Retrieval-Augmented Generation for Question Answering over Brazilian Portuguese Institutional Normative Documents

Avelar Rodrigues de Sousa[1], **Jean Carlos Rodrigues Sousa**[2], **Karielly de Carvalho**[2]
Manoel Messias Pereira Medeiros[3], **Marx Rodrigues de Moura**[3]

[1] Institute of Mathematics and Computer Science (ICMC)
University of São Paulo (USP)
São Carlos – SP – Brazil
avelars@usp.br

[2]Federal University of Piauí (UFPI)
Picos – PI – Brazil
{jean.rodrigues, karielly.carvalho}@ufpi.edu.br

[3]Federal Institute of Piauí (IFPI)
Piccos – PI – Brazil
{mmessias, marxrodriguesdemoura}@ifpi.edu.br

Abstract. *This paper evaluates a Retrieval-Augmented Generation (RAG) prototype for question answering over Brazilian Portuguese institutional normative documents. The study uses an IFPI document subset, 20 factual questions, 10 independent runs, four retrieval settings, and two generator scales:* `gemma3:12b` *and* `gemma3:27b`*. Results show that dense retrieval outperformed lexical baselines in end-to-end answer quality, while reranking improved retrieval metrics without necessarily improving generated answers. The larger generator achieved modest semantic gains but was substantially slower. Error analysis showed limitations in metadata-sensitive questions involving document numbers, dates, and administrative references. The findings provide exploratory evidence about standard RAG components in an institutional public-information scenario and highlight the need for explicit source grounding and exact normative verification.*

1. Introduction

In contexts where transparency and public accountability are essential, citizens, public servants, managers, and legal practitioners often face difficulties in locating information within extensive normative documents, such as laws, decrees, ordinances, and institutional regulations. These documents combine technical language, formal style, and fragmented organization, making access difficult when search depends mainly on exact lexical matching. Traditional retrieval methods such as BM25 and TF-IDF remain widely used, but may fail when relevant content is expressed through different formulations [Aquino et al. 2024, Karpukhin et al. 2020]. Improving natural-language access to such documents may therefore contribute to transparency, institutional accessibility, and more equitable access to public information [UNESCO 2021].

Retrieval-Augmented Generation (RAG) has emerged as a relevant paradigm for this scenario by combining a retrieval component with a language generation model.

In this framework, a user query is used to retrieve relevant passages from an external collection, and the retrieved content is then incorporated as grounding context for answer generation [Karpukhin et al. 2020, Lewis et al. 2020]. Prior work has shown that this integration can improve access to relevant information and help produce answers that are more closely anchored in textual evidence than purely parametric generation [Lewis et al. 2020, Ji et al. 2023].

Recent surveys have consolidated RAG as an important research direction, covering architectural variations, retrieval strategies, reranking methods, and evaluation practices [Fan et al. 2024, Gao et al. 2024, Gan et al. 2025]. In the Brazilian and Portuguese-language context, recent studies have begun to explore RAG and retrieval-based Question Answering (QA) in legal settings, including information extraction from Brazilian legal documents [Aquino et al. 2024], evaluation strategies based on implicit feedback [Albuquerque et al. 2024], domain-specific embedding models for Brazilian legal QA [Novais et al. 2025], and graph-based RAG (GraphRAG) pipelines for Portuguese-language legal documents [Oliveira et al. 2026]. These studies indicate the relevance of investigating RAG pipelines in national institutional contexts, especially for document collections written in Portuguese.

In this context, this paper reports an exploratory empirical study of a RAG prototype for Brazilian Portuguese institutional normative documents. Rather than proposing a new RAG architecture, the study examines how established retrieval and generation components behave in a controlled IFPI public-information scenario. The study provides: (i) a reproducible experimental pipeline for Portuguese normative PDF documents; (ii) a controlled comparison between lexical and dense retrieval strategies, with and without reranking; (iii) an analysis of the quality–latency trade-off between two generator scales; and (iv) a qualitative error analysis focused on metadata-sensitive normative questions. The results should be interpreted as evidence about the evaluated institutional collection, not as a general benchmark for Brazilian public normative documents.

