Indexação Semântica Modular para Recuperação Colaborativa de Conhecimento em Sistemas RAG

Emelyn C. Freire[1,2], Karolayne S. Azevedo[1,2], Sérgio N. Silva[3], Marcelo A. C. Fernandes[1,2,4]

[1]InovAI Lab, nPITI/IMD, UFRN, 59.078-900, Natal, RN, Brasil

[2]Leading Advanced Technologies Center of Excellence (LANCE), nPITI/IMD,UFRN

[3]Departamento de Engenharia Elétrica (DEE), UFCG, Campina Grande, PB, Brasil

[4]Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal, Brasil

emelyn.freire.116@ufrn.edu.br, karolayneazevsantos@gmail.com

sergionatan@dee.ufcg.edu.br, mfernandes@dca.ufrn.br

Resumo. *Sistemas de Recuperação Aumentada por Geração (Retrieval-Augmented Generation – RAG) dependem fortemente das estratégias de indexação e segmentação dos documentos para garantir respostas precisas e confiáveis. Este trabalho apresenta um estudo comparativo entre duas arquiteturas de indexação para sistemas RAG: uma abordagem tradicional monolítica, baseada em chunking uniforme, e uma abordagem modular com enriquecimento semântico, que incorpora segmentação estrutural e metadados contextuais. Os experimentos foram conduzidos sobre um corpus real de 20 dissertações de mestrado, utilizando dois modelos de embedding e diferentes valores de recuperação de contexto (K), totalizando 120 execuções em um protocolo experimental controlado. A avaliação foi realizada por meio de múltiplas métricas do framework RAGAS, abrangendo fidelidade, relevância da resposta, precisão e recall do contexto, similaridade semântica e correção da resposta. Os resultados mostram que a arquitetura modular supera consistentemente a abordagem monolítica em todas as métricas analisadas, apresentando ganhos de desempenho e menor variabilidade entre execuções. Esses achados indicam que decisões arquiteturais no processo de indexação exercem impacto direto na qualidade, robustez e confiabilidade das respostas geradas por sistemas RAG.*

Abstract. *Retrieval-Augmented Generation (RAG) systems rely heavily on document indexing and segmentation strategies to ensure accurate and reliable responses. This work presents a comparative study between two indexing architectures for RAG systems: a traditional monolithic approach based on uniform chunking, and a modular approach with semantic enrichment, which incorporates structural segmentation and contextual metadata. The experiments were conducted on a real corpus of 20 master's dissertations, using two embedding models and different context retrieval values (K), totaling 120 runs under a controlled experimental protocol. The evaluation was carried out using multiple metrics from the RAGAS framework, including faithfulness, answer relevancy, context precision and recall, semantic similarity, and answer correctness. The*

results show that the modular architecture consistently outperforms the monolithic approach across all analyzed metrics, demonstrating performance gains and lower variability between runs. These findings indicate that architectural decisions in the indexing process have a direct impact on the quality, robustness, and reliability of responses generated by RAG systems.

1. Introdução

Modelos de Linguagem de Grande Escala (Large Language Models – LLMs) têm ampliado significativamente as possibilidades de interação humano–computador em tarefas de acesso, síntese e interpretação de informação textual. Entretanto, apesar de sua elevada capacidade generativa, esses modelos apresentam limitações conhecidas relacionadas à confiabilidade factual, atualização do conhecimento e controle sobre as fontes utilizadas durante a geração das respostas, incluindo a ocorrência de alucinações e inferências não fundamentadas [Ji et al. 2023]. Nesse contexto, a abordagem de Recuperação Aumentada por Geração (*Retrieval-Augmented Generation* – RAG) tem se consolidado como uma solução eficaz para mitigar tais limitações, ao integrar mecanismos de recuperação de informação externa ao processo de geração textual [Lewis et al. 2020, Guu et al. 2020, Setty et al. 2024].

Em sistemas RAG, a qualidade das respostas geradas depende fortemente das etapas de indexação, segmentação e recuperação dos documentos que compõem o corpus de conhecimento. Estratégias inadequadas de *chunking* e organização dos dados podem introduzir ruído informacional, fragmentar conteúdos semanticamente relacionados e comprometer tanto a relevância quanto a consistência factual das respostas. Apesar da ampla adoção de pipelines RAG em aplicações práticas, grande parte das implementações ainda se baseia em arquiteturas monolíticas, caracterizadas por segmentação uniforme de documentos e uso limitado de metadados estruturais, o que restringe a capacidade de controle e análise do processo de recuperação [Karpukhin et al. 2020, Liu et al. 2024].

Nos últimos anos, abordagens mais sofisticadas têm explorado a incorporação de informações estruturais e semânticas aos processos de indexação, buscando preservar a organização lógica dos documentos e enriquecer os trechos recuperados com contexto adicional. Técnicas clássicas e contemporâneas de segmentação textual indicam que a preservação de unidades semânticas coerentes pode reduzir a fragmentação da informação e melhorar a qualidade da recuperação [Hearst 1997]. No entanto, ainda são escassos estudos experimentais que avaliem de forma sistemática o impacto dessas decisões arquiteturais sobre o desempenho global de sistemas RAG, especialmente sob múltiplas métricas de qualidade e diferentes configurações de recuperação [Gao et al. 2024]. Em particular, carece à literatura uma análise comparativa controlada entre arquiteturas monolíticas tradicionais e arquiteturas modulares com enriquecimento semântico, considerando não apenas valores médios de desempenho, mas também a estabilidade e a variabilidade dos resultados [Abo El-Enen et al. 2025].

