# A RAG Chatbot with Incremental Context Retrieval based on Local LLMs for Hospital Documents

**Murilo Vargas da Cunha**[1,2] **Marília Rosa Silveira**[1] **César Brasil Sperb**[1]
**Larissa Astrogildo Freitas**[1] **Ulisses Brisolara Corrêa**[1]
[1]Federal University of Pelotas (UFPEL), Pelotas, RS, Brasil
[2]Federal Institute of Rio Grande do Sul (IFRS), Rio Grande, RS, Brasil
{mvcunha, mrsilveira, cbsperb, larissa, ulisses}@inf.ufpel.edu.br

## Abstract

The adoption of LLMs in hospital environments demands solutions that ensure information security, computational efficiency, and rigorous control over sensitive institutional data. This work presents the development and evaluation of a chatbot based on RAG, using exclusively local LLMs, applied to internal documents of a university hospital in Portuguese, composed of Standard Operating Procedures and technical manuals. The methodology initially evaluates the quality of information retrieval through dense embedding models, measured by the Mean Reciprocal Rank (MRR) metric. Then, the generation stage is analyzed in two distinct scenarios: (i) RAG with fixed context, in which multiple chunks are provided simultaneously to the model, and (ii) Incremental page retrieval, in which chunks are sent sequentially according to the retrieval ranking. The generation assessment was conducted with four local LLMs — MedGemma3:27B, Gemma3:27B, Gpt-oss:20B, and Mistral Small 3.1 — using BERTScore as a quality metric. The results indicate that indiscriminate context increase in the fixed-context scenario degrades generation quality, even while increasing the probability of recovering the relevant chunk. In contrast, the incremental page retrieval technique showed improvements in BERTScore values, with the MedGemma3:27B model standing out with the best overall results. These findings demonstrate that adaptive context control is a critical factor in increasing the reliability and efficiency of RAG systems based on local LLMs in the healthcare domain.

## 1 Introduction

Large Language Models (LLMs) have been widely used in natural language processing (NLP) tasks in healthcare (Abo El-Enen et al., 2025; Lee et al., 2023). However, applications in the hospital setting involve highly sensitive documents, such as Standard Operating Procedures (SOPs), technical manuals, and internal regulations, which often cannot be sent to external services due to privacy restrictions, information security, and institutional compliance (Haltaufderheide and Ranisch, 2024; Li et al., 2023). In this scenario, the use of locally executed LLMs emerges as a viable alternative, as it allows data to remain entirely under the institution's control (Ng et al., 2024).

Nevertheless, these models are limited by their training data, which are often outdated and prone to generating inaccurate content, such as hallucinations, producing plausible but unfounded responses based on available data (Perković et al., 2024; Ji et al., 2023). This behavior represents a significant risk in hospital settings, where the accuracy and reliability of information are essential (Amugongo et al., 2025).

As an alternative to these implications, a technique has been widely adopted, Retrieval-Augmented Generation (RAG), in which the model generates responses conditioned on excerpts retrieved from an external document database (Neha et al., 2025; Oliveira et al., 2025). However, this strategy applied in local LLMs also presents additional limitations, especially related to the processing of large volumes of context, where the efficient incorporation of external knowledge during inference in long or dynamic contexts remains a challenge (Taguchi et al., 2025).

In this sense, fixed-context approaches tend to retrieve a predefined number of segments or chunks and concatenate them into a single input for the language model, which can introduce irrelevant or conflicting information. Furthermore, this strategy does not guarantee that the LLM will effectively utilize the retrieved data during generation, resulting in responses that may still be inaccurate or inconsistent (Asai et al., 2023).

This problem is particularly relevant in local LLMs, where a degradation in the quality of responses is observed as the size of the context in-



*Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*
*Vol. 2*, pages 68–77
April 13-16, 2026 

creases, whether due to attentional limitations, loss of focus, or a greater propensity for hallucinations (Levy et al., 2024; Li et al., 2024). Thus, the use of a fixed context may become inadequate in scenarios with extensive and heterogeneous documentary databases.

In order to mitigate this limitation, this work proposes an incremental page-by-page retrieval strategy, in which the retrieved chunks/pages are sent individually to the LLM, following the order of relevance ranking, interrupting the process as soon as a valid response is obtained. This approach avoids the cumulative sending of context and seeks to improve the effective use of retrieved information, reducing the interference of irrelevant segments. This method does not require fine-tuning of the model or additional inferences in LLM.

Given this context, this article presents the development and evaluation of a RAG-based chatbot using exclusively local LLMs, exploring the incremental page retrieval technique, applied to internal documents of a university hospital composed of SOPs and technical manuals in Portuguese. The evaluation considers both the quality of the retrieval and the generation of responses, contributing to the discussion of more effective and secure practices in the use of local LLMs in sensitive hospital environments.

