Anais do Congresso Brasileiro de Iniciação Científica
(https://periodicos.unisanta.br/COB)

ISSN 2317-1316

INFLUÊNCIA DO TAMANHO DOS SEGMENTOS TEXTUAIS GERADOS POR SEGMENTAÇÃO RECURSIVA NO DESEMPENHO DA RECUPERAÇÃO DE INFORMAÇÕES EM SISTEMAS RAG

INFLUENCE OF RECURSIVE TEXT CHUNK SIZE ON INFORMATION RETRIEVAL PERFORMANCE IN RAG SYSTEMS

Sophia Ferreira Lima Leal, Nelson Nascimento Junior

Instituto Federal de Educação Ciência e Tecnologia de São Paulo, Campus Cubatão - Curso Técnico em Informática Integrado ao Ensino Médio
E-mail para contato: sophia.leal@ifsp.edu.br, nnascjunior@gmail.com

RESUMO – Os modelos de linguagem podem apresentar dificuldades ao responder perguntas que exigem informações específicas ou atualizadas. Este trabalho avalia o efeito do tamanho dos segmentos na recuperação de informações do Projeto Pedagógico do Curso Técnico em Informática Integrado ao Ensino Médio do IFSP campus Cubatão. Foram comparados segmentos de 300, 700 e 1.100 caracteres, com sobreposição de 120 caracteres, usando 60 perguntas em oito categorias. Avaliaram-se a posição do primeiro resultado correto, a MRR e o tempo médio de busca. Segmentos de 1.100 caracteres obtiveram a maior MRR (0,5359) e o menor tempo médio (26,36 ms). Os segmentos de 300 e 1.100 caracteres alcançaram 40% de acertos em primeiro lugar. Os resultados limitam-se às condições avaliadas.

Palavras-chave: Recuperação de informações; Retrieval-Augmented Generation (RAG); Segmentação textual; Embeddings; Processamento de linguagem natural.

ABSTRACT – Language models may struggle to answer questions that require specific or up-to-date information. This study evaluates the effect of chunk size on information retrieval from the Pedagogical Project of the Technical Course in Informatics Integrated with High School at the Federal Institute of São Paulo (IFSP), Cubatão campus. Chunks of 300, 700, and 1,100 characters, with an overlap of 120 characters, were compared using 60 questions across eight categories. The evaluation considered the rank of the first correct result, Mean Reciprocal Rank (MRR), and average search time. Chunks of 1,100 characters achieved the highest MRR (0.5359) and the lowest average search time (26.36 ms). Chunks of 300 and 1,100 characters both achieved a 40% rate of correct results ranked first. The findings are limited to the conditions evaluated.

Keywords: Information retrieval; Retrieval-Augmented Generation (RAG); Text chunking; Embeddings; Natural language processing.

ISSN 2317-1316

1 INTRODUÇÃO

A geração aumentada por recuperação (*Retrieval-Augmented Generation* — RAG) é uma abordagem que integra modelos de linguagem a mecanismos de recuperação de informações externas, permitindo que a geração de respostas utilize conteúdos obtidos de uma base documental. Lewis et al. (2020) apresentaram essa arquitetura como uma forma de combinar o conhecimento paramétrico dos modelos com informações armazenadas em uma memória não paramétrica. Sua aplicação é relevante em tarefas que exigem consultas a documentos específicos, nas quais a qualidade das respostas depende, entre outros fatores, da recuperação de informações pertinentes à consulta realizada.

Nesse contexto, a recuperação de informações constitui uma etapa fundamental dos sistemas RAG. Uma das abordagens utilizadas é a recuperação densa, na qual consultas e passagens documentais são convertidas em vetores numéricos, chamados *embeddings*, que representam características semânticas do texto. O sistema compara esses vetores por meio de uma medida de similaridade e ordena os trechos conforme sua proximidade em relação à consulta. Assim, uma pergunta sobre “economizar energia” pode recuperar um trecho que fale em “reduzir o consumo de eletricidade”, sem repetir exatamente as mesmas palavras. Karpukhin et al. (2020) demonstraram a aplicação dessa abordagem na recuperação de passagens para sistemas de perguntas e respostas de domínio aberto. O desempenho dessa etapa também depende da preparação dos documentos disponibilizados para consulta, especialmente da forma como seu conteúdo é segmentado e indexado.

