Enhancing Epidemiological Insights with RAG for SIREVA-SUS Reports

Christian Freitas[1], **Ricardo Trainotti Rabonato**[1], **Lilian Berton**[1]

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brazil

{christian.freitas, trainotti.ricardo,lberton}@unifesp.br

Abstract. *This study introduces a Retrieval-Augmented Generation (RAG) framework designed to extract and generate epidemiological insights from SIREVA-SUS reports, Brazil's national surveillance system for bacterial pathogens. By integrating dense information retrieval with generative language models, this approach facilitates explainable question answering over extensive, unstructured epidemiological data, aiding healthcare professionals and researchers in detecting trends, outbreaks, and antimicrobial resistance patterns. The evaluation includes a comparison of various large language models (LLMs), such as the multilingual qwen and Portuguese fine-tuned models Sabia, port5, and Bertimbau. Additionally, the architecture examines multiple retrieval strategies, including vector, dense, sparse, hybrid, and reranker methods. Preliminary findings suggest that the system effectively retrieves relevant information with reasonable BERTScore. This work reinforces the importance of language-specific evaluation and opens new directions for deploying RAG systems in public health decision support.*

1. Introduction

Epidemiological surveillance plays a vital role in public health by enabling early detection, monitoring, and response to infectious diseases [Ihekweazu et al. 2020, Pruccoli et al. 2023, Aguilar-Vargas et al. 2022]. In any country, robust surveillance systems are essential not only for guiding vaccination policies and clinical treatment protocols but also for anticipating outbreaks and reducing health inequalities. Through the continuous collection, analysis, and dissemination of health data, epidemiological surveillance supports evidence-based decision-making and contributes to the prevention and control of disease transmission.

In Brazil, one of the key systems supporting this effort is SIREVA (Sistema Regional de Vacinas), operated within the framework of the Unified Health System (SUS) [Red SIREVA Network 2024]. SIREVA-SUS is part of a regional initiative coordinated by the Pan American Health Organization (PAHO) and developed in collaboration with national laboratories and health authorities. Its primary objective is to monitor bacterial agents responsible for invasive diseases, especially Streptococcus pneumoniae, Haemophilus influenzae, and Neisseria meningitidis.

SIREVA functions as a sentinel surveillance network that collects clinical isolates from hospitals and laboratories across the country. These isolates are analyzed to determine bacterial serotypes, antimicrobial susceptibility profiles, and other molecular

characteristics. The system provides valuable insights into vaccine coverage and antimicrobial resistance patterns, informing public health policies such as updates to vaccination schedules and antibiotic guidelines. By providing a reliable national repository of microbiological and epidemiological data, SIREVA-SUS serves as a cornerstone of Brazil's strategy to combat vaccine-preventable diseases and to ensure timely responses to changes in pathogen behavior.

Despite the relevance of the data provided by SIREVA-SUS, accessing specific and actionable information from the reports can be challenging. These reports are often published in PDF format, contain dense technical content, and lack a standardized structure for querying. Health professionals and decision-makers may struggle to efficiently retrieve targeted information, such as serotype distributions by region, temporal trends in antimicrobial resistance, or alerts on emerging pathogens. In this context, a Retrieval-Augmented Generation (RAG) approach, combining the precision of information retrieval with the flexibility of large language models, offers a promising solution. By enabling natural language queries over unstructured epidemiological documents, RAG systems can help health professionals extract relevant insights more efficiently and improve situational awareness.

The main objective of this work is to develop and evaluate a RAG system tailored to support the exploration of epidemiological data from SIREVA-SUS reports. The contributions of this work are:

- Facilitate access to epidemiological insights by enabling natural language queries over unstructured textual reports.
- Evaluate the effectiveness of different retrieval approaches and several LLMs, including multilingual and fine-tuned Portuguese, in processing and generating accurate responses based on Brazilian epidemiological content.
- Contribute to the advancement of technologies for the Portuguese language, particularly in the domain of health and epidemiological surveillance.

The paper is organized as follows: Section 2 presents the related work. Section 3 describes the methodology for RAG development. Section 4 presents the results and discussion. Finally, Section 5 presents the conclusions and directions for future work.

2. Related Works

In this section, we present some works that have explored RAG in health applications.