This article is organized as follows: Section 2 presents the background and related work. Section 3 details the architecture and implementation of the evaluated prototype. Section 4 reports and discusses the experimental results. Finally, Section 5 concludes the paper and outlines limitations and directions for future research.

2. Background and Related Work

RAG combines information retrieval and language generation. In a typical RAG pipeline, a user query is first used to retrieve candidate passages from an external document collection; these passages are then inserted into the prompt as grounding context for the generator. This design reduces exclusive dependence on the parametric knowledge of the language model and allows generated answers to be conditioned on explicit textual evidence [Lewis et al. 2020, Ji et al. 2023].

Retrieval strategies are commonly divided into lexical and dense approaches. Lexical methods, such as BM25 and TF-IDF, rely on term frequency, inverse document frequency, and exact or approximate word overlap between queries and documents. These methods are efficient and interpretable, but they may be limited when relevant passages use different wording from the user query. Dense retrieval, in contrast, maps queries and passages to vector representations and compares them in an embedding space. Dense

Passage Retrieval was introduced as an effective approach for open-domain question answering, showing that dense representations can outperform lexical baselines in some QA settings [Karpukhin et al. 2020].

A second component often used in RAG pipelines is reranking. In this setting, an initial retriever returns a broader set of candidate passages, and a cross-encoder or another relevance model reorders these candidates before the final context is built. Recent surveys have systematized RAG architectures, retrieval methods, reranking strategies, generation models, and evaluation practices, indicating that pipeline behavior varies substantially across domains, collections, and context-construction strategies [Gao et al. 2024, Fan et al. 2024, Gan et al. 2025].

For normative and legal-administrative documents, grounding is especially important because small errors in document numbers, dates, revoked rules, or administrative references may change the meaning of an answer. Prior work has also shown that irrelevant or noisy retrieved context may degrade generated responses, which reinforces the need to analyze not only retrieval metrics, but also the quality of the final generated answers [Yoran et al. 2024, Huang et al. 2025]. Benchmarks such as BEIR also highlight that retrieval effectiveness varies across tasks and domains, supporting the need for domain-specific empirical analysis [Thakur et al. 2021].

In the Brazilian and Portuguese-language context, applications of RAG and retrieval-based QA to legal and public-sector documents are still emerging. Prior work has investigated RAG for information extraction from Brazilian legal PDF documents, including the evaluation of chunking strategies, embedding models, and numbers of retrieved passages [Aquino et al. 2024]. Complementary research has proposed evaluation strategies for RAG applications based on implicit user feedback [Albuquerque et al. 2024]. More recent studies have explored domain-specific embedding models for Brazilian legal QA [Novais et al. 2025] and structurally aware GraphRAG pipelines for Portuguese-language legal documents [Oliveira et al. 2026]. These works reinforce the importance of retrieval quality, document structure, and grounding in legal and administrative QA systems.

Thus, there is still room for empirical studies that characterize how retrieval configurations and generator scales affect QA over Portuguese institutional normative documents. Unlike prior work focused on information extraction, feedback-based evaluation, embedding adaptation, or graph-based modeling, this study evaluates a closed IFPI collection under a controlled comparison of BM25, TF-IDF, dense retrieval, and dense retrieval with reranking, coupled with two generator scales. This positioning is deliberately practical and diagnostic: the goal is to measure the behavior, strengths, and failure modes of a standard assistive RAG pipeline for public normative information in Brazilian Portuguese.

3. Methodology

The prototype was developed in Python[1] and organized into two stages: indexing, responsible for document preparation and searchable collection construction, and experimentation, responsible for retrieval, answer generation, and evaluation. This separation improves reproducibility and allows the indexed collection to be reused across experi-

[1] https://www.python.org/

mental runs.