A segmentação textual, denominada *chunking*, consiste na divisão de documentos em unidades menores que podem ser recuperadas individualmente. Entre as estratégias possíveis, a segmentação por tamanho fixo permite estabelecer limites para a extensão dos trechos, enquanto a sobreposição entre segmentos consecutivos pode preservar parte do contexto compartilhado entre eles. Apesar de sua simplicidade operacional, essa abordagem pode produzir trechos que separam informações relacionadas ou agrupam conteúdos distintos. Assim, o tamanho dos segmentos representa uma variável relevante para a organização da base documental e para a recuperação posterior das informações.

Pesquisas recentes investigam estratégias alternativas para aprimorar essa etapa. Kiss, Nagy e Szilágyi (2025) propuseram o método Max–Min de segmentação semântica, que utiliza similaridade entre sentenças para formar segmentos semanticamente coerentes. Os autores

ISSN 2317-1316

avaliaram a abordagem em diferentes conjuntos de dados, incluindo uma tarefa de perguntas e respostas baseadas em RAG, e observaram melhorias em comparação com métodos de referência. Em outro estudo, Gomez-Cabello et al. (2025) compararam estratégias de segmentação fixa, semântica, por proposições e adaptativa em um sistema RAG aplicado ao apoio à decisão clínica. Nesse contexto, a segmentação adaptativa apresentou os melhores resultados de acurácia e relevância entre as estratégias avaliadas. Em contrapartida, Qu, Tu e Bao (2025) observaram que a segmentação semântica não proporcionou ganhos suficientemente consistentes para justificar seu custo computacional. Em conjunto, esses estudos demonstram que os resultados dependem das estratégias e das condições experimentais consideradas, não sendo possível pressupor que uma única abordagem seja a mais adequada para todas as aplicações.

Diante desse cenário, este trabalho investiga como diferentes tamanhos de segmentos textuais influenciam a recuperação de informações em um sistema aplicado ao Projeto Pedagógico do Curso (PPC) do Técnico em Informática Integrado ao Ensino Médio do IFSP, campus Cubatão. Esse documento reúne informações sobre componentes curriculares, ementas, objetivos, conteúdos programáticos e bibliografia, que precisam ser localizadas a partir de consultas em linguagem natural. Considera-se a hipótese de que a variação do tamanho dos segmentos pode alterar o desempenho da recuperação, pois modifica a quantidade de contexto disponível em cada unidade indexada. Essa hipótese é examinada experimentalmente, sem pressupor antecipadamente a superioridade de uma das configurações.

Para investigar essa questão, são comparados três tamanhos de segmentos: 300, 700 e 1.100 caracteres, produzidos por uma estratégia de segmentação recursiva, com sobreposição de 120 caracteres. As configurações são avaliadas por meio de um conjunto de 60 perguntas distribuídas em oito categorias, utilizando recuperação por similaridade do cosseno e limite de dez resultados por consulta. O critério de acerto considera a correspondência entre o componente curricular e o campo documental esperados. O desempenho é analisado pela posição do primeiro resultado correto, pela *Mean Reciprocal Rank*[1] (MRR) e pelo tempo médio

[1] Mean Reciprocal Rank (MRR) é a média do inverso da posição do primeiro resultado correto em cada consulta. Valores mais próximos de 1 indicam que os resultados corretos aparecem em posições mais altas; consultas sem resultado correto recebem valor zero.

ISSN 2317-1316

de busca. A avaliação concentra-se na etapa de recuperação, sem medir a qualidade das respostas geradas por um modelo de linguagem.

O objetivo deste trabalho é analisar a influência do tamanho dos segmentos textuais no desempenho da recuperação de informações em um sistema RAG aplicado ao Projeto Pedagógico do Curso (PPC) de um curso técnico integrado ao ensino médio, utilizando a estratégia de segmentação recursiva. Para isso, são comparadas três configurações de tamanho de segmentos, mantendo-se os demais parâmetros controlados. Busca-se identificar diferenças no desempenho da recuperação e fornecer subsídios para a escolha do tamanho dos segmentos em aplicações semelhantes de consulta a documentos curriculares.

2 MATERIAIS E MÉTODOS

2.1 Caracterização da pesquisa e ambiente computacional

A pesquisa caracteriza-se como experimental, de abordagem quantitativa, uma vez que investiga a influência do tamanho dos segmentos de texto (*chunks*) no desempenho da recuperação de informações em um sistema baseado em Retrieval-Augmented Generation (RAG). O estudo foi conduzido por meio de experimentos computacionais, nos quais diferentes configurações de segmentação foram submetidas ao mesmo conjunto de perguntas e avaliadas a partir de métricas de recuperação.

