

Ontology-Enriched GraphRAG: A Case Study in Portuguese Geoscience Question Answering

Tiago Rios da Rocha^{1,2}, João Cesar Netto¹, Karin Becker¹

¹ Universidade Federal do Rio Grande do Sul, Brazil
karin.becker@inf.ufrgs.br, netto@inf.ufrgs.br

² Instituto Federal do Rio Grande do Sul, Brazil
tiago.rios@ibiruba.ifrs.edu.br

Abstract. Retrieval-Augmented Generation (RAG) significantly enhances Large Language Models (LLMs) in Question-Answering (QA) tasks. However, traditional text-based retrieval may prove insufficient in specialized domains, such as geoscience, where critical information is highly interconnected and distributed across complex concepts, relations, and interpretative descriptions. While Graph-based Retrieval-Augmented Generation (GraphRAG) addresses part of this limitation by structuring evidence into entity-relation graphs, it remains susceptible to terminological inconsistencies, synonyms, and domain-specific variants. To bridge this gap, this paper presents a controlled empirical study evaluating ontology-enriched GraphRAG for Portuguese geoscience QA. We systematically compared four generation strategies: *LLM-only*, textual RAG, standard GraphRAG, and ontology-enriched GraphRAG. The ontology-driven approach grounds text-extracted entities in formal reference ontologies using a cascading matching strategy that starts with embedding-based semantic similarity, followed by URI-based, label-based, and alternative-label-based matching. Evaluated across semantic, lexical, contextual, and graph-level metrics, our results confirm that external retrieval is indispensable, as retrieval-based strategies outperform the *LLM-only* baseline. Textual RAG achieved the strongest lexical overlap, while ontology-enriched GraphRAG achieved the highest *Correctness*, *Completeness*, and *Contextual Recall*. GraphRAG provided stronger grounding and question alignment, as reflected by higher *Faithfulness* and *Answer Relevancy*. Beyond these answer- and context-level gains, the most substantial effect of ontology enrichment was structural, driving a substantial expansion in graph semantic coverage, graph size, and matching rate. These findings suggest that ontology grounding acts primarily as a powerful enabling mechanism for semantic representation; however, its impact on answer quality depends on retrieval, subgraph selection, and prompt construction.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Knowledge representation and reasoning*; • **Information systems** → **Question answering**; *Information retrieval*.

Keywords: Geoscience, GraphRAG, Ontology, Portuguese Texts, Question Answering, Retrieval-Augmented Generation

1. INTRODUCTION

Retrieval-Augmented Generation (RAG) is commonly used to improve Large Language Models (LLMs) in Question-Answering (QA) tasks by providing external evidence during generation [Lewis et al. 2021]. Most RAG pipelines retrieve independent text chunks using lexical or vector similarity. This strategy is effective when the answer is concentrated in a short passage, but it may be limited in scientific domains, where relevant information is often distributed across concepts, relations, and interpretative descriptions [Sharma et al. 2025].

Graph-based Retrieval-Augmented Generation (GraphRAG) addresses part of this problem by representing documents as graphs of entities and relations, enabling retrieval over structured evidence rather than isolated text chunks [Edge et al. 2025; Peng et al. 2025]. However, graphs extracted from text may still contain inconsistent terminology, synonyms, abbreviations, and domain-specific variants. Domain ontologies may help address this limitation by adding formal concepts and rela-

The authors acknowledge Petrobras and the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) for their financial support of the project ONTOKG through cooperation agreement No. SIGITEC 2025/00193-0. We also acknowledge the INCT TILD-IAR (Responsible Artificial Intelligence for Computational Linguistics, Information Processing and Dissemination) (CNPq/SECTICS/CAPES/FAPs grant 408490/2024-1), CNPq (Grant 308246/2026-8), and the support from Federal Institute of Education, Science, and Technology of Rio Grande do Sul.

tions that preserve domain semantics in specialized QA tasks. However, it remains unclear whether ontology enrichment effectively improves GraphRAG systems built from text-extracted entities and relations. More specifically, it is not known whether the additional semantic knowledge provided by ontologies translates into better answers, improved retrieval quality, or merely larger and semantically richer graphs [Cao et al. 2025].

