Analysis of the Chunking Process in RAG Architectures

Vitor Mateus R. do Amaral[1], Luiza Espinosa[1], Gabriel M. Lunardi[1], Adriano Q. de Oliveira, Thiago Lopes T. da Silveira[1], Leonardo R. Emmendorfer[1]

[1]Centro de Tecnologia - Universidade Federal de Santa Maria (UFSM)

{vitor.romancini, luiza.silveira}@acad.ufsm.br

{gabriel.lunardi, thiago.silveira}@ufsm.br

{adriano.q.oliveira@ufsm.br, leonardo.emmendorfer}@ufsm.br

Abstract. *This paper studies how text segmentation affects Retrieval Augmented Generation. We compare heuristic, semantic, and recursive strategies on a Portuguese institutional corpus, holding the pipeline fixed with bge base en v1.5 embeddings, a 500 token target size, and about 10% overlap. In the semantic variant, boundaries are detected by changes in similarity between sentence embeddings using a 95th percentile threshold. Evaluation covers intrinsic coherence and extrinsic metrics such as Factual Accuracy, Precision, Recall, F1-score, and MRR.*

1. Introduction

Chatbots based on RAG (Retrieval Augmented Generation) [Lewis et al. 2020] constitute a reference architecture for applications that require relevance in retrieval and factual fidelity in generation, including in university settings that range from student services to the automation of administrative routines [Figueiredo et al. 2023, Soares et al. 2025]. This article aims to evaluate, in a Portuguese institutional chatbot, the effect of different document segmentation strategies on the performance of RAG systems.

The contributions are: (i) a controlled experimental protocol that isolates the segmentation stage; (ii) a systematic comparison of three strategies (heuristic, semantic, and recursive) with calibrated threshold and overlap; (iii) a joint evaluation of retrieval and generation metrics, including Factual Accuracy, Precision, Recall, F1-score, and MRR (Mean Reciprocal Rank); and (iv) practical guidelines for selecting strategies in scenarios with structured documents. The main results indicate that semantic segmentation achieved the best Accuracy (84.6%) and F1-score (84.4%), the heuristic approach yielded the highest MRR (90.1%) with more favorable top k ranking, and the recursive strategy performed worse, suggesting that calibrated textual or semantic boundaries, with moderate overlap, maximize factual quality and pipeline efficiency.

2. Related Work

Segmentation in RAG architectures has been identified as a critical component. [Barnett et al. 2024] discusses improvements in RAG development, highlighting heuristic approaches (based on punctuation/textual structure) and semantic approaches (content-oriented), and suggests investigating their *trade-offs*, in addition to exploring multi-modal embeddings for tables, figures, and formulas.

Recent advances include **LateSplit** [Peng et al. 2025], which combines pre-retrieval *chunking* with post-retrieval refinement, and **HOPE** [Brådland et al. 2025], which systematically evaluates the impact of different segmentation strategies on RAG performance. In line with these studies, this work compares heuristic and semantic approaches on a structured Portuguese corpus and reports their effects on retrieval relevance and precision, as well as generation metrics.

3. Chunks

Chunking is the segmentation of text into smaller, coherent units that are sufficient to preserve context and fit within the LLM's context window. In RAG systems, this step directly affects retrieval and answer quality; therefore, the chosen strategy should take into account the document's structure and the intended use case.

In this study, with a focus on structured documents, we compare three strategies: (i) heuristic, which splits along textual cues (paragraphs, sentences, topics) without enforcing a fixed size; (ii) semantic, which relies on content coherence to maintain meaningful units [Barnett et al. 2024]; and (iii) recursive, which applies separators in order of strength until the desired size or shape is reached (for example, `["\n\n", "\n", ".", ""]`), the method adopted in university chatbot. Hybrid schemes and segment overlap can also be used to reduce context loss.

4. Experiment

The corpus comprises Portuguese institutional documents (norms, regulations, support materials) available in the GitHub repository[1], preprocessed by normalizing whitespace, removing redundant markers, and sentence tokenization. We built a dense-embedding index and used top-k passage retrieval for LLM-conditioned generation, keeping all non-segmentation components fixed to isolate segmentation effects.

4.1. Fragmentation methods

Heuristic splits at natural textual boundaries (paragraph and sentence ends, bullets, topic shifts); simple and efficient, but may misalign with semantic units. Semantic uses sentence (or short-window) embeddings; cosine distances between consecutive embeddings define boundaries via a single threshold at the 95th percentile. A small buffersize adds neighboring sentences to strengthen local context. Recursive applies an ordered list of separators from most to least structuring, re-splitting blocks that exceed the size limit until the target format is reached.

4.2. Methodological basis and data structure

Semantic fragmentation involved three steps: sentence embeddings, sequential distance computation, and percentile-based boundary detection. A data structure stored index, position, embeddings, and distances for easy inspection and processing. After sliding-window consolidation, local-context embeddings were generated using the bge-base-en-v1.5 model, and the distance series revealed cohesive stretches and peaks (Figure 1).