O sistema foi desenvolvido em Python 3.14.7, utilizando Streamlit 1.64.0 para a interface da aplicação, pandas 3.0.5 e NumPy 2.5.3 para a manipulação dos dados, e Sentence Transformers 6.0.1 para a geração de representações vetoriais dos textos. Foi utilizado o modelo *paraphrase-multilingual-MiniLM-L12-v2* para a geração dos *embeddings*. O cálculo da similaridade do cosseno foi realizado com o auxílio da biblioteca scikit-learn 1.9.1. O ambiente computacional também incluía PyTorch 2.14.0 e Transformers 5.17.0, utilizados como suporte à execução dos modelos de inteligência artificial envolvidos no processamento dos textos e na geração de *embeddings*.

A variável independente do experimento foi o tamanho dos segmentos textuais, com valores de 300, 700 e 1.100 caracteres. As variáveis de resposta foram a distribuição dos primeiros resultados corretos nas faixas de posição estabelecidas, a *Mean Reciprocal Rank*

ISSN 2317-1316

(MRR) e o tempo de execução da busca. Foram mantidos constantes o conjunto de perguntas, o modelo de *embeddings*, o critério de similaridade e o valor de Top-K[2], a fim de permitir a comparação entre as configurações.

2.2 Base de dados e conjunto de perguntas

A base documental utilizada nos experimentos foi o Projeto Pedagógico de Curso (PPC) do curso Técnico em Informática Integrado ao Ensino Médio. O documento contém informações curriculares e institucionais, incluindo ementas, objetivos, conhecimentos, bibliografia e informações estruturais do curso.

Para avaliar a recuperação das informações, foi utilizado um conjunto de 60 perguntas, distribuídas em oito categorias: áreas de integração, bibliografia, conhecimentos, conteúdo, ementa, informações estruturais, objetivos e consultas semânticas.

Para cada pergunta, foram estabelecidos previamente os critérios esperados de avaliação, correspondentes ao componente curricular e ao campo documental que deveriam ser recuperados. Esses critérios foram utilizados exclusivamente na etapa de avaliação dos resultados, após a execução da busca, não sendo empregados para orientar ou restringir a recuperação.

2.3 Configurações experimentais

O experimento investigou três configurações de segmentação textual, com tamanhos de *chunks* de 300, 700 e 1.100 caracteres. Em todas as configurações, foi utilizado um *overlap* de 120 caracteres entre segmentos consecutivos, com o objetivo de preservar parte do contexto textual nas fronteiras entre os segmentos.

Após a segmentação, os textos foram representados vetorialmente por meio do modelo paraphrase-multilingual-MiniLM-L12-v2, que gera *embeddings* para representar o conteúdo textual. A recuperação foi realizada comparando o vetor da pergunta com os vetores dos

[2] Top-K corresponde ao número máximo de segmentos que o sistema retorna para cada consulta, ordenados por similaridade com a pergunta. Neste estudo, Top-K = 10 significa que são considerados até dez segmentos por pergunta.

ISSN 2317-1316

segmentos documentais por meio da similaridade do cosseno. Com base nesses valores, os segmentos foram ordenados para que os mais semelhantes à consulta aparecessem primeiro. Portanto, a recuperação avaliada neste estudo corresponde à localização e classificação dos trechos potencialmente relevantes, e não à geração de uma resposta por um modelo de linguagem.

Para manter a comparabilidade entre as configurações, foram utilizadas as mesmas 60 perguntas, o mesmo modelo de *embeddings*, o mesmo critério de similaridade e o mesmo valor de Top-K, definido como 10. Dessa forma, cada consulta retornou até dez segmentos classificados segundo sua similaridade com a pergunta.

As configurações experimentais diferiram, portanto, no tamanho dos segmentos utilizados para representar o documento, permitindo investigar a influência dessa variável sobre a posição dos resultados relevantes e o tempo de recuperação.

2.4 Procedimentos de execução e coleta dos dados

Para cada configuração experimental, as 60 perguntas foram submetidas ao mecanismo de recuperação. Os resultados foram registrados para posterior análise, incluindo a pergunta consultada, os segmentos recuperados, suas posições no ranking, os componentes curriculares e campos documentais associados, os valores de similaridade e o tempo de execução da busca.

