Optimizing and Evaluating a Retrieval-Augmented Generation System for Normative Document Retrieval in Hospital Settings

Murilo Vargas da Cunha
Federal University of Pelotas (UFPEL)
Pelotas, RS, Brazil
Federal Institute of Rio Grande do Sul (IFRS)
Rio Grande, RS, Brazil
mvcunha@inf.ufpel.edu.br

Marilia Rosa Silveira
Federal University of Pelotas (UFPEL)
Pelotas, RS, Brazil
mrsilveira@inf.ufpel.edu.br

Brenda Salenave Santana
Federal University of Pelotas (UFPEL)
Pelotas, RS, Brazil
bssalenave@inf.ufpel.edu.br

Larissa Astrogildo Freitas
Federal University of Pelotas (UFPEL)
Pelotas, RS, Brazil
larissa@inf.ufpel.edu.br

Ulisses Brisolara Corrêa
Federal University of Pelotas (UFPEL)
Pelotas, RS, Brazil
ulisses@inf.ufpel.edu.br

ABSTRACT

This paper presents the development and evaluation of a chatbot designed to consult documents written in Portuguese on regulatory procedures in a hospital environment, which uses a Retrieval-Augmented Generation (RAG) pipeline to increase the factual accuracy and relevance of its underlying Large Language Model (LLM). Using the RAG technique will allow for more efficient and accurate retrieval of information contained in hospital manuals and institutional documents, helping workers quickly access internal guidelines and procedures. The objective is to optimize each system component (retrieval, re-ranking, and generation) to analyze the impact of each step in developing a RAG system for a low-resource language such as Portuguese. The methodology can be divided into the following stages: (1) the Golden Set Preparation, formed by a set of questions and answer data; (2) comparison of three embedding models for initial retrieval and of three re-ranking methods, including Cross-Encoder, Reciprocal Rank Fusion (RRF), and an LLM-based re-ranker, using metrics such as MRR, NDCG@10; and (3) comparison of two generative models (Gemini 1.5 Flash and GPT-4o-mini), using the metric BERTScore. The results indicate that the intfloat/multilingual-e5-small embedding model minimizes retrieval failures. In the re-ranking stage, the LLM-based re-ranker achieved the highest ranking accuracy, yet the computationally lighter RRF method emerged as an excellent cost-effective alternative. We conclude that an optimized architecture for both efficiency and performance combines the intfloat embedding, the RRF re-ranker, and the Gemini generator.

KEYWORDS

Chatbot, Retrieval-Augmented Generation, Low-resource languages, Large language models

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

1 INTRODUCTION

The recent advancement of large language models (LLMs) has transformed the way we interact with textual information, enabling fast and accurate responses to complex questions. This evolution has expanded the potential use of large document repositories in tasks such as legal and academic research, technical support, and the consultation of institutional regulations [6, 20]. In this context, traditional information retrieval techniques, previously employed primarily in search engines, have been integrated with generative models, resulting in more robust and effective architectures for document reading and comprehension [23].

However, despite the advancements in LLMs, there are still significant limitations that affect their reliability, such as hallucinations [13, 22]. Furthermore, these models operate with static knowledge, restricted to the data available during training, making them unsuitable for constantly evolving contexts. Their inability to perform accurately in highly specialized domains, such as the medical or hospital sectors, combined with the high computational cost associated with model retraining for knowledge updates, presents substantial barriers to these systems' safe and scalable adoption.

One of the most promising approaches in this context is Retrieval-Augmented Generation (RAG), which enhances response generation by incorporating information from relevant external sources [4, 28]. This mechanism helps mitigate limitations related to memory constraints or outdated knowledge by enabling dynamic queries to external knowledge repositories.

A RAG architecture is generally composed of two main modules widely recognized in the literature [8, 14]:

- Retrieval Component: Part of the system responsible for retrieving relevant information from a large corpus of documents or knowledge sources.
- Generation Component: Part responsible for producing a natural language response based on previously retrieved information.

In this context, Fan et al. [11] highlights that the retrieval component can have two types of retrievers: sparse and dense. According to the authors, sparse retrievers (such as TF-IDF) rely on keywords

and are simpler, while dense retrievers (such as BERT) use meaning (semantics) to allow for more flexible search. Information is usually retrieved in fragments (chunks), as this approach offers the best balance between efficiency and relevance [17]. Regarding the generation component, if the model is white-box (open-source), it can be fine-tuned and trained with specific data. If it is black-box (proprietary), optimization is performed through prompt engineering, injecting the retrieved context directly into the query to guide the response generation.

The generation step depends on the retrieval component, as inadequate context may lead to incorrect final answers. Therefore, this study emphasizes the importance of evaluating the retrieval stage within an RAG system. Moreover, the literature highlights a predominance of RAG systems developed for the English language, revealing that the architecture still faces significant challenges in low-resource languages such as Portuguese [3, 7]. These challenges include the scarcity of high-quality embeddings trained in Portuguese, the limited availability of public benchmarks and datasets, and linguistic particularities affecting retrieval and generation processes.

In light of this, the present work proposes a systematic evaluation of an RAG system applied to institutional documents in Portuguese, aiming to understand the impact of technical choices on overall performance and to contribute practical evidence to support the adoption of this approach in specialized domains and underrepresented languages.

To this end, we develop an RAG system for querying information related to procedures and tasks within a hospital setting. A Portuguese question-answering dataset was created through human curation, based on normative documents such as manuals and Standard Operating Procedures (SOPs), which describe routines and responsibilities across hospital departments. This study focuses on the quantitative analysis of the components involved in the retrieval stage, intending to inform the development of more effective solutions in domains with high specialization and limited linguistic resources.

This paper is organized as follows: Section 2 presents the related works. Section 3 describes the proposed methodology, divided into four main stages. Section 4 presents and discusses the obtained results. Finally, Section 5 provides the conclusions and outlines directions for future work.