## 2 Related Works

The use of RAG in healthcare applications has been extensively investigated as a strategy to mitigate hallucinations and increase the reliability of LLM-based systems. In hospital settings, the work of (Son et al., 2025) proposes a RAG chatbot to support operational queries in Electronic Patient Record (EMR) systems, employing contrast-learning-tuned multilingual embeddings and a commercial LLM via API. The results demonstrate high retrieval performance, with Top-K Accuracy exceeding 97%, focusing the analysis primarily on the retrieval stage.

In the context of patient education, the study by (Baur et al., 2025) presents a German-language RAG chatbot for orthopedics and traumatology, built from validated clinical guidelines and educational materials. The system combines semantic search with the Qdrant vector database and generation with OpenAI's GPT, being evaluated by automatic metrics and user studies, which indicate high acceptance and perceived quality. Another work that explores languages other than English is that of (Zhang et al., 2025a), in which the authors integrate RAG with medical knowledge graphs for clinical question-and-answer systems in Chinese, demonstrating consistent gains in accuracy in clinical benchmarks. These works reinforce the practical applicability of RAG in healthcare, without exploring adaptive mechanisms for context control.

Approaches focused on initial medical screening and guidance are also explored by (Nandi et al., 2024), who propose the MedMate chatbot, based on a RAG pipeline with BERT embeddings, retrieval via FAISS, and generation with LLaMA. The authors report an approximate accuracy of 76% and good alignment with medical recommendations, highlighting the potential of RAG to surpass traditional searches. The analysis, however, does not consider the impact of context size on the quality of the generation.

The work of (Kulshreshtha et al., 2025) develops a RAG-based medical chatbot using LLaMA 3.2-3B, LangChain, and FAISS, with a document-based database derived from MedlinePlus. The system injects retrieved snippets directly into the generator model's prompt and is evaluated using qualitative metrics of conciseness, accuracy, and relevance, as well as comparisons with commercial models. While demonstrating gains in reliability, the study does not analyze limitations associated with the use of extensive contexts.

Another study explores specialized biomedical models and semantic context enrichment; the authors of (Sinha et al., 2024) propose CMedRAGBot, which combines BioMistral-7B and PubMedBERT with conversational retrieval chains to support clinical triage. Although the study highlights the potential of RAG chatbots for clinical support and patient engagement, it does not systematically evaluate how the way context is delivered influences generation.

In contrast, the present work explicitly investigates the impact of context control on the quality of data generation in RAG systems applied to sensitive hospital documents, considering exclusively local LLMs.

## 3 Methodology

This section describes the methodology adopted for the development and evaluation of the chatbot. The methodology was structured in sequential steps, allowing for separate evaluation of information re-

trieval, response generation, and the proposed strategy for adaptive context control.

### 3.1 Dataset

The document collection consists of 107 internal normative documents from five hospital departments: Occupational Health and Safety Unit, Human Resources Division, Hospital Infection Control Service, Teaching and Research Management Department, and Information Technology and Digital Health Sector (Empresa Brasileira de Serviços Hospitalares (EBSERH), 2025).

The documents were segmented into smaller units (chunks) for indexing in the vector database. We adopted chunks of approximately 100 tokens, applying a soft limit: upon reaching this value, the cut was shifted to the next endpoint, preserving syntactic cohesion. To evaluate the generation of responses with LLM, the full text of the source page of each chunk was also stored in a dedicated field in the database containing approximately 400 tokens.

For the system evaluation, a hybrid dataset with 192 questions and answers was constructed, divided into two complementary subsets:

- **Synthetic dataset:** composed of 74 questions/answers automatically generated with Gemini 1.5 Flash, and manual identification of the identifier (ID) of the chunk containing the correct answer in the database. This subset allows for an objective and controlled evaluation of the retrieval step;
- **Expert dataset:** composed of 118 questions/answers developed by hospital professionals in a preliminary qualitative evaluation phase of the system.

Table 1 presents representative examples of questions and answers, illustrating the linguistic structure and informational granularity addressed in the evaluation.

### 3.2 Information Retrieval Assessment

The initial stage of the methodology evaluated the quality of information retrieval. This evaluation was performed using only the synthetic dataset of 74 questions, as it enables precise identification of whether the correct chunk was retrieved. The evaluated models were Gemma Embedding (embeddinggemma) (Vera et al., 2025), Qwen3 Embedding (qwen3embedding) (Zhang et al., 2025b), Granite Embedding (graniteembedding) (Awasthy et al., 2025), and multilingual-e5-small (intfloat/multilingual-e5-small) (Wang et al., 2024). Retrieval was performed using semantic similarity, measuring the cosine similarity between the questions and the chunks stored in the database, resulting in the selection of the 100 chunks with the best ranking. Performance was evaluated using the Mean Reciprocal Rank (MRR) metric. The evaluation process is shown in Figure 1.