Diante desse cenário, este trabalho apresenta um estudo comparativo entre duas arquiteturas de indexação para sistemas RAG: uma abordagem tradicional monolítica, baseada em *chunking* uniforme, e uma abordagem modular com enriquecimento semântico, que incorpora segmentação estrutural e metadados contextuais ao processo de indexação. A avaliação é conduzida sobre um corpus real de dissertações acadêmicas, utilizando dois

modelos distintos de *embedding* e diferentes valores de recuperação de contexto, com análise baseada em múltiplas métricas do framework RAGAS [Es et al. 2024].

Este trabalho está inserido em um projeto de maior porte, financiado pela Financiadora de Estudos e Projetos (FINEP), cujo objetivo é a preservação, modernização e ampliação do acesso aos acervos científicos e acadêmicos da Universidade Federal do Rio Grande do Norte, com ênfase em teses e dissertações de mestrado e doutorado. No âmbito desse projeto, está sendo desenvolvida a plataforma *AcadêmicoIA*, uma infraestrutura baseada em inteligência artificial e grandes modelos de linguagem que visa transformar a forma de interação com repositórios institucionais, permitindo consultas semânticas, respostas contextualizadas, resumos automáticos e maior acessibilidade aos conteúdos acadêmicos. Nesse contexto, o presente estudo contribui como uma etapa fundamental de investigação científica, avaliando arquiteturas de indexação e recuperação que servirão de base metodológica e tecnológica para a implementação da plataforma AcadêmicoIA em escala institucional.

As principais contribuições deste trabalho são: (i) a proposição e descrição detalhada de uma arquitetura modular de indexação com enriquecimento semântico para sistemas RAG; (ii) uma avaliação experimental sistemática e reprodutível comparando essa arquitetura com uma abordagem monolítica tradicional; e (iii) uma análise quantitativa do impacto das decisões arquiteturais na qualidade, consistência e robustez das respostas geradas. Os resultados obtidos demonstram que escolhas arquiteturais no processo de indexação exercem influência direta e significativa sobre o desempenho de sistemas RAG, reforçando a importância de abordagens semanticamente conscientes para aplicações colaborativas de acesso ao conhecimento.

2. Metodologia

A metodologia adotada neste trabalho foi estruturada para permitir uma comparação controlada e reprodutível entre duas arquiteturas de indexação aplicadas a sistemas de RAG: uma abordagem tradicional monolítica e uma abordagem modular com enriquecimento semântico. O processo experimental compreende a construção do corpus, a definição das abordagens de indexação, a configuração dos modelos de *embedding* e de linguagem, bem como a avaliação sistemática do desempenho por meio de métricas padronizadas.

O corpus utilizado é composto por 20 dissertações de mestrado obtidas do Repositório Institucional da UFRN, sendo 10 provenientes do Programa de Pós-Graduação em Engenharia Elétrica e Computação (PPgEEC) e 10 do Programa de Pós-Graduação em Enfermagem (PPgENF). Foram considerados apenas documentos publicados entre 2024 e 2025, todos redigidos em língua portuguesa e disponibilizados em formato PDF. Essa seleção visa garantir diversidade temática e contemporaneidade do conteúdo analisado.

Na abordagem monolítica, os documentos são processados de forma sequencial, com extração do texto bruto e segmentação em *chunks* de tamanho fixo (256 caracteres), utilizando sobreposição de aproximadamente 20% entre segmentos consecutivos. Nessa configuração, apenas metadados mínimos são associados aos *chunks*, caracterizando um pipeline linear e pouco flexível. A Figura 1 ilustra o fluxo de ingestão da abordagem tradicional monolítica. Nessa arquitetura, os documentos em formato PDF são processados de forma sequencial, com extração do texto bruto por meio da biblioteca PyMuPDF, seguida pela segmentação mecânica em *chunks* de tamanho fixo utilizando o *Recursi-*

veCharacterTextSplitter. O processo não considera a estrutura semântica ou hierárquica do documento, resultando em fragmentação uniforme do conteúdo textual. Os *chunks* gerados são então convertidos em representações vetoriais por modelos de *embedding* e armazenados diretamente em um banco vetorial, caracterizando um pipeline linear, rígido e fortemente acoplado.

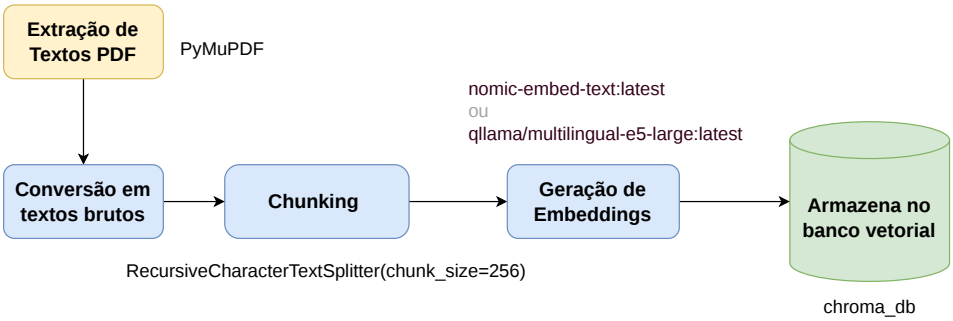

Figura 1. Fluxo de ingestão da abordagem tradicional monolítica para indexação em sistemas RAG.