O tempo de recuperação foi mensurado com a função time.time(), registrando-se o instante imediatamente anterior e posterior à chamada da função responsável pela busca. A diferença entre os instantes foi convertida para milissegundos. Assim, a medida corresponde ao tempo de execução da função de recuperação, não abrangendo necessariamente todas as etapas de preparação dos dados, geração de *embeddings* ou inicialização do modelo.

Também foram realizadas execuções repetidas, com comparação dos resultados completos obtidos. Nas repetições verificadas, não foram observadas diferenças nos resultados registrados. Esse procedimento permitiu verificar a consistência das saídas nas condições testadas.

ISSN 2317-1316

2.5 Critérios e métricas de avaliação

A avaliação da recuperação considerou a correspondência simultânea entre o componente curricular esperado e o campo documental esperado para cada pergunta. Um resultado foi considerado correto quando atendia aos dois critérios. Caso apenas um deles correspondesse ao esperado, o resultado não era contabilizado como acerto.

Para cada consulta, foi identificada a primeira posição do ranking em que apareceu um resultado que satisfazia ambos os critérios. A partir dessa posição, foram calculadas as proporções de acertos em três faixas distintas: primeiro lugar, posições de 2 a 5 e posições de 6 a 10.

A métrica **Acertos em 1º lugar (%)** representa a proporção de perguntas cujo primeiro resultado recuperado atende aos critérios de avaliação. A métrica **Acertos entre 2º e 5º (%)** corresponde à proporção de perguntas cujo primeiro resultado correto aparece entre a segunda e a quinta posições. Já **Acertos entre 6º e 10º (%)** representa a proporção de perguntas cujo primeiro resultado correto aparece entre a sexta e a décima posições.

As proporções foram calculadas em relação ao total de 60 perguntas, permitindo comparar a distribuição dos primeiros resultados corretos entre as configurações de segmentação.

Também foi utilizada a *Mean Reciprocal Rank* (MRR), que considera a posição do primeiro resultado correto em cada consulta. Para uma pergunta (q), o *Reciprocal Rank* é definido como o inverso da posição (r_q) do primeiro resultado correto. Quando nenhum resultado correto é encontrado entre os dez primeiros, o valor atribuído à consulta é zero. A MRR corresponde à média desses valores para todas as perguntas:

$$\boldsymbol{MRR = \frac{1}{N}\frac{\sum_{q=1}^{N} 1}{r_q}}$$

em que (N) representa o número total de perguntas e (r_q) corresponde à posição do primeiro resultado correto para a pergunta (q), considerando-se valor zero para a contribuição da consulta quando não há resultado correto no Top-10.

ISSN 2317-1316

A utilização conjunta das faixas de acerto e da MRR permite analisar tanto a distribuição dos primeiros resultados corretos quanto a posição em que esses resultados aparecem no ranking de recuperação.

2.6 Organização e análise dos resultados

Os dados coletados foram organizados em arquivos estruturados para permitir a comparação entre as três configurações de segmentação. A análise foi conduzida de forma quantitativa e descritiva, considerando as proporções de acerto por faixa de posição, os valores de MRR e os tempos de recuperação.

Os resultados das configurações de 300, 700 e 1.100 caracteres foram comparados mantendo-se constantes o conjunto de perguntas e os parâmetros de recuperação. Essa estratégia possibilitou examinar as diferenças observadas associadas aos tamanhos de *chunks*, sem alterar simultaneamente os demais parâmetros experimentais.

As conclusões foram limitadas às condições avaliadas, considerando o documento utilizado, o conjunto de perguntas, o modelo de *embeddings* e os parâmetros definidos. Dessa forma, os resultados permitem discutir o comportamento da recuperação nesse cenário experimental, sem pressupor que a configuração de melhor desempenho necessariamente apresentará os mesmos resultados em outras bases documentais ou sistemas.

3 RESULTADOS E DISCUSSÃO

Os experimentos foram realizados com o objetivo de comparar o desempenho da recuperação de informações em um documento de Projeto Pedagógico de Curso (PPC), utilizando três configurações de segmentação textual: 300, 700 e 1.100 caracteres por *chunk*. Para cada configuração, foram utilizadas as mesmas 60 perguntas, permitindo comparar a posição dos resultados relevantes e o tempo médio de busca.

Tabela 1: Comparação dos resultados obtidos nas configurações de segmentação textual

ISSN 2317-1316

Tamanho do segmento (caracteres)	Número de segmentos	Acertos em 1º lugar (%)	Acertos entre 2º e 5º lugar (%)	Acertos entre 6º e 10º lugar (%)	MRR	Tempo médio (ms)
300	996	40,0	28,3	13,3	*0,5319*	*29,03*
700	512	38,3	25,0	18,3	*0,5163*	*33,08*
1.100	453	40,0	30,0	16,7	*0,5359*	*26,36*

Fonte: Elaborado pelos autores com base nos resultados experimentais.