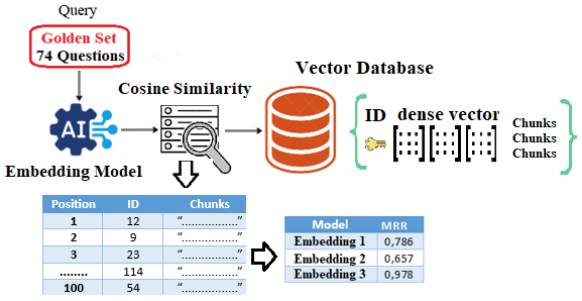


Figure 1: Information Retrieval Assessment Pipeline

### 3.3 Evaluation of Generation with Fixed Context

After defining the retrieval strategy, the generation stage was evaluated using the complete hybrid dataset of 192 questions, and the quality of the responses was assessed using the BERTScore metric, a semantic similarity metric based on contextual embeddings. The purpose of this stage was to evaluate how local LLMs behave when sent contexts containing a larger number of chunks in each inference. In this stage, the embedding model, qwen3embedding, was used, which showed better performance in the retrieval evaluation stage.

Four local LLMs were used, which were run on a machine equipped with an Nvidia RTX 5090 GPU: MedGemma3:27B, Gemma3:27B, Gpt-oss:20b, and Mistral Small 3.1. The impact of a fixed context size was analyzed using five independent scenarios, in which all questions in the dataset were submitted to the LLM using 1, 3, 5, 10, and 50 chunks/pages, respectively, corresponding to the top positions in the retrieval ranking. The objective of this step was to evaluate how increasing the context size affects the quality of the generation.

### 3.4 Incremental Page Retrieval Technique

As a main methodological contribution, this work presents a systematic analysis of the Incremental Page Retrieval technique, whose objective is to

| Question | Answer |
|---|---|
| Qual o tempo médio de duração da higienização simples das mãos com água e sabonete, e qual o tempo médio da fricção antisséptica das mãos com preparações alcoólicas? (*What is the average duration of simple hand hygiene with soap and water, and what is the average time of antiseptic hand rubbing with alcohol-based preparations?*) | A higienização simples das mãos com água e sabonete deve durar em média 40-60 segundos; A fricção antisséptica das mãos dura em média 30 segundos. (*Simple hand hygiene with soap and water should last an average of 40-60 seconds; antiseptic hand rubbing lasts an average of 30 seconds.*) |
| Qual o tipo de acomodação para pacientes com tuberculose ativa? (*What type of accommodation is available for patients with active tuberculosis?*) | Pacientes com tuberculose ativa devem ser acomodados preferencialmente em quarto privativo, com portas fechadas e saída restrita. A exceção ocorre em casos de coorte, onde pacientes com a mesma patologia podem dividir o quarto, exceto se forem pacientes com suspeita ou confirmação de tuberculose resistente. (*Patients with active tuberculosis should preferably be accommodated in a private room with closed doors and restricted access. An exception is made in cohort cases, where patients with the same pathology may share a room, except if they are patients with suspected or confirmed drug-resistant tuberculosis.*) |

Table 1: Example questions and answers

mitigate the quality degradation observed with the increase in cumulative context. In this approach, each chunk/page is sent individually to the LLM, avoiding the cumulative sending of context.

In the experiments, the language models were allowed to incrementally retrieve data up to the chunk/page at position 50, sending a single chunk/page at a time to the LLM, following the retrieval ranking order. The process is interrupted as soon as the model produces a response considered valid or when a maximum iteration limit is reached.

To enable this decision, the prompt explicitly instructs the LLM to indicate that the information was not found whenever the answer is not contained in the provided documents. Thus, an answer is classified as invalid when the model signals the absence of the information in the retrieved context. The identification of these answers was carried out using a mechanism based on regular expressions, designed to capture different linguistic variations used by the LLM to express negation or non-existence of the information. The response patterns were empirically defined from the observed outputs, allowing for the handling of textual variability and ensuring consistent detection of invalid answers.

Thus, experiments were conducted allowing incremental recovery up to the chunk at position 50, evaluating the efficiency of the method in terms of:

- Number of chunks needed to generate a valid response;
- Quality of the generated response (BERTScore).

This step was evaluated by combining the four LLM models (MedGemma3:27B, gemma3:27B, Mistral, and gpt-oss:20b) with the four embedding models (embeddinggemma, qwen3embedding, graniteembedding, and intfloat/multilingual-e5-small), totaling multiple experimental scenarios. The objective was to compare the retrieval efficiency between different LLM × embedding combinations using simple and interpretable metrics associated with the distribution of chunks used. The flowchart of this methodology is presented in Figure 2.

## 4 Results and Discussion

This section presents the results obtained in the experiments described in the methodology, followed by a comparative analysis and discussion of their implications.