Já a abordagem modular com enriquecimento semântico emprega um pipeline composto por módulos independentes, responsáveis pela extração de metadados, detecção da estrutura documental e validação de qualidade dos segmentos. O *chunking* é realizado de forma semântica, respeitando limites estruturais como capítulos, seções e parágrafos, e cada *chunk* é enriquecido com metadados contextuais, incluindo informações bibliográficas e estruturais do documento. A Figura 2 apresenta o fluxo de ingestão da abordagem modular com enriquecimento semântico. Diferentemente da abordagem monolítica, o pipeline é organizado em módulos independentes responsáveis pela extração de texto, identificação da estrutura documental, enriquecimento com metadados e validação de qualidade dos segmentos. O processo de *chunking* respeita limites estruturais do documento, evitando cortes arbitrários e preservando a coesão semântica. Cada *chunk* é enriquecido com metadados contextuais e estruturais, que são combinados com os metadados globais do documento antes da geração dos *embeddings*, resultando em unidades textuais mais informativas para a etapa de recuperação vetorial.

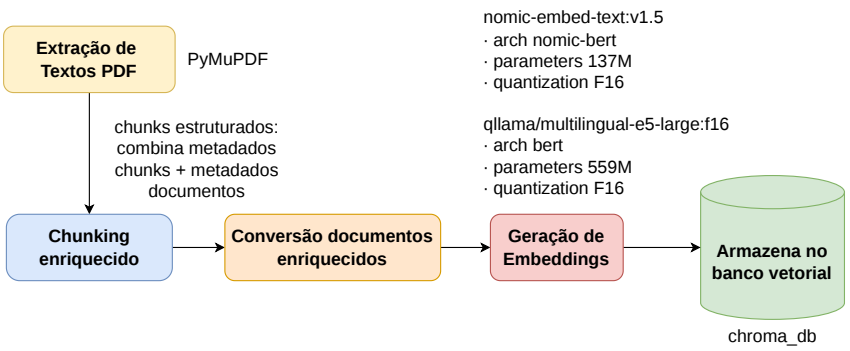

Figura 2. Fluxo de ingestão da abordagem modular com enriquecimento semântico para indexação em sistemas RAG.

A Figura 3 detalha o fluxo interno de enriquecimento semântico e *chunking* adotado na abordagem modular. O processo inicia-se com a extração do texto e dos metadados dos documentos em formato PDF, seguida por uma varredura página a página para identificação de seções pré-textuais e da seção de referências, que são tratadas ou descartadas conforme regras específicas. Em seguida, o texto passa por uma etapa de limpeza e normalização, removendo artefatos da conversão do PDF, como quebras artificiais, hífens e elementos redundantes. O conteúdo limpo é então dividido em parágrafos, que constituem a unidade básica para o *chunking*. Parágrafos com tamanho inferior a um limiar mínimo são descartados, enquanto parágrafos maiores que o tamanho máximo permitido são subdivididos por meio de uma estratégia híbrida baseada no *RecursiveCharacterTextSplitter*, preservando sobreposição controlada. Cada *chunk* gerado é enriquecido com metadados estruturais e descritivos, incluindo informações de seção, posição no documento e atributos bibliográficos, resultando em unidades textuais semanticamente coesas e contextualizadas para a etapa de indexação vetorial.

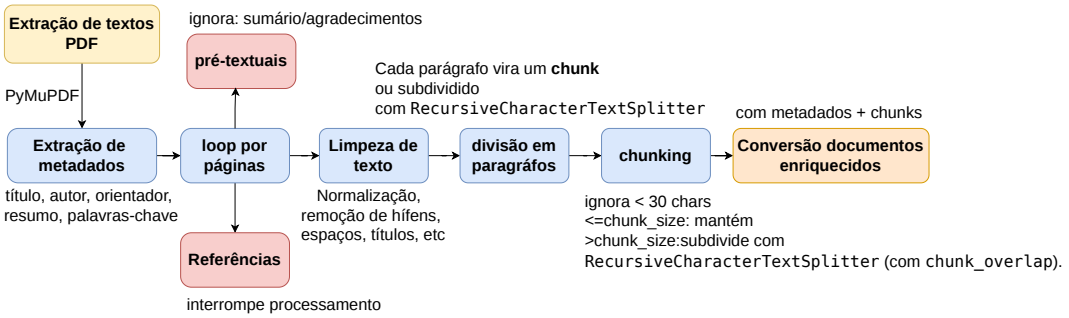

Figura 3. Fluxo de enriquecimento semântico e *chunking* adotado na abordagem modular, detalhando as etapas de extração, limpeza, segmentação estrutural e enriquecimento com metadados.

Para a geração das representações vetoriais, foram utilizados dois modelos de *embedding* executados localmente, permitindo avaliar o impacto do encoder na qualidade da recuperação. Os *embeddings* gerados foram armazenados em bancos vetoriais independentes, garantindo isolamento completo entre as diferentes configurações experimentais. O processo de recuperação considera dois valores de *top-*$K \in \{5, 10\}$, que determinam o número de *chunks* retornados para compor o contexto fornecido ao modelo de linguagem. A Figura 4 apresenta o fluxo de inferência adotado pelo sistema de RAG. O processo inicia-se com o envio da pergunta pelo usuário por meio da interface de aplicação, que é encaminhada ao backend para processamento. Em seguida, a consulta é convertida em uma representação vetorial utilizando o mesmo modelo de *embedding* empregado na indexação dos documentos, garantindo consistência semântica entre consulta e corpus. A partir desse vetor, é realizada uma busca semântica no banco vetorial ChromaDB [Chroma 2024, Chase 2023], recuperando os K trechos mais relevantes. Os trechos recuperados são então organizados e formatados em um bloco de contexto unificado, etapa que inclui a remoção de redundâncias e a padronização do conteúdo. Esse contexto é combinado com a pergunta original e com instruções do sistema para compor o *prompt* final, que é submetido ao modelo de linguagem. O LLM gera a resposta condicionada exclusivamente ao contexto fornecido, e o resultado é retornado ao usuário como resposta final da API.