A review of the use of RAG in healthcare was conducted by [Amugongo et al. 2025], addressing the limitations of LLMs, such as outdated training data, hallucinated content, and lack of transparency. The study examines available datasets, RAG methodologies, and evaluation frameworks, identifying gaps in the current literature. Key findings indicate that proprietary models like GPT-3.5/4 are the most commonly used for RAG applications in healthcare, but standardized evaluation frameworks are lacking. Additionally, ethical considerations related to RAG implementation in clinical settings remain largely unaddressed. The paper underscores the need for further research to ensure responsible and effective adoption of RAG in medical applications.

Different RAG applications in healthcare have been proposed, such as BioRAGent [Ateia and Kruschwitz 2025], an interactive web-based RAG system designed for

biomedical question answering. Building on the authors' participation in BioASQ 2024, BioRAGent showcases how few-shot learning with LLMs can be effectively applied in professional search environments. EMERGE [Zhu et al. 2024] was designed to enhance multimodal Electronic Health Record (EHR) predictive modeling. It addresses limitations in existing models by integrating clinical notes, time-series data, and knowledge graphs (KGs) for improved medical context. i-MedRAG [Xiong et al. 2024b] is an iterative RAG framework designed to enhance medical question answering by allowing LLMs to ask follow-up queries based on previous retrieval attempts. [Gilson et al. 2024] propose a RAG system for ophthalmology-specific applications.

MIRAGE [Xiong et al. 2024a] is a benchmark designed to evaluate RAG systems for medical question answering. MIRAGE systematically tests 41 different RAG configurations using 7,663 medical QA questions across five datasets, leveraging the MedRAG toolkit. MedRAG enhances the accuracy of six LLMs by up to 18%, bringing GPT-3.5/Mixtral to GPT-4-level performance. The study also reports the discovery of log-linear scaling properties and the "lost-in-the-middle" effects in medical RAG.

Regarding epidemiological content, we found only the work of [Ziletti and D'Ambrosi 2024], which presents an end-to-end methodology combining text-to-SQL generation with RAG to enhance epidemiological queries using Electronic Health Records (EHR) and claims data. By integrating a medical coding step into the text-to-SQL process, the proposed method significantly improves performance over simple prompting.

3. Materials and Methods

3.1. Methodology

The development of the code used in this study was carried out in Python and implemented on the Google Colab platform, using an NVIDIA A100 GPU with 40GB of RAM. The code and data collected for this study will be made available at: `https://github.com/`, after paper acceptance.

3.2. Datasets

SIREVA is a regional surveillance program based on a network of sentinel hospitals and laboratories, implemented by PAHO since 1993 [Pan American Health Organization 2024]. It provides prospective information on the distribution of serotypes and the susceptibility of some bacterial to antibiotics, as well as epidemiological data to estimate the burden of these diseases and to formulate more effective vaccines. These reports focus primarily on bacterial pathogens that cause invasive diseases, such as *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Neisseria meningitidis*. A typical SIREVA report includes the following sections:

1. **Executive Summary:** highlights of the year, key changes in serotype prevalence, resistance trends, and geographic spread.
2. **Methodology:** description of how isolates were collected (from which sentinel hospitals, labs, etc.), and the laboratory techniques used (e.g., serotyping, PCR, MIC testing).
3. **Epidemiological Results:** number of isolates collected, patient age groups, geographic distribution, and clinical sample types (e.g., blood, cerebrospinal fluid).

4. **Serotype Distribution:** tables and graphs showing which serotypes of *S. pneumoniae* were most common, comparison with previous years, stratified by age group and region.
5. **Antimicrobial Resistance (AMR):** MIC (Minimum Inhibitory Concentration) data, resistance percentages for key antibiotics such as penicillin, ceftriaxone, and erythromycin, often displayed in tables by serotype or region.
6. **Discussion and Public Health Implications:** vaccine coverage gaps, rise in non-vaccine serotypes, resistance threats, and policy recommendations.

An example of information from a SIREVA report is presented in Table 1.

Table 1. Example of information presented in the SIREVA 2023 report: invasive isolates of *H. influenzae* by age group and sex. Out of a total of 298 samples, 219 refer to culture and 79 refer to real-time PCR.