2 RELATED WORKS

The literature on RAG is predominantly focused on the English language. In this section, we review studies that explore the application of RAG and conversational systems in low-resource languages, with particular attention to the use of different evaluation metrics.

An important step in the development of RAG systems is the quantitative evaluation of information retrieval. For instance, the study by [21] employs Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) to assess the overall system performance. The work of Wijaya and Purwarianti [26], which focuses on a tutoring system for programming, uses only the MAP metric to evaluate the impact of different chunk sizes on retrieval performance. Similarly, the study by Nai et al. [19] adopts the NDCG@10 (Normalized Discounted Cumulative Gain) metric to conduct an extensive comparison of retrieval effectiveness across multiple embedding models, including BM25, Cohere, and various OpenAI models. A good metric for evaluating ranked results should be able to calculate how many relevant documents were retrieved and how close they are to the top of the ranking. The MRR metric used in the studies, which calculates the inverse of the position of the first relevant document, is a good example. This is also true in some cases where it's important to differentiate a highly relevant document from marginally relevant documents. In these cases, the NDCG can be used, which is based on the idea that highly relevant documents are more valuable than marginally relevant ones, and both are more valuable if they appear at the top of the ranking.

Despite the diversity of metrics and analytical focuses, a common point among these studies is the absence of a re-ranking step. The re-ranking mechanism is used to refine the results of the information retrieval step by leveraging dense models or specialized classifiers, with the goal of improving the selection of the most relevant documents to be submitted for final answer generation. Studies such as Yang et al. [27], Magar et al. [16], and Shen et al. [24] demonstrate that incorporating a re-ranker can substantially improve the accuracy of generated responses by ensuring that only the most relevant passages are included in the context provided to the generative model.

In Dieu et al. [9], re-ranking is performed using a BERT cross-encoder model, which evaluates and prioritizes the chunks before the answer generation step. However, the study does not quantitatively evaluate this effect on answer generation metrics. For example, the study by Alonso et al. [2] introduces MedExpQA, a multilingual benchmark for evaluating language models in Medical Question Answering (Medical QA) tasks. The system applies the Reciprocal Rank Fusion (RRF) technique to combine and re-rank the lists of candidates retrieved by multiple retrievers (BM25 and MedCPT). Nevertheless, the authors do not assess the specific impact of this technique on retrieval quality, which limits our ability to understand the direct effect of this method on the quality of generated answers.

Similarly, the study by Alshammary et al. [3], which proposes a Question Answering (QA) system for the Arabic language, also incorporates a filtering and fact-checking process that functions as an implicit re-ranking mechanism, selecting the most relevant documents ("Gold Hadiths") for answer generation. However, the study does not provide a metric-based analysis of the effectiveness of this refinement step.

On the other hand, some studies align more directly with our research proposal by employing re-ranking techniques and quantitatively evaluating their effects on both the retrieval and generation stages. The studies by Aguzzi et al. [1] and Hernandez-Salinas et al. [12] implement an architecture based on an Ensemble Retriever, which combines different retrieval strategies to obtain relevant documents. Both works integrate the results of a sparse search (BM25) with those of a dense vector search using Reciprocal Rank Fusion (RRF) to re-rank the final list of documents. The effectiveness of this approach is evaluated in both cases using the RAGAS framework, focusing on metrics such as Context Precision and Context Recall.

These studies highlight the growing importance of re-ranking as an optimization step within the RAG pipeline, as well as the

relevance of evaluating the retrieval stage to validate system performance. It is important to note, however, that a gap remains in the application and evaluation of such systems within the specific domain of hospital procedures, a niche that the present work aims to explore.

3 METHODOLOGY

Our methodology can be broadly divided into the following stages: Golden Set Preparation, Embedding Evaluation, Re-ranking Evaluation, and Answer Generation Evaluation. The flowchart presented in Figure 1 illustrates the basic pipeline of the implemented methodology for the RAG system.

3.1 Golden Set Preparation

In the Golden Set preparation phase, internal normative 76 documents were selected from four specific hospital departments: the Occupational Health and Worker Safety Unit, the Human Resources Division, the Hospital Infection Control Service, and the Teaching and Research Management Department [10]. Based on these documents, 90 question–answer pairs were automatically generated using the Gemini 1.5 Flash model, drawing from chunks of up to 1000 tokens from the original texts. The purpose of using document segments for question generation was to produce more focused and less generic questions, as well as to increase the number of questions per document.

Following manual review, 74 question–answer pairs were deemed valid and used as the foundation for all subsequent evaluation stages. The Table 1 shows 3 examples of questions and answers generated.

After generating the golden set, the documents needed to be divided into smaller segments to facilitate representation during the retrieval stage. This step involved splitting the documents into chunks of approximately 100 tokens, with the cut defined to occur at the next full stop in order to preserve the semantics and syntax of each chunk [15]. This strategy proved to be the best among others tested, such as creating chunks per line or page, because 100 tokens represent approximately a paragraph, which is typically the amount of text required to obtain a response for this type of document used in the experiment. To enrich the information available to the language model, the full text of the page from which each chunk originated was also stored in a dedicated column in the database.

The next step was to manually identify the chunks containing the answers to the 74 golden questions. This way, each question was linked to the chunk ID with the corresponding correct answer in the retrieval vector database, ensuring the system could accurately identify the related content.

3.2 Embedding Evaluation

In this stage, three embedding models were tested to convert both text chunks and questions into high-dimensional vector representations. The evaluated models were: all-MiniLM-L6-v2, paraphrase-multilingual-MiniLM-L12-v2, and intfloat/multilingual-e5-small. This transformation is essential for storing and performing semantic searches. To retrieve information, cosine similarity was employed to calculate the similarity between the question embedding and the embeddings of the document chunks stored in the database [25]. The quality of retrieval was measured using two key ranking metrics: Mean Reciprocal Rank (MRR), which evaluates the position of the first correct result, and Normalized Discounted Cumulative Gain at 10 (NDCG10), which assesses the overall ranking quality of the top-10 results.