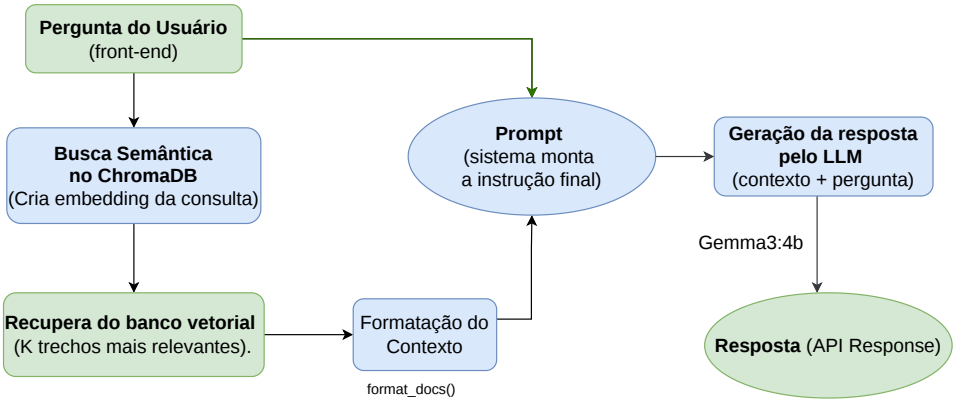

Figura 4. Fluxo de inferência do sistema RAG, desde a consulta do usuário até a geração da resposta pelo modelo de linguagem.

A avaliação do sistema foi conduzida a partir de cinco perguntas de validação, elaboradas para testar diferentes tipos de recuperação, como extração de metadados textuais, entidades nomeadas, relacionamentos semânticos e valores numéricos. Cada pergunta foi executada três vezes em todas as combinações de abordagem, modelo de *embedding* e valor de K, totalizando 120 execuções. O desempenho foi mensurado por meio de seis métricas do framework RAGAS, abrangendo fidelidade, relevância da resposta, precisão e *recall* do contexto, similaridade semântica e correção da resposta, possibilitando uma análise abrangente e comparativa das arquiteturas investigadas. A Figura 5 apresenta o fluxo completo da metodologia de avaliação adotada neste trabalho. O protocolo experimental considera duas abordagens de indexação (monolítica e modular), dois modelos de *embedding* e dois valores de K para recuperação de contexto, resultando em múltiplas configurações de bancos vetoriais isolados. Para cada configuração, as perguntas de validação são submetidas ao fluxo de inferência RAG, no qual o modelo de linguagem gera a resposta condicionada aos contextos recuperados. As respostas produzidas são então avaliadas em duas etapas complementares: inicialmente, o framework RAGAS avalia as respostas e chama um modelo de linguagem que atua como juiz semântico, analisando a consistência e a correção conceitual das respostas em relação aos contextos e aos *ground truths*; em seguida, o framework RAGAS converte essas avaliações qualitativas em métricas numéricas padronizadas. Esse processo garante uma avaliação sistemática, comparável e reprodutível entre todas as combinações experimentais consideradas.

3. Resultados e Análises

Esta seção apresenta os resultados experimentais obtidos a partir da comparação entre duas metodologias de indexação: a Abordagem 1, correspondente à indexação tradicional monolítica, e a Abordagem 2, baseada em uma arquitetura modular com enriquecimento semântico. Os experimentos consideram ainda dois modelos de *embedding*, denotados por M1 (qllama/multilingual-e5-large:F16) e M2 (nomic-embed-text:v1.5), bem como dois valores de recuperação de contexto ($K \in \{5, 10\}$). A análise é conduzida com base nas seis métricas do framework RAGAS, permitindo avaliar de forma sistemática a qualidade da recuperação de contexto e das respostas geradas pelos sistemas RAG sob diferentes configurações experimentais.

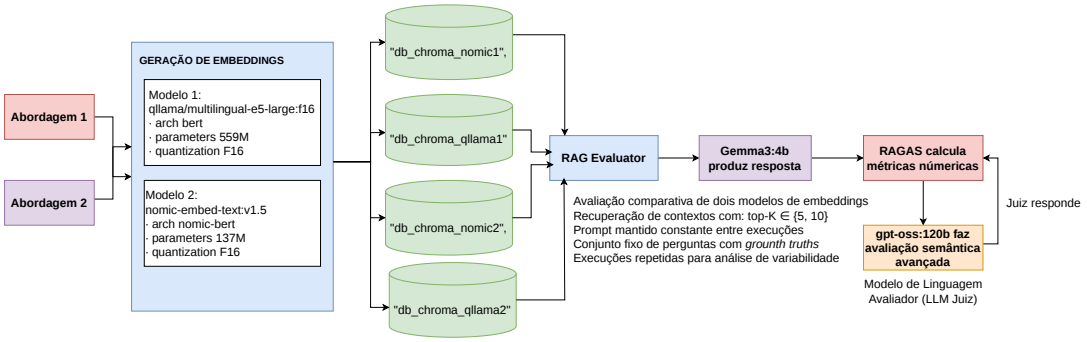

Figura 5. Fluxo da metodologia de avaliação dos sistemas RAG, incluindo a geração de respostas, a avaliação semântica por um LLM juiz e o cálculo das métricas quantitativas pelo framework RAGAS.

