Orclever Proceedings of
Research and Development

https://doi.org/10.56038/oprd.v5i1.516

OR CLEVER
Science & Research Group

Conference Article

Enhancing Retrieval-Augmented Generation Accuracy with Dynamic Chunking and Optimized Vector Search

Derya TANYILDIZ[1], Serkan AYVAZ[2] , Mehmet Fatih AMASYALI[3]

[1] Computer Engineering Department, Yildiz Technical University, Turkey,
Orcid ID: https://orcid.org/0009-0009-2802-5262, e-mail: derya.tanyildiz@std.yildiz.edu.tr
[2] Centre for Industrial Software, University of Southern Denmark, Denmark,
Orcid ID: https://orcid.org/0000-0003-2016-4443, e-mail: seay@mmmi.sdu.dk
[3] Computer Engineering Department, Yildiz Technical University, Turkey,
Orcid ID: https://orcid.org/0000-0002-0404-5973, e-mail: amasyali@yildiz.edu.tr

[*] Correspondence: deryatanyildiz96 @gmail.com; Tel.: +90 538 893 55 38

Received 15 October 2024
Received in revised form 26 November 2024
In final form 28 December 2024

4th International Conference on Design, Research and Development
(RDCONF 2024)
December 19 - 20, 2024

Reference: Tanyıldız, D., Ayvaz, S., & Amasyalı, M. F. (2024). Enhancing retrieval-augmented generation accuracy with dynamic chunking and optimized vector search. Orclever Proceedings of Research and Development, 5*(1)*, 215-225.

Abstract

Retrieval-Augmented Generation (RAG) architectures depend on the integration of efficient retrieval and ranking mechanisms to enhance response accuracy and relevance. This study investigates a novel approach to improving the response performance of RAG systems, leveraging dynamic chunking for contextual coherence, Sentence-Transformers (all-mpnet-base-v2) for high-quality embeddings, and cross-encoder-based re-ranking for retrieval refinement. Our evaluation utilizes RAGAS metrics to assess key performance metrics, including faithfulness, relevancy, correctness, and context precision. Empirical evaluations highlighted the significant impact of index choice on the performance. Our proposed approach integrates the FAISS HNSW index with re-ranking, resulting in a balanced architecture that improves response fidelity without compromising efficiency. These insights underscore the importance of advanced indexing and retrieval techniques in bridging the gap between large-scale language models and domain-specific information needs. The findings provide a robust framework for future research in optimizing RAG systems, particularly in scenarios requiring high-context preservation and precision.

Keywords: *Retrieval-Augmented Generation (RAG), Cross Encoders, Vector Database, Dynamic Chunking, Semantic Search*

1. Introduction

In recent years, the proliferation of large-scale pre-trained language models (LLMs) has significantly advanced the field of natural language processing (NLP). These models, with their capacity to generate coherent and contextually relevant text, have found applications across diverse domains [1] . However, their reliance on static, fixed knowledge representations often limits their ability to retrieve and utilize the most current and domain-specific information. To address this challenge, Retrieval-Augmented Generation (RAG) architectures have emerged as a promising solution, integrating external retrieval mechanisms with LLMs to provide accurate and contextually relevant responses [2]. Despite their potential, the performance of RAG systems is highly contingent on the quality of the retrieval and ranking processes, making it critical to optimize these components for enhanced response accuracy and relevance.

The standard RAG pipeline comprises two main stages: retrieval and generation. The retrieval stage identifies relevant pieces of information from a vast corpus, while the generation stage synthesizes this information into coherent responses. A key limitation in traditional RAG implementations lies in their reliance on static chunking strategies, where textual data is split into uniform segments. While this approach simplifies retrieval, it often leads to a loss of contextual integrity, especially for queries requiring nuanced understanding. Moreover, the retriever-reader imbalance—where the retriever is computationally intensive but lacks contextual awareness—further hampers the system's performance. This paper addresses these limitations by proposing an optimized RAG architecture that incorporates dynamic chunking, advanced embedding techniques, and state-of-the-art indexing methods in vector databases.

This research makes several key contributions to the field of RAG and NLP. First, it introduces a dynamic chunking strategy that addresses the limitations of static chunking by preserving contextual integrity. Second, it demonstrates the effectiveness of advanced embedding techniques using Sentence-Transformers for improving retrieval accuracy. Third, it provides a detailed comparison of indexing methods in FAISS, with empirical evidence supporting the superiority of the HNSW index, specifically in the context of the Turkish language. Finally, the integration of a cross-encoder-based re-ranking mechanism ensures high-quality responses, paving the way for more reliable and context-aware RAG systems. By evaluating key components, the research highlights the applicability and performance of known methodologies in addressing real-world information retrieval challenges for Turkish data. The findings contribute to understanding how these techniques can be effectively adapted and integrated, offering a practical framework for future studies in this domain.