In this step, for each embedding model evaluation, chunks of approximately 100 tokens were generated and stored in the database along with their vector representations. Afterward, each model was subjected to an evaluation of the 74 golden questions, where the cosine similarity between the question and chunks was measured, resulting in the selection of the 100 best-ranked chunks. Finally, the position of the chunk database ID with the response was evaluated using MRR and NDCG@10 analysis. The evaluation pipeline is shown in Figure 2.

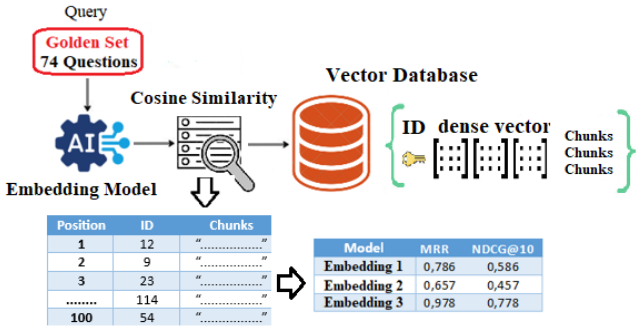

Figure 2: Pipeline Embedding Evaluation.

3.3 Re-ranking Evaluation

Considering that using 100 chunks in the context results in high token consumption, it is essential to adopt a re-ranking mechanism capable of refining the prioritization of the most relevant chunks. This approach makes it possible to limit the number of chunks submitted to the LLM, while increasing the likelihood of including the chunk containing the correct answer in the final context.

In the reclassification phase [18], the three embedding models were tested for initial retrieval, but with the application of the re-ranking model, the results were very similar. Therefore, this article presents only the results with the best-performing embedding model from the previous step. To assess the impact of re-ranking on retrieving the correct document for a given question, three methods were compared for re-ranking the top 100 initially retrieved chunks. In this phase, MRR and NDCG@10 were again used as evaluation metrics.

- Cross-Encoder, which processes the question and chunk jointly to perform a deeper relevance analysis;
- Reciprocal Rank Fusion (RRF), a hybrid approach that combines rankings from semantic and lexical searches (BM25);
- LLM-based Re-ranker, which leverages Gemini to re-rank candidate chunks through a structured prompt.

In the same way as the previous stage, the three re-ranking models were submitted to the 74 gold questions to observe which one presented the best positioning of the chunks with the answer to each question. A pipeline for this evaluation step is shown in the figure 3.

