XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGENEERING

Graph RAG vs. Vector RAG: A Performance Comparison in Response Generation

Cecília F. V. Couto[1], Nelson F. F. Ebecken[1]

[1]*Dept. of Civil Engineering, COPPE, Federal University of Rio de Janeiro*
Av. Athos da Silveira Ramos, 149, CT, Room B101, Rio de Janeiro, RJ 21941-909, Brazil
cecilia.couto@coc.ufrj.br, nelson@ntt.ufrj.br

Abstract. Large Language Models (LLMs) are powerful tools capable of generating coherent and human-like text across a wide range of domains. Due to their extensive pretraining on large textual datasets, they exhibit strong generalization abilities. However, LLMs remain limited by the static knowledge embedded during training, making them prone to outdated information, conceptual gaps, and hallucinations. An emerging alternative to address these issues is Retrieval-Augmented Generation (RAG), which enhances text generation by incorporating external contextual data at inference time. This study evaluates the effectiveness of contextual augmentation by comparing three models: a Vector RAG, a Graph RAG, and a baseline LLM without external context. All models were tested on financial topics using a multiple-choice questionnaire composed of 60 questions divided into four levels of difficulty: easy, medium, hard, and sensemaking. Financial data was sourced from the Management Discussion and Analysis (MD&A) sections of 10-K reports from major companies. The results demonstrate that both RAG models significantly outperformed the baseline LLM, particularly in more complex or interpretive questions. Surprisingly, the Vector RAG model performed best overall, suggesting that in some scenarios, vector-based augmentation may be sufficient.

Keywords: RAG, Vector RAG, Graph RAG, LLM

1 Introduction

Large Language Models (LLMs) have revolutionized the field of artificial intelligence due to their ability to generate natural language in a coherent and contextualized way, playing a central role in tasks such as question answering, text summarization, and information retrieval [8].

These models have undergone significant development in a short period, becoming more robust and achieving levels of fluency and coherence increasingly close to human language. Due to their great technical capacity, LLMs are becoming increasingly popular among researchers and users, turning into one of the main study topics and an alternative to traditional search engines [4].

However, despite their growing popularity, the widespread use of LLMs still faces some challenges. Even with the great development they have undergone, the occurrence of hallucinations in these models remains a significant limitation [3]. In addition, traditional LLMs rely on the knowledge used in their training, which represents temporal and conceptual limitations.

To overcome these limitations, some alternatives have emerged, such as Retrieval-Augmented Generation (RAG) models. This approach combines text generation with external information retrieval, allowing the model to access dynamic databases during text generation, which can significantly reduce hallucinations [2].

In conventional RAG models, vector-based approaches store information in a high-dimensional semantic space. Through similarity techniques, such as searching for closest embeddings, these models can retrieve relevant text fragments quickly and efficiently. The main advantage of this approach is its simplicity and scalability, allowing integration with large databases and low computational cost during retrieval [1].

Despite their efficiency, Vector RAG models have significant limitations. Because they rely exclusively on vector similarity, they may fail to capture more complex relationships between entities or concepts. This is especially problematic for questions that require deeper reasoning or contextual interpretation, where explicit connections between different pieces of information are essential [1].

Graph RAG emerges as an alternative to the vector model, proposing the organization of information in graph structures. In this approach, entities and relationships extracted from the data are represented as nodes and edges,

forming a semantic graph. This structure facilitates the explicit mapping of connections between concepts, which can enrich the context provided to the model and favor more contextualized and coherent responses, especially in complex tasks [1].

Despite its potential, Graph RAG involves a more complex implementation process that is sensitive to methodological choices. Graph construction requires steps such as triple extraction, definition of relevant entities, and neighborhood-based context retrieval, which can directly impact the quality of responses. Furthermore, the performance of this model is highly dependent on how the graph is constructed and how connections are explored, which can pose a challenge in practical applications [7].

To evaluate the impact of the data storage method on model performance, this work proposes a comparison between three models: a vector-based RAG, a graph-based RAG, and a context-free LLM.

Therefore, the objective of this work was to evaluate the relevance of providing auxiliary context to LLMs for text generation. To this end, the RAG models were applied using context extracted from the Management Discussion and Analysis (MD&A) sections of the Form 10-K reports of 9 large companies. These models were then tested with a questionnaire, and their results were compared with those of a simple context-free LLM.