This issue is especially relevant in geoscience, where questions often depend on relations among sedimentary basins, lithologies, depositional systems, stratigraphic units, seismic facies, and reservoir properties. In this domain, ontologies such as GeoCore and GeoReservoir provide formal concepts that can support semantic grounding and more consistent interpretation of geological terminology [Garcia et al. 2020; Cicconeto et al. 2022]. This makes geoscience a suitable domain for investigating whether ontology enrichment can effectively enhance GraphRAG-based question answering.

This paper presents a controlled empirical study of ontology-enriched GraphRAG for Portuguese geoscience question answering. Four answer-generation strategies are compared: LLM-only, textual RAG, GraphRAG, and ontology-enriched GraphRAG. The goal is to investigate whether ontology grounding improves graph-based retrieval and whether such improvements translate into better answer generation. The experimental design was built to analyze the task from multiple perspectives, exploring LLMs with different capacities and model profiles, different retrieval strategies, and enhanced ontology matching. The outputs were assessed with complementary semantic, lexical, contextual, and graph-level metrics.

This study is guided by four research questions: (RQ1) how LLM-only, textual RAG, GraphRAG, and ontology-enriched GraphRAG compare in Portuguese geoscience QA; (RQ2) how graph-based retrieval compares with textual RAG in answer quality and evidence grounding; (RQ3) whether ontology enrichment improves GraphRAG in answer quality, retrieved-context quality, or graph-level semantic coverage; and (RQ4) how the ontology matching strategy affects the matching rate.

The main contribution of this study is a controlled empirical evaluation of ontology-enriched GraphRAG for Portuguese geoscience QA, analyzing its impact on answer quality, retrieved-context quality, and graph-level semantic coverage. The results confirm that external retrieval is essential for this task, as retrieval-based approaches consistently outperformed the LLM-only strategy. Textual RAG achieved the strongest lexical overlap, while GraphRAG provided higher evidence grounding, as reflected by its *Faithfulness* and *Relevancy* score. Ontology-enriched GraphRAG achieved the highest *Correctness*, *Completeness*, and *Contextual Recall*. Its most significant effect, however, was observed at the graph level: the proposed ontology-matching strategy substantially increased graph semantic coverage, graph size, and matching rate. These findings suggest that ontology enrichment primarily serves as an enabling mechanism for semantic coverage, whereas its benefits to question answering depend on subsequent retrieval and subgraph selection stages.

The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 describes the materials and methods. Section 4 discusses the experimental results. Finally, Section 5 summarizes the main findings and outlines future research directions.

2. RELATED WORK

Retrieval-Augmented Generation (RAG) combines a retriever and a generator, allowing LLMs to answer questions using external evidence rather than relying solely on parametric knowledge [Lewis et al. 2021]. Most RAG systems retrieve independent text chunks using lexical or vector similarity. While effective when answers are concentrated in specific passages, this approach may be limited when relevant information is distributed across multiple entities, concepts, or documents.

GraphRAG addresses this limitation by organizing evidence through entities and relations, enabling retrieval over nodes, paths, or subgraphs rather than isolated text chunks [Peng et al. 2025]. A study [Edge et al. 2025] demonstrated that graph-based retrieval can improve QA over large corpora by exploiting entity-centric graph structures and community summaries. Similarly, KG²RAG [Zhu et al. 2025] combines semantic retrieval with graph-based expansion of evidence, improving support for

multi-hop question answering. LEGO-GraphRAG [Cao et al. 2025] showed that GraphRAG performance depends strongly on retrieval design, particularly on how subgraphs and reasoning paths are selected. Together, these studies suggest that graph structures can improve evidence organization and retrieval, but that answer quality depends on the relevance of the retrieved graph objects.

A complementary direction is ontology-grounded retrieval. Ontologies provide formal concepts and relations that can reduce ambiguity and improve semantic consistency in specialized domains. However, integrating ontologies into retrieval pipelines requires ontology matching, i.e., aligning textual entities with ontology concepts [Otero-Cerdeira et al. 2015]. OG-RAG [Sharma et al. 2025] uses ontology-grounded hypergraphs to guide retrieval and reported improvements in factual recall, correctness, and reasoning quality. Nevertheless, OG-RAG does not investigate ontology enrichment within a GraphRAG pipeline built from text-extracted entities and relations, nor does it analyze the impact of different ontology-matching strategies on graph construction and answer generation.

These studies have explored various retrieval strategies to improve QA systems. However, the interaction between ontology enrichment and GraphRAG remains insufficiently understood. This work addresses this gap through a controlled evaluation of ontology-enriched GraphRAG for Portuguese geoscience QA, analyzing its effects on answer quality, retrieved-context quality, and graph-level semantic coverage.