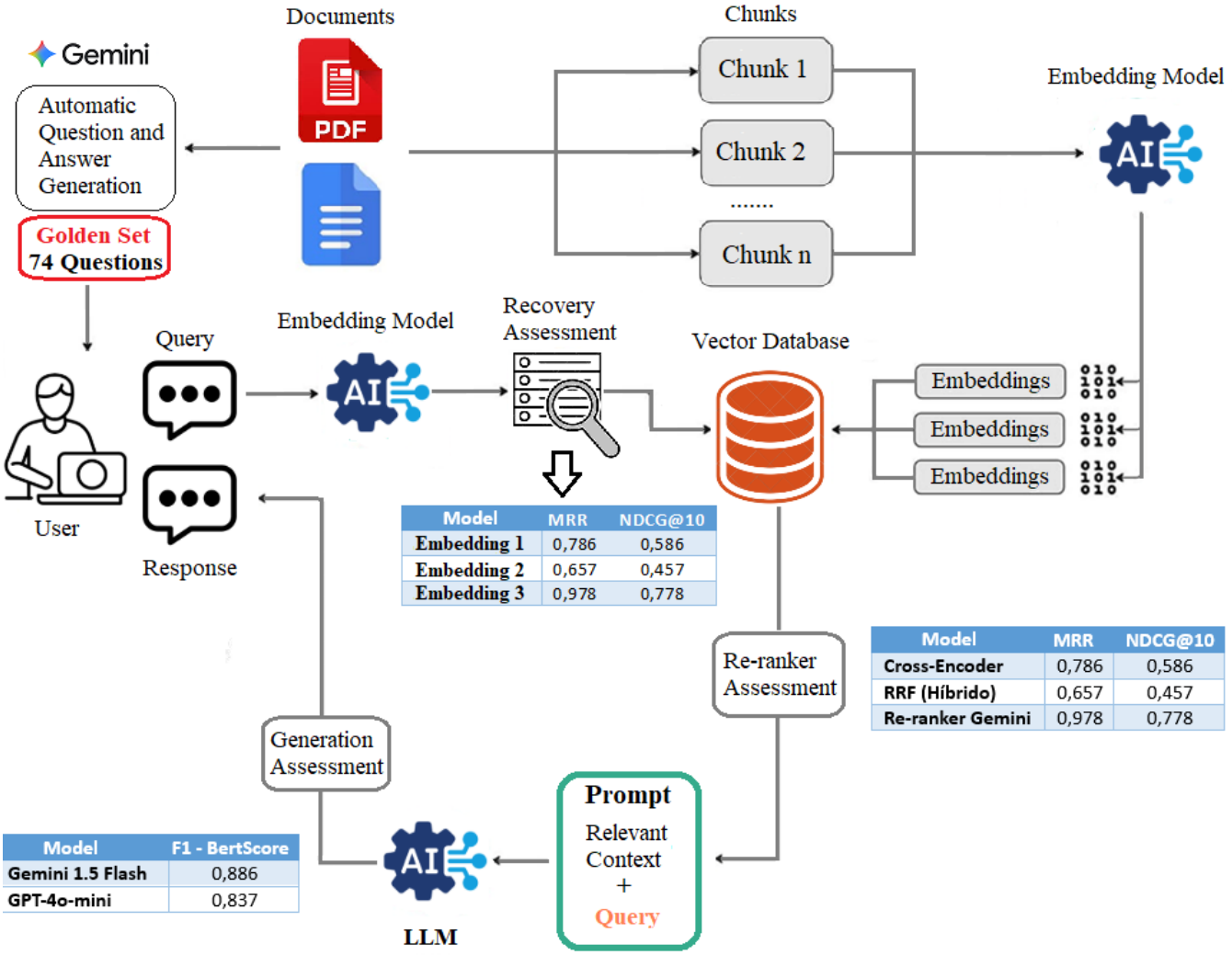

Figure 1: Methodology Flowchart. Diagram of the RAG pipeline architecture, highlighting data flow, evaluation stages, and performance metrics for each component.

3.4 Answer Generation Evaluation

The final phase focused on the actual generation of responses, based on the retrieved and re-ranked documents. At this stage, for each of the 74 questions, the 50 best-ranked chunks after applying the RRF re-ranking model were selected and sent along with a guidance prompt for the LLM to answer the question. To this end, we evaluated the quality of the responses by comparing the performance of two state-of-the-art language models: Gemini 1.5 Flash and GPT-4o-mi. The final answers were assessed by calculating the semantic similarity between the system-generated response and the reference answer using BERTScore (F1-Score) [5].

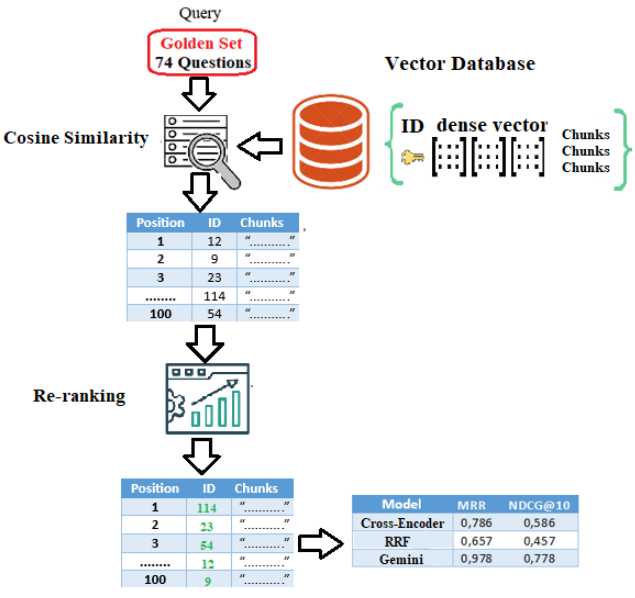

Figure 3: Pipeline Embedding Evaluation.

Table 1: Example questions and answers

Question	Answer
Quais escalas são utilizadas para avaliar o risco de quedas em pacientes adultos e crianças, respectivamente, e como são identificados os pacientes adultos com risco alto de quedas? *(What scales are used to assess the risk of falls in adult and child patients, respectively, and how are adult patients at high risk of falls identified?)*	Para crianças é utilizada a Escala Humpty Dumpty e para os adultos a Morse Fall Scale (MSF). Os pacientes adultos, que apresentem Risco Alto para quedas são identificados com uma pulseira amarela. *(The Humpty Dumpty Scale is used for children and the Morse Fall Scale (MSF) for adults. Adult patients at high risk for falls are identified with a yellow bracelet.)*
De acordo com o documento, qual a solução recomendada para a nebulização em caso de coleta de escarro induzido, especificando a composição e advertindo sobre uma possível substância a ser evitada? *(According to the document, what is the recommended solution for nebulization in case of induced sputum collection, specifying the composition and warning about a possible substance to be avoided?)*	Preparar a solução salina 3% (5 ml de Soro Fisiológico 0,9% + 0,5 ml de NaCl 20%) e realizar nebulização durante 15 minutos. Não utilizar solução preparada com água destilada e NaCl devido ao risco de broncoespasmo. *(Prepare the 3% saline solution (5 ml of 0.9% Saline + 0.5 ml of 20% NaCl) and nebulize for 15 minutes. Do not use a solution prepared with distilled water and NaCl due to the risk of bronchospasm.)*
Quais são os Equipamentos de Proteção Individual (EPIs) obrigatórios para manipulação de quimioterápicos, especificando o tipo de luva e o número de pares necessários? *(What are the mandatory Personal Protective Equipment (PPE) for handling chemotherapy drugs, specifying the type of glove and the number of pairs required?)*	Avental cirúrgico com proteção contra produtos químicos; Luvas para proteção química (exceto durante uso de soluções como álcool, clorexidina e outras de atividades médicas assistenciais); Luvas de procedimento (durante atendimento com uso de soluções como álcool, clorexidina, administração de quimioterápicos e outras atividades médicas assistenciais, exceto para higienização de mãos); Luvas cirúrgicas (dois pares). *(Surgical apron with protection against chemicals; Gloves for chemical protection (except during the use of solutions such as alcohol, chlorhexidine and other medical assistance activities); Procedure gloves (during care using solutions such as alcohol, chlorhexidine, administration of chemotherapy and other medical assistance activities, except for hand hygiene); Surgical gloves (two pairs))*

4 RESULTS AND DISCUSSION

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

4.1 Embedding Model Performance in Initial Retrieval

The evaluation of the initial retrieval stage is critical, as a failure to retrieve the correct document at this point prevents the system from providing an accurate answer—regardless of the effectiveness of subsequent stages. The comparative analysis of the three embedding models revealed a substantial disparity in their robustness. The intfloat/multilingual-e5-small model demonstrated clear superiority, failing to retrieve the correct chunk for only 1 out of the 74 questions (1.4%). In contrast, the paraphrase-multilingual-MiniLM-L12-v2 model showed a failure rate of 8.1% (6 questions), while the baseline model, all-MiniLM-L6-v2, failed in 9.5% of the cases (7 questions).