Conforme apresentado na Tabela 1, o tamanho dos chunks influenciou os resultados da recuperação e o tempo médio de busca. A configuração de 300 caracteres gerou 996 segmentos, enquanto as configurações de 700 e 1.100 caracteres produziram 512 e 453 segmentos, respectivamente. Essa diferença demonstra que o aumento do tamanho dos segmentos reduziu a quantidade de unidades disponíveis para recuperação no documento analisado.

Em relação à posição dos resultados relevantes, as configurações de 300 e 1.100 caracteres apresentaram 40,0% de acertos em primeiro lugar, enquanto a configuração de 700 caracteres obteve 38,3%. Na faixa entre a segunda e a quinta posição, o maior percentual foi observado na configuração de 1.100 caracteres, com 30,0%, seguida pelas configurações de 300 e 700 caracteres, com 28,3% e 25,0%, respectivamente. Já entre a sexta e a décima posição, a configuração de 700 caracteres apresentou o maior percentual, de 18,3%. Esse resultado indica que parte dos resultados relevantes dessa configuração apareceu em posições mais baixas do conjunto recuperado.

O *Mean Reciprocal Rank* (MRR), métrica que considera a posição do primeiro resultado correto, apresentou valores de 0,5359 para *chunks* de 1.100 caracteres, 0,5319 para *chunks* de 300 caracteres e 0,5163 para *chunks* de 700 caracteres. Dessa forma, a configuração de 1.100 caracteres obteve o maior valor de MRR, embora a diferença em relação à configuração de 300 caracteres tenha sido pequena.

Quanto ao tempo médio de busca, a configuração de 1.100 caracteres apresentou o menor valor, de 26,36 ms, seguida pela configuração de 300 caracteres, com 29,03 ms, e pela configuração de 700 caracteres, com 33,08 ms. Portanto, nos experimentos realizados, os segmentos de 1.100 caracteres combinaram o maior MRR com o menor tempo médio de busca. Esses tempos correspondem à execução da busca, não incluindo todas as etapas de preparação do sistema.

ISSN 2317-1316

Os resultados devem ser interpretados considerando que cada tamanho de segmento apresenta características próprias. Gomez-Cabello et al. (2025), ao investigarem estratégias de *chunking* para sistemas RAG aplicados ao apoio à decisão clínica, identificaram vantagens da segmentação adaptativa em seu contexto de estudo. Em contrapartida, Qu, Tu e Bao (2025) observaram que dividir os textos de acordo com seu significado nem sempre melhora os resultados da busca o suficiente para compensar o processamento adicional exigido por essa técnica. Essas conclusões não são necessariamente contraditórias, pois os estudos investigaram estratégias e contextos distintos, evidenciando que a eficácia da segmentação depende das características dos documentos, das tarefas e dos sistemas avaliados.

Kiss, Nagy e Szilágyi (2025) também investigaram o impacto do tamanho dos segmentos na recuperação de informações em aplicações RAG. Entretanto, os autores utilizaram uma abordagem de segmentação semântica, enquanto este trabalho adota a segmentação recursiva. Essa diferença contribui para a discussão sobre como as estratégias de segmentação e os limites dos segmentos podem influenciar a recuperação de informações. Nesse contexto, Reimers e Gurevych (2019) apresentaram o Sentence-BERT, método para produzir representações vetoriais de sentenças que possibilitam comparar sua similaridade semântica. Karpukhin et al. (2020), por sua vez, demonstraram a aplicação de recuperação densa de passagens em tarefas de perguntas e respostas, enquanto Lewis et al. (2020) formalizaram a arquitetura RAG, que combina recuperação de documentos com geração de respostas. Esses trabalhos fundamentam as técnicas de representação vetorial e recuperação utilizadas em sistemas dessa natureza.

Além disso, Bossenz et al. (2025) avaliaram estratégias de segmentação e *embeddings* para a recuperação de documentos administrativos em um contexto hospitalar, mostrando a relevância da escolha dessas configurações para o desempenho do sistema. Em conjunto, os estudos reforçam a importância de avaliar empiricamente as estratégias de segmentação e representação textual, sem pressupor que uma configuração seja a mais adequada para todos os cenários.