A Figura 6 apresenta os resultados da métrica de Fidelidade (*Faithfulness*), que avalia a consistência factual entre as respostas geradas pelo modelo de linguagem e os contextos recuperados. Observa-se que a Abordagem 2 supera consistentemente a Abordagem 1 em todas as configurações avaliadas, independentemente do modelo de *embedding* utilizado. Em particular, a combinação da Abordagem 2 com o modelo M1 alcança valores próximos ou iguais a 1.0, especialmente para $K = 5$, indicando que as respostas geradas são quase integralmente suportadas pelo contexto recuperado.

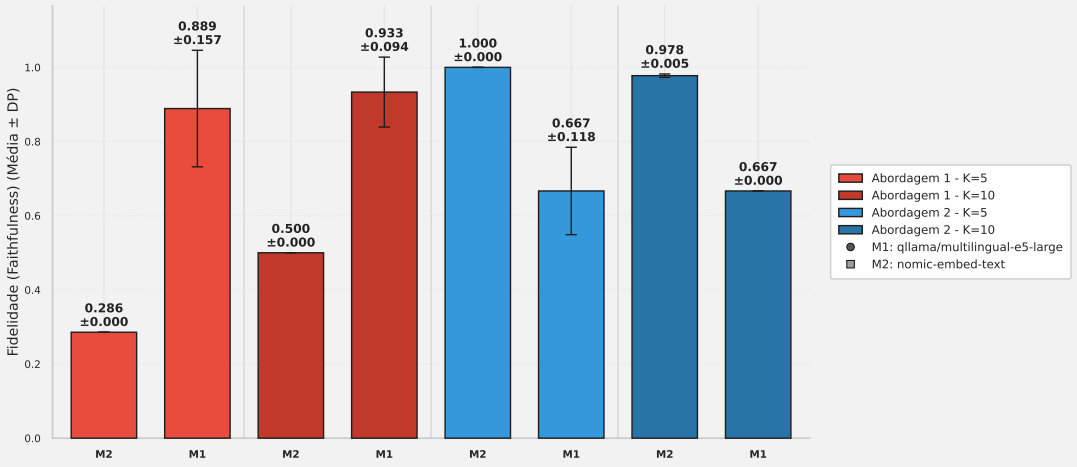

Figura 6. Resultados da métrica de Fidelidade (Faithfulness) para as diferentes configurações experimentais, comparando a Abordagem 1 e a Abordagem 2, com os modelos de embedding M1 e M2 e valores de recuperação de contexto $K \in \{5, 10\}$.

Além do ganho em média, nota-se uma redução significativa da variabilidade dos resultados na Abordagem 2, refletida por desvios padrão menores em comparação à Abordagem 1. Esse comportamento sugere maior estabilidade no processo de geração de respostas quando estratégias de *chunking* semântico e metadados estruturados são empregadas, reduzindo a ocorrência de inferências não fundamentadas e aumentando a confiabilidade do sistema RAG.

A Figura 7 apresenta os resultados da métrica de Relevância da Resposta (*Answer Relevancy*) para as diferentes configurações experimentais consideradas. Essa métrica

avalia o grau de alinhamento entre a resposta gerada pelo sistema RAG e a pergunta original, penalizando respostas excessivamente verbosas, pouco focadas ou semanticamente desalinhadas.

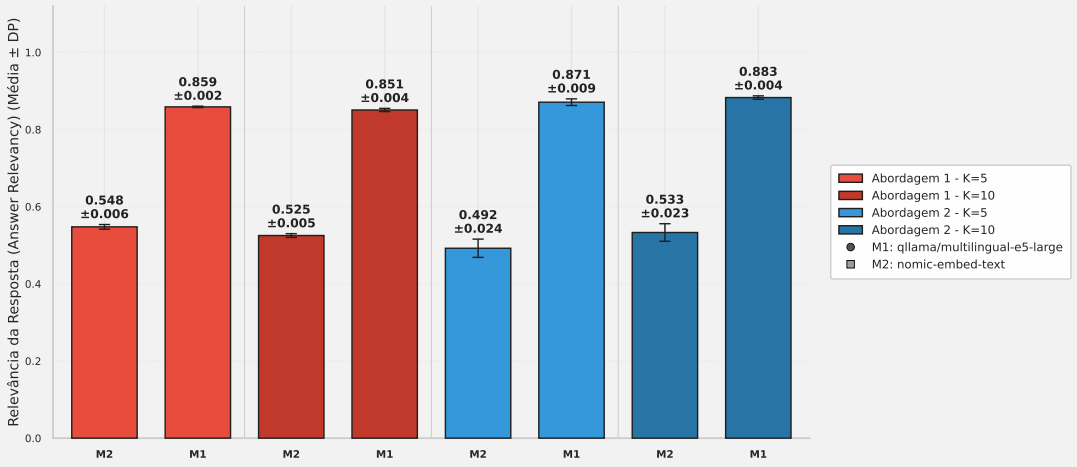

Figura 7. Resultados da métrica de Relevância da Resposta (Answer Relevancy) para as diferentes configurações experimentais, comparando a Abordagem 1 e a Abordagem 2, com os modelos de embedding M1 e M2 e valores de recuperação de contexto $K \in \{5, 10\}$.

