STM JOURNALS

Journal of
Advanced Database
Management & Systems

ISSN: 2393-8730
Volume 12, Issue 2, 2025
May—August
DOI (Journal): 10.37591/JoADMS

https://journals.stmjournals.com/joadms

Review JoADMS

Retrieval Augmented Generation for Question Answering in Financial Documents

S. Sharon Benita[1,*], V. Srividhya[2]

Abstract

In recent years, the integration of Question Answering (QA) with the Retrieval Augmented Generation (RAG) system has transformed to interact with numerous documents. It uses Natural Language Processing (NLP) techniques to improve accuracy and relevant responses derived from huge documents. RAG integrates the advantages of the retrieval and generation process, which allows systems to generate natural responses and extract context from multiple sources. The main reason to use RAG is that it can help Large Language Model (LLM). Several personalized pieces of information are used in the RAG architecture. In this work, the financial domain is used to handle the financial documents to answer complex questions that efficiently retrieve knowledge of specific terms and context to generate the answers. Collections of financial documents are used to create questions with answers based on the information provided and to compare the performance of the system with known ground truth answers. The ROUGE score is used to evaluate the performance of the RAG. It is used to evaluate the accuracy between the generated responses and matched reference answers for a range of questions. Finally, incorporating RAG into question answering frameworks can improve user interface, confidence, and accuracy in automated solutions in the finance domain.

Keywords: Retrieval augmented generation, natural language processing, embedding, large language model, ROUGE score

INTRODUCTION

Question Answering is a subfield of Natural Language Processing (NLP), which enables systems to answer questions based on structured and unstructured data. Traditional Question Answering relies on predefined rules, keyword-based retrieval, and structured databases to provide answers. QA through retrieval augmented generation ensures an accurate and context-aware answer by fetching relevant information before responding. Transformer models like T5 and BERT introduced self-attention mechanisms, allowing them to better understand context and handle unstructured text [1]. These systems often struggle with complex questions due to limited contextual understanding. To enhance traditional QA and transformer-based QA, Retrieval Augmented Generation has emerged as a powerful technique that combines retrieval and generation for more accurate and context-aware answers [2]. It allows Large Language Model access to new knowledge sources and answers questions. This method of knowledge injection enhances pre-trained LLMs by augmenting questions with additional knowledge. Figure 1 shows the RAG Architecture.

***Author for Correspondence**
S. Sharon Benita
E-mail: 23pca019@avinuty.ac.in

[1]Student, Master of Computer Applications, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India
[2]Associate Professor Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

Received Date: April 28, 2025
Accepted Date: June 20, 2025
Published Date: September 27, 2025

Citation: S. Sharon Benita, V. Srividhya. Retrieval Augmented Generation for Question Answering in Financial Documents. Journal of Advanced Database Management & Systems. 2025; 12(2): 62–68p.

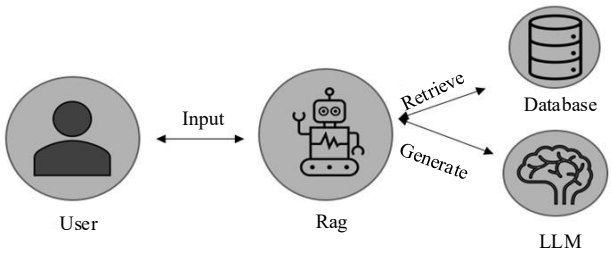

Figure 1. RAG architecture.

In recent years, the growth of LLMs represents a critical turning point in Generative AI. RAG has been developed to address the hallucination problem. It can increase productivity across various domains. This work focuses on financial document question answering. It uses the 5 years of company documents. Accurate text extraction from PDFs is crucial in RAG applications for effective retrieval and generation. It aims to improve financial documents accessibility by reducing hallucinations in LLM-generated responses [3].

The problem statement is to develop a RAG-based system to answer questions and retrieve financial documents using an embedding model and a generative AI model.

The objectives are to collect and preprocess financial documents, including chunking for efficient retrieval. It integrates an optimized embedding model for information retrieval and stores the data in a vector database. An LLM model is used to generate answers based on user questions.