### 4.1 Evaluation of Embedding Models in Information Retrieval

The results of the comparative analysis of the four embedding models are presented in Table 2, considering the synthetic dataset of 74 questions, in which the relevant chunk is previously known. The evaluation was conducted using the MRR metric, the median MRR, and the frequency of perfect retrieval (MRR = 1).

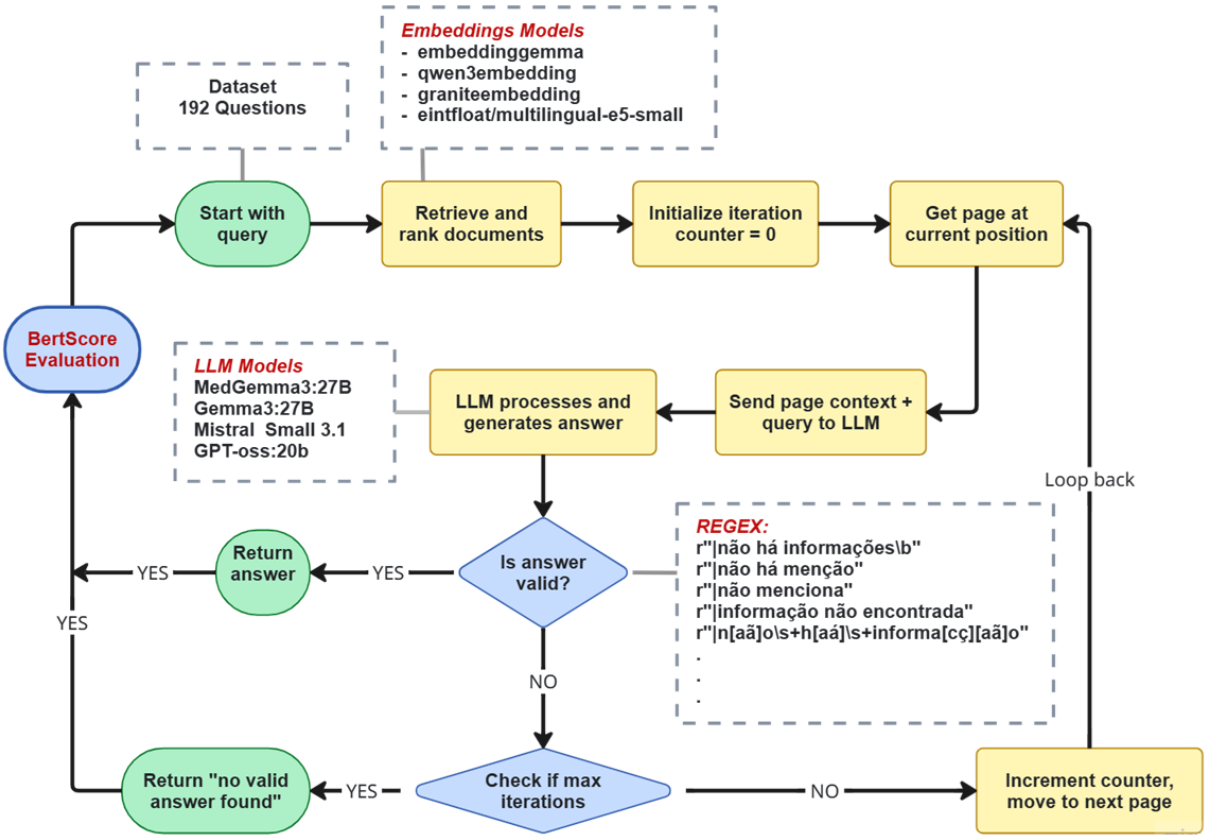


Figure 2: Flowchart of the proposed incremental retrieval pipeline

| Model | Average | Median | Freq MRR = 1 (%) |
|---|---|---|---|
| qwen3embedding | 0.8666 | 1 | 79.73 |
| embeddinggemma | 0.8348 | 1 | 72.97 |
| graniteembedding | 0.7695 | 1 | 66.22 |
| intfloat/multilingual | 0.6782 | 1 | 52.70 |

Table 2: MRR metric statistics by model

Among the models tested, qwen3embedding showed the best overall performance, with an average MRR of 0.8666 and a perfect recovery rate of 79.73%, indicating a greater ability to position the relevant chunk at the top of the ranking. In comparison, embeddinggemma showed intermediate performance, with an average MRR of 0.8348 and perfect recovery in 72.97% of queries. Graniteembedding and multilingual-e5-small obtained the lowest values, with average MRRs of 0.7695 and 0.6782, and perfect recovery rates of 66.22% and 57.70%.

Overall, the results indicate that the qwen3embedding and embeddinggemma models performed better, as the relevant chunk is retrieved in the first position in most queries.