Section 2 presents the methodology of this work. Section 3 presents and discusses the results obtained by each of the models adopted in this article. Section 4 provides the final conclusions.

2 Methodology

To achieve the proposed objective, three models were used: a context-free LLM, a RAG vector model, and a RAG graph model. All models were based on Llama 3.2 1B Instruct. This version was selected because it combines reasonable quality with reduced computational requirements. Since this study was conducted on a standard personal computer, employing larger models was not feasible. Nevertheless, the smaller scale does not invalidate the comparative analysis proposed, as the focus lies on evaluating retrieval strategies rather than the absolute performance of the LLM.

The Llama 3.2 1B Instruct model is part of Meta's model family. With 1 billion parameters, this model has an autoregressive architecture based on the optimized transformer architecture. It is optimized for dialogue tasks in several languages, such as information retrieval and summarization [6].

For the context-free model, the LLM was used in its original form, without modifications. It was used to enable a fairer comparison with the other RAG models. As this version does not receive additional information, it allows a clearer assessment of the effects of providing external context.

For the Vector RAG and Graph RAG models, the context provided was extracted from the Management Discussion and Analysis (MD&A) sections of the Form 10-K reports of large US-listed companies. This section provides an overview of past performance and future prospects of companies, as well as discussing various aspects of the business [5].

Reports from 9 companies (Apple, Microsoft Corporation, Alphabet, Amazon, Tesla, Meta Platforms, NVIDIA Corporation, Berkshire Hathaway, and Johnson & Johnson) over the past 10 years were used, totaling 90 documents. These reports were obtained through the public SEC API, accessible via the SEC API service [9].

For the Vector RAG, the content of the reports was segmented into sentences with at least 5 words, and preprocessing was applied to remove highly similar sentences. The remaining sentences were encoded using the all-MiniLM-L6-v2 embedding model and the vectors were indexed using the FAISS library, using inner product similarity. This process for Vector RAG context preparation typically took only a few seconds

For the Graph RAG model, a knowledge graph was created from automatically extracted triples from the report sentences, in the format (entity, relation, value). The entity was identified as a relevant noun phrase, the relation as the first verb of the sentence, and the value as a numeric or percentage expression. These triples formed a directed graph, where nodes represent entities and values, and edges indicate relationships between them, also containing the original sentence as evidence. Finally, the nodes were converted into semantic vectors using the all-MiniLM-L6-v2 model, allowing similarity-based retrieval during context recovery. The construction of this knowledge graph, however, required approximately 1 hour of processing time.

It is important to emphasize that the entire graph construction process was designed to be fully automatic, without the need for human intervention. This decision was made to ensure that the methodology could be applied at scale and remained reproducible, rather than relying on manual adjustments. While the automatic extraction of entities, relationships, and values can introduce some noise, a robust pipeline was implemented to mitigate such issues and preserve relevant information.

During the application of the RAG models, for the vector version, the three most relevant sentences based on similarity with each question were selected. In the Graph RAG, context retrieval occurred through a combination of semantic similarity and graph structure exploration. First, the question was encoded into an embedding vector and compared with the graph nodes using cosine similarity. The ten most similar nodes were then selected, along

with nodes containing keywords from the question. For each of these nodes, sentences stored in their direct and second-degree neighbors were retrieved. The resulting sentences were re-evaluated in relation to the question, and the three most relevant ones were selected as the final context for the model.

During preliminary experiments, different amounts of retrieved sentences were tested as context for the models. When too many sentences were provided, the LLM tended to become distracted or confused, generating less accurate answers due to the excess of information. On the other hand, when too few sentences were supplied, the context became too weak to effectively support the response. Empirically, using three sentences proved to be the optimal balance, offering enough information to guide the model while avoiding cognitive overload. For this reason, the context size was fixed at three sentences in both the Vector RAG and Graph RAG configurations.

The choice of prompt provided to the models was a crucial factor for the success of the test. It was observed that the Llama model, even when used with context, showed a significant bias in responses, especially in questions of higher complexity and difficulty.