The first Section introduces this work. The next Section gives an overview of existing research. The Section following that describes the overall research methodology. Then the next Section discusses the performance evaluation metrics of the ROUGE score. In the last Section, recommendations for future improvements are provided along with the conclusion of the work.

LITERATURE REVIEW

Financial document analysis has improved with the integration of Natural Language Processing (NLP) and with Transformer models. This literature review explores significant research articles on analysis enhancement, using an emphasis on their applicability in the financial sector. The paper by Abbasiantaeb and Momtazi explored the application of transformer-based models, notably for long financial records [4]. It addresses problems such as the need for domain-specific knowledge. It also creates summaries that resemble those of humans. The models' promising ROUGE scores on validation datasets, which are obtained through multi-stage fine-tuning, demonstrated the effectiveness of transformer topologies in processing long financial texts. Jayakumar *et al.* explored recurrent Neural Network (RNN) architectures to produce and grade questions and responses obtained from financial documents [5]. The objective of their approach was to improve the quality and relevance of generated QA pairs by utilizing financial Named Entity Recognition (NER). The study by Phogat *et al*. provides insights into the application of RNNs in financial document analysis, which emphasizes their potential for automating QA generation, and offers insights into their use in financial document analysis [6]. The paper by Shah *et al.* suggested techniques for employing LLMs to answer multi-document financial questions [7]. They used knowledge graph-based methods and semantic tagging to improve context retrieval from a variety of financial documents. The paper by Srivastava *et al*. evaluated LLMs' aptitude for mathematical thinking in responding to questions on financial documents [8]. Numerous financial tabular QA datasets were used in their comprehensive trials with different models and prompting strategies. According to the study by Singh *et al*., these methods performed better than conventional retrieval augmented generation models, and LLMs' capacity to synthesize data from a variety of financial texts [9]. Financial document analysis has remarkably advanced with the incorporation of Transformer models and LLMs. The studies under evaluation demonstrate possibilities as well as difficulties in implementing these technologies in the financial industry. To overcome these constraints

and fully use the potential of LLMs and Transformer designs in financial applications, further study is necessary.

RESEARCH METHODOLOGY

This research improves financial question answering by using Retrieval Augmented Generation (RAG). It involves several processes, each playing a vital role in ensuring precise and effective information retrieval and response generation. Figure 2 illustrates the research design of the work.

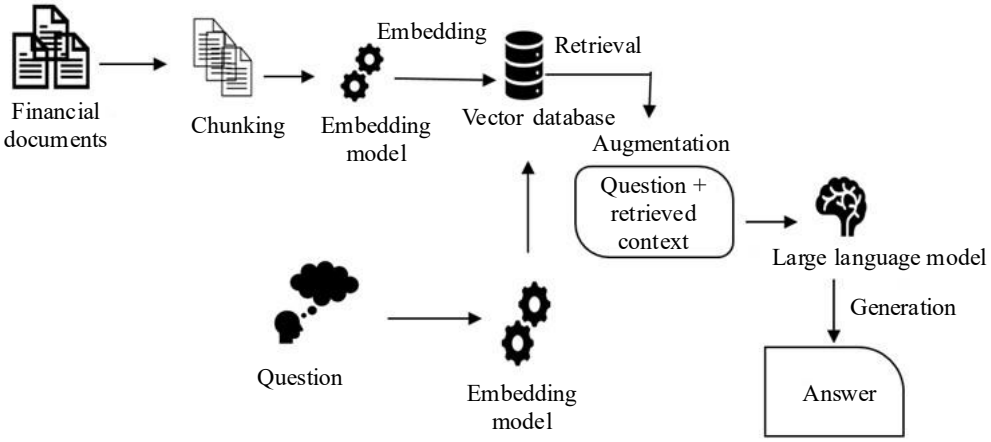

Figure 2. Research design.

Data Preparation and Processing

The documents used in this work consist of financial documents of NVIDIA, and it uses the 5 years of company documents from 2020–2024. Preprocessing text data is a crucial stage in Natural Language Processing. It includes tokenization, stop word removal, and chunking.

The collected documents are pre-processed into chunks. Chunking is the process of dividing large documents into smaller, meaningful segments for effective retrieval. Chunking enables the RAG model to retrieve the most relevant information more accurately. Chunking makes handling massive documents more feasible.