This superiority was further supported by the quantitative analysis of the Mean Reciprocal Rank (MRR) metric, which evaluates the position of the first correct result. As shown in Table 2, intfloat/multilingual-e5-small achieved a mean MRR of 0.6662, considerably higher than the scores of the other two models, both around 0.52. Even more telling is the median value, which was 1.0 for intfloat/multilingual-e5-small, indicating that in more than half of the queries, the correct answer was ranked at the very top.

Table 2: MRR metric statistics by model

Model	Average	Median	Freq MRR = 1 (%)
all-MiniLM-L6-v2	0.5225	0.5	40.5405
paraphrase	0.5200	0.5	40.5405
infloat	**0.6662**	**1.0**	**58.1081**

Beyond the position of the first correct result, the overall ranking quality was assessed using the NDCG@10 metric. The results of this analysis, presented in Table 3, further reinforce the advantage of the intfloat/multilingual-e5-small model, which achieved a mean NDCG@10 score of 0.6782, outperforming both paraphrase-multilingual-MiniLM-L12-v2 (0.5625) and all-MiniLM-L6-v2 (0.5445).

Given these results, the choice of intfloat/multilingual-e5-small is justified not only by its ability to minimize critical retrieval failures but also by its greater effectiveness in positioning relevant documents at the top of the list, as demonstrated by the MRR and NDCG@10 metrics. Therefore, this model was selected as the basis for all subsequent evaluations.

Table 3: NDCG@10 metric statistics by model

Model	Average	Median	Freq NDCG@10 = 1 (%)
all-MiniLM-L6-v2	0.5445	0.6309	36.4865
paraphrase	0.5625	0.6220	37.8378
infloat	**0.6782**	**1.0000**	**52.7027**

4.2 Comparative Analysis of Re-ranking Methods

After defining the most effective embedding model, the research focused on optimizing the re-ranking step. In this phase, the 100 most relevant chunks retrieved by the initial search were reordered using three distinct methods: a Cross-Encoder model, the hybrid Reciprocal Rank Fusion (RRF) method, and a re-ranking approach with the Gemini LLM. The Mean Reciprocal Rank (MRR) metric analysis revealed a clear performance hierarchy, as shown in Table 4. The Gemini Re-ranker proved to be the most potent method, achieving the highest average MRR, with 0.8612, and the highest frequency of perfect hits (MRR=1), securing first place. The RRF method also demonstrated notably strong performance, with an average MRR of 0.7857.

Table 4: MRR metric statistics by reordering model

Model	Average	Median	Freq MRR = 1 (%)
CrossEncoder	0.6584	1.0	54.0541
RRF	0.7857	1.0	67.5676
Gemini	**0.8612**	**1.0**	**78.3784**

To provide a more granular analysis of per-query performance, the heatmap in Figure 4 visualizes the MRR scores for each of the 74 questions across the three evaluated models. Dark blue indicates a perfect score of 1.0, while lighter colors (green and yellow) represent progressively lower scores. This visualization reinforces the conclusions drawn from the statistical analysis.

The heatmap shows that the CrossEncoder model performs well on several questions, but in some columns, the MRR drops considerably (lighter tones), indicating more unstable results. The RRF model also performs consistently well on most questions, often dark (MRR=1), with slight fluctuation. The Gemini Re-ranker model, on the other hand, presents several dark columns, indicating good performance (MRR=1) on many questions.

Additionally, the quality of the ranking was further evaluated using the NDCG@10 metric. As shown in Table 5, the results confirm the superiority of the Gemini-based re-ranker, which achieved a mean NDCG@10 score of 0.8826. The RRF method followed with a strong performance, reaching a mean of 0.8157, once again outperforming the Cross-Encoder, which obtained a lower average of 0.6906.

The discussion of these results allows us to conclude that Re-ranker with the Gemini model is the technically superior choice for maximizing ranking accuracy, benefiting from LLM's advanced reasoning capabilities to capture contextual nuances. However, RRF's impressive performance is one of the most relevant findings of this study. Being a computationally lighter and more straightforward method, its ability to outperform a sophisticated approach like Cross-Encoder positions it as an excellent cost-effective alternative for practical applications, where factors such as latency and computational cost are critical.

Table 5: NDCG@10 metric statistics by reordering model

Model	Average	Median	Freq NDCG@10 = 1 (%)
CrossEncoder	0.6906	0.9599	50.0000
RRF	0.8157	1.0000	64.8649
Gemini	**0.8826**	**1.0000**	**75.6757**

4.3 Response Generation Quality Assessment

The final phase of the RAG system evaluation focused on the semantic quality of the response actually delivered to the user. This analysis was divided into two sequential investigations: first, the impact of the reordering method on response quality and, second, a direct comparative analysis between different state-of-the-art language models.

a) Impact of Reordering Method on Response Quality

A central question in RAG system architecture is whether the superiority of a reordering method directly translates into an improvement in the quality of the final response. To investigate this point, responses were generated with Gemini 1.5 Flash using the chunks from the two best reranking methods from the previous phase: the Gemini Reranker and the RRF method as context.

The results, measured by BERTScore (F1-Score), were remarkably close and revealing. The context provided by the RRF method resulted in an average F1-score of 0.8585, slightly higher than the average of 0.8527 obtained with the Gemini Re-ranker context. The results can be seen in Table 6.