### 4.2 Results of the Generation Evaluation with Fixed Context

This stage presents the results of the generation evaluation using a fixed context, in which a predefined number of retrieved chunks/pages, ranked by cosine similarity, is sent to the language model, regardless of the actual information needed to answer the question. The experiments were conducted in five scenarios, varying the number of chunks/pages sent to the model (1, 3, 5, 10, and the first 50 chunks/pages from the retrieval ranking). The quality of the responses was evaluated using the BERTScore (F1) metric.

The results are presented in Table 3 and show a consistent behavior among all the models evaluated: the quality of the generation progressively decreases as the context size increases.

| Model | 1 Page | 3 Pages | 5 Pages | 10 Pages | 50 Pages |
|---|---|---|---|---|---|
| MedGemma3:27B | 0.8058 | 0.7991 | 0.7694 | 0.6841 | 0.6575 |
| Gemma3:27B | 0.8020 | 0.7958 | 0.7704 | 0.6956 | 0.6634 |
| Gpt-oss:20b | 0.7674 | 0.7601 | 0.7539 | 0.7098 | 0.6401 |
| Mistral Small 3.1 | 0.7947 | 0.7820 | 0.7388 | 0.6726 | 0.6541 |

Table 3: BertScore(F1) results fixed context

The degradation becomes more pronounced start-

ing at 10 pages, and even more evident in the extreme scenario of 50 pages, where all models show a substantial drop in performance. In the case of gemma3:27B, for example, increasing the context from 1 to 50 chunks results in an approximate 17% reduction in the BERTScore. A similar trend is observed for the other models, indicating that this behavior is not specific to a particular architecture.

These results suggest the occurrence of a context dilution effect, in which the excessive inclusion of irrelevant information or information only partially related to the question hinders the model's ability to identify the truly useful sections for generating the answer. Although increasing the number of chunks/pages raises the probability that the correct section is present in the context, this benefit is outweighed by the negative impact of information overload.

Thus, the results of this stage provide a clear empirical justification for the proposed adaptive context control strategy presented in this work, which aims to balance informational coverage and generation quality, avoiding the unnecessary sending of large volumes of context to the language model.

### 4.3 Results of Incremental Page Retrieval

This subchapter presents and discusses the results of the Incremental Page Retrieval strategy, proposed in this work as an alternative to fixed context delivery. To better summarize and understand the results obtained, Table 4 is presented, comparing different combinations of LLMs and embeddings, showing performance in generation and retrieval quality metrics. The BERTScore (F1), percentages of responses found with 1, up to 3, and up to 5 chunks, as well as the maximum retrieval ranking required to obtain a valid response, are presented. Finally, the percentage of cases not found indicates the coverage and efficiency of each configuration.

The results show that the vast majority of questions were answered using only the first chunk/page retrieved, regardless of the language model or embedding employed. In the case of Gpt-oss:20b, the proportion of questions answered with only one chunk ranged from 79.69% to 91.15%, with qwen3embedding showing the highest percentage. Similar behavior was observed for gemma3:27B, whose values ranged from 77.08% to 90.10%, again with better performance associated with qwen3embedding. Mistral small 3.1 showed slightly lower, but still high, percentages, ranging from 70.31% to 84.38%.

However, there are cases that required the system to perform a deeper search in the ranking, for example, the pair (Mistral S3.1 + intfloat/multilingual) that went up to the chunk in position 50 to answer a question, suggesting that the incremental retrieval system is robust enough to recover information that the initial ranking places in unfavorable positions, which would be lost in a RAG system with a fixed and low $k$ (e.g., $k = 5$).

Another factor observed is that even when using the same embedding model and, therefore, the same retrieval ranking, the different LLM models arrive at the answer in distinct maximum chunk positions. This behavior indicates that, for certain questions, the models consider different contexts as sufficient throughout the ranking, despite the ordering of the retrieved documents being identical.

Table 4 also reports the average BERTScore (F1). The incremental retrieval strategy consistently maintains or improves response quality compared to fixed-context configurations using multiple chunks. Medgemma3:27B, for example, presented the best overall results, reaching 0.8124 with embeddinggemma and 0.8116 with qwen3embedding. As for most queries, the models considered that the first chunk contained sufficient information to generate a response, this reinforces the hypothesis that sending large volumes of context is unnecessary in most cases where there is a good information retrieval step.

Nevertheless, it is important to emphasize that a central aspect of the proposed methodology is the joint evaluation of the quality of generation and the behavior of incremental retrieval. By simultaneously analyzing the BERTScore of the generated responses and the distribution of the number of chunks needed to answer each question, it becomes possible to identify scenarios where the model prematurely interrupts retrieval, incorrectly assuming that the relevant information has been found. In these cases, although the chunk distribution indicates high efficiency, often with responses generated from the first chunk, the BERTScore shows semantic degradation in relation to the answer key, characterizing inaccurate or misleading responses.

