CITE
II INTERNATIONAL CONGRESS ON EMERGING TECHNOLOGIES
Where today's ideas create tomorrow's solutions
November 14 and 15, 2025

SEMANTIC RETRIEVAL FOR SPECIALIZED DOMAINS: A RAG-LLM FRAMEWORK APPLIED TO ECONOMIC THEORY

Emanuel de Jesus Santos da Silva [1]
Eduardo Palhares Júnior [2]
Erick da Silva Farias [3]
Alexandre Lopes Martiniano [4]
Wenndisson da Silva Souza [5]

[1] Instituto Federal do Amazonas - IFAM, Campus Zona Leste – Manaus, AM – Brasil, leunamedj@gmail.com
[2] Instituto Federal do Amazonas - IFAM, Campus Distrito Industrial – Manaus, AM – Brasil, eduardo.palharesjr@ifam.edu.br
[3] Instituto Federal do Amazonas - IFAM, Campus Zona Leste – Manaus, AM – Brasil, edsfrlinux@gmail.com
[4] Instituto Federal do Amazonas - IFAM, Campus Distrito Industrial – Manaus, AM – Brasil, alexandre.martiniano@ifam.edu.br
[5] Instituto Federal do Amazonas - IFAM, Campus Distrito Industrial – Manaus, AM – Brasil, wenndisson.souza@ifam.edu.br

DOI:10.5281/zenodo.17918335

Abstract

Retrieval-Augmented Generation (RAG) presents a promising approach to enhance transformer-based language models, yet its implementation faces challenges in response relevance, accuracy, and scalability. This study addresses these challenges by proposing and evaluating a RAG architecture applied to the specialized domain of economic theory, specifically using Adam Smith's "The Wealth of Nations" as a case study. The proposed system integrates the LangChain framework , a Neo4j graph database for information storage , semantic embeddings generated via the Ollama platform , and the Llama 3.2 (1B) model for final response generation. We conducted a comparative analysis of responses generated with and without the RAG architecture. The results were assessed qualitatively and quantitatively, using ROUGE-L and BERTScore metrics. The findings demonstrate that RAG-supported responses are significantly more precise and conceptually aligned with the source text. While ROUGE-L scores were modest, as expected due to textual variations , BERTScore results indicated high semantic similarity. This confirms that the RAG architecture substantially improves the conceptual fidelity of LLMs in specialized question-answering tasks, proving valuable for domains where accuracy is critical.

Keywords: Retrieval-Augmented Generation (RAG); Large Language Models (LLMs); semantic retrieval; specialized domains; economic theory.

1. INTRODUCTION

In recent years, artificial intelligence (AI) has seen significant advancements, especially in the field of language models, such as transformers. One area that has stood out is Retrieval-Augmented Generation (RAG), which combines information retrieval techniques with text generation. These models are designed to provide more precise and contextualized answers to complex questions, a crucial capability for the evolution of human-machine interactions. However, as noted by Proença (2024), implementing RAG in question-answering systems still faces a series of significant challenges, which are essential for ensuring the relevance, accuracy, and reliability of the provided responses.

The current challenges faced by AI models in the area of question answering are primarily related to the dynamic and vast nature of human knowledge. When dealing with complex or specialized questions, systems must not only understand the content of the question but also have access to relevant and updated data sources to generate appropriate answers. According to Wang et al. (2024), RAG models need to be able to retrieve data from broad and varied sources in contexts of multiple disciplines or constantly changing information, which poses a challenge in terms of both efficiency and precision.

Another major challenge is adapting these models to different types of queries and the diversity of contexts in which questions may be asked. A question that might seem simple to a human, due to shared context, can be extremely difficult for an AI model to understand and answer adequately. This occurs because traditional AI models often lack the ability to integrate contextual and cultural nuances vital for an accurate response. For RAG models, the ability to query external sources and integrate this information cohesively is fundamental, but as Rezaei et al. (2024) highlight, it remains a developing field.

