Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 1889–1898
https://doi.org/10.54941/ahfe1005756

Advanced Chunking and Search Methods for Improved Retrieval-Augmented Generation (RAG) System Performance in E-Learning

Danter Daniel[1], Mühle Heidrun[1], and Stöckl Andreas[2]

[1]506 Data & Performance GmbH, Linz, Austria
[2]Digital Media Lab, University of Applied Sciences Upper Austria, Hagenberg, Austria

ABSTRACT

This study evaluates two search methodologies—Hybrid Search and Semantic Search—within a Retrieval-Augmented Generation (RAG) framework for an E-Learning use case. The goal was to enhance the accuracy and efficiency of Large Language Models (LLMs), such as GPT-4, through advanced Prompt Engineering Techniques and optimized retrieval processes. Thus, efficient search and chunking methods are essential for improving the quality of system-generated answers. Using the Evaluation framework for Retrieval-Augmented Generation (RAG) pipelines, the Ragas framework as our testing framework, we measured five key metrics: answer correctness, context recall, context precision, faithfulness, and answer relevancy. The dataset utilized in this study comprises question-answer pairs, with the answers established as ground truth, derived from educational sources such as textbooks, research papers, and lectures with over 215 pages of highly complex theoretical and practical learning material. In order to evaluate the chunking and search methodologies the Ragas testing framework dataset covers 57 questions out of the used educational material related to generative AI concepts and prompt engineering techniques. These source documents were pre-processed into smaller, manageable chunks and indexed using both vector embeddings and keyword-based indexing, aimed at facilitating efficient retrieval and improving response accuracy. The ground truth constituted the benchmark for assessing the performance of the Ragas testing framework. The AI model used for embeddings, OpenAI's text-embedding-ada-002, generated high-dimensional representations to capture deep semantic meanings. The study tested three chunking strategies (Token-Based, Recursive, and BERT-based) and compared the search methods using statistical analyses like ANOVA and paired t-tests. The results show that Hybrid Search consistently outperformed Semantic Search across all metrics. However, the effect size (Cohen's $d = -0.11$) indicated that the practical difference was negligible. Token-Based Chunking underperformed in Context Recall compared to BERT-based and Recursive Chunking. These findings offer valuable insights for optimizing RAG systems in E-Learning, with future directions focusing on continuously improving chunking techniques and integrating long-context LLMs for enhanced scalability and accuracy.

Keywords: Advanced chunking, Semantic search, Hybrid search, Retrieval-augmented generation (RAG), E-learning, Large language models (LLMS), Generative AI, Prompt engineering techniques

INTRODUCTION

Retrieval-Augmented Generation (RAG) systems are a promising approach in E-Learning for their ability to retrieve and generate relevant information grounded on vector databases in terms of valid data knowledge. Traditional Large Language Models (LLMs) are limited by context windows, but RAG systems extend capabilities by accessing external, up-to-date, and valid information, which is especially important in (E-)Learning environments because knowledge in an educational context must be valid and based on ground truth.

This is particularly beneficial in E-Learning, where educational material is vast and fragmented, considering LLMs – even with larger context windows like Google Gemini's Pro-Versions – are limited to their partly not transparent training data.

Background on RAG Systems

RAG systems integrate the retrieval of external documents and the generation capabilities of Large Language Models (LLMs), such as GPT-4. This dual approach significantly enhances the relevance and accuracy of generated responses, a necessity for environments dealing with large-scale and evolving datasets, such as E-Learning datasets. Traditional LLMs operate within limited context windows, which restricts their ability to incorporate extensive external knowledge (Bubeck et al., 2023). However, RAG systems offer a way to extend these capabilities by allowing models to access up-to-date and domain-specific information from external sources (Kandpal et al., 2022).

In essence, a RAG system retrieves pertinent documents related to a given query and uses these as input for a language model to generate an informed response (Zhao et al., 2024). Studies by Guu et al. (2020) demonstrate that by incorporating both dense and sparse retrieval methods, RAG systems can significantly improve the quality of responses, especially when dealing with large datasets (Yu, et al., 2024), also used for E-Learning documents. Gao et al. (2023) further elaborate that RAG frameworks are particularly valuable in situations where accuracy and knowledge specificity are crucial, such as in educational settings. By combining semantic and keyword-based search methods, RAG systems effectively filter and prioritize relevant information (Yu et al., 2024).