Table 6: BERTScore F1 statistics for re-ranking models

Metric	RRF	Re-ranker Gemini
Average (BERTScore F1)	**0.8585**	0.8527
Median	**0.8672**	0.8619
Standard Deviation	**0.0676**	0.0803

The most important conclusion of this phase is that Re-ranker Gemini's superior accuracy in the ranking task did not imply a significant improvement in the quality of the final response. Both methods provided a good enough context for LLM to generate high-quality responses. These results suggest that the RRF method, being computationally more efficient and faster, presents itself as an excellent cost-benefit alternative for the final system architecture, without sacrificing the quality of the response perceived by the user.

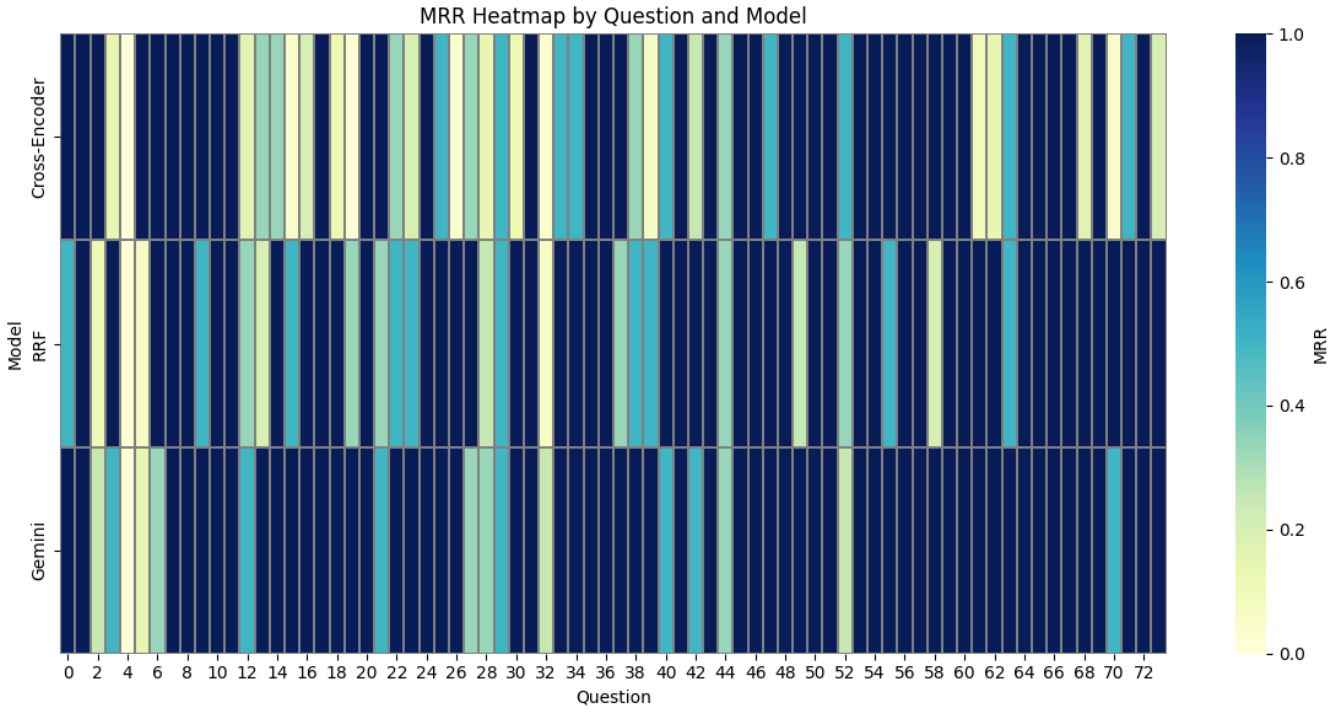

Figure 4: Mean Reciprocal Rank (MRR) heatmap by question and model.

b) Comparative Analysis of Generator Models

To enrich the analysis, the final step compared the performance of Gemini 1.5 Flash with another state-of-the-art model, gpt-4o-mini. To isolate the generation capability of each model, both received the same context, provided by the RRF reorderer.

Analysis of the results in Table 7 shows that, although both models performed very well, Gemini 1.5 Flash demonstrated a slight advantage. In addition to the superior averages, Gemini 1.5 Flash had a smaller standard deviation (0.0676 versus 0.0706), suggesting not only higher average quality answers but also more consistent and predictable performance across the entire question set.

Table 7: BERTScore F1 statistics for generation models

Metric	Gemini 1.5 Flash	GPT-4o-mini
Average (BERTScore F1)	**0.8585**	0.8376
Median	**0.8672**	0.8538
Standard Deviation	**0.0676**	0.0706

To further illustrate the comparison, the bar chart in Figure 5 visualizes the performance difference (Delta F1 Score) for each of the 74 questions. The green bars indicate the questions on which Gemini 1.5 Flash outperformed GPT-4o-mini (positive difference), while the red bars represent the cases where GPT-4o-mini achieved a higher score (negative difference). A visual inspection of the chart reveals that the number and magnitude of the green bars are substantially larger, corroborating the statistical data that Gemini 1.5 Flash achieved a superior average performance. It is therefore concluded that for the task of generating responses in this project, the Gemini 1.5 Flash proved to be marginally superior, establishing itself as the best performing option among the models evaluated.

5 CONCLUSIONS AND FUTURE WORK

The systematic evaluation of the RAG pipeline demonstrated the effectiveness of this methodological approach in optimizing the balance between response accuracy and system computational efficiency. Tests conducted across the three distinct stages—information retrieval, document reordering, and response generation demonstrated that granular analysis of each component was crucial for identifying the most cost-effective model combination.