Os resultados indicam que a Abordagem 2 apresenta desempenho superior em termos de relevância da resposta quando comparada à Abordagem 1, independentemente do modelo de *embedding* utilizado. O maior desempenho é observado na combinação da Abordagem 2 com o modelo M1, especialmente para $K = 10$, indicando que o enriquecimento semântico permite recuperar contextos mais diretamente relacionados à consulta, mesmo com um número reduzido de trechos. Observa-se também que o aumento de K tende a melhorar a relevância média das respostas em ambas as abordagens; contudo, na Abordagem 1 esse aumento vem acompanhado de maior variabilidade, sugerindo a introdução de ruído informacional. Em contraste, a Abordagem 2 mantém desvios padrão mais controlados, o que indica maior estabilidade e consistência na geração de respostas relevantes. Esses resultados reforçam que a preservação da estrutura semântica dos documentos e o uso de metadados contextuais contribuem significativamente para respostas mais focadas e alinhadas às perguntas em sistemas RAG.

A Figura 8 apresenta os resultados da métrica de Precisão do Contexto (*Context Precision*) para as diferentes configurações experimentais avaliadas. Essa métrica mede a proporção de trechos recuperados que são efetivamente relevantes para responder à consulta, refletindo o nível de ruído informacional presente no conjunto de contextos fornecido ao modelo de linguagem.

Os resultados evidenciam uma diferença significativa entre as abordagens no que se refere à precisão do contexto recuperado. A Abordagem 2 apresenta valores consistentemente superiores aos da Abordagem 1, especialmente quando combinada com o modelo M1, indicando que a recuperação baseada em enriquecimento semântico é capaz de selecionar trechos mais relevantes e reduzir a inclusão de contexto irrelevante. Observa-se ainda que, na Abordagem 1, o aumento do valor de K resulta em queda acentuada da precisão, sugerindo maior introdução de ruído à medida que mais trechos são recuperados.

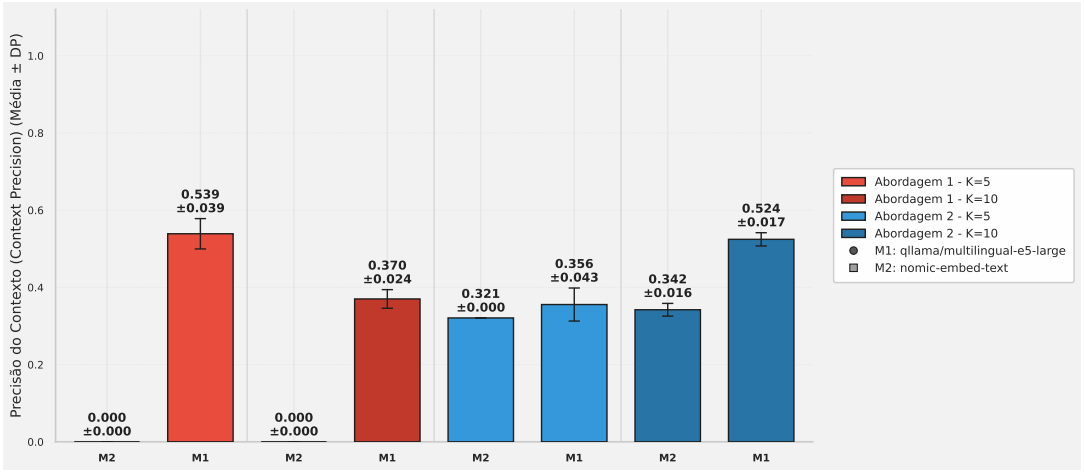

Figura 8. Resultados da métrica de Precisão do Contexto (Context Precision) para as diferentes configurações experimentais, comparando a Abordagem 1 e a Abordagem 2, com os modelos de embedding M1 e M2 e valores de recuperação de contexto $K \in \{5, 10\}$.

Em contraste, a Abordagem 2 mantém níveis elevados de precisão mesmo para $K = 10$, demonstrando maior robustez ao aumento do número de contextos. Esse comportamento indica que o *chunking* semântico e o uso de metadados estruturados contribuem para um ranking mais consistente de trechos relevantes, mitigando os efeitos negativos normalmente associados ao aumento de K em sistemas RAG.

A Figura 9 apresenta os resultados da métrica de Recall do Contexto (*Context Recall*) para as diferentes configurações experimentais analisadas. Essa métrica avalia o grau de cobertura dos trechos relevantes recuperados pelo sistema, indicando se o conjunto de contextos fornecido ao modelo de linguagem contém as informações necessárias para responder adequadamente à consulta.