Furthermore, there is the problem of explainability and trust in the generated responses. In many situations, it is crucial that users can trust the information provided by AI systems, especially in areas like health, education, and security. RAG models, by relying on external data to formulate answers, may struggle to clearly explain the reasoning behind their responses. This lack of transparency and the difficulty in tracing the origin of information can generate distrust, which Zhao et al. (2023) identify as a significant obstacle to the widespread adoption of these technologies.

Finally, another central challenge is scalability and the management of large volumes of data. With the exponential growth of information available on the web and in specialized databases, RAG models must be constantly updated and improved to ensure they can access the most relevant sources efficiently. Integrating these systems with dynamic databases, which may

include everything from scientific articles to real-time data, requires robust infrastructure and advanced information processing strategies. Thus, the current challenges are multiple and complex, but as Gao et al. (2023) suggest, they also offer opportunities for innovations that could significantly transform how we interact with technology to answer our questions.

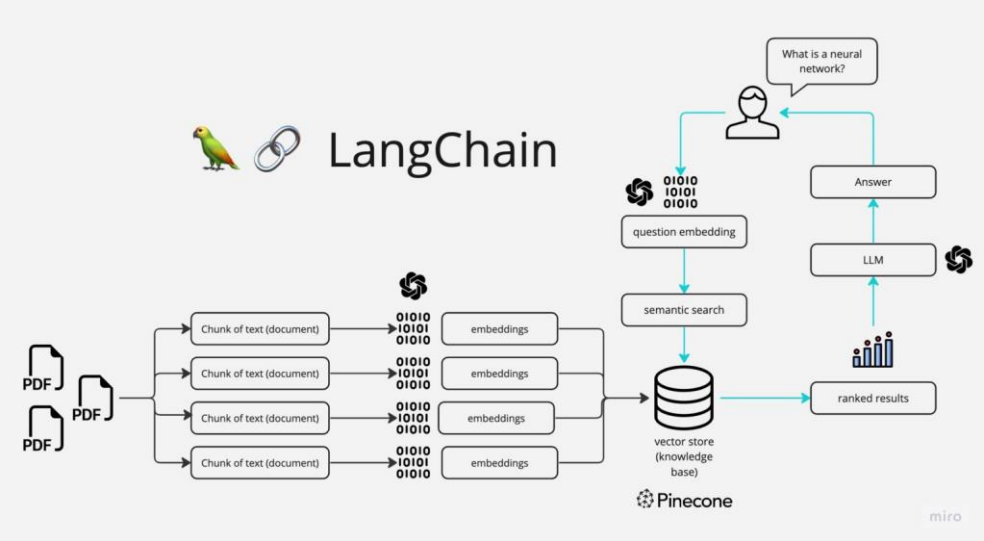
2. METHODOLOGY

The construction of the proposed system is based on the combination of different technologies that allow the integration of information retrieval mechanisms with generative language models. The central structure relies on the RAG architecture, implemented using the LangChain framework, which orchestrates the communication between components . Information storage and querying are handled by the Neo4j graph database, while semantic search is enabled by embeddings generated by the Ollama platform. The final responses are produced by the Llama 3.2 (1B) model, adjusted to handle natural language interactions and enrich responses based on the retrieved documents.

2.1 RAG Architecture

The RAG (Retrieval-Augmented Generation) architecture using LangChain, a software framework that helps build LLM-based applications, enhances AI model responses by integrating external information. This structure facilitates the implementation of RAG, allowing connection to databases, embedding vectors, and structured documents. As noted by Vidivelli et al. (2023), this increases the precision and reliability of the generated responses . This architecture is demonstrated in Figure 1.

Figure 1 - RAG Model Architecture.



Source: Adapted from LangChain (2024).

2.2 Graph-Oriented Database (Neo4j)

Neo4j is a graph-oriented database designed to store and query data that can be represented as graphs. It was created to handle complex relationships between data, making it ideal for applications requiring dynamic interactions between entities, such as social networks, recommendation systems, and fraud detection (KHAN, 2023). Currently, Neo4j remains one of the most widely used graph database platforms, adopted across various industries to solve problems related to interconnected and complex data. The tool supports various query languages, including its own Cypher language, which is designed to be intuitive and easy to learn (NEO4J, 2024).