Age	Male	Male %	Female	Female %	Total	Total %
$\leq$ 12 months	38	65.5	20	34.5	58	19.5
12–23 months	16	61.5	10	38.5	26	8.7
24–59 months	15	45.5	18	54.5	33	11.1
Subtotal (1)	69	59.0	48	41.0	117	39.3
5–14 years old	16	53.3	14	46.7	30	10.1
15–29 years old	9	75.0	3	25.0	12	4.0
30–49 years old	30	62.5	18	37.5	48	16.1
Subtotal (2)	55	61.1	35	38.9	90	30.2
50–59 years old	17	60.7	11	39.3	28	9.4
$\geq$ 60 years old	33	52.4	30	47.6	63	21.1
Subtotal (3)	50	54.9	41	45.1	91	30.5
Total	**174**	**58.4**	**124**	**41.6**	**298**	**100.0**

3.3. RAG

RAG is a text generation architecture based on large language models (LLMs) that combines document retrieval with neural generation to produce more accurate and contextualized responses [Lewis et al. 2020, Fan et al. 2024]. The goal of RAG is to overcome LLM limitations, as these models, despite generating high-quality natural language, often fabricate information ("hallucinations") when they lack updated or specific knowledge. RAG consists of two main modules:

- **Retrieval module:** an information retrieval system (usually based on embeddings and vector search) finds the most relevant documents from an external database (e.g., scientific texts, medical records, or a knowledge base).
- **Generation module:** a generative model (such as BERT) conditions its response based on the retrieved documents, generating a content-grounded answer.

The RAG system was implemented using a modular pipeline, allowing experimentation with different retrieval strategies. This work expands the evaluation to include multiple large language models (LLMs), such as `sabia-7b`, `port5`, `bertimbau`, `sabia-3.1`, `sabia-3`, `sabiazim-3`, and `qwen-7b` multilingual.

3.3.1. Corpus Preparation

The input documents for this study are SIREVA-SUS epidemiological reports from 2020, 2021, 2022 and 2023 in PDF format. These documents were converted to plain text using [Docling Project 2024] and segmented into chunks that end when a table finishes. Each chunk was indexed along with associated metadata, including the corresponding year.

3.3.2. Retrieval Component

The retrieval component is crucial to the effectiveness of the RAG system, as it identifies and supplies the most relevant text snippets from the epidemiological reports to the generation module. To evaluate different information retrieval approaches, we implemented and tested multiple strategies, using Qdrant as our vector database and similarity search engine. Qdrant is optimized for Approximate Nearest Neighbor (ANN) searches over high-dimensional vector collections and performs cosine-distance similarity searches.

Our primary strategy, dense retrieval, uses vector embeddings to capture the semantic meaning of texts. We employed the pre-trained `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2` model to generate embeddings for both the indexed text fragments ("chunks") of the reports and the user queries. Qdrant then performs a cosine-distance similarity search (via `similarity_search_with_score` or `similarity_search` in our LangChain implementation) between the query vector and the indexed chunk vectors, returning the k most similar chunks (typically $k = 3$). We also extended this approach with metadata filtering, refining results by applying document filters (e.g., `source="sireva_2023.pdf"`) to combine semantic relevance with factual constraints.

In contrast, our sparse retrieval pipeline relies on BM25 term-based scoring. The query is tokenized into words, and BM25 scores are computed against each indexed document. The top k documents with the highest BM25 scores are then retrieved, a method particularly effective for exact lexical matches and keyword searches.

To further improve precision, we implemented a two-stage retrieval pipeline with a re-ranking step. In the first stage, a larger set of candidate chunks (e.g., $k = 5$) is retrieved using dense retrieval. In the second stage, these candidates are re-ranked by a more powerful cross-encoder model (`cross-encoder/ms-marco-MiniLM-L-6-v2`), which evaluates query–document relevance more granularly and produces a final ranking from which the most relevant chunk is selected.

Additionally, we explored Qdrant's native client interface for dense retrieval, configuring HNSW parameters (`hnsw_ef=128`, `exact=false`) to fine-tune our ANN search. For every strategy, the k most relevant chunks (or the top re-ranked chunk) are concatenated to form the context fed into the Sabiá LLM for response generation, enabling a direct comparison of retrieval techniques in terms of relevance, accuracy, and computational efficiency.