3. MATERIALS AND METHODS

The experimental design compares four answer generation strategies over a Portuguese QA benchmark in the geoscientific domain: *LLM-only*, textual RAG, GraphRAG, and ontology-enriched GraphRAG. This comparison was designed to isolate three effects: external retrieval, graph-based retrieval, and ontology grounding. We explored LLMs with distinct capacities and model profiles. The analysis combines semantic metrics, lexical metrics, and graph-level indicators, with the final ontology-enriched setting using an embedding-first cascading matching strategy.

Figure 1 summarizes the experimental workflow. Each benchmark instance provides a document context, a question, and an expected answer. The document context is processed to generate textual chunks (1). In the RAG setting, only the retrieved textual chunks are provided to the LLM as contextual evidence. In the Graph-RAG setting, the chunks are leveraged during retrieval to identify relevant entities, relationships, and connected contexts, enabling the LLM to reason over both textual evidence and graph-based semantic connections (2). In the ontology-enriched Graph-RAG setting, this graph is additionally expanded with ontology concepts through a matching process (3). The question is then answered using the selected strategy (4), and the generated answer is evaluated against the expected answer and the retrieved evidence (5). The code and supplementary material is available in a public repository¹. The supplementary material includes the pseudo-algorithms and parameters used in the experimental pipeline, as well as extended tables with detailed results and token budget.

3.1 QA Dataset and Ontology

The Portuguese QA benchmark released by [Paula et al. 2024] was used as the evaluation dataset. The original benchmark spans multiple domains. Our study focuses on its geoscientific subset, comprising 312 questions. A manual curation step removed duplicated questions and instances with potentially inconsistent answers. GeoCore and GeoReservoir [Garcia et al. 2020; Cicconeto et al. 2022] were used as domain ontologies. GeoCore provides a general conceptualization of geological entities based on the Basic Formal Ontology (BFO), while GeoReservoir extends this foundation with concepts related to reservoir characterization and depositional systems.

3.2 Tools and Answer Generation Strategies

We adopted LLMs with distinct capacities and model profiles, namely GPT-4o, GPT-4.1, GPT-5-mini, GPT-5.2, GPT-5.4, and GPT-o3, and executed them on the Petrobras high-computing infrastructure.

¹<https://github.com/tiagoriosrocha/ontology-enriched-graphrag>

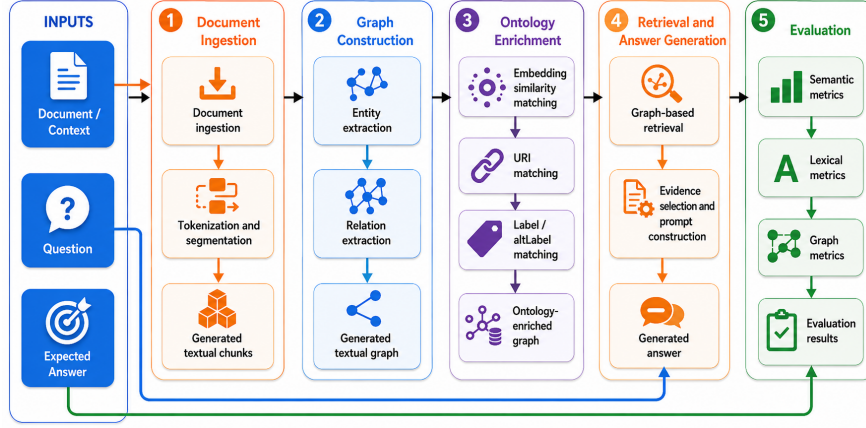


Fig. 1: Experimental workflow for ontology-enriched GraphRAG evaluation.

Cognee² was used to implement the RAG, GraphRAG, and ontology-enriched GraphRAG pipelines. Cognee was chosen because it automates the ingestion of unstructured text and seamlessly maps it into structured graph topologies, supporting the integration of domain-specific ontologies. This capability facilitates the implementation of a consistent and reproducible experimental pipeline.

Four answer generation strategies were evaluated. In the *LLM-only* strategy, the question is sent directly to the language model without external context. This strategy serves as a baseline for assessing the limits of parametric knowledge. In textual RAG, relevant passages are retrieved from the indexed context and provided to the model together with the question. In GraphRAG, retrieval uses a graph built from entities and relations extracted from the documents. In ontology-enriched GraphRAG, the extracted entities are additionally aligned with ontology concepts, and the enriched graph is used during retrieval and answer generation.