The code and experimental materials used in this study are publicly available in a GitHub repository: `https://github.com/avelando/IFPI-wise`

Figure 1 summarizes the overall workflow adopted in this study. Stage 1 is executed once to preprocess the PDF collection, segment the documents, and build the searchable representations. Stage 2 is repeatedly executed to evaluate retrieval settings, context construction, answer generation, and final metrics under different generator scales and random seeds.

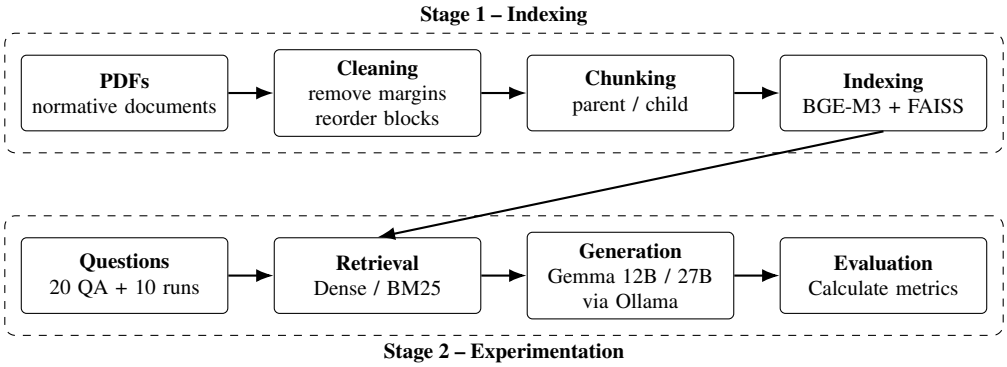


Figure 1. Overview of the two-stage methodology used in the prototype. Stage 1 builds the searchable index, while Stage 2 evaluates retrieval, answer generation, and final metrics across methods and runs.

Experimental Setup. The experiments were conducted from Google Colaboratory[2], with execution connected to a local server for model inference. This configuration made it possible to maintain an interactive experimental workflow while using local computational resources during the answer generation stage. The local infrastructure used in the experiments included an Intel Core i5-10400F processor, 64 GB of RAM, SSD storage, and an NVIDIA RTX 3060 GPU with 12 GB of VRAM.

Document Collection and Preprocessing. The indexed collection is composed of normative and institutional documents in PDF format. During preprocessing, each file is read page by page, and textual blocks are extracted from the document layout. In order to reduce recurrent noise commonly found in institutional files, blocks located in the upper and lower margins of the page are removed before the final textual representation is assembled. The remaining blocks are ordered according to their position on the page and then merged into a continuous textual sequence for each document. At this stage, metadata such as source file name, total number of pages, and character count are preserved.

This preprocessing reduces the interference of repeated visual elements such as headers, footers, and administrative marks during semantic indexing. The indexed corpus contains 248 documents and exhibits substantial variation in length, ranging from 1 to 281 pages, with a mean of 37.56 pages and a median of 16 pages per document. This heterogeneity motivated the adoption of a hierarchical chunking strategy in the subsequent indexing stage. Figure 2 shows the page-length distribution recorded during index construction.