We considered the following retrieval strategies:

1. **Vector Retrieval (`vector`):** uses embeddings (vector representations) for both documents and the query, computing cosine similarity or distance for ranking.
2. **Keyword or Metadata Filtering (`filtered`):** applies simple filters such as exact keyword matches, structured criteria (e.g., year, location, pathogen type), or regular expressions. This method does not handle synonyms or semantic similarity, relying solely on exact matches or rule-based logic.
3. **Combination of Methods (`hybrid`):** combines vector-based and sparse retrieval or semantic and keyword filtering. Results from both are aggregated and re-ranked

using combined scoring or learning-to-rank approaches.

4. **Sparse Retrieval with BM25/TF-IDF (`sparse`):** employs traditional Term Frequency–Inverse Document Frequency (TF-IDF). Documents are represented as sparse vectors with term weights from the vocabulary. This approach does not capture semantic similarity or handle synonyms.
5. **Neural Re-ranking (`reranker`):** after initial retrieval (vector or sparse), a reranker uses a deeper model (e.g., BERT) to re-analyze each query-document pair, producing a refined score and reordering the results. This method is computationally expensive and depends on having strong candidate documents from the initial retrieval.

3.3.3. Generator Module

The generator module in this study was designed to support experimentation with multiple large language models (LLMs), primarily focused on the Portuguese language. Among these, the `sabia-7b` model developed by Maritaca AI stands out. It is a 7-billion-parameter model based on the LLaMA-1-7B architecture, pre-trained on 7 billion tokens from the Portuguese subset of ClueWeb22 and further trained with an additional 10 billion Portuguese tokens. This dual-phase training process was designed to capture both linguistic and cultural nuances of Brazilian Portuguese, enhancing the model's performance in comprehension and generation tasks. Sabiá-7B supports input sequences of up to 2048 tokens and has shown competitive results in Portuguese benchmarks, such as Poeta.

In addition to `sabia-7b` [AI 2024], this study evaluates other variants from the Maritaca family, including `sabia-3.1`, `sabia-3`, and `sabiazim-3`, as well as external models such as `qwen-7b` [Bai et al. 2023], `port5` [Carmo et al. 2020], and `Bertimbau` [Souza et al. 2020]. This comparative approach enables a more comprehensive analysis of RAG performance across different LLMs, particularly in the context of generating answers to epidemiological queries in Portuguese.

The inclusion of instruction-tuned and smaller-sized models, such as `sabiazim-3` and `qwen-7b`, allows for evaluating the impact of fine-tuning and model scale on response quality. These models were selected to assess their capability to generate contextually accurate and semantically coherent answers based on retrieved content from SIREVA-SUS reports.

3.3.4. Evaluation Strategy

To evaluate the system, we defined a set of real-world epidemiological questions (e.g., "What were the most common *S. pneumoniae* serotypes in 2022 for children under 5?"). For each retrieval strategy, we computed the following metrics:

- **BLEU and ROUGE**: used to evaluate the surface-level similarity between the generated responses and expert-written reference answers, based on n-gram overlap.
- **BERTScore**: applied to assess the semantic similarity between the generated and reference responses using contextual embeddings, offering a more robust metric

for evaluating meaning preservation.

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \tag{1}$$

- **Semantic Textual Similarity (STS)**: used to measure the degree of semantic alignment between the generated answer and the ground truth, regardless of exact lexical match.

These metrics provide complementary perspectives on the quality of the generated outputs, balancing lexical fidelity (BLEU, ROUGE) with semantic adequacy (BERTScore, STS).

4. Results and Discussion

This section presents the results obtained from applying the RAG architecture to SIREVA-SUS epidemiological reports from 2020 to 2023. The system was evaluated based on its ability to answer epidemiological queries in Portuguese with factual accuracy, relevance, and completeness.

4.1. Example Queries and Answers

Table 2 presents a set of queries employed to validate the RAG system, while Table 3 provides the corresponding answers. The Portuguese versions of the queries and answers are highlighted in blue, as these were the actual texts used in the system.

We conducted experiments using 80 queries formulated by the paper's author. Due to the length of the responses generated for each query, they are made available via an anonymous Google Drive link[1].