All strategies were executed over the same benchmark instances, using the same questions and expected answers. We ran each strategy across all six LLMs. The automatic evaluation was implemented using DeepEval³, a robust LLM-as-a-judge framework that provides research-validated metrics for benchmarking RAG and GraphRAG architectures. We selected GPT-5.4 as the judge, since it is the strongest LLM.

3.3 Graph Construction and Ontology Matching

In the GraphRAG and ontology-enriched GraphRAG strategies, entities and their relations are extracted from textual chunks by Cognee with LLM support, making graph quality partly dependent on the underlying model. In the GraphRAG without ontology, the graph is built from information extracted from the text. In the ontology-enriched strategy, the extracted textual entities are matched against ontology concepts. This task concerns ontology matching, which identifies semantic correspondences between entities across different knowledge representations [Otero-Cerdeira et al. 2015].

The default Cognee implementation supports only URI-based ontology matching. We propose an ontology-matching approach that follows an embedding-first, cascading strategy. A term is first matched using embedding-based semantic similarity, aiming to capture semantic correspondences that may not be recovered by exact string matching. If no semantic match is found, the term is then compared against ontology URIs, followed by ontology labels and alternative labels. The default Cognee implementation supports only URI-based matching; the embedding-based, label-based, and alternative-label-based stages were implemented in this work. Hence, our design prioritizes semantic coverage through embeddings while retaining lexical matching as fallback mechanisms.

²<https://docs.cognee.ai/>

³<https://deepeval.com/>

For each matched ontology node, the pipeline stores its URI, labels, alternative labels, definitions or descriptions, and relevant RDF/OWL properties, including structural, domain and range relations, restrictions, and incoming ontology links when available. Figure 2 illustrates this process for a geoscientific QA instance.

3.4 Answer Generation and Evaluation

For each question, one answer was generated in each model-strategy pair. For every execution, the generated answer, the retrieved context, and, when applicable, graph evidence and ontology matching information were stored. This setup makes it possible to relate answer quality to the retrieved evidence, the graph structure, and the ontology concepts used during retrieval. The analysis combines answer-level, retrieval-level, lexical, and graph-level metrics, enabling evaluation of the effects of retrieval, graph structure, and ontology grounding from complementary perspectives.

Answer quality was assessed using Answer *Relevancy*, *Correctness*, and *Completeness*, which measure question alignment, factual agreement with the reference answer, and coverage of the expected information, respectively. Retrieved evidence quality was assessed using *Faithfulness*, *Contextual Relevancy*, and *Contextual Recall*, which evaluate answer grounding, context relevance, and context coverage [Es et al. 2024]. All semantic metrics were computed using DeepEval’s framework, returning scores in the [0,1] interval. In addition, lexical metrics were computed using Exact Match, Token F1, and ROUGE-L. Finally, graph-level indicators, including graph size, number of valid ontology matches, and matching rate, were used to analyze the structural effect of ontology enrichment.

In the graph-based strategies, answer generation does not use the entire graph as input to the LLM. In this work, Cogneer was used in graph-completion mode. When a question is submitted, Cogneer first retrieves, with the support of the LLM, the graph objects most relevant to the question, such as nodes, relations, and associated textual or ontology metadata. In all experiments, this retrieval step used a fixed setting of the 15 graph nodes most aligned with the question. This value was chosen as a trade-off between graph coverage and prompt size, providing enough evidence for high contextual recall while avoiding excessive expansion of the retrieved context. These selected objects are then converted into contextual evidence and used to assemble the prompt for answer generation. This retrieval step is important as the final answer depends not only on the number of ontology matches in the graph, but also on which objects are selected, how they are filtered, and how they are presented in the prompt. As a consequence, the graph-based strategies produced prompts that were approximately 5.6 times larger than those generated by textual RAG. Further details on the token budget are available in the Supplementary Material.