This design is particularly potentially suitable for E-Learning use cases, where the educational material is often vast and fragmented across multiple sources. RAG thereby ensures that a learner receives relevant, accurate, and contextually appropriate information, enhancing understanding and retention.

Problem Statement

The challenges associated with information retrieval in large, complex educational datasets are grounded in the process that prepares these external sources of information through data-preprocessing techniques. The process, used for RAG system preparation encompasses three core components: constructing a data index, developing a retrieval system, and generating

answers. At the data processing stage, document parsing plays a pivotal role in ensuring accurate extraction of information from text documents. It is essential to retrieve coherent and relevant snippets, which presents a significant challenge. These snippets, also called chunks, are essentially for the accurate preparation of longer documents (Liu et al., 2024). Therefore, chunking optimization is crucial, as it segments the text into smaller, more manageable chunks. While chunks are fundamental and chunking methods are important, another aspect of optimizing retrieval results from the index refers to the search method. In this context, the relevance lies not only in using accurate chunking methods but also in combining them with the most effective retrieval strategies to achieve the best results in answering questions (Tan et al., 2024). This study is focused to investigating these issues.

ADVANCED CHUNKING TECHNIQUES

Guu et al. (2020) found that integrating BERT-based retrieval models into language model pre-training allows the system to access semantically relevant text segments, thereby improving performance in retrieval-augmented generation systems. This improvement in retrieval accuracy is particularly significant when dealing with educational material that is often hierarchical and interconnected. For example, a section on a scientific concept may span several paragraphs, and breaking it arbitrarily could lead to a loss of essential context (Tan et al., 2024).

Recursive chunking divides text into progressively smaller sections using hierarchical separators, iterating the process until the desired chunk size is achieved. This method enhances text's structure-aware segmentation, preserving context and meaning, making it particularly suitable for handling documents with varied structures (Pinecone, 2023).

Token-chunking is a method that divides text into fixed-length segments based on the number of tokens, rather than sentences or paragraphs. This approach ensures uniform chunk sizes, facilitating more consistent processing in large language models, particularly when managing long-form content (Finardi et al., 2022).

Furthermore, semantic-based chunking is integral to Hybrid Search, where the combination of keyword and semantic search techniques requires chunking strategies that can capture both syntactic and contextual nuances of the text. Finardi et al. (2024) observed that integrating Recursive Chunking with semantic understanding enhanced the system's ability to recall relevant information while minimizing the retrieval of irrelevant or tangential content (Tan et al., 2024).

In educational applications, semantic chunking (Stöckl et al., 2024) improves the precision of retrieved information, the system's faithfulness, and context recall. By maintaining semantic coherence across chunks, the system can retrieve sections more closely aligned with the learner's intent and educational goals, facilitating a more effective learning experience. This approach reduces the risk of fragmented or misleading responses, which can occur when chunks are split without regard for their semantic content.

EXPERIMENTAL DESIGN

This section provides an overview of the research methodology, including the search methods, chunking techniques, and evaluation metrics. The experimental design contains 57 questions, that were processed using Hybrid and Semantic Search, each paired with one of three chunking methods: Token-Based, Recursive, and Bert-based. The results are evaluated using five key metrics in terms of the Ragas testing framework: faithfulness, answer relevance, context recall, context precision, and answer correctness.

Dataset

For this study, a custom-built dataset was designed to evaluate RAG systems in the context of E-Learning within the Ragas testing framework. Inspired by datasets like WikiEval (2023), which use a question-answer design, the dataset contains 215 pages of educational materials. These materials were collected from a variety of sources, including textbooks, research papers, academic articles, reports from institutions like UNESCO, practical handbooks and lectures.

This interdisciplinary collection covers detailed case studies, research reports, practical guides, and discussions on topics on generative AI and prompt techniques. From these documents, 57 question-and-answer pairs were derived as ground truth to help learners acquire the educational content. Designed for E-Learning, these questions aim to enhance interactive knowledge transfer by testing the effectiveness of various prompt-engineering approaches. The data is structured to provide practical applications of these technologies in real-world generative AI and its optimization. The documents were pre-processed into smaller, manageable chunks and indexed using both vector embeddings and keyword-based indexing. This dual-indexing method allows for the retrieval of both semantically relevant information and exact matches for educational queries. The dataset is hierarchically structured, enabling the RAG system to retrieve information at varying levels of granularity, which is particularly important in E-Learning where responses may need to include general overviews or detailed explanations depending on the learner's needs. Additionally, the question-dataset includes metadata on topics, difficulty levels, and intended learning outcomes, further improving retrieval precision and the educational relevance of generated responses.