[2] `https://colab.research.google.com/`

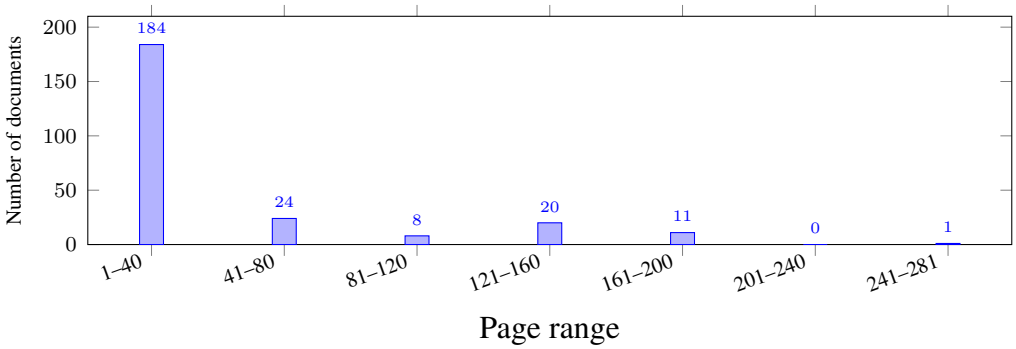


Figure 2. Histogram of document lengths by page range in the indexed collection.

All documents and questions were processed directly in Brazilian Portuguese. No translation or cross-lingual adaptation was applied. This design choice preserves the original linguistic characteristics of the collection and avoids introducing external transformations between the source documents and the generated answers.

Hierarchical Indexing. After preprocessing, each document is segmented using a hierarchical strategy composed of larger parent segments and smaller child segments, reconciling two competing needs: preserving broader local context during answer generation while maintaining smaller textual units for more precise retrieval.

In the implemented configuration, parent segments have 1,500 characters with 200-character overlap, and child segments have 400 characters with 50-character overlap. The overlap values reduce the risk of separating related information across adjacent chunks, which is relevant in normative documents where definitions, dates, and administrative references may appear close to one another. The dense indexing stage embeds the child segments with the `BAAI/bge-m3`[3] model and stores them in a FAISS[4] vector index for retrieval, while the parent segments are preserved for contextual reconstruction during answer generation. The indexed representations are stored once and reused across experimental runs, which makes repeated evaluation more stable and computationally practical.

Retrieval and Answer Generation. At query time, the prototype supports four retrieval configurations: Dense without reranking, Dense with reranking, BM25, and TF-IDF. The dense retriever operates over the hierarchical index and returns up to 10 candidate passages. The lexical baselines were built over the same indexed collection of parent segments, allowing a controlled comparison between lexical and dense retrieval under a shared document representation. In the Dense with reranking setting, the 10 initial candidates are reordered by the cross-encoder relevance model (cross-encoder/ms-marco-MiniLM-L-6-v2)[5]. In all settings, the final context is formed from up to 4 highest-ranked passages, which are concatenated and presented to the generator together with document-level source identifiers.

The retrieval depth of 10 candidates provides enough passages for the reranking stage, while limiting the final context to 4 passages reduces prompt length and avoids

[3] `https://huggingface.co/BAAI/bge-m3`
[4] `https://faiss.ai/index.html`
[5] `https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2`

including potentially irrelevant information.

The generation stage is controlled by a prompt in Brazilian Portuguese that instructs the model to answer only from the retrieved context, avoid unsupported information, and produce concise factual answers in complete sentences. The prompt also includes short style examples and uses the same decoding configuration across runs, with low-temperature generation and fixed context window parameters. In the current implementation, source grounding is provided at the document-file level rather than at the page or paragraph level.

Answer generation was performed through Ollama[6] using two language models under the same experimental protocol: `gemma3:12b`[7] and `gemma3:27b`[8]. These models were selected to represent two generator scales in an operational and comparative sense: `gemma3:12b` represents a Small Language Model (SLM) setting, whereas `gemma3:27b` represents a Large Language Model (LLM) setting. The purpose of this comparison is not to redefine the prototype itself, but to verify how the same retrieval pipeline behaves when coupled with generators of different scales. In both cases, the retrieval procedure remains fixed and only the generative model varies at inference time.

Evaluation Protocol. The evaluation dataset was manually constructed and verified by the authors exclusively for investigative evaluation of the prototype, rather than as a general-purpose benchmark. The final dataset contains 20 factual questions derived from IFPI normative and institutional documents publicly available on the official Conselho Superior (Higher Council, CONSUP) website[9]. Each instance is organized as a structured record containing the reference document, the natural-language question, and an expected answer. This structure makes it possible to compare generated answers against document-grounded reference responses under controlled conditions. In the implementation, all question instances are validated to contain these three mandatory fields before execution. Because the dataset is small, manually built, and restricted to a single institutional collection, the results should be interpreted as an exploratory case study rather than as evidence of broad generalization across Brazilian public normative documents.