2. Materials and Methods

This section presents the methodologies utilized to enhance the Retrieval-Augmented Generation (RAG) architecture, focusing on embedding techniques, dynamic text segmentation, vector database optimization, re-ranking mechanisms, and rigorous performance evaluation. The integrated framework aims to maximize retrieval accuracy and response quality through advanced and complementary components.

2.1. Data and Dataset Configuration

For this study, we utilized the Metin/WikiRAG-TR dataset, a curated corpus designed to evaluate the effectiveness of RAG architectures in retrieving and generating contextually accurate responses [3]. To ensure the evaluation focused on the model's ability to retrieve and leverage relevant context, a subset of 15 targeted questions was carefully selected from the dataset. These questions were specifically designed such that generating a correct answer would be infeasible without access to the corresponding context. This configuration ensured that the responses relied heavily on the retrieved context, minimizing the possibility of context-independent, generic answers from the model.

The selection process prioritized questions that required precise semantic understanding and alignment with the context. This experimental setup aimed to test the core strengths of the RAG framework:

1. **Context Recall:** Evaluating how effectively the retrieval module identifies and extracts the most relevant chunks.
2. **Answer Faithfulness:** Ensuring that the generated responses are grounded in the provided context.
3. **Model Accuracy:** Measuring the overall correctness of the answers in relation to the original dataset.

Through this approach, the dataset provided a robust foundation for assessing and improving the RAG architecture's ability to integrate retrieval with generative capabilities.

2.2. Architecture Design for Optimized RAG Model with FAISS

The proposed architecture aims to enhance the response accuracy and retrieval efficiency of Retrieval-Augmented Generation (RAG) systems by incorporating advanced embedding techniques, dynamic chunking, and optimized vector database indexing. The integration of these components ensures that the system delivers high-quality and contextually relevant responses [4]. Figure 1 illustrates an overview of the proposed system architecture with the enhanced RAG model.

2.2.1 Document Management Pipeline

The document management pipeline consists of several interconnected steps:

Data Ingestion and Chunking: In the phase, the input documents are first processed by a document loader, where they are prepared for downstream operations. Instead of traditional static chunking methods, dynamic chunking is employed in the proposed approach. This method segments the text based on semantic coherence, preserving contextual relationships and ensuring that relevant information is not fragmented across chunks. By maintaining the integrity of semantic units, dynamic chunking significantly improves retrieval accuracy [5].

Embedding with Sentence-Transformers: The dynamically chunked text is embedded using the Sentence-Transformers model (all-mpnet-base-v2), a state-of-the-art model optimized for generating dense vector representations [6]. These embeddings capture semantic similarity and enable the system to retrieve conceptually relevant documents rather than relying solely on surface-level keyword matching.

Storage and Retrieval in FAISS Vector Database: The embeddings are stored in a FAISS (Facebook AI Similarity Search) vector database, which facilitates efficient and scalable similarity searches. Two indexing methods are implemented and evaluated: IndexIVFFlat and IndexHNSWFlat [7].

IndexIVFFlat provides a balance between computational efficiency and retrieval accuracy by partitioning the vector space into clusters. On the other hand, IndexHNSWFlat constructs a hierarchical navigable small-world graph for nearest neighbor search, achieving superior performance in both accuracy and speed. The HNSW indexing method is known to outperform IVFFlat, particularly in tasks requiring precise semantic retrieval, as demonstrated in performance evaluations [8].

2.2.2. LLM Prompting and Answer Generation Pipeline

This pipeline outlines how the system generates responses to user queries by leveraging the document embeddings:

Query Processing and Retrieval: User queries are processed and transformed into embeddings using the same Sentence-Transformers model to ensure compatibility with stored document embeddings. A similarity search is performed in the FAISS vector database to retrieve the top_k1 relevant chunks. These chunks represent the most contextually aligned segments of data corresponding to the user query.

Re-Ranking with Cross-Encoder: The top_k1 chunks retrieved from the vector database undergo a re-ranking process using a cross-encoder. This cross-encoder model evaluates the relevance of each chunk in relation to the user query, scoring them based on semantic and contextual alignment. The highest-scoring chunks (top_k2) are selected for further processing, ensuring that only the most relevant information is passed to the next stage [9].