Evaluation Metrics

To assess the performance of the RAG system, the Rags framework (Retrieval-Augmented Generation Assessment) was used for the evaluation of the study results. This framework employs a series of metrics that evaluate both the retrieval and generation components of the system. The first and most important metric is 'answer correctness'. This metric measures how closely the generated answer aligns with a predefined reference or ground truth. It incorporates both semantic and factual correctness, ensuring that the answer is not only accurate but also consistent with the context provided by the query. 'Context recall' measures how effectively the retrieval system identifies relevant chunks of context from the dataset. It calculates the

proportion of relevant information retrieved that is necessary to answer the query. A higher recall score indicates that the system retrieved a more complete set of relevant documents or context chunks. The third metric is 'context precision'. While 'context recall' focuses on retrieving all relevant information, 'context precision' measures the relevance of the retrieved context, penalizing the inclusion of irrelevant or unrelated content.

This metric assesses the accuracy of the retrieved information. The fourth metric is 'faithfulness'. This metric evaluates the factual consistency of the generated answer with the retrieved context. An answer is deemed faithful if all the claims made can be directly inferred from the retrieved context. The faithfulness score is calculated by determining the proportion of statements in the answer that align with the context, ensuring that no extraneous or 'hallucinated' information is introduced. The last metric is 'answer relevance', which assesses how well the generated answer addresses the query. Even if the answer is factually correct, it is penalized if it does not fully meet the user's needs or includes unnecessary information. The aim is to ensure that answers are concise, accurate, and contextually appropriate (Es et al., 2022).

By using these metrics, the evaluation process ensures that the system retrieves relevant information and generates responses that are accurate, contextually aligned, and pedagogically useful. The combination of the question-answer dataset and these comprehensive evaluation metrics enables a thorough assessment of the RAG system's performance in relation to our e-learning use case.

Experimental Procedure

Each search method (Hybrid Search and Semantic Search) was tested with three different chunking strategies (Token-Based, Recursive, and BERT-Based). The performance of each combination was measured using the defined evaluation metrics: faithfulness, answer relevancy, context recall, context precision, and answer correctness.

ANOVA Test: An ANOVA test was conducted to assess the statistical significance of the differences between the chunking methods within each search method (hybrid and semantic). This test evaluated whether there were significant differences in performance across the chunking methods for each search type.

Paired t-test for Search Methods: A paired t-test was performed to statistically compare the aggregated means of the metrics between the Hybrid Search and Semantic Search methods. This test evaluated whether there was a statistically significant difference in overall performance between the two search methods, treating the metrics as paired data across both methods.

RESULTS

In the following section, we present the results of our study. Table 1 summarizes the performance of the Hybrid Search and Semantic Search methods across the three chunking techniques: Token-Based, Recursive, and BERT-based. For each chunking method, the mean values are provided for the

five evaluation metrics: faithfulness, answer relevancy, context recall, context precision, and answer correctness.

Table 1. Average performance metrics for hybrid and semantic search methods across different chunking strategies.

Metric	Hybrid Recursive Mean	Hybrid Token Mean	Hybrid BERT Mean	Semantic Recursive Mean	Semantic Token Mean	Semantic BERT Mean
faithfulness	0.8711	0.9316	0.9071	0.9105	0.8595	0.8882
answer relevancy	0.8939	0.9255	0.8683	0.9307	0.8648	0.8653
context recall	0.7850	0.7567	0.7821	0.8625	0.5882	0.8192
context precision	0.5674	0.5611	0.5769	0.6492	0.6428	0.6156
answer correctness	0.5312	0.4756	0.4660	0.5074	0.4281	0.4518

Hybrid Search Performance

Figure 1
illustrates the performance of the Hybrid Search method across the three chunking techniques (Token-Based, Recursive, and BERT-based).

The bar chart presents the mean values of each chunking method across the five evaluation metrics: faithfulness, answer relevancy, context recall, context precision, and answer correctness.