2.3 Llama 3.2 (1B) Model

The Llama 3.2 (1B) model is a large-scale language model (LLM) developed by Meta, designed for advanced natural language processing tasks. With versions ranging from 1 to 3 billion parameters, Llama 3.2 (1B) was chosen for this work due to its good balance of performance, computational cost, and ability to generate contextualized responses. It can handle context windows of up to 128,000 tokens, making it suitable for working with extensive inputs integrated with documents of different natures. It also features multilingual support, including Portuguese, making it particularly suitable for applications in this language, as is the case in this study.

The use of Llama 3.2 (1B) was enabled through the Ollama platform, which allows LLMs to be run locally. This ensures greater control over the execution environment, eliminates dependency on external APIs, and reduces cloud computing costs. In the implemented system, the model is responsible for the final generation of responses, using information previously retrieved from the vectorized database as context. Its configuration was adjusted to promote more direct responses through optimized prompts and token limits to avoid prolixity and maintain objectivity.

In addition to its efficiency in generating responses, Llama 3.2 (1B) stands out for its adaptability in different RAG-based workflows. In this study, it is integrated into the LangChain architecture, which manages the query, retrieval, and generation cycle, allowing the model to act in a contextualized manner. The Llama model's ability to interpret and articulate information from multiple sources makes it a key element in ensuring cohesion and accuracy in user interactions.

2.4 Semantic Embeddings with Ollama

The retrieval of relevant documents in this work is based on vector representations of texts, known as semantic embeddings. These numerical vectors capture contextual similarity relationships between terms, allowing textual queries to be compared with documents more accurately than purely lexical approaches. In the RAG architecture, embeddings are fundamental for finding the most relevant documents that will serve as the basis for the final response generation by the language model.

The embeddings are generated using the Ollama platform, which enables the local use of LLMs for semantic vectorization. Local execution eliminates the need for external API calls, reducing latency and increasing system autonomy. The platform provides dense and contextually rich vector embeddings, which are stored in the Neo4j database and later used in the semantic retrieval step.

The implemented system follows a structured flow: the user inputs a question via a web interface, and this input is transmitted to the chat_with_bot function. This function triggers the query_neo4j method, which is responsible for generating the query embedding, calculating its similarity with the previously vectorized documents in Neo4j, and retrieving the most relevant content. The result of this process is the documents with the highest semantic proximity to the query, which will be used as context for generating the final answer.

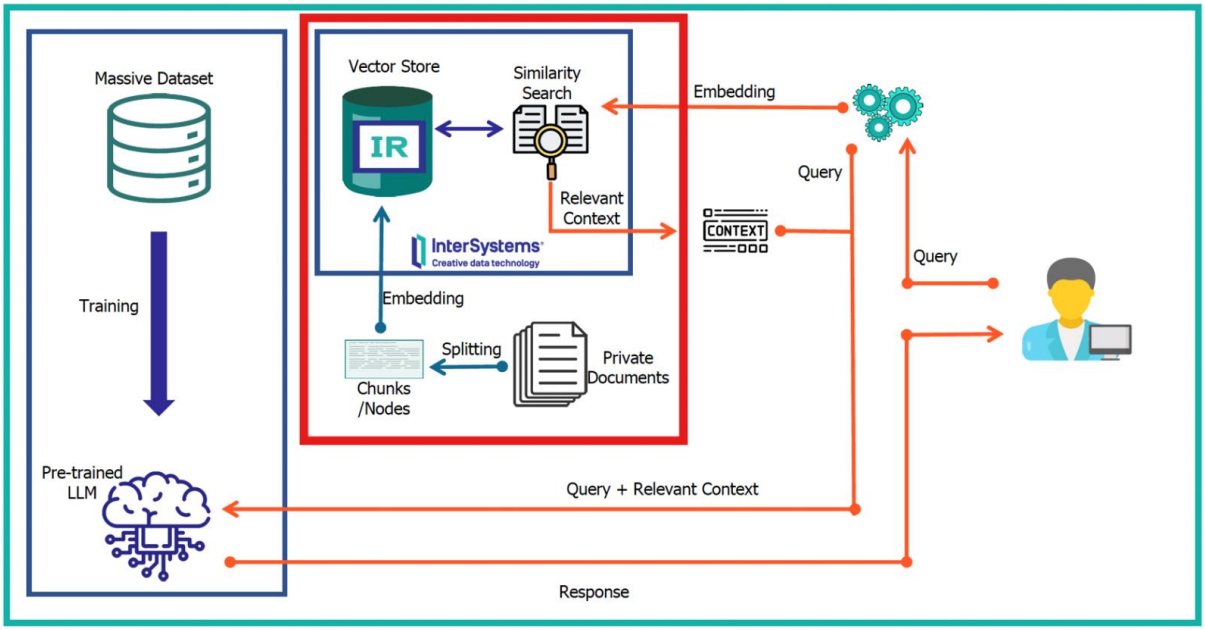
The metric used to compare the query and document vectors is cosine similarity, defined by the equation:

$$silimarity\,(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A·B represents the dot product of A and B, while ||A|| and ||B|| are their respective magnitudes . The similarity value ranges from -1 to 1, with values closer to 1 indicating greater semantic similarity. This metric allows the system to identify documents most aligned with the query's meaning, even with textual variations.

After calculating the similarities, the system selects the five most relevant documents based on the score obtained. These documents are then sent to the generative model (Llama 3.2 (1B)), which uses them as context to craft a response. This combination of embedding-based retrieval and generation with LLMs characterizes RAG architecture, as illustrated in Figure 2.

Figure 2 - LLM Models and RAG Applications.



Source: Paiva, H. (2024).

2.5 Evaluation Metrics for Generated Responses

The quality of the responses generated by the system is evaluated using two metrics widely used in the natural language processing literature: ROUGE-L and BERTScore . Both aim to measure the degree of similarity between the model-generated response and a reference response, but they adopt distinct and complementary approaches.

- **ROUGE-L:** The ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) metric is based on identifying the Longest Common Subsequence (LCS) between the words of the generated response and the reference. This metric is useful for capturing the sequential structure of the response but does not consider deeper semantic relationships between words. Phrase reformulations, synonym use, or reordering of terms can result in low scores, even when the essential content is preserved.
- **BERTScore:** BERTScore, on the other hand, uses pre-trained models like BERT to generate semantic embeddings for the words of the generated response and the reference . Based on these embeddings, semantic similarity between the texts is calculated. This metric is especially suitable for evaluating responses generated by language models, as it considers the meaning of the words, not just their order or superficial exactness. For this reason, it is common for F1 scores obtained by BERTScore to be significantly higher than those from ROUGE-L, even when evaluating the same responses. This difference will be revisited and discussed in the results section.

3. STANDARD FORMATTING

This section presents the results obtained from applying the Llama 3.2 (1B) model in two distinct scenarios: with and without the use of the RAG architecture . The objective is to compare the quality of the responses generated in each approach, based on questions elaborated from the content of Adam Smith's book The Wealth of Nations. Each question is analyzed individually, considering the response generated by the model without access to external sources (pure model) and the response generated with the support of the RAG architecture, which incorporates semantic retrieval via embeddings.

3.1 Qualitative Analysis: Question 1 (Concept of Value)

The first question sought to explore the model's understanding of the concept of value, a recurring theme in Adam Smith's work.

- **Answer without RAG:**

  "*The concept of value is related to the importance or utility attributed to a good or service. In classical economics, value can be understood in different ways: value in use and value in exchange. Value in use refers to the utility of a good, while value in exchange refers to what it can be exchanged for in the market. Adam Smith discussed these ideas when analyzing the paradox of water and diamonds, where water has high use value and low exchange value, and the diamond the opposite.*"

- **Answer with RAG:**

  "The concept of value, according to Adam Smith, is linked to the labor necessary to produce a good. He distinguishes between value in use, which is the utility of a good, and value in exchange, which is the ability to be exchanged for other goods in the market. Labor is the real measure of exchange value."