Table 2. Example of queries used to evaluate the RAG system

ID	Query
1	Na página 2 do documento, qual é a designação oficial completa da instituição referida como IAL?
	What is the full official designation of the institution referred to as IAL on page 2 of the document?
2	Sobre *Haemophilus influenzae*, do total de 298 amostras, quantas se referem à cultura?
	Regarding *Haemophilus influenzae*, out of the total 298 samples, how many refer to culture?
3	Do total de 380 amostras de *N. meningitidis*, quantas se referem à cultura?
	Out of the total 380 *N. meningitidis* samples, how many refer to culture?
4	Do total de 1891 amostras de *S. pneumoniae*, quantas se referem à PCR em tempo real?
	Out of the total 1891 *S. pneumoniae* samples, how many refer to real-time PCR?
5	Na vigilância de *Neisseria meningitidis*, qual foi a porcentagem de casos registrados para a faixa etária de 5 a 14 anos?
	In the surveillance of *Neisseria meningitidis*, what was the percentage of reported cases for the 5–14 age group?

4.2. Metrics

Figure 1 shows the median BERTScore results across different retrieval strategies and LLMs. We observe the following:

- The `qwen-7b` and `sabiazim-3` models achieved the highest BERTScore medians (0.73 and 0.70), demonstrating strong semantic alignment with reference answers across all retrieval methods.

[1] `https://docs.google.com/spreadsheets/d/1cmvcYBGctakqC8LJ94ss2zH1Nf45_S0B6KcwVPmYDYs/edit?usp=sharing`

Table 3. Answers corresponding to the queries from Table 2

ID	Answer
1	A instituição conhecida como IAL é o Instituto Adolfo Lutz. The institution known as IAL is the Adolfo Lutz Institute.
2	Na *Haemophilus influenzae*, 219 se referem à cultura e 79 à PCR em tempo real. For *Haemophilus influenzae*, 219 refer to culture and 79 to real-time PCR.
3	124 se referem à cultura e 256 à PCR em tempo real. 124 refer to culture and 256 to real-time PCR.
4	Para *S. pneumoniae*, 1426 amostras se referem à cultura e 465 à PCR em tempo real. For *S. pneumoniae*, 1426 samples refer to culture and 465 to real-time PCR.
5	Na faixa etária de 5 a 14 anos, a porcentagem de casos registrados (porc) foi 12,63. In the 5–14 age group, the percentage of reported cases was 12.63.