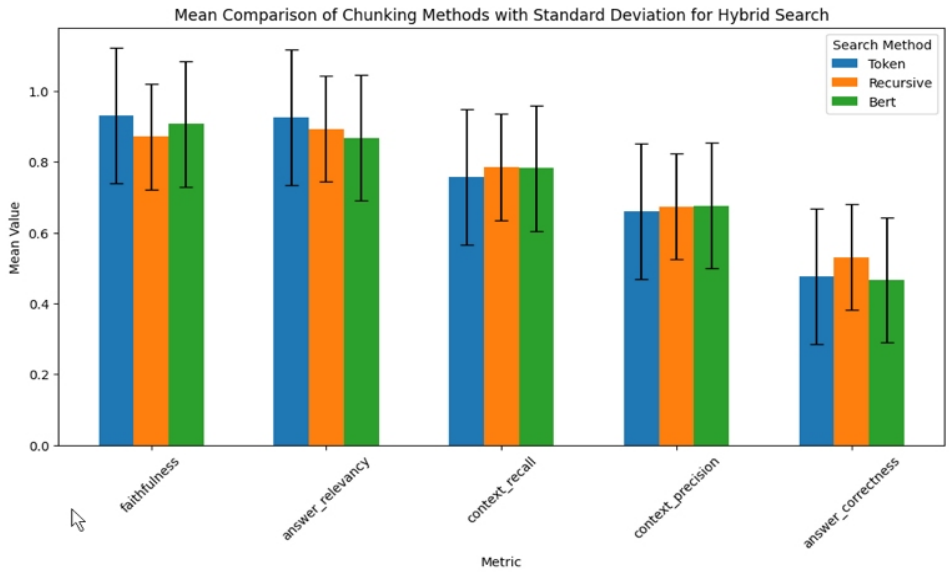

Figure 1: Mean comparison of chunking methods – deviation hybrid search.

We conducted an ANOVA test to evaluate the statistical significance of the differences in performance across the three chunking techniques. The results for each evaluation metric are illustrated in Table 2 below.

Table 2. Statistical analysis of performance metrics.

Metric	F-statistic	p-value	Significance
faithfulness	0.8403	0.4335	ns
answer relevancy	1.0177	0.3636	ns
context recall	0.0848	0.9188	ns
context precision	0.0491	0.9521	ns
answer correctness	1.336	0.2657	ns

None of the p-values are below the significance threshold of 0.05, indicating no statistically significant differences between the chunking methods for any of the metrics.

Semantic Search Performance

Similarly, the performance of the Semantic Search method across the three chunking techniques (Token-Based Chunking, Recursive Chunking, and BERT-Based Chunking) is illustrated in Figure 2. The bar chart presents the mean values of each chunking method across the five evaluation metrics: faithfulness, answer relevancy, context recall, context precision, and answer correctness.

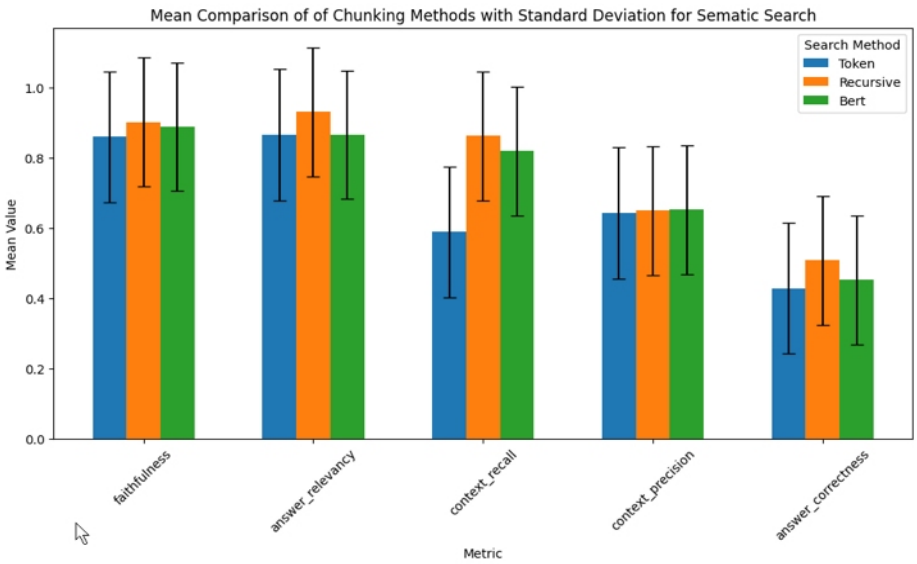

Figure 2: Mean comparison of chunking methods – deviation semantic search.