Response Generation with a Large Language Model (LLM): The selected chunks, now re-ranked for relevance, are provided as context to a fine-tuned large language model (e.g., Turkish-Llama-8b-DPO-v0.1 [10] or Turkish-Llama-8b-Instruct-v0.1 [11]). The LLM synthesizes this context with the user query to generate a coherent and accurate response, as discussed in [12]. The output is designed to maximize faithfulness, answer relevancy, and correctness, aligning with predefined evaluation metrics.

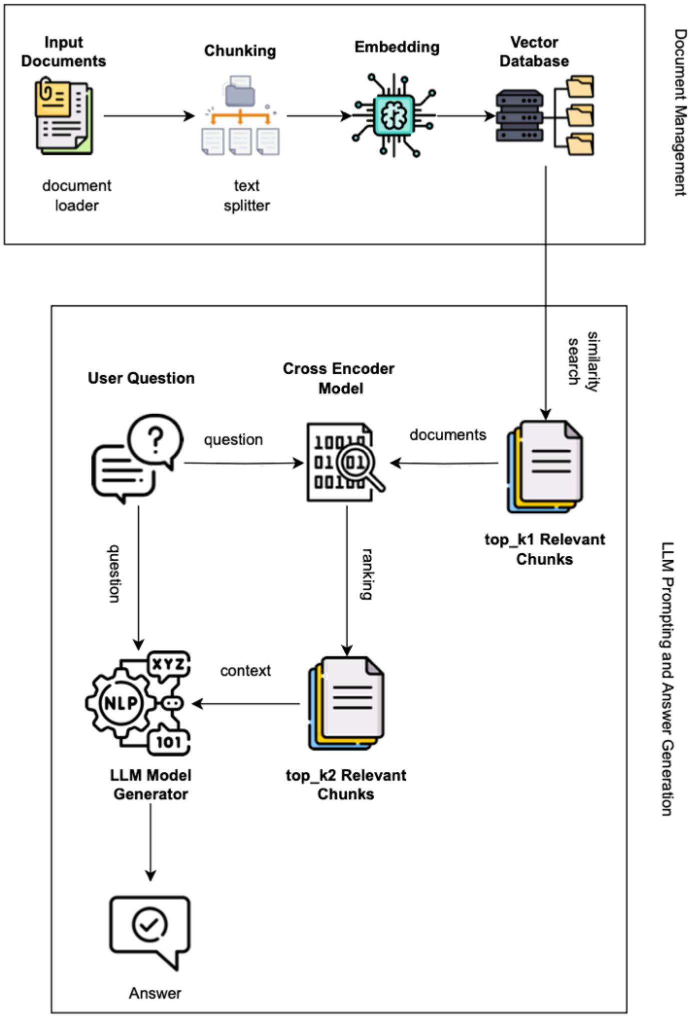

Figure 1: System Architecture of the Enhanced RAG Model

2.3. RAGAS Metrics: Evaluation Framework for Retrieval-Augmented Generation

Retrieval-Augmented Generation Assessment Scores (RAGAS) is a comprehensive evaluation framework designed to assess the performance of Retrieval-Augmented Generation (RAG) systems [13]. The framework employs a diverse set of metrics that collectively quantify the quality, accuracy, and relevance of responses generated by the model. These metrics provide critical insights into the strengths and weaknesses of the RAG architecture, guiding optimization efforts for improved performance. Below, the key components of RAGAS are defined and contextualized in relation to their significance in evaluating RAG systems [14].

Faithfulness measures the degree to which the model's response aligns with the information provided in the retrieved context. This metric ensures that the generated output does not introduce hallucinations or inaccuracies that deviate from the supporting evidence. High faithfulness scores indicate that the model reliably bases its answers on factual content retrieved during the query process.

$$Faithfulness = \frac{Number\ of\ statements\ supported\ by\ the\ retrieved\ context}{Total\ number\ of\ statements\ in\ the\ response}$$

Answer relevancy evaluates how closely the generated response addresses the user's query. This metric assesses the semantic alignment between the query and the response, ensuring that the output remains contextually appropriate and directly answers the posed question. A decline in this metric may suggest that the retrieval mechanism or chunk selection is not optimally aligned with user intent.

$$Answer\ Relevancy\ = \frac{Relevance\ Score\ of\ the\ response\ to\ the\ query}{Maximum\ possible\ relevancy\ score}$$