Each question was evaluated under four retrieval settings and two generator scales. The full experimental protocol was executed across 10 independent rounds, each associated with a different random seed derived from a common base configuration, resulting in a total of 1,600 answer generations. This repeated-run design was adopted to reduce the influence of stochastic variation in the generation stage and to allow the reporting of mean values and standard deviations across runs.

For each question and retrieval configuration, the prototype stores the generated answer, the recovered sources, the number of retrieved sources, the recovered context, and the end-to-end latency. The evaluation combines lexical, semantic, and retrieval-oriented metrics in order to capture different aspects of system behavior. ROUGE and BLEU assess lexical overlap with the reference answer, BERTScore estimates semantic similarity, token-overlap F1 balances precision and recall after normalization, and Exact Match (EM)

[6] `https://ollama.com/`

[7] `https://ollama.com/library/gemma3:12b`

[8] `https://ollama.com/library/gemma3:27b`

[9] `https://sites.google.com/ifpi.edu.br/ifpi-resolucoes-do-consup/p\%C3\%A1gina-inicial`

serves as a stricter auxiliary indicator. Retrieval effectiveness is primarily measured with Hit@k, Mean Reciprocal Rank (MRR), and the rate at which the target source appears in the final context provided to the generator. Because each question is associated with a single target document, Recall@k is numerically equivalent to Hit@k and is therefore not emphasized in the analysis. Results are aggregated both at the question level and at the method level, and partial outputs are saved after each round to improve experimental traceability. As a preliminary sanity check, the protocol also includes one out-of-scope question to inspect abstention behavior when the retrieved context does not support an answer.

The evaluation is based on automatic metrics and qualitative inspection by the authors. No independent human evaluation was conducted, which limits conclusions about perceived usefulness, factual reliability, and adequacy for real users. This limitation is particularly relevant in normative-document QA, where factual grounding, exact references, and abstention behavior are central requirements.

The present study focuses on a public-information access scenario in which the system is intended to assist users in locating and summarizing normative content in natural language. It is not designed as a substitute for formal legal or administrative verification, particularly in cases involving exact document identifiers, dates, revoked rules, or other metadata-sensitive information.

4. Results and Analysis

The following subsections report the aggregated results of the evaluated configurations, covering answer quality, retrieval effectiveness, robustness across runs, and qualitative error patterns.

4.1. Comparative Results Across Retrieval Methods and Generators

Table 1 summarizes the mean results for each retrieval setting under both generators, including lexical, semantic, token-overlap, exact-match, and latency metrics. Overall, dense retrieval achieved the strongest answer-quality results for both models, indicating a better balance between lexical overlap and semantic adequacy than the lexical baselines in the evaluated IFPI question set.

As shown in Table 1, Dense without reranking achieved the best end-to-end answer quality for both generators, leading in ROUGE-1, BLEU, BERTScore F1, and Token F1. The only exception was EM for `gemma3:27b`, where BM25 obtained the highest value. Although Dense + Reranking remained competitive, it did not surpass Dense without reranking in the main generation metrics. Given the small number of questions, these differences should be interpreted as descriptive evidence rather than as statistically validated superiority.

The comparison between generators also shows a clear quality–latency trade-off. The `gemma3:27b` model achieved modest gains in BLEU and BERTScore F1, but was approximately five to six times slower than `gemma3:12b`. Under the evaluated infrastructure and question set, `gemma3:12b` appears more suitable for interactive use, while `gemma3:27b` may be preferable only when small quality gains justify the additional computational cost.