Thus, we conclude that the intfloat/multilingual-e5-small embedding model is the ideal choice for the initial retrieval stage, minimizing critical recall errors. In the reordering stage, although the Gemini Re-ranker offers the highest accuracy, the RRF hybrid method emerges as a superior cost-benefit alternative, as its efficiency does not compromise the quality of the final generated response. Finally, in the generation stage, Gemini 1.5 Flash proved marginally superior to gpt-4o-mini, establishing itself as the best-performing and most consistent option among the models evaluated for this project. The final architecture choice, therefore, prioritizes efficiency, combining "intfloat/multilingual-e5-small" with the RRF re-ranker and the Gemini 1.5 Flash generator for a robust, accurate, and computationally feasible system.

For future work, several research possibilities emerge to expand and improve the system. One planned step is a qualitative evaluation

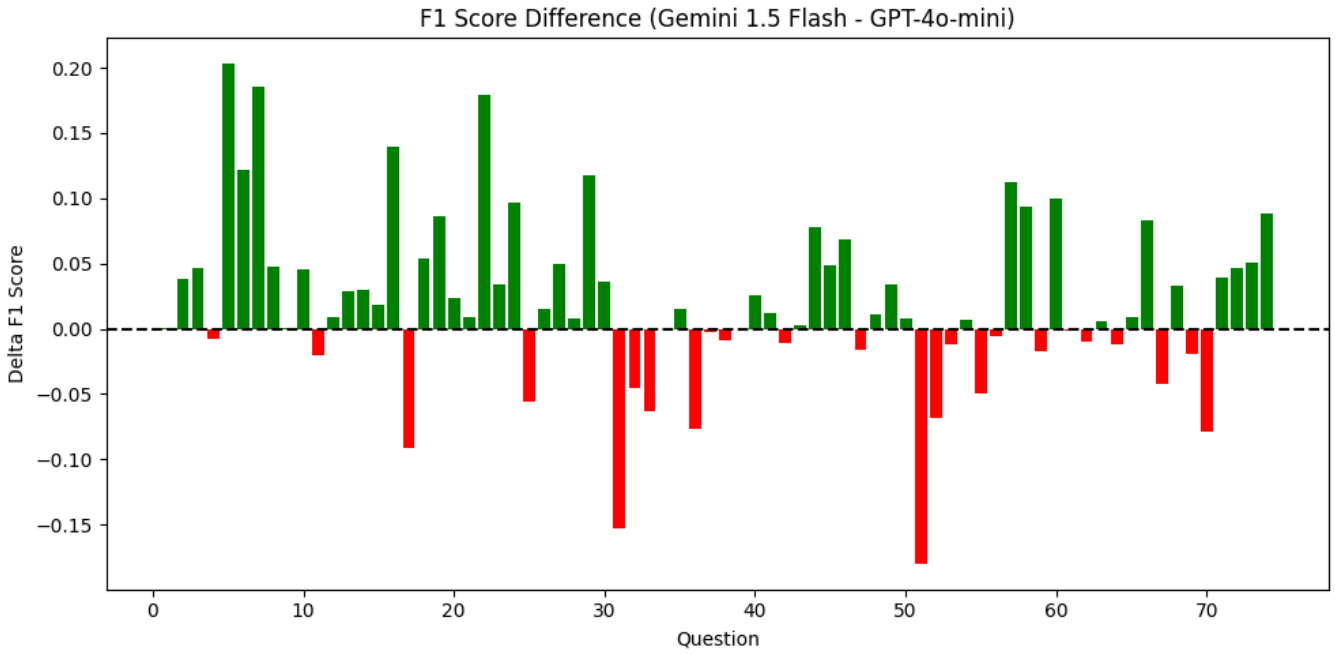

Figure 5: F1 Score Difference between Gemini 1.5 Flash and GPT-4o-mini by question id.

of the chatbot directly with hospital professionals. This interaction will allow for feedback on the usability of the interface, the contextual relevance of responses, and overall user satisfaction, validating the system's effectiveness in a real-world scenario. Additionally, the plan is to test the generation stage with lighter language models that can be executed locally. This approach aims to mitigate cost and latency issues and, crucially, ensure data privacy and security.

6 ACKNOWLEDGMENTS

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

[1] G. Aguzzi, M. Magnini, G. P. Salcuni, S. Ferretti, and S. Montagna. 2024. Applying Retrieval-Augmented Generation on Open LLMs for a Medical Chatbot Supporting Hypertensive Patients. In *Proc. of the 3rd AIxIA Workshop on Artificial Intelligence for Healthcare (HC@AIxIA 2024), co-located with the 23rd Int. Conf. of the Italian Association for Artificial Intelligence (AIxIA 2024)*, Vol. 3880. CEUR-WS.org, Bolzano, Italy, 189–201.

[2] I. Alonso, M. Oronoz, and R. Agerri. 2024. MedexpQA: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine* 155 (2024), 102938. doi:10.1016/j.artmed.2024.102938

[3] M. Alshammary, M. N. Uddin, and L. Khan. 2024. RFPG: Question-Answering from Low-Resource Language (Arabic) Texts using Factually Aware RAG. In *Proceedings of the 2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC 2024)*. IEEE, 107–116. doi:10.1109/CIC62241.2024.00023

[4] Patrice Béchard and Orlando Marquez Ayala. 2024. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. *arXiv preprint arXiv:2404.08189* (2024). https://arxiv.org/abs/2404.08189

[5] H.M. Caseli and M.G.V. Nunes (Eds.). 2024. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português* (3 ed.). BPLN. https://brasileiraspln.com/livro-pln/3a-edicao Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao.

