial performance could be impacted by various factors, including macro-economic conditions, changes in consumer spending, and the competitive landscape.* The company's financial strategy and management decisions will also play a significant role in determining future results.* It's important to refer to the official financial statements and reports released b adidas for the most accurate and up-to-date information.

Table 7: Sample Input and Output of RAG with Tree of Thoughts prompt

The above prompts were tested on multiple listed companies financial documents and results are summarized in Table 8, 9 10. The results demonstrate that RAG with Tree of thoughts Prompting consistently outperformed the other approaches in extracting the financial information. This suggests that providing a few example question-answer pairs significantly enhances the LLM's ability to understand and reason about the information, leading to more accurate and reliable responses.

	Normal Question	Few Shot	Chain of Thought	Tree of Thought	Ground Truth
Adidas	23,250	21,457	21,457	21,500	21,427
Bayer	49,873	47,545	46,989	47,122	47,637
Continental	44,000	40,757	38,101	42,024	41,420
Volkswagen	321,040	319,291	317,871	322,721	322,284

Table 8: Performance of Different RAG Pipeline Configurations for Extracting Net Sales in million euros

4 CONCLUSION

The paper presents the efficiency of Retrieval-Augmented Generation (RAG) pipelines in extracting insightful information from the financial documents of publicly listed companies on the Frankfurt Stock

	Normal Question	Few Shot	Chain of Thought	Tree of Thought	Ground Truth
Adidas	-	10,444	10,466	9,900	10,184
Bayer	-	46930	46819	47396	47,637
Continental	-	-	-	-	-
Volkswagen	-	62897	62886	61534	61,022

Table 9: Performance of Different RAG Pipeline Configurations for Extracting Gross Profit in million euros. Normal question extracted in terms of percentage growth which is not considered as right format of extraction.

	Normal Question	Few Shot	Chain of Thought	Tree of Thought	Ground Truth
Adidas	1,700	1,358	1,874	2,700	1,358
Bayer	9,136	9,254	9,280	9,811	9,529
Continental	3,809	4,144	4,128	4,095	4,079
Volkswagen	38106	38183	38131	36150	36,513

Table 10: Performance of Different RAG Pipeline Configurations for Extracting EBITDA in million euros

Exchange. Our study covers many aspects of design of an RAG pipeline: the choice of vector database, whether an image or a text-based approach, and prompting strategy.

Our results demonstrate that PGVector performs better in vector database operations than Milvus and Qdrant, but it is efficient enough to be a good choice for real-time analysis of financial documents. Furthermore, text-based pipelines were stronger in production than their corresponding image-based pipelines and indicate the strength of LLMs for text processes and understandings.

Among the prompting strategies tested, ToT was the most effective because it produced higher accuracy results compared with Chain of Thought, Few-Shot prompting, and the traditional question-based approach. The improvements show that nudging LLMs to consider multiple lines of thought when doing so with ToT prompting is much more promising in extracting meaningful insights from financial documents.

The current study provides promising insights into the design and optimization of RAG pipelines for financial document analysis and therefore warrants more exploration. This may be done by further researching if the extracted insights can become more precise and relevant by adding news and announcements coming out from companies, market-based sentiment data, or other applicable external information in the RAG pipeline. In addition, research would be needed into how to include picture-based RAG techniques for specific scenarios when visual information is relevant in order to fully optimize the system's performance.

This technology combines powerful vector databases, text-based RAG pipelines, Tree of Thoughts prompts, and contextual information to create robust RAG systems that yield actionable insights from financial records. Using complex vector databases, text-based RAG pipelines, and Tree of Thoughts prompts, as well as other contextual information, this strategy will allow us to create more effective and comprehensive RAG systems that can provide actionable insights from financial records.

References

Basyal, L., & Sanghvi, M. (2023). Text summarization using large language models: a comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*.

Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cao, Y., Chen, Z., Pei, Q., Dimino, F., Ausiello, L., Kumar, P., ... Ndiaye, P. M. (2024). Risklabs: Predicting financial risk using large language model based on multi-sources data. *arXiv preprint arXiv:2404.07452*.

Fasha, M., Rub, F. A., Matar, N., Sowan, B., Al Khaldy, M., & Barham, H. (2024). Mitigating the owasp top 10 for large language models applications using intelligent agents. In *2024 2nd international conference on cyber resilience (iccr)* (pp. 1–9).

Feng, D., Dai, Y., Huang, J., Zhang, Y., Xie, Q., Han, W., ... Wang, H. (2023). Empowering many, biasing a few: Generalist credit scoring through large language models. *arXiv preprint arXiv:2310.00566*.

Filippo, C., Vito, G., Irene, S., Simone, B., & Gualtiero, F. (2024). Future applications of generative large language models: A data-driven case study on chatgpt. *Technovation*, *133*, 103002.

Haider, K. (2024). Document data extraction 101: Understanding the basics. *astera*.

Investing.com. (n.d.). Stock quotes. Retrieved from `https://in.investing.com`

K, D. (2024). Credit analysis essentials: Fundamentals, importance techniques. *proschoolonline*.

Kalyan, K. S. (2023). A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 100048.

Kim, A., Muhn, M., & Nikolaev, V. (2024). Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Kumaraswamy, S. (2024). Financial statement analysis: A summary of your business performance. *happay*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... others (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Lin, Y., Hulsebos, M., Ma, R., Shankar, S., Zeigham, S., Parameswaran, A. G., & Wu, E. (2024). Towards accurate and efficient document analytics with large language models. *arXiv preprint arXiv:2405.04674*.

NSIM. (n.d.). Financial statement analysis. *NSIM*.

Nucci, A. (2024). Leveraging large language models in finance. *aisera*.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., ... others (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Segal, T. (2023). 5 cs of credit: What they are, how they're used, and which is most important. *Investopedia*.

Siddiqui, W. (2023). The four pillars of financial analysis: A deep dive into key financial statements. *LinkedIn*.

Tuovila, A. (2024). Financial analysis: Definition, importance, types, and examples. *Investopedia*.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., ... others (2021). Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 international conference on management of data* (pp. 2614–2627).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models, 2023. *URL https://arxiv.org/pdf/2305.10601. pdf*.