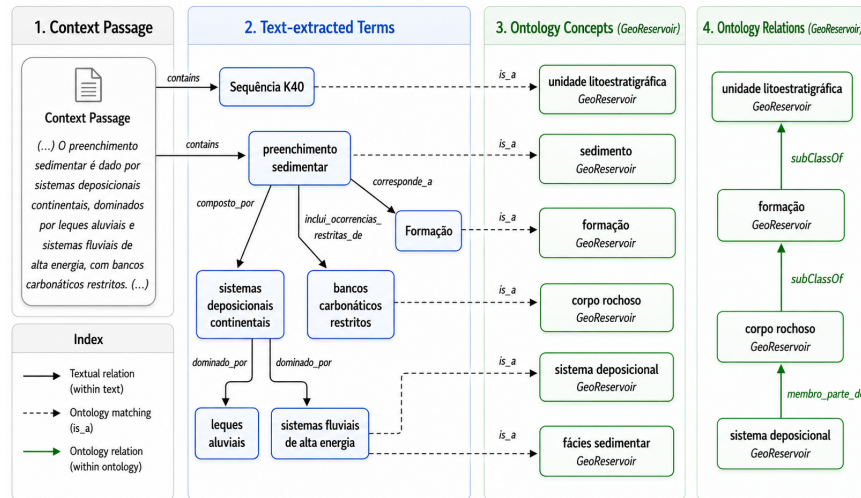


Fig. 2: Example of ontology-enriched graph construction.

4. RESULTS AND DISCUSSION

This section presents the experimental results and discusses their implications for the proposed research questions. We first compare the answer-generation strategies (RQ1), then analyze the quality of the retrieved evidence (RQ2), and finally assess the impact of ontology enrichment and ontology matching on GraphRAG (RQ3 and RQ4).

The metrics were computed from the detailed per-answer evaluation instances. Each record corresponds to a generated answer for a given test set instance and experimental strategy. Results were first computed separately for each model-strategy pair and subsequently aggregated by answer generation strategy. Due to space constraints, the per-model results are available in the public repository as supplementary material.

4.1 Performance by Answer Generation Strategy

Table I shows the mean performance of the answering strategies, including answer-level, lexical, and context-related metrics. The main result is the large gap between *LLM-only* and retrieval-based methods. Although *LLM-only* achieved high *Answer Relevancy*, its *Correctness*, *Completeness*, and lexical scores were much lower, indicating that parametric knowledge alone was insufficient to recover the expected geoscientific content.

Table I: Mean performance by answer generation strategy.

Strategy	Relev.	Faith.	Correct.	Comple.	EM	ROUGE-L	F1	Ctx.Rec.	Ctx.Rel.
LLM-only	0.9594	0.7059	0.1709	0.1827	0.0000	0.0736	0.0719	—	—
Textual RAG	0.9670	0.9671	0.9385	0.9675	0.0432	0.4521	0.4415	0.9712	0.2728
GraphRAG	0.9690	0.9758	0.9301	0.9598	0.0240	0.4136	0.4029	0.9700	0.2769
GraphRAG + Ont.	0.9671	0.9718	0.9412	0.9706	0.0264	0.4197	0.4114	0.9976	0.2716

Textual RAG achieved the best average lexical scores, including Exact Match, ROUGE-L, and *Token F1*, suggesting that many expected answers were explicitly present in retrievable passages. GraphRAG achieved the highest *Answer Relevancy*, *Faithfulness*, and *Contextual Relevancy*, indicating stronger adherence to retrieved evidence and slightly more focused retrieved contexts. Ontology-enriched GraphRAG achieved the highest *Correctness*, *Completeness*, and *Contextual Recall*, but did not improve lexical metrics, suggesting that ontology grounding contributes more to semantic coverage and evidence availability than to lexical overlap with the reference answer. This behavior may be explained by the characteristics of the dataset, where many questions appear to depend on information contained in localized passages rather than on multi-hop reasoning across entities and relations. In such settings, direct passage retrieval may be sufficient to recover the expected answer, limiting the potential advantage of graph-based retrieval and reasoning.

At the model level, no single LLM dominated all evaluation criteria. The detailed data are provided in the supplementary material. Considering only the retrieval-based strategies (Textual RAG, GraphRAG, GraphRAG + Ont.), GPT-5.2 achieved the highest average *Answer Relevancy* (0.9813) and *Faithfulness* (0.9847), while GPT-o3 obtained the highest *Correctness* (0.9463), ROUGE-L (0.5260), *Token F1* (0.5107), and *Contextual Recall* (0.9853). GPT-5-mini achieved the highest *Completeness* (0.9713) and *Contextual Relevancy* (0.2787). These results show that model choice affects the balance between answer correctness, grounding, lexical overlap, and retrieved-context focus.