Table 1. Answer-quality and latency results by retrieval setting and generator. Best values within each model are shown in bold; for latency, lower values are better. BS-F1 = BERTScore F1, Token F1 = token-overlap F1, and Lat. = latency in seconds.

Model	Retrieval setting	ROUGE-1	BLEU	BS-F1	Token F1	EM	Lat. (s)
gemma3:12b	BM25	0.497	0.225	0.899	0.517	0.060	**3.792**
	TF-IDF	0.425	0.156	0.887	0.456	0.050	3.911
	Dense + Reranking	0.570	0.311	0.914	0.616	0.100	4.365
	Dense	**0.648**	**0.390**	**0.926**	**0.679**	**0.200**	4.042
gemma3:27b	BM25	0.460	0.261	0.901	0.466	**0.100**	**19.899**
	TF-IDF	0.454	0.220	0.894	0.472	0.050	19.995
	Dense + Reranking	0.545	0.307	0.916	0.586	0.065	23.406
	Dense	**0.632**	**0.421**	**0.932**	**0.670**	0.085	21.949

4.2. Retrieval Effectiveness

Because retrieval is independent of the generator, retrieval-oriented metrics are identical for `gemma3:12b` and `gemma3:27b`. Table 2 reports Hit@k, MRR, and the rate at which the target source appears in the final context. Recall@k is omitted because each question is associated with a single target document, making it numerically equivalent to Hit@k in this evaluation design.

Table 2. Retrieval metrics by retrieval setting. Retrieval is generator-independent in the current pipeline, so the values are identical for both generators.

Setting	H@1	H@3	H@5	H@10	MRR	Src.@Ctx.
BM25	0.350	0.550	0.550	0.650	0.454	0.550
TF-IDF	0.650	0.750	0.750	0.750	0.700	0.750
Dense + Reranking	**0.800**	**0.800**	**0.900**	**0.950**	**0.831**	**0.850**
Dense	0.700	0.750	0.800	**0.950**	0.760	0.800

Dense + Reranking achieved the strongest retrieval effectiveness, with the best or tied-best values in most ranking metrics. However, these gains did not translate into better generated answers, since the best end-to-end results were obtained by Dense without reranking. This suggests that better document ranking alone is not sufficient to guarantee better generation under the current context-construction strategy. One possible explanation is that the target document may be retrieved and ranked highly while the final selected passages still omit the specific evidence needed for exact answer generation.

4.3. Robustness Across Runs

Table 3 reports the mean $\pm$ standard deviation of the round-level averages across the 10 runs for the best-performing Dense setting without reranking in each model. Smaller standard deviations indicate greater stability across repeated executions; for latency, lower values are preferable.

The results indicate stable behavior across repeated executions, with small standard deviations in the main metrics. The `gemma3:12b` model achieved the highest

Table 3. Mean ± standard deviation of the best-performing Dense setting without reranking across 10 runs. Best mean values are shown in bold; for latency, lower values are better.

Generator	ROUGE-1	BLEU	BERTScore F1
`gemma3:12b`	0.648 ± 0.003	0.390 ± 0.007	0.926 ± 0.001
`gemma3:27b`	0.632 ± 0.009	0.421 ± 0.014	0.932 ± 0.001
Generator	**Token F1**	**EM**	**Latency (s)**
`gemma3:12b`	0.679 ± 0.003	0.200 ± 0.000	4.042 ± 0.020
`gemma3:27b`	0.670 ± 0.006	0.085 ± 0.024	21.949 ± 0.268

ROUGE-1, Token F1, and EM values, while also presenting the lowest latency. In contrast, `gemma3:27b` obtained the highest BLEU and BERTScore F1 values, but with substantially higher latency. These results suggest that the observed trends are not driven by isolated favorable runs, although the small question set still limits statistical generalization.

4.4. Error Analysis

Errors are concentrated in questions requiring exact normative identification, especially those involving numbers, dates, and document references. In such cases, the system often produces answers that are semantically plausible but factually incorrect, showing that semantic similarity alone is insufficient to guarantee correct normative grounding.