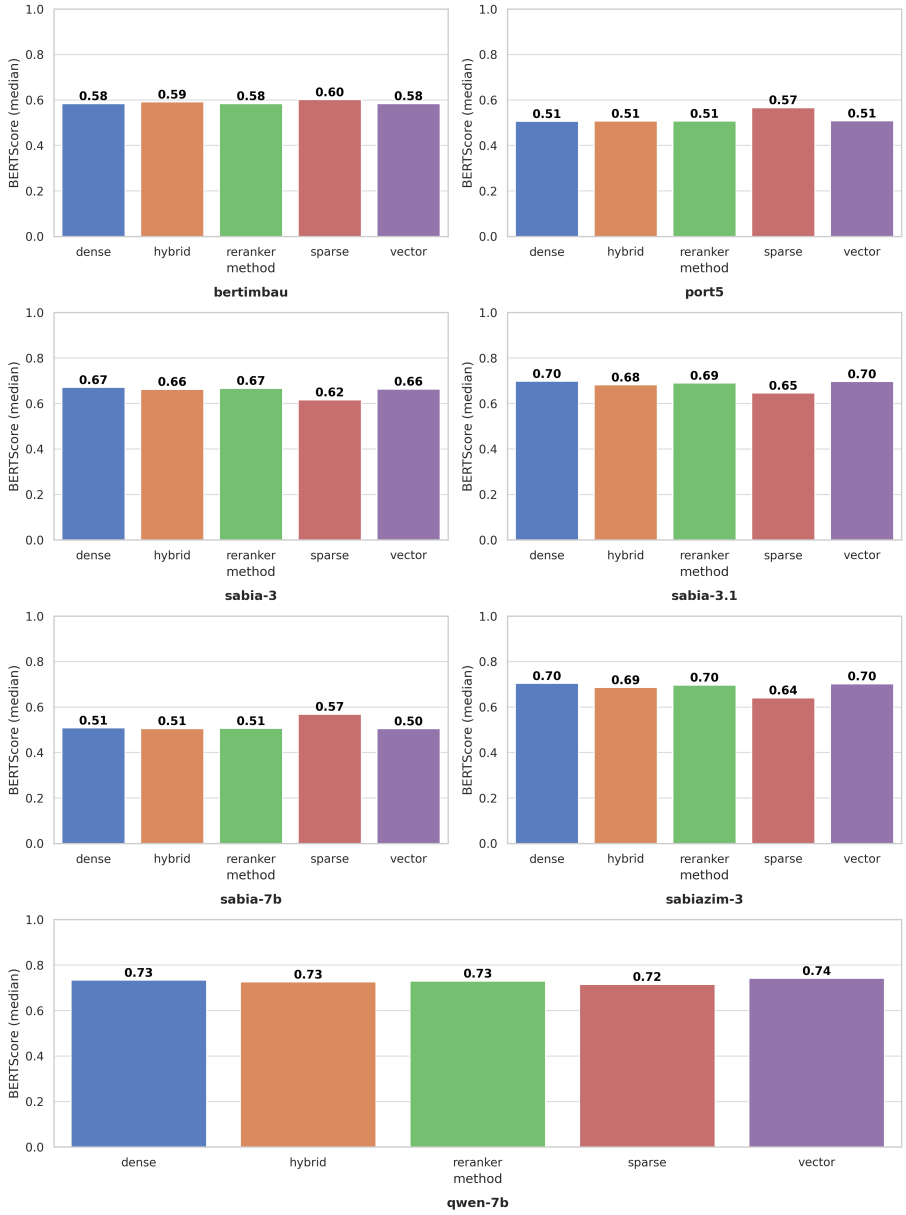

Figure 1. Median values for the BERTScore across retrieval methods and LLMs.

- `sabia-3.1` and `sabia-3` also performed well, with BERTScore medians reaching 0.69 and 0.66, particularly under dense and vector-based retrieval.

- `bertimbau` showed moderate performance, with a median of 0.58.
- `port5` and `sabia-7b` presented the lowest BERTScore values (around 0.50–0.51), suggesting lower semantic alignment when compared to more recent or instruction-tuned models.
- Across all models, the `sparse` retrieval method frequently produced competitive results, while `hybrid` and `reranker` exhibited more modest or inconsistent outcomes.

Figure 2 illustrates the BERTScore evaluation across all LLMs utilized in the RAG. The models `qwen-7b`, `sabia 3.0`, `sabia 3.1`, and `sabiazin-3` are instruction-tuned language models, whereas `sabia 7b`, `port5`, and `bertimbau` are not. Due to this distinction, the non-tuned models exhibited lower scores. Notably, the `qwen-7b` model outperformed the others, achieving the highest scores despite being a multilingual model. The `qwen-7b` may have been exposed to a larger number of scientific documents and government reports, enabling a more robust learning process regarding medical terms and epidemiological analyses. Models specialized in Portuguese, may have a bias toward less varied language, which could impact the similarity measured by BERTScore.

The LLMs were also evaluated using BLEU, ROUGE-L, and STS, with the results presented in Figure 3. Among these metrics, `qwen-7b` and `sabiazin-3` achieved the highest scores, indicating stronger performance in semantic similarity and linguistic fluency. Despite their frequent use in text evaluation, BLEU and ROUGE-L can present lower values in certain contexts. BLEU relies on n-gram precision, meaning slight variations in wording significantly reduce scores, even if the meaning is preserved. In generative tasks, where models provide paraphrased responses, BLEU tends to underestimate quality. ROUGE-L focuses on overlapping sequences between generated and reference texts. If responses maintain coherence but vary structurally, ROUGE-L may fail to reflect actual semantic correctness. Unlike BLEU and ROUGE-L, STS captures meaning-based alignment, making it a better fit for tasks involving abstract reasoning and epidemiological insights.

Regarding the retrieval approach, Figure 4 showcases the comparison between the explored techniques, with dense and vector-based methods achieving the highest BERTScore. This indicates that similarity-based retrieval proved to be more effective in preserving semantic coherence in responses generated by LLMs within epidemiological contexts.