This behavior can be observed in the Gpt-oss:20b model, which presents a favorable distribution of the number of chunks used, with a high frequency of responses generated from the first chunks in the ranking. However, this behavior is accompanied by lower BERTScore values than the other models evaluated.

Table 4: Average performance results of LLMs considering different embedding models

| Model (LLM + Embedding) | BertScore(F1) | % 1 Chunk | % ≤ 3 Chunks | % ≤ 5 Chunks | Chunk Máx | % Not Found |
|---|---|---|---|---|---|---|
| MedGemma 3 27B + Emb Gemma | 0.8124 | 87.50% | 94.79% | 95.31% | 19º Position | 1.04% |
| MedGemma 3 27B + Qwen 3 8B | 0.8116 | 87.50% | 96.35% | 97.39% | 21º Position | 1.04% |
| MedGemma 3 27B + Granite 278M | 0.8009 | 83.33% | 94.26% | 95.30% | 26º Position | 0.00% |
| MedGemma 3 27B + intfloat/multilingual | 0.7916 | 76.56% | 89.59% | 94.79% | 21º Position | 0.52% |
| Gemma 3 27B + Emb Gemma | 0.8087 | 88.54% | 95.31% | 96.35% | 19º Position | 0.00% |
| Gemma 3 27B + Qwen 3 8B | 0.8090 | 90.10% | 94.79% | 96.87% | 47º Position | 0.00% |
| Gemma 3 27B + Granite 278M | 0.7942 | 86.46% | 94.27% | 95.83% | 26º Position | 0.52% |
| Gemma 3 27B + intfloat/multilingual | 0.7924 | 77.08% | 87.50% | 92.19% | 44º Position | 0.52% |
| GPT OSS 20B + Emb Gemma | 0.7678 | 87.50% | 95.83% | 96.35% | 16º Position | 0.00% |
| GPT OSS 20B + Qwen 3 8B | 0.7635 | 91.15% | 96.88% | 98.96% | 7º Position | 0.00% |
| GPT OSS 20B + Granite 278M | 0.7489 | 86.46% | 95.31% | 96.88% | 17º Position | 0.00% |
| GPT OSS 20B + intfloat/multilingual | 0.7520 | 79.69% | 93.24% | 97.40% | 38º Position | 0.00% |
| Mistral S3.1 + Emb Gemma | 0.7988 | 82.81% | 94.27% | 95.31% | 16º Position | 0.00% |
| Mistral S3.1 + Qwen 3 8B | 0.7970 | 84.38% | 93.75% | 96.35% | 35º Position | 0.00% |
| Mistral S3.1 + Granite 278M | 0.7880 | 79.69% | 91.67% | 92.71% | 26º Position | 0.52% |
| Mistral S3.1 + intfloat/multilingual | 0.7887 | 70.31% | 85.94% | 90.62% | 50º Position | 0.52% |

To corroborate this observation, a two-dimensional analysis is presented in Figure 3, which reveals that efficiency in incremental retrieval is not linearly linked to semantic quality. While the Gemma:3 27B model demonstrates superiority on both axes, it is observed that gpt-oss:20b, despite reaching the stopping criterion (valid response) early in 91.15% of cases, presents an average BERTScore 5.6% lower than MedGemma:3 27B under the same embedding conditions. This behavior indicates that the validation of a 'valid response' in the incremental process can be achieved by simpler models with less precise content, reinforcing the importance of evaluating the generated responses with metrics such as the BertScore for quality control in RAG systems with dynamic retrieval.

Finally, the analysis of the results reinforces the limitations of traditional approaches based on fixed top-K and positions Incremental Page Retrieval as an effective adaptive strategy, particularly suitable for local LLM scenarios where the retrieval step is not as efficient and the chunk with the context of the correct answer is located further down in the ranking.

## 5 Conclusions and Future Work

This work presented the development and evaluation of a RAG-based chatbot using exclusively local LLMs, applied to institutional documents of a university hospital in Portuguese. The results showed that, although information retrieval has high performance, the indiscriminate increase in context sent to the model compromises the quality of the generation, as indicated by the consistent reduction in the BERTScore. This behavior reinforces the need for strategies that dynamically control the use of context in RAG systems, especially in sensitive domains such as healthcare.

As a main contribution, the incremental page retrieval technique was proposed and evaluated, in which chunks are sent individually to the LLM following the retrieval ranking, interrupting the process as soon as a valid answer is obtained. The results demonstrate that most questions can be answered with only the first or second chunk/page, significantly reducing the volume of context, the number of model calls, and the computational cost, without compromising the semantic quality of the answers. The joint analysis of the BERTScore and the chunk distribution proved fundamental in identifying hallucination behaviors and qualitative differences between the evaluated models.