To minimize this bias and maximize the performance of all models used, the following prompt was adopted for the context-based RAG models (both vector and graph versions):

Listing 1. Prompt used for RAG models with context.

```
prompt = f"""
You are a knowledgeable financial assistant. Below is a multiple-choice question,
    followed by a context extracted from corporate financial documents.

Carefully read the context. **Use it to answer the question if it provides relevant
information**. If the context seems incomplete or unrelated, rely on your broader
    financial knowledge.

Choose **only one letter (A, B, C, or D)** that best answers the question.

---

### Context:
{context.strip()}

### Question {i}:
{question}

{formatted_options}

Answer (only one letter - A, B, C, or D):
"""
```

In {context.strip()}, the selected context for the question is inserted. In {question}, the question itself is inserted, and {formatted_options} contains the answer choices. The prompt for the context-free Llama follows a similar logic, except that no additional context is provided to the model. The prompt used for this version is shown below:

Listing 2. Prompt used for LLM model without context.

```
prompt = f"""
You are a financial assistant. For each question, select the correct answer based on
     your financial knowledge.
Choose **only one letter (A, B, C, or D)** that best answers the question.

### Question {i}
{question}

{formatted_options}

Answer (A, B, C, or D):
"""
```

To evaluate the performance of each of the analyzed models, a questionnaire focusing on the financial area was developed, containing 60 multiple-choice questions with four options each. The test was divided into four sections of 15 questions, organized by difficulty level: easy, medium, hard, and sensemaking.

The "sensemaking" category questions were formulated based on information from the MD&A sections of the reports used as context, aiming to assess the models' ability to answer questions that require broad understanding.

The multiple-choice questions used in this work were generated with the assistance of ChatGPT. This choice was motivated by the lack of publicly available and domain-specific benchmarks that matched the financial scope of the study while also being suitable for automation. The use of ChatGPT ensured the creation of a consistent and reproducible set of questions, which was essential for enabling a fair comparison across the evaluated models. It is acknowledged that employing a generative model to create evaluation data may introduce some bias; however, since the same questionnaire was applied uniformly to all models, the relative validity of the comparison remains preserved. Furthermore, the use of LLMs to generate synthetic evaluation data has become increasingly common in recent NLP research, particularly in scenarios where domain-specific resources are limited.

To provide a clearer understanding of question types and difficulty levels, representative examples from each category are presented below:

Nível Easy

1. What is a bond?
   A. A long-term insurance plan
   B. A form of equity
   C. A financial obligation issued by entities to raise capital
   D. A certificate of employment
   (Correct Answer: C)

Nível Medium

1. What is a credit score used for?
   A. Determining voting eligibility
   B. Assessing loan and creditworthiness
   C. Calculating income tax
   D. Evaluating insurance rates only
   (Correct Answer: B)

Nível Hard

1. What does EBITDA stand for?
   A. Earnings Before Interest, Taxes, Depreciation, and Amortization
   B. Earnings Before Income, Tax, and Debt Analysis
   C. Equity Balance Interest and Tax Adjustment
   D. Estimated Business Income After Depreciation
   (Correct Answer: A)

Nível Sensemaking

1. Which valuation models does Johnson & Johnson use to determine fair value for performance share units?
   A. Black-Scholes and Monte Carlo models
   B. Gordon Growth and CAPM
   C. Binomial and CAPM
   D. Dividend Discount Model
   (Correct Answer: A)

The responses provided by the models went through a normalization process before evaluation, to eliminate formatting variations such as capitalization, punctuation, or additional symbols. After normalization, the responses were compared with the answer key to calculate accuracy.

Figure 1 provides a brief summary of the methodology applied in this work.

This work was implemented in Python and executed in the Google Colab environment. Open-source libraries were used for text preprocessing, triple extraction, embedding creation, and vector indexing, including spaCy, sentence-transformers, and FAISS.