Qualitative Comparison: The answer without RAG presents a general explanation of the concept of value, mixing economic definitions with a modern interpretation of classical theory. Although it correctly mentions the concepts of use value and exchange value, the explanation moves away from Adam Smith's original emphasis on the role of labor as the measure of exchange value . On the other hand, the response with RAG demonstrates greater fidelity to the work's content, explicitly citing the relationship between labor and value—a central element of Smith's theory.

3.2 Qualitative Analysis: Question 2 (Origin of Value)

The second question aimed to investigate whether the model could identify the origin of the concept of value according to Adam Smith's theoretical approach.

- **Answer without RAG:**

  "The concept of value arises from the interaction between supply and demand in the market. Throughout the history of economic thought, various authors have proposed different explanations for the emergence of value. In the classical tradition, value was associated with the labor necessary to produce a good. In more modern approaches, the utility perceived by consumers and the opportunity costs involved in production are also considered."

- **Answer with RAG:**

  "According to Adam Smith, the value of goods arises from the quantity of labor employed in their production. Labor is the real measure of the exchange value of any commodity."

Qualitative Comparison: The answer without RAG presents a broader and more anachronistic explanation of the origin of value, including contemporary concepts like utility and opportunity costs that are not part of Adam Smith's original theory. In contrast, the RAG response is concise, direct, and conceptually aligned with the work's content, focusing on the relationship between labor and exchange value.

3.3 Qualitative Analysis: Question 3 (Factors of Production)

The third question aimed to assess whether the model could identify the factors that, according to classical theory, contribute to the formation of the value of goods.

- **Answer without RAG:**

  "The factors of production that contribute to the concept of value include labor, capital, and land. In classical economic theory, labor was considered the main factor in generating value, but over time, the role of capital (machines, tools) and land (natural resources) came to be recognized."

- **Answer with RAG:**

  "According to Adam Smith, the exchange value of a commodity is determined by the labor required to produce it. Labor is the fundamental factor that generates value, especially in primitive societies, where labor was the only factor of production."

Qualitative Comparison: The answer without RAG includes elements like capital and land, which deviates from the specific focus of Adam Smith's work, where labor is emphasized as the primary source of value . The RAG response, however, remains faithful, highlighting the central role of labor, especially in the context of primitive societies . This demonstrates that RAG helped align the model to the correct theoretical perspective within the question's scope

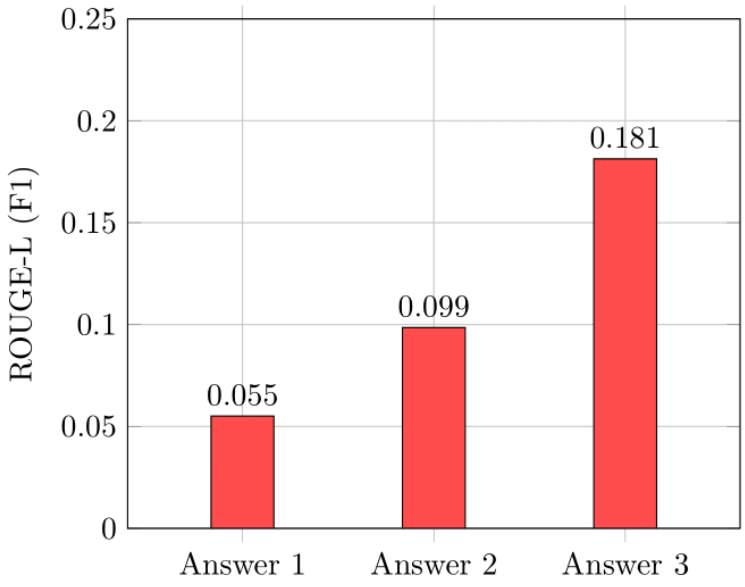
3.4 Quantitative Evaluation

In addition to the qualitative comparison, a quantitative evaluation was performed using two complementary metrics: ROUGE-L and BERTScore. The analyzed values refer exclusively to the responses generated with the RAG architecture. ROUGE-L measures literal textual overlap, while BERTScore evaluates semantic similarity based on embeddings.