There are various chunking approaches, recursive chunking ensures better structure and context retention, and it is commonly used in Natural Language Processing. This step is essential to ensuring an accurate and efficient retrieval process. The questions have been sourced from hugging face, an open-source platform [10].

Document Embedding and Retrieval

The chunked documents are converted into a vector representation using the embedding model called Sentence Transformers. It captures the semantic meaning of the text, allowing the system to perform similarity searches effectively. A sentence transformer model, all-miniLM-L6-v2 [11] maps the sentences and paragraphs to a 384-dimensional dense vector space.

This model supports natural Language Processing tasks that involve understanding the meaning of text. After generating the vector embedding, they are stored in the vector database optimized for fast and effective retrieval. There are many vector databases available. In this work, Chroma DB is used as it provides a highly optimized vector storage mechanism and supports real-time retrieval and integration.

The question phase converts inputs into vectors for database searches, enhancing retrieval efficiency. The vector embeddings are stored in a Vector Database; a structured index designed for retrieval based on similarity measures. LangChain facilitates interaction between the embedding model, vector database, and retrieval process. It provides a data-preprocessing pipeline that uses ChromaDB.

Implementation of LLM Model

The LLM model is designed to deliver precise and context-aware responses by using retrieved financial data. An LLM requires context to generate an accurate response. This work uses a large language model, Mistral 7B, to extract meaningful insights from embedded and sourced information. It is an open-source model used for high-performance text generation, retrieval, augmented generation, and domain-specific applications. It also provides freely available tools and other resources. It is built on a dense transformer architecture and delivers state-of-the-art reasoning [12].

This model is highly advanced and trained on a vast amount of textual data to effectively interpret and generate human-like text. It also enables RAG by integrating an external knowledge source for an accurate response. Financial question answering using LLM has been extensively studied, with many successful implementations. It integrates LLM with vector databases like Chroma DB, and real-time, context-aware financial retrieval is achieved.

RESULTS AND ANALYSIS

Setting Up the Environment

This section provides a detailed process for configuring the Retrieval Augmented Generation development environment on the Kaggle platform. Using Kaggle eliminates the need for manual Python and virtual environment installation, as it provides a pre-configured, cloud-based Jupyter Notebook environment.

It concentrates on setting up the Kaggle notebook, installing necessary dependencies, and streamlining the workspace to execute RAG effectively. It uses GPUs to speed up model processing and inference.

Accessing the Hugging Face API is a crucial step. After creating a Hugging Face account and logging in, it navigates to the settings section and generates new API tokens from the access token page. To authenticate and access the Hugging Face model, datasets, and other resources, this token is necessary.

Evaluation

For the evaluation of the RAG system, the two key components are: assessing the ability of the model to retrieve relevant context and the capability to generate accurate answers on the context. In this work, the evaluation is performed by comparing the model-generated answer with the ground truth answer. The question and ground truth answer are sourced from a Hugging Face dataset.

The ROUGE is commonly used as a metric for evaluating the quality of text generation models [13]. It measures the similarity between a machine-generated answer and with ground truth answer by comparing overlapping words, phrases, and sequences. Among the various ROUGE metrics, ROUGE 1, ROUGE 2, and ROUGE L are the most used for Evaluation.

- *ROUGE 1:* It measures the match of unigram (single words) between the generated answer and the reference text. It primarily evaluates word-level recall and precision.

$$\text{ROUGE 1}=\frac{|\,W(G)\cap W(R)\,|}{|W(R)|} \tag{1}$$

- *ROUGE 2:* It extends this evaluation by considering bigrams (two-word sequences), providing a more refined assessment.

$$\text{ROUGE 2}=\frac{|\,P(G)\cap P(R)\,|}{|P(R)|} \tag{2}$$

- *ROUGE L:* It evaluates the longest common subsequence (LCS) between the generated and reference answer, capturing sentence fluency and structure.

$$\text{ROUGE L}=\frac{|\,LCS(G,R)\,|}{|W(R)|} \tag{3}$$

Where,
G= Generated answer,
R= Reference (ground truth) answer,
W(G)= Words in generated answer,
P(G)= Phrase (bigrams) in the generated answer,
P(R)= Phrase in the reference answer, and
LCS(G, R)= Longest common subsequence (generated answer and reference answers).