3 Results and Discussion

The Vector RAG, Graph RAG and context-free LLM models were subjected to a questionnaire on the financial area. All models received the same questions in the same order to avoid any bias. During the questionnaire

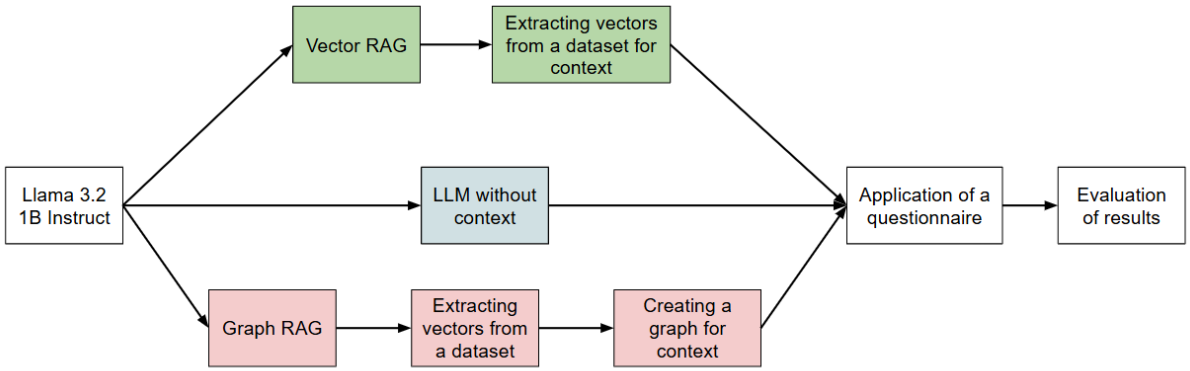


Figure 1. Brief summary of the methodology applied in this work.

application, the questions were asked separately, simulating a conversation.

In the Vector RAG model, for each question, the three sentences most similar to the question were selected to be provided as context to the model. To illustrate the selected context, the following sentences were retrieved for the first easy-level question, "What does inflation refer to?":

- "For example, inflationary pressures have increased across the markets in which we operate."
- "Inflation rates continue to have an effect on worldwide economies and, consequently, on the way companies operate."
- "As the year progressed, inflationary pressures increased across the markets in which we operate."

In the Graph RAG, three sentences were selected for each question using the methodology presented in Section 2. Also considering the question "What does inflation refer to?", the following sentences were provided as context:

- "The Company has accounted for operations in Argentina, Venezuela and Turkey (beginning in the fiscal second quarter of 2022) as highly inflationary, as the prior three-year cumulative inflation rate surpassed 100%."
- "During 2023, the Company recorded a charge of approximately $130 million related to operations in Argentina due to the application of highly inflationary accounting."
- "During the years ended December 31, 2022 and 2021, we also realized gains of $64 million and $128 million, respectively, in connection with converting our holdings of digital assets into fiat currency."

A total of 18.475 sentences were extracted from the dataset used in this work to compose the context provided to the Vector RAG. These sentences were also used to build a graph with 5.909 nodes and 6.581 edges for the Graph RAG.

The results obtained by the Vector RAG, Graph RAG, and baseline LLM in the financial questionnaire are shown in Table 1.

Table 1. Model performance across difficulty levels

Model/Mode	**Easy**	**Medium**	**Hard**	**Sensemaking**	**Average**
Vector RAG	100% (15/15)	93.33% (14/15)	86.67% (13/15)	86.67% (13/15)	91.67% (55/60)
Graph RAG	100% (15/15)	86.67% (13/15)	86.67% (13/15)	86.67% (13/15)	90% (54/60)
LLM baseline	93.33% (14/15)	73.33% (11/15)	66.67% (10/15)	46.67% (7/15)	70% (42/60)

Before analyzing the results, it is worth noting that Llama is a very advanced model with great generative capability. Therefore, even without context, it already shows strong potential to achieve good results in the questionnaire. Thus, any improvement from the RAG models represents significant progress.

By observing the results obtained in the questionnaire, it becomes evident that providing context to the model, whether in vector or graph form, significantly improves LLM performance.

In all question levels, the RAG models outperformed the context-free LLM, especially in the sensemaking level, where accuracy jumped from 46.67% (baseline) to 86.67% in both RAG models. This result highlights the impact of providing additional knowledge in questions requiring contextual inference.

Comparing the RAG model performances, both achieved excellent results, but the vector version outperformed the graph-based version, contrary to theoretical expectations.