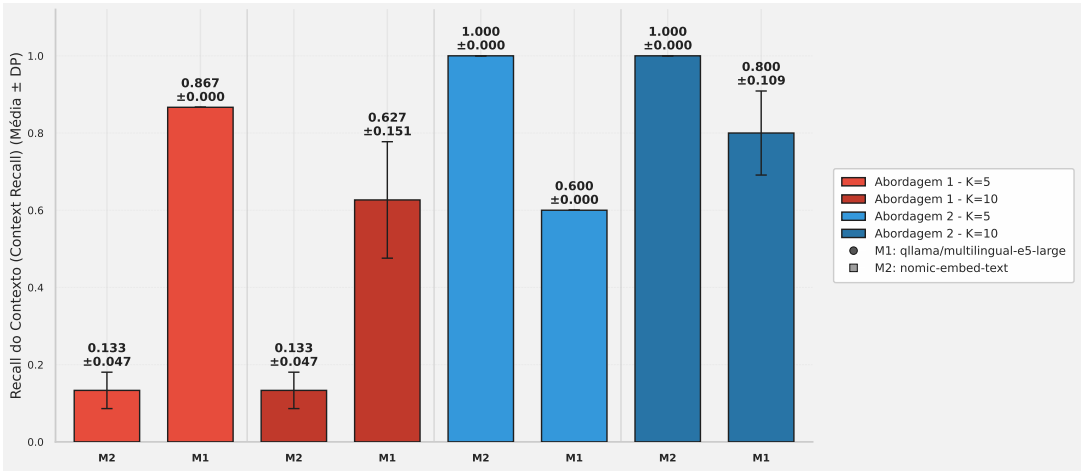

Figura 9. Resultados da métrica de Recall do Contexto (Context Recall) para as diferentes configurações experimentais, comparando a Abordagem 1 e a Abordagem 2, com os modelos de embedding M1 e M2 e valores de recuperação de contexto $K \in \{5, 10\}$.

A análise dos resultados mostra que a Abordagem 2 apresenta desempenho signifi-

cativamente superior em termos de recall do contexto quando comparada à Abordagem 1. Observa-se que, ao utilizar o modelo M2, a Abordagem 2 atinge valores máximos de recall, iguais a 1.0, tanto para $K = 5$ quanto para $K = 10$, indicando recuperação completa dos trechos relevantes necessários para a resposta. Em contraste, a Abordagem 1 apresenta valores de recall mais baixos e maior variabilidade, mesmo quando K é aumentado. Esse comportamento sugere que a segmentação uniforme tende a fragmentar informações semanticamente relacionadas, reduzindo a probabilidade de recuperação completa. A Abordagem 2, por sua vez, beneficia-se do *chunking* semântico e do enriquecimento com metadados estruturados, que permitem agrupar informações correlatas em unidades textuais mais coesas, aumentando a cobertura do contexto mesmo com um número reduzido de trechos recuperados.

A Figura 10 apresenta os resultados da métrica de Similaridade Semântica da Resposta (*Answer Semantic Similarity*) para as diferentes configurações experimentais avaliadas. Essa métrica quantifica o grau de alinhamento semântico entre a resposta gerada pelo sistema RAG e a resposta de referência (*ground truth*), sendo robusta a variações lexicais e paráfrases.

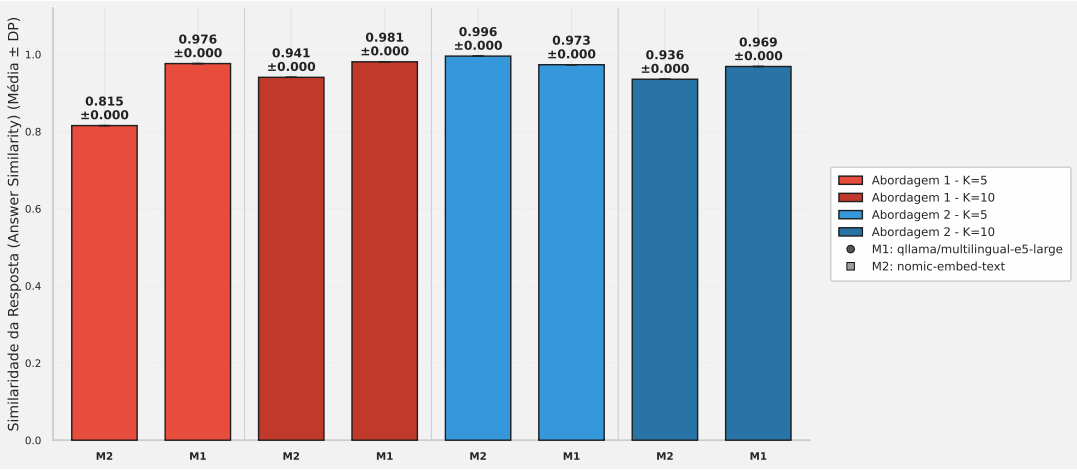

Figura 10. Resultados da métrica de Similaridade Semântica da Resposta (Answer Semantic Similarity) para as diferentes configurações experimentais, comparando a Abordagem 1 e a Abordagem 2, com os modelos de embedding M1 e M2 e valores de recuperação de contexto $K \in \{5, 10\}$.

A partir dos resultados obtidos, observa-se que ambas as abordagens alcançam valores elevados de similaridade semântica, refletindo a capacidade geral dos sistemas RAG em produzir respostas semanticamente alinhadas às referências. No entanto, a Abordagem 2 apresenta desempenho mais consistente, com médias elevadas e menor variabilidade, especialmente quando combinada com o modelo M1. Observa-se que, na Abordagem 1, embora os valores médios sejam altos, há maior dispersão entre execuções, sugerindo que a fragmentação do contexto pode introduzir instabilidade no processo de geração. A Abordagem 2, ao preservar unidades textuais semanticamente coesas por meio do *chunking* semântico e do enriquecimento com metadados estruturados, reduz a necessidade de inferências adicionais pelo modelo de linguagem, resultando em respostas mais estáveis e semanticamente próximas ao *ground truth*, mesmo para valores maiores de K.