5. Conclusion

This work contributed to advancing Portuguese-language RAG models, especially in epidemiological applications. We performed a comparative analysis of different information retrieval strategies and different LLMs applied to the task of automatic knowledge extraction from SIREVA-SUS epidemiological reports. The application of RAG systems in this domain achieved high BERTScore and demonstrated potential to assist healthcare professionals in swiftly accessing historical and technical information, particularly in contexts that demand rapid decision-making and epidemiological surveillance. Nonetheless, challenges related to evaluation methodologies, response segmentation, and ensuring contextual accuracy remain to be addressed.

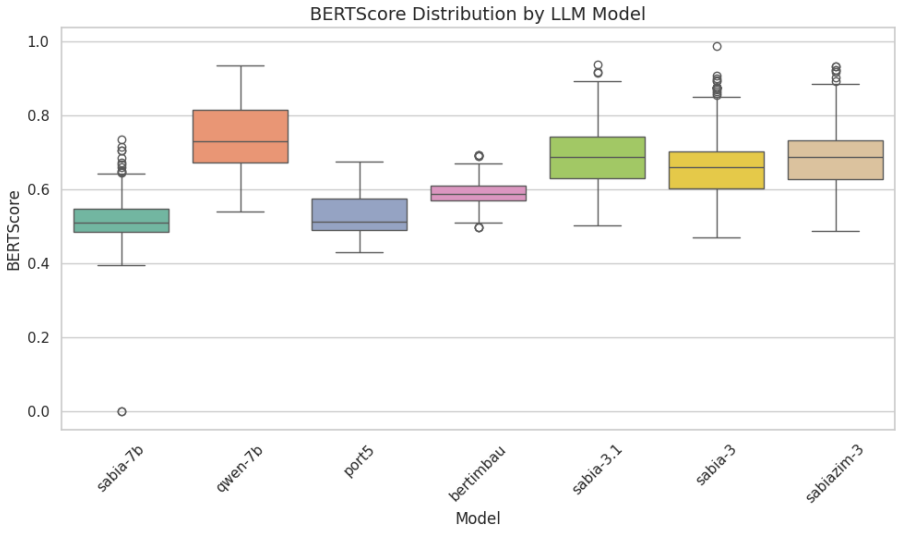

Figure 2. Comparison of BertScore LLMs.

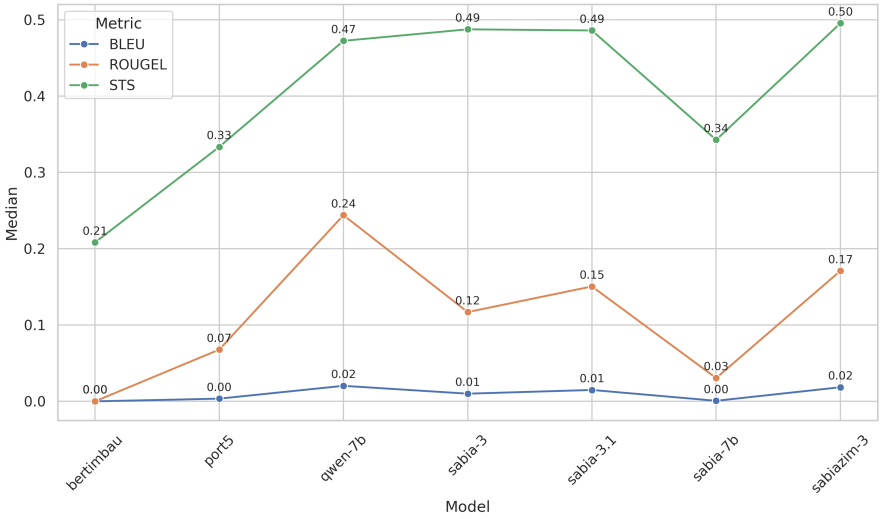

Figure 3. Comparison of LLMs by BLEU, ROUGE-L and STS.

For future research, several approaches can be explored to enhance this work. One potential direction is testing additional LLMs, such as models from the GPT family. Another improvement involves integrating automatic summarization techniques, utilizing models like T5 or PEGASUS, to generate more concise and informative responses. Additionally, the adoption of human evaluation protocols is crucial, as they provide a more accurate assessment for question-answering tasks involving extensive documents. Lastly, developing a publicly available benchmark with real queries and reference answers based on SIREVA-SUS reports would facilitate further research and foster advancements in this field.

6. Ackowledgement

We thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant 2021/10599-3.