3.4.1 ROUGE-L Results

Figure 3 presents the ROUGE-L F1-score values for each of the three responses generated with the RAG architecture.

Figure 3 - ROUGE-L (F1-score) performance on RAG-generated answers.



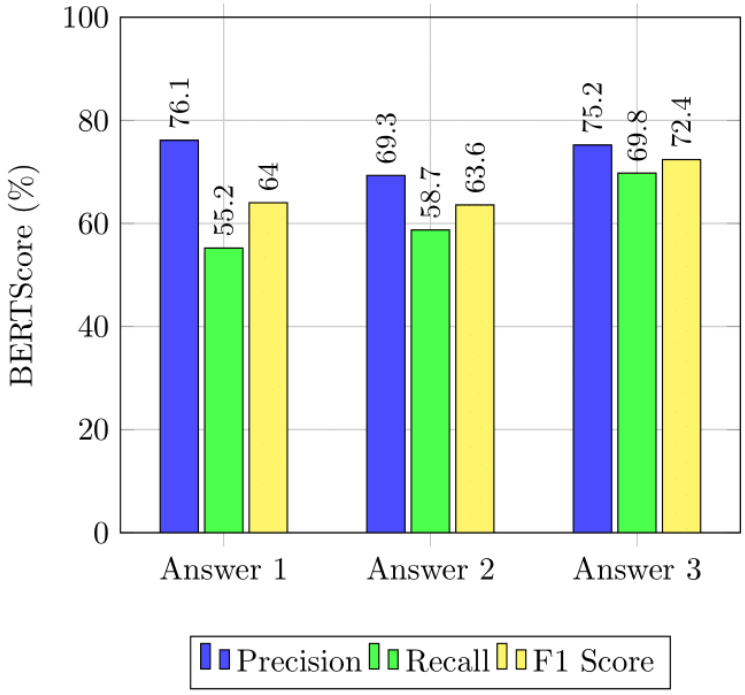
Source: The authors.

The ROUGE-L values observed were relatively low. This result is expected for this metric, which privileges exact word overlap and order. Legitimate reformulations or synonym use, common in RAG-generated responses, are not recognized as valid matches, which penalizes the final score.

3.4.2 BERTScore Results

Figure 4 presents the Precision, Recall, and F1 scores obtained via BERTScore for the three RAG-generated responses.

Figure 4 - BERTScore (Precision, Recall, F1-Score) performance on RAG-generated responses.



Source: The autors.

The ROUGE-L values observed were relatively low. This result is expected for this metric, which privileges exact word overlap and order. Legitimate reformulations or synonym use, common in RAG-generated responses, are not recognized as valid matches, which penalizes the final score

3.5 Discussion of Results

The analysis of the three questions revealed consistent differences between the responses generated by the pure model (without RAG) and those produced with the RAG architecture. In general, the RAG responses proved to be more precise, concise, and conceptually aligned with the original content of Adam Smith's work.

This trend was confirmed both qualitatively and quantitatively. Qualitatively, semantic retrieval allowed the model to anchor itself directly in relevant content, reducing ambiguity and avoiding extrapolations. Quantitatively, the BERTScore values reinforced this conclusion, indicating high semantic similarity to the references. The ROUGE-L scores, though lower, are in line with the expected behavior of that specific metric.

These results indicate that integrating retrieval mechanisms and generative models can significantly improve response quality in document-based Q&A tasks. However, the RAG approach's effectiveness depends on the quality and semantic coverage of the database used.