A Figura 11 apresenta os resultados da métrica de Correção da Resposta (*Answer*

Correctness) para as diferentes configurações experimentais avaliadas. Essa métrica representa uma avaliação agregada da qualidade das respostas geradas pelo sistema RAG, combinando aspectos de consistência factual, alinhamento semântico e adequação em relação à resposta de referência.

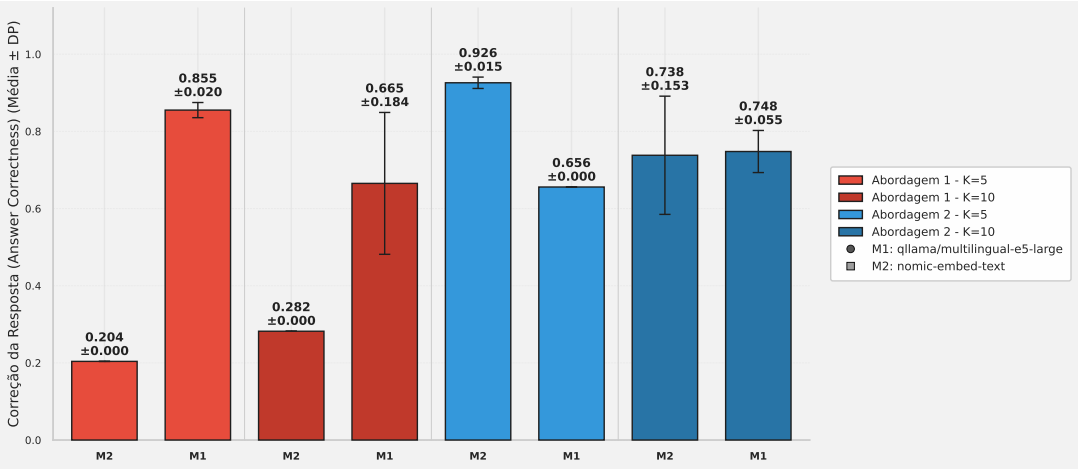

Figura 11. Resultados da métrica de Correção da Resposta (Answer Correctness) para as diferentes configurações experimentais, comparando a Abordagem 1 e a Abordagem 2, com os modelos de embedding M1 e M2 e valores de recuperação de contexto $K \in \{5, 10\}$.

A comparação entre as abordagens revela que existe uma separação clara entre as duas abordagens no que se refere à correção global das respostas. A Abordagem 2 apresenta desempenho substancialmente superior à Abordagem 1, especialmente quando combinada com tanto com modelo M1 como M2 e para $K = 5$ e $K = 10$, com baixa variabilidade entre execuções. Em contraste, a Abordagem 1 apresenta valores médios significativamente inferiores e maior dispersão, indicando maior ocorrência de respostas parcialmente corretas ou inconsistentes. Observa-se ainda que o aumento do valor de K não resulta em ganhos expressivos de correção na Abordagem 1, enquanto a Abordagem 2 mantém desempenho elevado mesmo com mais contexto, evidenciando robustez ao acréscimo de trechos recuperados. Esses resultados confirmam que o enriquecimento semântico e o *chunking* estruturado exercem impacto direto na qualidade final das respostas, tornando a Abordagem 2 mais adequada para aplicações práticas de sistemas RAG.

4. Conclusões

Este trabalho apresentou um estudo comparativo entre duas arquiteturas de indexação para sistemas de Recuperação Aumentada por Geração (RAG): uma abordagem tradicional monolítica, baseada em *chunking* uniforme, e uma abordagem modular com enriquecimento semântico, que incorpora segmentação estrutural e metadados contextuais. A avaliação experimental, conduzida de forma sistemática com diferentes modelos de *embedding*, valores de recuperação de contexto e métricas do framework RAGAS, demonstrou que a arquitetura modular supera consistentemente a abordagem monolítica em todas as métricas analisadas, apresentando ganhos tanto em desempenho médio quanto em estabilidade das respostas geradas. Os resultados evidenciam que decisões arquiteturais no processo de indexação exercem impacto direto na qualidade, consistência e robustez de

sistemas RAG, sendo a preservação de unidades textuais semanticamente coesas e o uso de metadados estruturados fatores determinantes para a redução de ruído informacional e aumento da confiabilidade das respostas. Embora o estudo tenha sido conduzido em um corpus específico de documentos acadêmicos, os achados indicam que abordagens semanticamente conscientes de indexação constituem um caminho promissor para o desenvolvimento de sistemas RAG mais robustos e adequados a aplicações colaborativas de acesso ao conhecimento, abrindo espaço para investigações futuras em outros domínios e cenários de uso.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Referências

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

Chase, H. (2023). Langchain: Building applications with large language models. *arXiv preprint arXiv:2308.00000*.

Chroma (2024). Chroma: The open-source embedding database. `https://www.trychroma.com`. Accessed: 2025-02.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In Aletras, N. and De Clercq, O., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

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

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. (2020). Retrieval augmented language model pre-training. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Hearst, M. A. (1997). Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

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

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

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

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Setty, S., Thakkar, H., Lee, A., Chung, E., and Vidra, N. (2024). Improving retrieval for rag based question answering models on financial documents. *arXiv preprint arXiv:2404.07221*.