Answer correctness quantifies the accuracy of the generated response, focusing on its factual and logical validity. Unlike faithfulness, which pertains to alignment with the context, correctness evaluates the overall veracity of the output. Improvements in this metric typically reflect enhancements in re-ranking mechanisms or better context integration during generation.

$$Answer\ Correctness\ = \frac{Number\ of\ factually\ correct\ statements\ in\ the\ response}{Total\ number\ of\ statements\ in\ the\ response}$$

Context recall measures the proportion of relevant information from the available context retrieved during the retrieval phase. This metric is crucial for ensuring that the system captures all necessary information to generate a comprehensive response. A high context recall score indicates that the vector database and indexing methods effectively identify and retrieve relevant data.

$$Context\ Recall\ = \frac{Number\ of\ relevant\ chunks\ retrieved}{Total\ number\ of\ relevant\ chunks\ in\ the\ dataset}$$

Context Precision assesses the specificity of the retrieved context in relation to the query. It indicates the extent to which retrieved chunks contain only relevant information without extraneous or unrelated content. While higher precision suggests focused retrieval, excessive narrowing may inadvertently omit critical contextual elements.

$$Context\ Precision\ = \frac{Number\ of\ relevant\ chunks\ retrieved}{Total\ number\ of\ chunks\ retrieval}$$

3. Result

3.1 Evaluation of Model Performance

This section examines the performance of the proposed Retrieval-Augmented Generation (RAG) model using FAISS for semantic retrieval. The evaluation compares two FAISS indexing methods: IndexIVFFlat, which clusters the vector space for efficient retrieval, and IndexHNSWFlat, which constructs a hierarchical navigable small-world graph for precise nearest-neighbour searches. Performance metrics include faithfulness, answer correctness, context recall, context precision, and answer relevancy.

Dynamic chunking and cross-encoder re-ranking are also evaluated to assess their contribution to retrieval quality. Results demonstrate that these techniques significantly enhance faithfulness and correctness by preserving semantic coherence and prioritizing contextually aligned responses. However, trade-offs were observed in context precision and answer relevancy, likely due to overlap or redundancy introduced during retrieval.

3.2. Comparison of Indexing Methods: IVF vs. HNSW

IndexHNSWFlat consistently outperformed IndexIVFFlat across most metrics, particularly in context recall (+5.7%), attributed to its ability to capture finer semantic distinctions. Conversely, IndexIVFFlat showed a -17.1% drop in context recall, reflecting partial information loss due to its clustering approach. Both methods experienced a minor decline in context precision (-6.5%), likely resulting from the introduction of dynamic chunking.

3.3. Analysis of the Impact of Dynamic Chunking and Re-Ranking on RAGAS Metrics

Dynamic chunking and cross-encoder re-ranking significantly influenced RAGAS metrics, as summarized in Table 1:

- Faithfulness improved the most with dynamic chunking, achieving a +221.1% increase using HNSW indexing.
- Answer Correctness saw notable gains with both indexing methods (+36.5% for IVF and +28.8% for HNSW) due to re-ranking's ability to prioritize relevant chunks.
- Answer Relevancy decreased with FAISS indexing (-28.6% with IVF, -30.4% with HNSW), reflecting a trade-off in prioritizing retrieval accuracy.

Table 1: Quantitative Analysis of Metric Improvements

Metric	Baseline (NO RAG)	RAG Only	RAG + IVF (Dynamic + Re-Ranking)	RAG + HNSW (Dynamic + Re-Ranking)	% Improvement (IVF vs. Baseline)	% Improvement (HNSW vs. Baseline)
Faithfulness	0.19	0.30	0.53	0.61	+178.9%	+221.1%
Answer Relevancy	0.56	0.73	0.40	0.39	-28.6%	-30.4%
Answer Correctness	0.52	0.51	0.71	0.67	+36.5%	+28.8%
Context Recall	0.70	0.71	0.58	0.74	-17.1%	+5.7%
Context Precision	0.93	0.93	0.87	0.87	-6.5%	-6.5%

The integration of dynamic chunking, cross-encoder re-ranking, and FAISS indexing significantly enhances faithfulness and answer correctness, key metrics for generating high-quality responses. While context precision and answer relevancy saw minor declines, the improvements in core metrics validate the effectiveness of these techniques for retrieval-augmented systems. These results underscore the importance of leveraging advanced retrieval and chunking strategies to optimize performance in semantic search and response generation tasks.