4. FINAL CONSIDERATIONS

This work presented an architecture based on Retrieval-Augmented Generation (RAG) to improve response generation in document-based question-answering tasks. The proposal integrated semantic information retrieval, using embeddings generated with the Ollama platform, with a Llama 3.2 (1B) language model, allowing the system to combine contextualized search with natural language text generation.

The system's evaluation was conducted based on questions extracted from Adam Smith's The Wealth of Nations. The qualitative analyses showed that using the RAG architecture provided more concise, precise, and conceptually aligned answers to the work's content. This finding was reinforced by quantitative metrics, especially BERTScore, which indicated strong semantic similarity to the references.

The results demonstrate that integrating semantic retrieval mechanisms and generative models can substantially elevate the quality of responses generated by LLMs. The proposed architecture proved effective in aligning text generation with the content of specific documents, which is particularly relevant in contexts where conceptual accuracy is essential, such as education, health, law, and scientific research .

Future work can explore several directions. One possibility is to expand the vectorized document base, incorporating multiple sources and organizing data into more robust structures, such as knowledge graphs. Another line of investigation involves comparing different vectorization and semantic retrieval techniques, evaluating their impact on the architecture's performance. There is also potential to apply the proposed approach in specific domains, evaluating its robustness in more specialized tasks, and incorporating human feedback mechanisms to refine the generated responses.

The authors thank Samsung Eletrônica da Amazônia Ltda., through the Aranouá Project, and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), for the financial support via the Programa de Excelência Acadêmica (PROEX). This work is a result of the Research and Development (P&D) project 001/2021, established with the Instituto Federal do Amazonas (IFAM) and FAEPI, with funding from Samsung.

REFERENCES

ESMAEELI, Shahriar et al. RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:2309.15217, 2024. Disponível em: https://arxiv.org/abs/2309.15217. Acesso em: 11 nov. 2025.

GAO, Y., et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997, 2023. Disponível em: https://arxiv.org/abs/2312.10997. Acesso em: 11 nov. 2025.

KHAN, A. Knowledge graphs querying. SIGMOD Rec., v. 52, n. 2, p. 18-29, 2023.

LEWIS, Patrick et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS), v. 33, p. 9459-9474, 2020. Disponível em: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf. Acesso em: 11 nov. 2025.

NEO4J. How to Implement Advanced Retrieval RAG Strategies With Neo4j. Neo4j Developer Blog, 2024. Disponível em: https://neo4j.com/blog/developer/advanced-rag-strategies-neo4j/. Acesso em: 11 nov. 2025.

NEO4J. Neo4j e Microsoft anunciam colaboração em soluções de GenAI e dados. Inforchannel, 26 mar. 2024. Disponível em: https://inforchannel.com.br/2024/03/26/neo4j-e-microsoft-anunciam-colaboracao-em-solucoes-de-genai-e-dados/. Acesso em: 11 nov. 2025.

PROENÇA, V. L. Automatização da revisão de literatura científica com geração aumentada por recuperação. 2024. Trabalho de Conclusão de Graduação (Graduação) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2024.

REZAEI, M. R., et al. At-rag: An adaptive rag model enhancing query efficiency with topic filtering and iterative reasoning. arXiv preprint arXiv:2410.12886, 2024. Disponível em: https://arxiv.org/abs/2410.12886. Acesso em: 11 nov. 2025.

WANG, Y., et al. Enhancing retrieval-augmented generation with self-consistent reasoning. arXiv preprint arXiv:2407.19393, 2024. Disponível em: https://arxiv.org/abs/2407.19393. Acesso em: 11 nov. 2025.

YIN, Cong et al. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. Applied Sciences, v. 15, n. 8, p. 4234, 2024. Disponível em: https://www.mdpi.com/2076-3417/15/8/4234. Acesso em: 11 nov. 2025.

ZHAO, H., et al. Explainability for Large Language Models: A Survey. arXiv preprint arXiv:2309.01029, 2023. Disponível em: https://arxiv.org/abs/2309.01029. Acesso em: 11 nov. 2025.