Two recurring error types were observed: metadata substitution, in which the system retrieves or generates a numerically plausible but incorrect identifier, and context drift, in which the answer remains topically related to the target document but shifts toward a nearby concept. These effects are most visible when the reference answer is short and highly specific, since even small deviations strongly affect EM and token-overlap F1.

Table 4 presents a representative failure case for the best-performing dense retrieval setting. This example highlights the difficulty of metadata-sensitive questions involving exact normative references, even when the generated answer remains topically related to the target document.

Table 4. Failure qualitative example for the best dense retrieval setting.

Failure
Question: *Qual resolução foi revogada pela Resolução Normativa nº 236 de 2025?* (Which resolution was revoked by Normative Resolution No. 236 of 2025?)
Expected answer: *Foi revogada a Resolução Normativa nº 133/2022 - CONSELHO SUPERIOR, de 29 de abril de 2022.* (Normative Resolution No. 133/2022 - Higher Council, of April 29, 2022, was revoked.)
`gemma3:12b`: *A Resolução Normativa nº 236 de 2025 revogou a Resolução Normativa nº 206/2024 do Conselho Superior.* (Normative Resolution No. 236 of 2025 revoked Normative Resolution No. 206/2024 of the Higher Council.)
`gemma3:27b`: *A Resolução Normativa nº 236 de 2025 revogou a Resolução nº 15/2021 - CONSELHO SUPERIOR, de 19 de maio de 2021, e a Resolução Normativa nº 206/2024 - CONSELHO SUPERIOR, de 13 de maio de 2024.* (Normative Resolution No. 236 of 2025 revoked Resolution No. 15/2021 - Higher Council, of May 19, 2021, and Normative Resolution No. 206/2024 - Higher Council, of May 13, 2024.)

A further representative failure involved the question *"Qual é a forma de acesso*

ao curso FIC em Aquicultor?" ("What is the mode of admission to the FIC course in Aquaculture?"). The expected answer was *"por meio de sorteio"* ("through a lottery system"), but the generated answers tended to drift toward adjacent admission requirements, such as incomplete elementary education. This pattern indicates that the errors are not restricted to numerical metadata; they also emerge when retrieved passages contain semantically related but administratively different information.

A complementary out-of-scope sanity check was conducted to inspect whether the system could abstain when the retrieved context did not support the question. As shown in Table 5, `gemma3:27b` abstained in this illustrative case, whereas `gemma3:12b` produced unsupported answers based on incidental retrieved content. Because this check uses only one out-of-scope question, it should not be interpreted as a systematic evaluation of abstention behavior.

Table 5. Illustrative out-of-scope sanity check with a question unrelated to the indexed collection.

Out-of-Scope Sanity Check
Question: *Qual é a capital do Japão?* (What is the capital of Japan?)
Expected answer: *Não encontrei essa informação no contexto fornecido.* (I could not find this information in the provided context.)
gemma3:12b: The smaller model failed to abstain and produced unsupported answers across all retrieval settings, typically drifting to incidental mentions of Teresina as the capital of the state of Piauí. A representative answer was: *"Teresina é a capital do Estado do Piauí."* ("Teresina is the capital of the state of Piauí.")
gemma3:27b: The larger model correctly abstained across all retrieval settings, consistently returning: *"Não encontrei essa informação no contexto fornecido."* ("I could not find this information in the provided context.")

Overall, dense retrieval proved more effective than lexical baselines for the evaluated IFPI collection, although stronger retrieval ranking did not necessarily lead to better generated answers. The prototype performed better on descriptive factual questions, but failures in metadata-sensitive cases indicate that standard RAG systems for normative documents should be used with caution when exact legal or administrative verification is required.

5. Conclusions

This paper presented an exploratory empirical evaluation of a RAG prototype for QA over Brazilian Portuguese institutional normative documents. Evaluated on an IFPI subset, the system showed potential for supporting natural-language access to public normative information in a socially relevant context of transparency and institutional accessibility.