4. Discussion and Conclusion

This study focused on enhancing the performance of Retrieval-Augmented Generation (RAG) systems through advanced retrieval techniques, including FAISS-based indexing, dynamic chunking, and cross-encoder re-ranking. The findings validate the significant impact of these methods on key metrics such as faithfulness, context recall, and answer correctness, highlighting their potential for optimizing RAG architectures.

The adoption of the HNSW indexing method in FAISS proved superior to the traditional IVF approach, particularly in improving context recall and context precision. HNSW's hierarchical structure enabled the retrieval of semantically richer and more relevant document chunks, demonstrating its efficacy in capturing fine-grained semantic relationships critical for high-accuracy retrieval.

Dynamic chunking further contributed to performance improvements by segmenting text into semantically coherent chunks, ensuring the preservation of contextual information. This approach minimized information loss and enhanced the quality of retrieved content, directly benefiting downstream response generation.

The integration of a cross-encoder re-ranking mechanism refined the retrieval process by prioritizing contextually aligned chunks, resulting in higher response correctness and overall quality. The combined use of dynamic chunking and re-ranking, alongside FAISS-based indexing, significantly improved the faithfulness and answer correctness of generated responses, as evidenced by the RAGAS evaluation framework.

The study underscores the critical role of optimized vector search methods and re-ranking techniques in enhancing the accuracy and relevance of RAG systems. While trade-offs in context precision and answer relevancy were observed, the overall improvements in faithfulness and correctness validate the utility of these techniques for retrieval-augmented models.

Future research could explore additional vector indexing strategies, hybrid retrieval methods, and advanced re-ranking mechanisms to further improve retrieval efficiency and quality. Expanding RAG architectures to support real-time data processing or multimodal applications also represents promising avenues for future exploration.

In conclusion, this research demonstrates the effectiveness of combining HNSW indexing, dynamic chunking, and cross-encoder re-ranking in optimizing RAG models. These findings provide a strong foundation for further advancements in retrieval-augmented systems, with significant implications for AI and Natural Language Processing applications.

References

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, and Y. Du, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[2] Z. Jiang, X. Ma, and W. Chen, "Longrag: Enhancing retrieval-augmented generation with long-context LLMs," *arXiv preprint arXiv:2406.15319*, 2024.

[3]"Metin/WikiRAG-TR," Hugging Face, 2024. [Online]. Available: https://huggingface.co/datasets/Metin/WikiRAG-TR

[4] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997v5, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2312.10997

[5] I. S. Singh, R. Aggarwal, I. Allahverdiyev, A. Akalin, K. Zhu, and S. O'Brien, "ChunkRAG: Novel LLM-Chunk Filtering Method for RAG Systems," arXiv preprint arXiv:2410.19572v4, Nov. 2024. [Online]. Available: https://arxiv.org/abs/2410.19572

[6] W. Song, T. Tan, Y. Qin, X. Lu, and T. Liu, "MPNet: Masked and Permuted Pre-training for Language Understanding," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [Online]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[7] M. Douze, J. Johnson, M. Lomeli, A. Guzhva, G. Szilvasy, L. Hosseini, C. Deng, P.-E. Mazaré, and H. Jégou, "The FAISS Library," arXiv preprint arXiv:2401.08281v2, Sep. 2024. [Online]. Available: https://arxiv.org/abs/2401.08281

[8] X. Ma, T. Teofili, and J. Lin, "Anserini Gets Dense Retrieval: Integration of Lucene's HNSW Indexes," arXiv preprint arXiv:2304.12139v1, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2304.12139

[9] H. Déjean, S. Clinchant, and T. Formal, "A Thorough Comparison of Cross-Encoders and LLMs for Reranking SPLADE," arXiv preprint arXiv:2403.10407v1, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2403.10407

[10] "ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1," Hugging Face, 2024. [Online]. Available: https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1

[11] "ytu-ce-cosmos/Turkish-Llama-8b-Instruct-v0.1," Hugging Face, 2024. [Online]. Available: https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-Instruct-v0.1

[12] H. T. Kesgin, M. K. Yuce, E. Dogan, M. E. Uzun, A. Uz, E. İnce, Y. Erdem, O. Shbib, A. Zeer, and M. F. Amasyali, "Optimizing Large Language Models for Turkish: New Methodologies in Corpus Selection and Training," in *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2024, pp. 1-6.

[13] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," arXiv preprint arXiv:2309.15217v1, Sep. 2023. [Online]. Available: https://arxiv.org/abs/2309.15217

[14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint arXiv:2312.10997v5*, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2312.10997