Performance Analysis

According to the performance evaluation, the model significantly reduces hallucinations while generating responses, improving factual accuracy. The Mistral 7B model demonstrated strong capabilities in financial data interpretation, enabling effective document retrieval. Table 1 presents a comparative analysis of the ground truth answer and model generated answer for 10 sample questions. Table 2 presents the result of the average ROUGE score. Figure 3 shows the visual representation of Average ROUGE Score.

Table 1. Ground truth vs. generated answer.

Q. No.	Questions	Ground truth answer	Generated answer
1	What area did NVIDIA initially focus on before expanding to other computationally intensive fields?	NVIDIA initially focused on PC graphics	NVIDIA initially focused on PC graphics.
2	What significant invention did NVIDIA create in 1999?	NVIDIA invented the GPU in 1999.	NVIDIA invented the GPU in 1999.
3	What was the percentage change in revenue for the Compute & Networking segment from fiscal year 2022 to 2023?	The revenue for Compute & Networking increased by 36% from fiscal year 2022 to 2023.	The revenue for the Compute & Networking segment decreased by 36% from fiscal year 2022 to 2023.
4	What is the anticipated total capital investment range for fiscal year 2024 related to property and equipment?	Approximately $ 1.10 billion to $ 1.30 billion	The anticipated total capital investment range for fiscal year 2024 related to property and equipment is approximately $ 1.10 billion to $ 1.30 billion.
5	What was the total amount of net cash provided by operating activities for fiscal year 2023?	$ 5,641 million	The total amount of net cash provided by operating activities for fiscal year 2023 was $ 5,641 million
6	How much did cash, cash equivalents, and marketable securities total at the end of fiscal year 2023?	$ 13.30 billion	$ 13.30 billion
7	What year was NVIDIA reincorporated in Delaware?	NVIDIA was reincorporated in Delaware in April 1998.	NVIDIA was reincorporated in Delaware in April 1998
8	What was the percentage change in revenue for the Compute & Networking segment from fiscal year 2022 to 2023?	$ 1.36 billion	$ 1.36 billion
9	How much is authorized for the repurchase of additional shares of common stock as of January 29, 2023?	$ 7.23 billion	$ 7.23 billion
10	What degree does Jen-Hsun Huang hold from Stanford University?	19%	19%

Table 2. Result of average ROUGE score.

Question(Q)	Average ROUGE score		
	ROUGE 1	***ROUGE 2***	***ROUGE L***
Q1–Q10	0.87	0.72	0.85

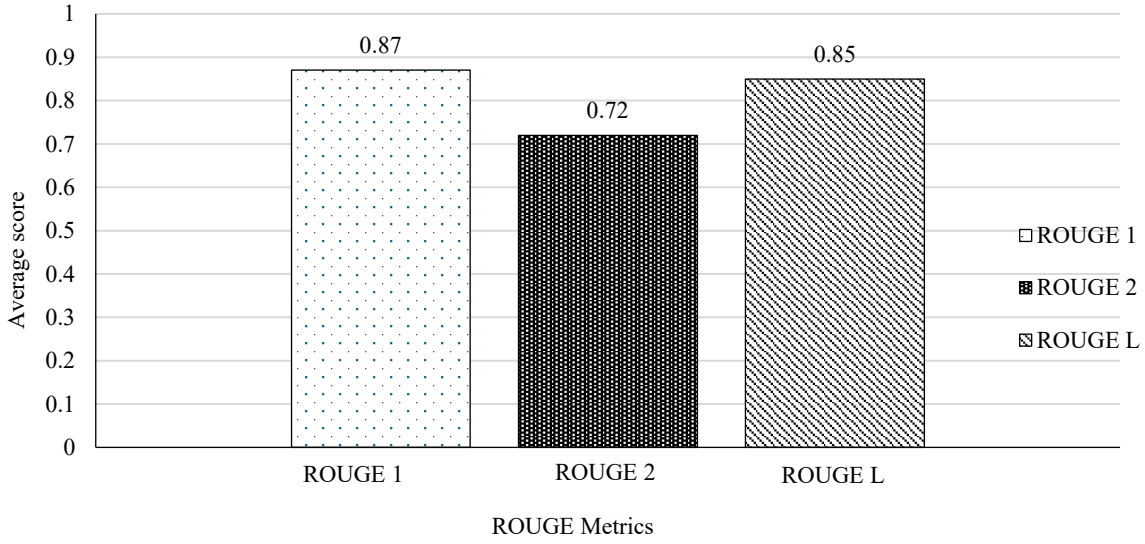

Figure 3. Visual representation of average ROUGE score.