Another noteworthy factor was the improvement in BertScore(F1) results after applying the incremental page retrieval technique, with particular emphasis on the performance of the MedGemma:3 27B model in generating responses, which proved superior to the other models. This can be attributed to the fact that its training was performed with text data and medical images. In addition to the empirical gain in response quality, the proposed strategy introduces flexibility to retrieve subsequent chunks when necessary, allowing the system to retrieve relevant information positioned deeper in the ranking, without incurring the cumulative context expansion observed in fixed-top-K approaches.

Although the results obtained are indicative of the behavior of the evaluated models, some limitations should be considered. Firstly, the retrieval

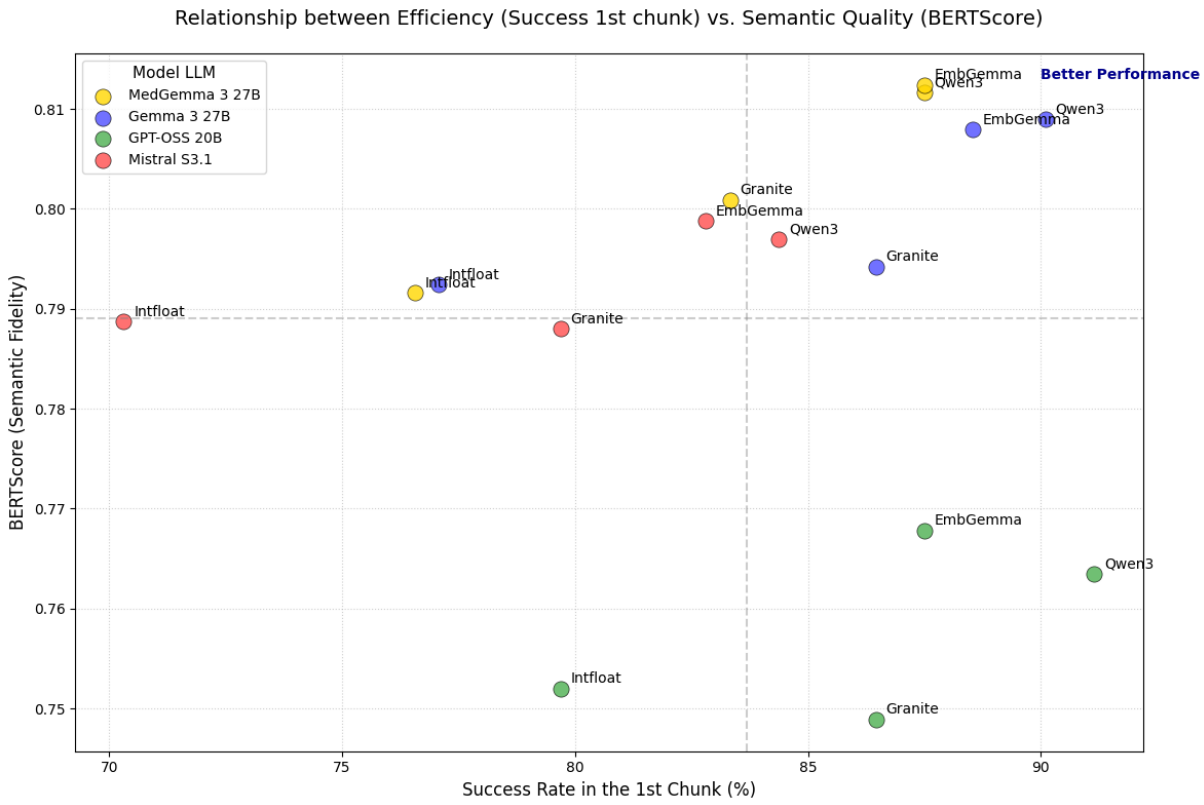


Figure 3: Scatter plot correlating First-Chunk Success and BERTScore for each pair.

assessment is based on a synthetic dataset composed of 74 questions, which, despite allowing for the precise identification of the relevant chunk, may not fully reflect the linguistic and semantic diversity observed in real-world usage scenarios. Similarly, context progression strategies, in which multiple chunks are incrementally accumulated to compose a more complete response, were not evaluated. Therefore, questions whose answer explicitly depends on the combination of information distributed across more than one chunk may not be fully covered by the adopted method.

Future work will involve investigating the progressive context retrieval technique, in which the number of chunks is expanded only when the model explicitly indicates the absence of relevant information, seeking to mitigate cases where the response depends on multiple complementary chunks. Additionally, a qualitative evaluation with hospital end-users, including healthcare professionals, is planned to analyze aspects such as usefulness, clarity, reliability, and adequacy of responses in real-world use. Finally, future extensions include the analysis of other LLM models and embeddings, as well as the integration of automatic mechanisms for detecting uncertainty and hallucination.

## Limitations

Although hybrid retrieval approaches such as BM25-based methods, rerankers, and commercial LLM baselines have shown strong performance in RAG systems, these configurations were not explored in the present study due to the scope and space limitations of this work. The primary focus of this paper is the evaluation of RAG pipelines using locally deployed LLMs in healthcare-related datasets, motivated by privacy and data protection constraints.