4.2 Influence of Retrieved Context

The last two columns of Table I report the context-related metrics for the retrieval-based strategies. All retrieval methods achieved high *Contextual Recall*, indicating that the retrieved evidence usually contained information needed to answer the questions. However, *Contextual Relevancy* remained low across all retrieval configurations, suggesting that the retrieved contexts also included non-relevant information. Ontology-enriched GraphRAG obtained the highest *Contextual Recall*, but the lowest *Contextual Relevancy* among the retrieval-based strategies. This result suggests a trade-off between evidence focus and evidence coverage. Ontology grounding appears to increase evidence coverage, thereby improving the probability that the retrieved context contains the information needed to answer

Table II: Effect of ontology on GraphRAG.

Metric	GraphRAG	GraphRAG + Ont.	Δ (%)
Correctness	0.9301	0.9412	1.1989
Faithfulness	0.9758	0.9718	-0.4035
Completeness	0.9598	0.9706	1.1243
Ctx. Relevancy	0.2769	0.2716	-1.9068
Ctx. Recall	0.9700	0.9976	2.8430

Table III: Default matching vs. adapted strategy.

Metric	Default	Adapted	Δ (%)
Graph nodes	3347	11427	241.41
Edges	5750	19759	243.63
Valid matches	361	5399	1395.57
Matching rate (%)	10.79	47.25	337.91

the question. At the same time, the retrieved context may become less focused, resulting in lower contextual relevancy.

These findings indicate that ontology enrichment alone does not solve the retrieval problem and that subgraph selection, evidence filtering, and context compression remain important directions for improvement. This observation is consistent with the answer-level results, where improvements in contextual recall were accompanied by gains in correctness and completeness.

4.3 Effect of Ontology Enrichment in GraphRAG and Ontology Matching

Table II compares GraphRAG with and without ontology enrichment using answer/context-level metrics. The Δ column reports the percentage difference relative to the GraphRAG baseline. The results show a consistent but modest improvement in answer-quality metrics after ontology enrichment. It increased *Correctness*, *Completeness*, and *Contextual Recall*, while *Faithfulness* and *Contextual Relevancy* decreased slightly. The largest relative gain occurred in *Contextual Recall*, indicating that ontology concepts helped increase the coverage of evidence needed to answer the question. However, the decrease in *Contextual Relevancy* suggests that the retrieved evidence became less focused after ontology enrichment. Overall, the results indicate that ontology grounding improves the coverage of the retrieved context, but its impact on answer quality remains modest. This observation is consistent with the retrieval-level analysis presented in Section 4.2, where ontology enrichment increased evidence coverage while reducing contextual focus.

To assess the ontology-matching effect at the graph level, Table III compares graph properties and matches considering name-based matching only (Cognée default) and the adopted cascading strategy, which starts with embedding-based matching and then applies URI and label-based matching attempts. The graph-level values reported in the table are aggregate totals over the complete dataset: nodes, edges, and valid matches are summed across all evaluated QA instances, while the matching rate is computed from these aggregated totals.

The structural effect of ontology matching was substantially stronger than its effect on answer quality. Compared with the default name-based matching, the cascading strategy increased the number of valid ontology matches from 361 to 5,399 and raised the matching rate from 10.79% to 47.25%. Similar increases were observed in the number of graph nodes and edges, indicating a significant expansion of the graph’s semantic coverage. Despite these structural gains, answer-level improvements were not uniform across all metrics. These findings suggest that increasing ontology coverage alone is insufficient to substantially improve QA performance. Although ontology enrichment expands the available evidence space, the LLM does not operate over the entire graph. Instead, the retrieval step selects only a subset of nodes, relations, and ontology concepts to compose the prompt.

We conclude that the impact of ontology enrichment depends not only on matching coverage but also on the effectiveness of object ranking, subgraph selection, evidence filtering, and prompt construction. A larger, semantically richer graph increases the likelihood of relevant information being available, but answer quality ultimately depends on whether the retrieved subgraph contains the concepts that are useful for a specific question. These results suggest that future improvements are more likely to come from retrieval and subgraph-selection strategies than from additional ontology matching alone.