[1] Available at:https://github.com/Vitormateusromancini/Repositorio-ERAMIA.git

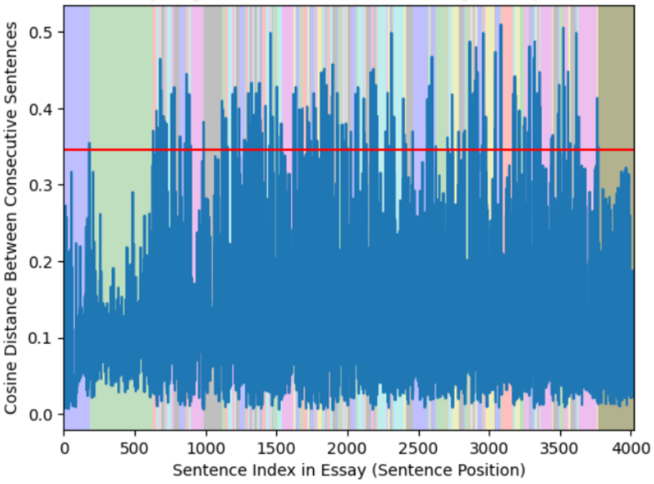

Figure 1. Distances between embeddings of consecutive sentences; peaks above the 95th percentile indicate semantic boundaries.

4.3. Hyperparameters and implementation

We adopted a 500-token block size to balance context preservation and retrieval efficiency, as shorter chunks reduce coherence and longer ones lower precision. A 10% overlap was used to smooth topic transitions without adding redundancy. For the semantic method, boundaries correspond to the 95th percentile of cosine distances between sentence embeddings, calibrated empirically to avoid over- or under-segmentation. All runs used bge-base-en-v1.5 embeddings and cosine distance, with small buffer windows to reinforce local context.

4.4. Evaluation protocol

Intrinsic metrics include intra-block coherence (mean inter-sentence distance) and size variation. Extrinsic metrics in the RAG pipeline include *Recall@k* and *MRR/nDCG* for retrieval (with $k \in \{5, 10\}$), plus factuality and adequacy for generation (LLM-assisted annotation with human sampling). Methods are compared on the same query set using the Wilcoxon signed-rank test at a 5% significance level.

5. Results and Discussion

The results are presented along two dimensions: (i) a quantitative analysis using Factual Accuracy, Precision, Recall, F1-score, and MRR, and (ii) a qualitative analysis based on inspection of chatbot answers. Table 1 summarizes the indicators obtained in the RAG pipeline.

Table 1. Evaluation metrics in the RAG chatbot

Fragmentation method	Factual Accuracy	Precision	Recall	F1-score	MRR
Heuristic	79.8%	83.5%	76.2%	79.7%	90.1%
Semantic	84.6%	86.9%	82.1%	84.4%	88.3%
Recursive	65.4%	69.2%	61.0%	64.9%	58.7%

Semantic fragmentation shows superior performance in Factual Accuracy (84.6%) and F1 (84.4%), indicating a better balance between precision and coverage due to the preservation of thematic boundaries. The heuristic method attains the best MRR (90.1%), suggesting more favorable top ranking of relevant passages, although with a slight drop in Recall and F1-score relative to the semantic approach. The recursive strategy yields the lowest scores across metrics, which is consistent with higher fragmentation and consequent loss of context across blocks.

Qualitatively, the semantic approach reduces omissions and contradictions in normative items and better anchors answers to the correct sections of the document. The heuristic method produces fast and well ranked answers but is more sensitive to cuts at non semantic boundaries. The recursive method tends to create blocks that are too short and misaligned with topics, impairing both retrieval, through lower Recall, and generation, through reduced factual accuracy.

In summary, for structured corpora, semantic segmentation with a calibrated threshold and moderate window and overlap offers better factual fidelity and F1, whereas the heuristic method can be preferred when the goal is to maximize top-k ranking and reduce latency. The recursive approach requires more careful tuning of block size and overlap to mitigate excessive fragmentation. In future work we will follow guidelines for user studies with chatbots [Barbosa et al. 2022].

Acknowledgments

The authors thank CNPq for supporting the Universal project 402086/2023-6, as well as FAPERGS for the ARD/ARC project – process 24/2551-0000645-1.

References

Barbosa, M., Valle, P., Nakamura, W., Guerino, G., Finger, A., Lunardi, G., and Silva, W. (2022). Um estudo exploratório sobre métodos de avaliação de user experience em chatbots. In *Anais da VI Escola Regional de Engenharia de Software*.

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., and Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI*.

Brådland, H., Goodwin, M., Andersen, P.-A., Nossum, A. S., and Gupta, A. (2025). A new hope: Domain-agnostic automatic evaluation of text chunking.

Figueiredo, L. O., Lopes, A. M. Z., Validorio, V. C., and Mussio, S. C. (2023). Desafios e impactos do uso da inteligência artificial na educação. *Educação Online*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *International Conference on Neural Information Processing Systems (NeurIPS)*.

Peng, Z., Liu, X., and Yang, G. (2025). Latesplit: Lightweight post-retrieval chunking for query-aligned text segmentation in rag systems. In *2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*.

Soares, T., Costa, R., Soares, E., Calderon, I., Lunardi, G., Valle, P., Guedes, G., and Silva, W. (2025). Machine learning-assisted tools for user experience evaluation: A systematic mapping study. In *Anais do XXI Simpósio Brasileiro de Sistemas de Informação*.