Experiments involving hybrid retrieval strategies, BM25-based ranking, rerankers, and comparisons with commercial LLMs were previously investigated in our earlier study (da Cunha et al., 2025), where a broader benchmarking of RAG configurations was conducted. Therefore, the present work complements that study by focusing specifically on privacy-preserving RAG architectures using local LLMs in sensitive data environments.

## Acknowledgments

This work was supported by Instituto Federal do Rio Grande do Sul (IFRS), Empresa Brasileira e Serviços Hospitalares (EBSERH) and Hospital Es-

cola da UFPEL. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We would like to thank the FAPERGS - Brasil for Financial Support, Award Agreement 22/2551-0000598-5. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

## References

Mohamed Abo El-Enen, Sally Saad, and Taymoor Nazmy. 2025. A survey on retrieval-augmentation generation (rag) models for healthcare applications. *Neural Computing and Applications*, 37(33):28191–28267.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLoS Digital Health*, 4(6).

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Self-reflective retrieval augmented generation. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, and 3 others. 2025. Granite embedding models. *Preprint*, arXiv:2502.20204.

David Baur, Jörg Ansorg, Christoph-Eckhard Heyde, and Anna Voelker. 2025. Development and evaluation of a retrieval-augmented generation chatbot for orthopedic and trauma surgery patient education: Mixed-methods study. *JMIR AI*, 4:e75262.

Murilo Vargas da Cunha, Marilia Rosa Silveira, Brenda Salenave Santana, Larissa Astrogildo Freitas, and Ulisses Brisolara Corrêa. 2025. Optimizing and evaluating a retrieval-augmented generation system for normative document retrieval in hospital settings. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 385–393. SBC.

Empresa Brasileira de Serviços Hospitalares (EBSERH). 2025. Estrutura administrativa — hu-ufsc. https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-sul/hu-ufsc/governanca/estrutura-administrativa. Acesso em: 13 jul. 2025.

Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *NPJ digital medicine*, 7(1):183.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):38.

Agrim Kulshreshtha, Aditya Choudhary, Tejas Taneja, and Seema Verma. 2025. Enhancing healthcare accessibility: A rag-based medical chatbot using transformer models. In *2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, pages 1–4. IEEE.

Peter Lee, Carey Goldberg, and Isaac Kohane. 2023. *The AI revolution in medicine: GPT-4 and beyond*. Pearson.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353.

Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. 2023. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6):e333–e335.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333.

Avhishek Nandi, Barnali Paul, Piyali Datta, and Deepsubhra Guha Roy. 2024. Medmate: A contextual approach for disease diagnosis using retrieval-augmented generation. In *International Conference on Recent Advances in Artificial Intelligence & Smart Applications*, pages 429–441. Springer.

Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. 2025. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9):226.

Karen Ka Yan Ng, Izuki Matsuba, and Peter Chengming Zhang. 2024. Rag in health care: A novel framework for improving communication and decision-making by addressing llm limitations. *NEJM AI*.

S. S. T. Oliveira, D. Fazzioni, and D. O. C. Ferreira. 2025. Grandes modelos de linguagem. In *In: Kudo, T. N. et al.* Cegraf UFG, Goiânia. E-book (254 p.). ISBN 978-85-495-1096-9.

G. Perković, A. Drobnjak, and I. Botički. 2024. Hallucinations in llms: Understanding and addressing challenges. In *Proceedings of the 47th MIPRO ICT and Electronics Convention (MIPRO 2024)*, pages 2084–2088, Opatija, Croatia. IEEE.

Kushagra Sinha, Vaibhav Singh, Ankit Vishnoi, Parul Madan, and Yadvendra Shukla. 2024. Healthcare diagnostic rag-based chatbot triage enabled by biomistral-7b. In *2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)*, pages 333–338. IEEE.

Namrye Son, Inchul Kang, Inhu Kim, Keehyuck Lee, Sejin Nam, and Donghyoung Lee. 2025. Development and evaluation of a retrieval-augmented generation-based electronic medical record chatbot system. *Healthcare Informatics Research*, 31(3):218–225.

Chihiro Taguchi, Seiji Maekawa, and Nikita Bhutani. 2025. Efficient context selection for long-context qa: No tuning, no iteration, just adaptive-$k$. *arXiv preprint arXiv:2506.08479*.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *Preprint*, arXiv:2509.20354.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Dongfang Zhang, Haoze Du, Xiaolei Wang, Mingdong Zhu, Xiaoxiao Pang, Dongqing Wei, and Xianfang Wang. 2025a. Cmedragbot: A chinese medical chatbot based on graph rag and large language models. *Interdisciplinary Sciences: Computational Life Sciences*, pages 1–16.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.