[6] O. Cederlund, S. Alawadi, and F. M. Awaysheh. 2024. LLMRAG: An Optimized Digital Support Service using LLM and Retrieval-Augmented Generation. In *Proceedings of the 9th International Conference on Fog and Mobile Edge Computing (FMEC 2024)*. Malmö, Sweden, 54–62. doi:10.1109/FMEC62297.2024.10710181

[7] Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. arXiv:2407.01463 [cs.CL] https://arxiv.org/abs/2407.01463

[8] S. Devi, G. Dhar, C. Bharadwaj, and A. M. 2024. Retrieval Augmented MedLM. In *Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI 2024)*. IEEE, Singapore, Singapore, 1220–1221. doi:10.1109/CAI59869.2024.00217

[9] A. N. T. Dieu, H. T. Nguyen, and C. T. D. Cong. 2024. The Enhanced Context for AI-Generated Learning Advisors with Advanced RAG. In *Proceedings of the 2024 18th International Conference on Advanced Computing and Analytics (ACOMPA)*. Ben Cat, Vietnam, 94–101. doi:10.1109/ACOMPA64883.2024.00021

[10] 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.

[11] W. Fan, Y. Ding, laur. Ning, S. Wang, H. lauri, D. Yin, and Q. Li. 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*. ACM, Barcelona, Spain, 6491–6501.

[12] Luis-Bernardo Hernandez-Salinas, Juan Terven, E. A. Chavez-Urbiola, Diana-Margarita Córdova-Esparza, Julio-Alejandro Romero-González, Amadeo Arguelles, and Ilse Cervantes. 2024. IDAS: Intelligent Driving Assistance System Using RAG. *IEEE Open Journal of Vehicular Technology* 5 (2024), 1139–1165. doi:10.1109/OJVT.2024.3447449

[13] 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 (Dec 2023), 38. doi:10.1145/3571730

[14] Q. R. Lauro, S. Shankar, S. Zeighami, and A. Parameswaran. 2025. RAG without the lag: Interactive debugging for Retrieval-Augmented Generation pipelines. arXiv:2504.13587 [cs.CL] https://arxiv.org/abs/2504.13587 arXiv

preprint arXiv:2504.13587.

[15] F. Magalhães. 2024. Estratégias de chunking para Retrieval-Augmented Generation (RAG): Uma análise detalhada da abordagem do SemDB. https://medium.com/intelligence-factory/estrat%C3%A9gias-de-chunking-para-retrieval-augmented-generation-rag-uma-an%C3%A1lise-detalhada-da-84cfa10ce01a [Accessed: Apr. 28, 2025].

[16] Rahul Magar, Corwin Behnke, Harshita Jhamtani, and Partha Talukdar. 2023. COSTA: Contextualizing Source Texts for Generative Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. https://arxiv.org/abs/2305.06983

[17] C. Merola and J. Singh. 2025. Reconstructing context: Evaluating advanced chunking strategies for Retrieval-Augmented Generation. arXiv:2504.19754 [cs.CL] https://arxiv.org/abs/2504.19754 arXiv preprint arXiv:2504.19754.

[18] P. Mishra, A. Mahakali, and P. S. Venkataraman. 2024. SEARCHD - Advanced Retrieval with Text Generation using Large Language Models and Cross Encoding Re-ranking. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*. IEEE, Bari, Italy, 975–980. doi:10.1109/CASE59546.2024.10711642

[19] R. Nai, E. Sulis, I. Fatima, and R. Meo. 2024. Large Language Models and Recommendation Systems: A Proof-of-Concept Study on Public Procurements. In *Proc. of the 29th Int. Conf. on Applications of Natural Language to Information Systems (NLDB 2024), Part II*. Turin, Italy, 280–290. doi:10.1007/978-3-031-70242-6_27

[20] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A Comprehensive Overview of Large Language Models. arXiv:2307.06435 [cs.CL] https://arxiv.org/abs/2307.06435

[21] S. Obaid and N. Z. Bawany. 2024. SeerahGPT: Retrieval Augmented Generation based Large Language Model. In *Proceedings of the 18th International Conference on Open Source Systems and Technologies (ICOSST 2024)*. Lahore, Pakistan, 1–7. doi:10.1109/ICOSST64562.2024.10871159

[22] 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)*. IEEE, Opatija, Croatia, 2084–2088. doi:10.1109/MIPRO60963.2024.10569238

[23] Mubashar Raza, Zarmina Jahangir, Muhammad Bilal Riaz, Muhammad Jasim Saeed, and Muhammad Awais Sattar. 2025. Industrial applications of large language models. *Scientific Reports* 15 (2025), 13755. doi:10.1038/s41598-025-98483-1

[24] Sheng Shen, Yuandong Tian, Lingfei Kong, Wei Chen, Shiyu Yan, Dinesh Sahoo, and Wayne Xin Zhao. 2023. RAGFusion: Towards Improved Retrieval-Augmented Generation with Fusion-in-Decoder. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://arxiv.org/abs/2305.10854

[25] K. S. K. Subramanyam and S. Sangeetha. 2020. SECNLP: A survey of embeddings in clinical natural language processing. *Journal of Biomedical Informatics* 101 (2020), 103323. doi:10.1016/j.jbi.2019.103323

[26] O. C. Wijaya and A. Purwarianti. 2024. An Interactive Question-Answering System Using Large Language Model and Retrieval-Augmented Generation in an Intelligent Tutoring System on the Programming Domain. In *Proceedings of the 2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*. Singapore, 1–6. doi:10.1109/ICAICTA63815.2024.10763263

[27] Duyu Yang, Yichong Zhang, Yuxuan Zhang, Lijie Zhang, Wenyuan Liu, Rui Xie, Zhiyuan Liu, Duyu Tang, and Ming Zhou. 2023. Faithful RAG: Towards Faithful Retrieval-Augmented Generation via Denoising Re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. https://arxiv.org/abs/2305.13852

[28] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu. 2025. Evaluation of Retrieval-Augmented Generation: A Survey. In *Proceedings of the Big Data*, W. Zhu, H. Xiong, X. Cheng, L. Cui, Z. Dou, J. Dong, S. Pang, L. Wang, L. Kong, and Z. Chen (Eds.). Springer, Singapore, 102–120.