CONCLUSION

The evolution of RAG systems is a novel approach to improve large language models (LLMs) by connecting their outputs in real time, for relevant information. The work outlines the key steps for building RAG systems that use financial documents as the data source.

This work focuses on RAG systems in the financial domain. It is specifically for question answering on financial documents. It leverages NVIDIA's annual documents (2020–2024) for financial question answering. It integrates the Mistral 7B language model with Chroma DB for retrieval. This system can effectively extract and generate contextually relevant responses while reducing hallucinations.

The ROUGE score analysis confirms the system's ability to generate high-quality responses, achieving an average ROUGE 1 of 0.87, ROUGE 2 of 0.72, and ROUGE L of 0.85. Future work focuses on expanding the system's capabilities by integrating multiple embedding models and retrieval techniques to refine accuracy further. This method is used in other applications such as medicine, agriculture, public policy, law, education, retail, supply chain, environmental science, business intelligence, and research.

REFERENCES

1. Biancofiore GM, Deldjoo Y, Noia TD, Di Sciascio E, Narducci F. Interactive question answering systems: Literature review. ACM Comput Surv. 2024 May 8; 56(9): 1–38.
2. Usbeck R, Röder M, Hoffmann M, Conrads F, Huthmann J, Ngonga-Ngomo AC, Demmler C, Unger C. Benchmarking question answering systems. Semant Web. 2019 Jan 21; 10(2): 293–304.
3. Nassiri K, Akhloufi M. Transformer models used for text-based question answering systems. Appl Intell. 2023 May; 53(9): 10602–35.
4. Abbasiantaeb Z, Momtazi S. Text‐based question answering from information retrieval and deep neural network perspectives: A survey. Wiley Interdiscip Rev: Data Min Knowl Discov. 2021 Nov; 11(6): e1412.

5. Jayakumar H, Krishnakumar MS, Peddagopu VV, Sridhar R. RNN based question answer generation and ranking for financial documents using financial NER. Sādhanā. 2020 Dec; 45: 269.
6. Phogat KS, Puranam SA, Dasaratha S, Harsha C, Ramakrishna S. Fine-tuning Smaller Language Models for Question Answering over Financial Documents. arXiv preprint arXiv:2408.12337. 2024 Aug 22.
7. Shah S, Ryali S, Venkatesh R. Multi-Document Financial Question Answering using LLMs. arXiv preprint arXiv:2411.07264. 2024 Nov 8.
8. Srivastava P, Malik M, Gupta V, Ganu T, Roth D. Evaluating LLMs' Mathematical Reasoning in Financial Document Question Answering. arXiv preprint arXiv:2402.11194. 2024 Feb 17.
9. Singh K, Kaur S, Smiley C. FinQAPT: Empowering Financial Decisions with End-to-End LLM-driven Question Answering Pipeline. In Proceedings of the 5th ACM International Conference on AI in Finance. 2024 Nov 14; 266–273.
10. Hugging Face. (2025). Hugging Face – The AI community building the future. [online] Hugging Face. Available from: https://huggingface.co/
11. Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Improving text embeddings with large language models. arXiv preprint arXiv:2401.00368. 2023 Dec 31.
12. Kukreja S, Kumar T, Purohit A, Dasgupta A, Guha D. A literature survey on open source large language models. In Proceedings of the 2024 7th International Conference on Computers in Management and Business. 2024 Jan 12; 133–143.
13. Chen A, Stanovsky G, Singh S, Gardner M. Evaluating question answering evaluation. In Proceedings of the 2nd workshop on machine reading for question answering. 2019 Nov; 119–124.