Em conjunto, os resultados indicam que o desempenho da recuperação não aumentou linearmente com o tamanho dos *chunks*. Embora a configuração de 1.100 caracteres tenha apresentado os maiores valores observados de MRR e o menor tempo médio de busca, a configuração de 300 caracteres obteve o mesmo percentual de acertos em primeiro lugar e resultados próximos de MRR. Já a configuração de 700 caracteres apresentou o menor MRR e

ISSN 2317-1316

o maior tempo médio de busca. Essas diferenças descrevem o comportamento observado nos experimentos e não permitem afirmar superioridade estatística entre as configurações, uma vez que não foi apresentada uma análise inferencial.

Como limitações, destaca-se a utilização de um único PPC e de um conjunto de 60 perguntas, o que restringe a generalização dos resultados. Além disso, foram comparados apenas tamanhos fixos de segmentação. Trabalhos futuros poderão incluir documentos de diferentes áreas, ampliar o conjunto de perguntas e avaliar estratégias semânticas e adaptativas, bem como diferentes modelos de *embeddings* e mecanismos de recuperação, a fim de verificar se os resultados se mantêm em outros contextos documentais. Além disso, a avaliação se restringiu à recuperação dos trechos, não contemplando a qualidade das respostas eventualmente geradas a partir deles.

4 CONCLUSÃO

O experimento permitiu comparar o desempenho da recuperação de informações em um sistema RAG com *chunks* de 300, 700 e 1.100 caracteres, utilizando um PPC e 60 perguntas. A configuração de 1.100 caracteres apresentou o maior valor de MRR (0,5359) e o menor tempo médio de busca (26,36 ms), enquanto as configurações de 300 e 1.100 caracteres obtiveram o mesmo percentual de acertos em primeiro lugar (40,0%). A configuração de 700 caracteres apresentou o menor MRR (0,5163) e o maior tempo médio de busca (33,08 ms). Esses resultados indicam que, nas condições avaliadas, o aumento do tamanho dos *chunks* não produziu uma melhora linear no desempenho, e as diferenças observadas não permitem afirmar superioridade estatística entre as configurações. Como trabalhos futuros, propõe-se ampliar a avaliação para outros documentos e conjuntos de perguntas, além de investigar estratégias de segmentação semântica e adaptativa.

Agradece-se ao Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) pelo apoio à realização deste trabalho, por meio do Programa Institucional de Bolsas de Iniciação Científica e Tecnológica (PIBIFSP).

5 DECLARAÇÃO DE USO DE IA

Durante a elaboração deste trabalho, foram utilizadas as ferramentas Perplexity AI, para

ISSN 2317-1316

auxiliar no mapeamento e na localização de referências bibliográficas; e ChatGPT (modelo GPT-6), para revisão gramatical, para auxiliar na formatação e adequação da linguagem à norma-padrão.

O conteúdo foi revisado, editado e validado pelos autores, que assumem total responsabilidade pelo texto final.

6 REFERÊNCIAS

BOSSENZ, J. et al. Evaluation of Chunking and Embedding Strategies for Local Document Retrieval Using an Open-Source LLM in a Hospital. In: *German Medical Data Sciences 2025: GMDS Illuminates Health*. 2025. p. 91–100. DOI: https://doi.org/10.3233/SHTI251383.

GOMEZ-CABELLO, C. A. et al. Comparative Evaluation of Advanced Chunking for Retrieval-Augmented Generation in Large Language Models for Clinical Decision Support. *Bioengineering*, v. 12, n. 11, art. 1194, 2025. DOI: https://doi.org/10.3390/bioengineering12111194.

KARPUKHIN, V. et al. Dense Passage Retrieval for Open-Domain Question Answering. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020. p. 6769–6781. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.550.

KISS, C.; NAGY, M.; SZILÁGYI, P. Max–Min semantic chunking of documents for RAG application. *Discover Computing*, v. 28, art. 117, 2025. DOI: https://doi.org/10.1007/s10791-025-09638-7.

LEWIS, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: *Advances in Neural Information Processing Systems*. v. 33, p. 9459–9474, 2020. Disponível em: https://arxiv.org/abs/2005.11401.

QU, R.; TU, R.; BAO, F. S. Is Semantic Chunking Worth the Computational Cost? In: *Findings of the Association for Computational Linguistics: NAACL 2025*. 2025. p. 2155–2177. DOI: https://doi.org/10.18653/v1/2025.findings-naacl.114.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019. p. 3982–3992. DOI: https://doi.org/10.18653/v1/D19-1410.