Analyzing the performance more granularly, we can see that the only difference between the RAG models occurred at the "Medium" level, where Vector RAG answered one more question correctly than Graph RAG. To investigate the cause of this difference, it is necessary to analyze the context recovered by each model for that specific question. For the question "What is capital gains tax applied to?", Vector RAG retrieved the following context:

- "Taxable investment gains on equity securities sold during the year, which is the difference between sales proceeds and the original cost basis of the securities sold, were $3.2 billion in 2019 and $3.3 billion in 2018."
- "Taxable investment gains on equity securities sold, which is generally the difference between sales proceeds and the original cost basis of the securities sold, were $3.6 billion in 2021, $6.2 billion in 2020 and $3.2 billion in 2019."
- "Taxable investment gains on equity securities sold, which is generally the difference between sales proceeds and the original cost basis of the securities sold, were $5.0 billion in 2023, $769 million in 2022 and $3.6 billion in 2021."

In contrast, Graph RAG retrieved the following context:

- "As of the end of fiscal years 2023 and 2022, we had a valuation allowance of $1.48 billion and $907 million, respectively, related to capital loss carryforwards, state, and certain other deferred tax assets that management determined not likely to be realized due, in part, to jurisdictional projections of future taxable income, including capital gains."
- "This included a one-time non-cash tax benefit of $5.93 billion for the release of valuation allowance on certain deferred tax assets."
- "Our property/casualty business produced pre-tax underwriting gains of $204 million in 2014 and $171 million in 2013."

The question had alternative answers as options: "A. Employee wages", "B. Business profits", "C. Profits from the sale of investments", or "D. Retirement contributions", with the letter C being the correct option.

When comparing the contexts, it was observed that the Graph RAG context was less directly related and contained irrelevant information than the Vector RAG context, which, despite being more repetitive, was more relevant to the question.

This discrepancy in the quality of the context retrieved for this specific question may explain why the Graph RAG failed to answer correctly, while the Vector RAG did.

This reinforces the idea that, despite its theoretical ability to organize complex relationships, the Graph RAG context retrieval process may still be susceptible to retrieving less relevant information in specific cases, especially when the triple extraction method and graph navigation do not perfectly align with the question's needs.

4 Conclusions

This study aimed to compare the performance of RAG models with a baseline version to assess the importance of providing context to LLMs for obtaining better results and reducing hallucinations.

The results indicate that providing context can significantly improve the performance of LLMs, especially in more complex questions or those that require deeper data interpretation (sensemaking).

Comparing the RAG models, both achieved excellent results. Contrary to expectations, the vector version performed slightly better than the graph version. This, combined with the longer context preparation time for Graph RAG, suggests that despite being theoretically superior, the graph version may not always be the best choice, depends on the context data and application goals.

Furthermore, the experiments highlighted that carefully limiting the amount of retrieved context is crucial for RAG performance. In this work, using three sentences per query proved to be the optimal setting, ensuring that the models benefited from external knowledge without being misled by excessive or irrelevant information.

Graph RAG performance directly depends on how the graphs are created and how context is retrieved. This is a complex and challenging process that involves several stages, such as keyword selection, ontology definition, and identification of the most important nodes.

It's worth noting that Vector RAG's improved performance doesn't invalidate Graph RAG's excellent results.

However, it can be concluded that, in some use cases, vector-based context organization may be sufficient to achieve good results.

We conclude that providing context, whether organized in vectors or graphs, contributes to more accurate LLM responses. In practice, this means that users can improve the accuracy of LLMs simply by adding well-formulated context to their questions.

Finally, despite the advances, this study has limitations, such as the use of a small LLM (1B) and possible inaccuracies in automatic triple extraction. For future work, we intend to test larger LLMs, explore more advanced semantic extraction techniques, and expand the dataset to include other sections of corporate reports.

Acknowledgements. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been here in included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[2] Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*, 2024.

[3] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, 2023.

[4] C Kulkarni. The evolution of large language models in natural language understanding. *J. Artif. Intell. Mach. Learn. Data Sci*, 1:49–53, 2023.

[5] Jaywon Lee and Jongwon Park. The impact of audit committee financial expertise on management discussion and analysis (md&a) tone. *European Accounting Review*, 28(1):129–150, 2019.

[6] Meta. Llama 3.2 1B Instruct. Available at: https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct, 2024. Accessed in July 2025.

[7] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024.

[8] Mengchao Ren. Advancements and applications of large language models in natural language processing: A comprehensive review. *Applied and Computational Engineering*, 97:55–63, 2024.

[9] SEC-API. SEC API – Real-Time Access to EDGAR Filings. Available at: https://sec-api.io/. Accessed in July 2025.