5. CONCLUSION

This study compared *LLM-only*, textual RAG, GraphRAG, and ontology-enriched GraphRAG for question answering over Portuguese geoscientific texts. The results confirm that external retrieval is essential for this task: retrieval-based approaches consistently outperformed *LLM-only*. Textual RAG achieved the strongest lexical overlap, while GraphRAG obtained the highest *Answer Relevance*, *Faithfulness*, and *Contextual Relevancy*, indicating stronger grounding in retrieved evidence. Ontology-enriched GraphRAG achieved the highest *Correctness*, *Completeness*, and *Contextual Recall*, showing that ontology grounding improved answer quality and evidence coverage.

The most significant impact of ontology enrichment was still observed at the graph level. The proposed embedding-first cascading matching strategy increased ontology matches by more than fifteen-fold and substantially expanded graph semantic coverage. However, these structural gains did not produce uniform improvements across all metrics. In particular, ontology-enriched GraphRAG improved *Correctness*, *Completeness*, and *Contextual Recall*, but reduced *Contextual Relevancy*. This finding suggests that ontology enrichment should be viewed as an enabling mechanism rather than a direct determinant of QA performance. While ontology grounding improves semantic coverage and evidence coverage, its effectiveness ultimately depends on retrieval, subgraph selection, evidence filtering, and prompt construction.

Two limitations should be noted. First, the dataset is mostly extractive, with answers concentrated in localized passages, which may limit the benefits of relational or multi-hop graph reasoning. Second, graph quality depends on Cogneer’s LLM-supported entity and relation extraction.

Future work should investigate ontology-aware retrieval and ranking strategies, as well as methods for selecting the most relevant concepts and relations for each question. It should also include ablation studies with a glossary-enriched RAG baseline to determine whether ontology gains come from formal structure, additional textual metadata, or their combination. More broadly, our results indicate that the main challenge is no longer only increasing graph coverage, but effectively transforming additional semantic knowledge into useful evidence for answer generation.

REFERENCES

- CAO, Y., GAO, Z., LI, Z., XIE, X., ZHOU, S. K., AND XU, J. Lego-graphrag: Modularizing graph-based retrieval-augmented generation for design space exploration. *Proc. VLDB Endow.* 18 (10): 3269–3283, June, 2025.
- CICCONETO, F., VIEIRA, L. V., ABEL, M., DOS SANTOS ALVARENGA, R., CARBONERA, J. L., AND GARCIA, L. F. Georeservoir: An ontology for deep-marine depositional system geometry description. *Computers Geosciences* vol. 159, pp. 105005, 2022.
- EDGE, D., TRINH, H., CHENG, N., BRADLEY, J., CHAO, A., MODY, A., TRUITT, S., METROPOLITANSKY, D., NESS, R. O., AND LARSON, J. From local to global: A graph rag approach to query-focused summarization, 2025.
- ES, S., JAMES, J., ESPINOSA ANKE, L., AND SCHOCKAERT, S. RAGAs: Automated evaluation of retrieval augmented generation. In *EACL*, N. Aletras and O. De Clercq (Eds.), 2024.
- GARCIA, L. F., ABEL, M., PERRIN, M., AND DOS SANTOS ALVARENGA, R. The geocore ontology: A core ontology for general use in geology. *Comput. Geosci.* 135 (C), Feb., 2020.
- LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., KÜTTLER, H., LEWIS, M., TAU YIH, W., ROCKTÄSCHEL, T., RIEDEL, S., AND KIELA, D. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.
- OTERO-CERDEIRA, L., RODRÍGUEZ-MARTÍNEZ, F. J., AND GÓMEZ-RODRÍGUEZ, A. Ontology matching: A literature review. *Expert Systems with Applications* 42 (2): 949–971, 2015.
- PAULA, F., MICHELIN, C. R. L., AND MOREIRA, V. Evaluation of question answer generation for Portuguese: Insights and datasets. In *EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen (Eds.). pp. 5315–5327, 2024.
- PENG, B., ZHU, Y., LIU, Y., BO, X., SHI, H., HONG, C., ZHANG, Y., AND TANG, S. Graph retrieval-augmented generation: A survey. *ACM Trans. Inf. Syst.* 44 (2), Dec., 2025.
- SHARMA, K., KUMAR, P., AND LI, Y. OG-RAG: Ontology-grounded retrieval-augmented generation for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.
- ZHU, X., XIE, Y., LIU, Y., LI, Y., AND HU, W. Knowledge graph-guided retrieval augmented generation. In *Proceedings of NAACL*, 2025.