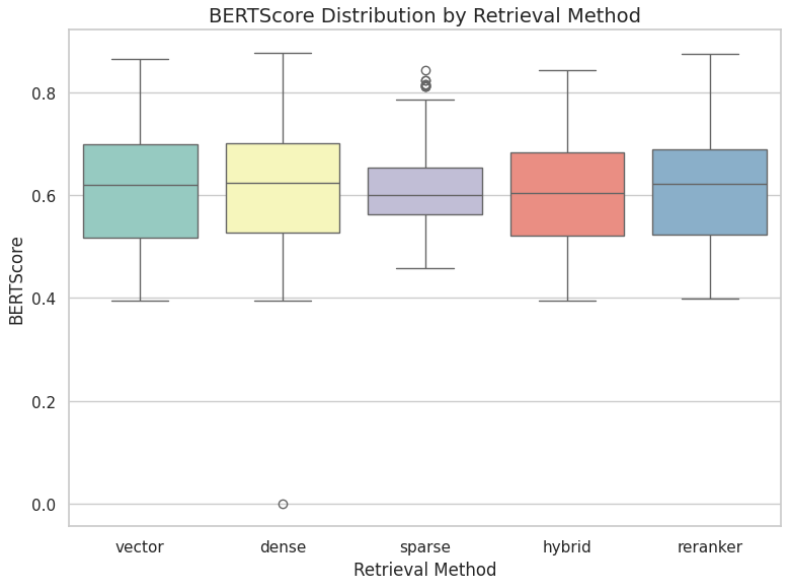

Figure 4. Comparison of Retrieval approaches.

References

Aguilar-Vargas, F., Solorzano-Scott, T., Baldi, M., Barquero-Calvo, E., Jiménez-Rocha, A., Jiménez, C., Piche-Ovares, M., Dolz, G., León, B., Corrales-Aguilar, E., et al. (2022). Passive epidemiological surveillance in wildlife in costa rica identifies pathogens of zoonotic and conservation importance. *PLoS One*, 17(9):e0262063.

AI, M. (2024). Sabiá: Large language models for portuguese. `https://huggingface.co/maritaca-ai`. Accessed: 2024-06-06.

Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., and Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.

Ateia, S. and Kruschwitz, U. (2025). Bioragent: A retrieval-augmented generation system for showcasing generative query expansion and domain-specific search for scientific q&a. In *European Conference on Information Retrieval*, pages 1–5. Springer.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Docling Project (2024). Docling - toolkit for document parsing and structuring. `https://github.com/doclingproject/docling`. Accessed: 2024-06-06.

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.

Gilson, A., Ai, X., Arunachalam, T., Chen, Z., Cheong, K. X., Dave, A., Duic, C., Kibe, M., Kaminaka, A., Prasad, M., et al. (2024). Enhancing large language models with domain-specific retrieval augment generation: A case study on long-form consumer health question answering in ophthalmology. *arXiv preprint arXiv:2409.13902*.

Ihekweazu, C., Yinka-Ogunleye, A., Lule, S., and Ibrahim, A. (2020). Importance of epidemiological research of monkeypox: is incidence increasing? *Expert review of anti-infective therapy*, 18(5):389–392.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Pan American Health Organization (2024). SIREVA: Regional System for Vaccines. `https://www.paho.org/en/sireva`. Accessed: 2024-05-17.

Pruccoli, G., Castagno, E., Raffaldi, I., Denina, M., Barisone, E., Baroero, L., Timeus, F., Rabbone, I., Monzani, A., Terragni, G. M., et al. (2023). The importance of rsv epidemiological surveillance: a multicenter observational study of rsv infection during the covid-19 pandemic. *Viruses*, 15(2):280.

Red SIREVA Network (2024). Brasil – SIREVA. `https://redsirevanetwork.com/brasil/`. Accessed: 2024-05-17.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024a). Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Xiong, G., Jin, Q., Wang, X., Zhang, M., Lu, Z., and Zhang, A. (2024b). Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.

Zhu, Y., Ren, C., Wang, Z., Zheng, X., Xie, S., Feng, J., Zhu, X., Li, Z., Ma, L., and Pan, C. (2024). Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3549–3559.

Ziletti, A. and D'Ambrosi, L. (2024). Retrieval augmented text-to-sql generation for epidemiological question answering using electronic health records. *arXiv preprint arXiv:2403.09226*.