The results showed that dense retrieval achieved the strongest answer-quality results among the evaluated settings, while Dense + Reranking achieved the strongest retrieval ranking effectiveness. This contrast indicates that better retrieval metrics do not necessarily translate into better generated answers under the current context-construction strategy. The comparison between `gemma3:12b` and `gemma3:27b` also revealed a quality–latency trade-off: the larger model obtained modest gains in BLEU and BERTScore F1, whereas the smaller model remained substantially faster and achieved slightly higher ROUGE-1 and token-overlap F1 in the best dense setting.

The evaluation across 10 runs showed stable behavior, while the error analysis highlighted the main remaining limitations of the prototype, especially in questions requiring exact normative identification, such as those involving document numbers, dates, revoked rules, and highly specific administrative references. The study is limited by the small evaluation set of 20 questions, by its focus on a single institutional collection, and by the absence of independent human evaluation. Therefore, the results should be interpreted as exploratory and diagnostic rather than as broadly generalizable evidence. In addition, the out-of-scope sanity check provided only preliminary evidence that abstention behavior may differ across generator scales, reinforcing the need for a broader evaluation of unsupported queries.

Overall, the evaluated pipeline provides preliminary evidence that dense retrieval combined with controlled generation can support access to institutional normative documents in Portuguese, especially for descriptive factual questions. Future work may advance the prototype in two complementary directions: from a system perspective, by extending it into a user-facing application with page-level source inspection, document navigation, feedback collection, and regular index updates; and from an evaluation perspective, by expanding the question set, including a broader and more diverse document collection, adopting independent human evaluation, and defining more systematic procedures for assessing grounding, abstention, and answer reliability. More broadly, the study suggests that established Artificial Intelligence methods can support socially relevant public-information applications when combined with careful evaluation, explicit grounding, and clear recognition of their limitations.

References

Albuquerque, A., Wensing, I., Joppi Filho, N., and Dorneles, C. (2024). Avaliação de aplicações de geração aumentada de recuperação por meio de feedback implícito. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 253–259, Porto Alegre, RS, Brasil. SBC.

Aquino, I., dos Santos, M. M., Dorneles, C., and Carvalho, J. T. (2024). Extracting information from brazilian legal documents with retrieval augmented generation. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 280–287, Porto Alegre, RS, Brasil. SBC.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. (2024). A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501, New York, NY, USA. Association for Computing Machinery.

Gan, A., Yu, H., Zhang, K., Liu, Q., Yan, W., Huang, Z., Tong, S., and Hu, G. (2025). Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey. *arXiv preprint arXiv:2504.14891*.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.

Huang, Y., Xu, Y. E., Sun, K., Yan, V., Sun, A., Khan, H., Nguyen, J., Chen, J., Kachuee, M., Lin, Z., Liu, Y., Colak, A., Kumar, A., Yih, W.-t., and Dong, X. L. (2025). Confrag: Confidence-guided retrieval-augmenting generation.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781, Online. Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Novais, A., Ferreira, D., Silva, J., Brakes, M., Presa, J., and Oliveira, S. (2025). Enhancing legal question answering in brazilian portuguese through domain-specific embedding models. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 523–533, Porto Alegre, RS, Brasil. SBC.

Oliveira, V. T. d., Silva, D. O. d., Souza, M. d. A., Lima, M. R., Oliveira, S. S. T. d., and Rosa, T. C. (2026). Retrieval-augmented generation and knowledge graphs in Portuguese-language legal documents. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 1–10, Salvador, Brazil. Association for Computational Linguistics.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models.

UNESCO (2021). To recovery and beyond: 2021 report on public access to information (sdg 16.10.2). Technical report, UNESCO, Paris.

Yoran, O., Wolfson, T., Ram, O., and Berant, J. (2024). Making retrieval-augmented language models robust to irrelevant context.