An ANOVA test was conducted to assess the statistical significance of performance differences between the three chunking methods (Token-Based Chunking, Recursive-Based Chunking, and BERT-Based Chunking) for semantic search across the five evaluation metrics. The results are shown in Table 3:

Table 3. Statistical significance of performance differences between chunking methods.

Metric	F-statistic	p-value	Significance
faithfulness	0.9183	0.4013	ns
answer relevancy	1.4853	0.2294	ns
context recall	8.0223	0.0005	***
context precision	0.0146	0.9855	ns
answer correctness	1.8922	0.1539	ns

The results indicate that context recall shows a statistically significant difference between the chunking methods (p-value < 0.05). Further analysis using Tukey's HSD test reveals the specific group differences responsible for this result. For the remaining metrics—faithfulness, answer relevancy, context precision, and answer correctness—the p-values are greater than 0.05, indicating no statistically significant differences between the chunking methods for these metrics.

Comparison Between Hybrid and Semantic Search

To determine which search method—Hybrid Search or Semantic Search—performs better across the evaluation metrics, we calculated the aggregated mean for each metric (faithfulness, answer relevancy, context recall, context precision, and answer correctness) across the three chunking methods (Recursive, Token-Based, and BERT-Based) for both search types.

The following bar chart, as shown in Figure 3 visually compares the aggregated mean values for Hybrid Search and Semantic Search across the five metrics: faithfulness, answer relevancy, context recall, context precision, and answer correctness.

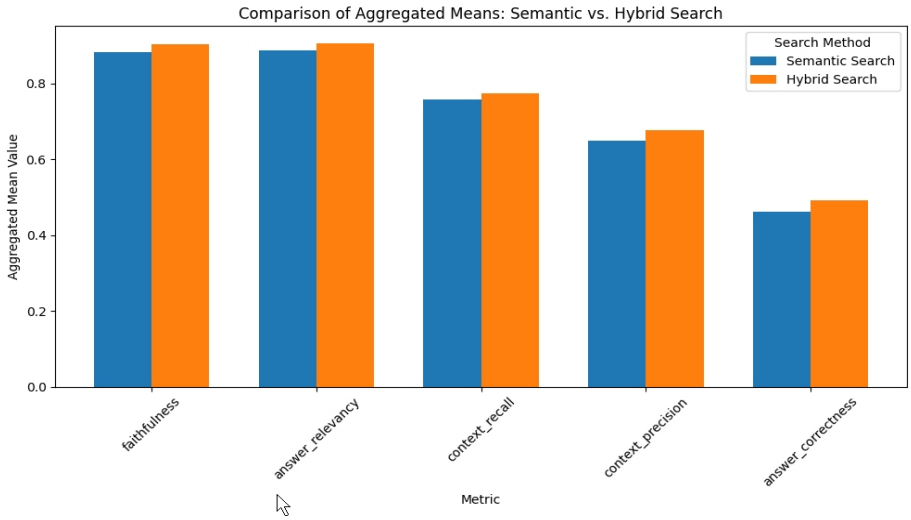

Figure 3: Mean Comparison of Aggregated Means: Sematic vs. Hybrid Search.

To compare the two search methods, Hybrid Search and Semantic Search, across the aggregated metrics, a paired t-test was conducted to evaluate whether the difference in mean performance between the two methods is statistically significant. The t-test resulted in a t-statistic of –6.4489 and a p-value of 0.0035, indicating a statistically significant difference between the methods. The paired t-test demonstrates that the difference in performance between Hybrid Search and Semantic Search is statistically significant ($p = 0.0035$), with Hybrid Search showing higher aggregated mean performance across the metrics. However, the calculated effect size (Cohen's $d = -0.11$) suggests that the magnitude of this difference is small, implying that while Hybrid Search may perform better overall, the practical significance of this advantage is minimal.

CONCLUSION

Hybrid Search has a statistical edge over Semantic Search across all five measured evaluation metrics. However, the small effect size (Cohen's $d = -0.11$) suggests that the practical difference between the two methods may be limited in real-world applications. Nonetheless, Hybrid Search remains preferable in tasks where accuracy, precision, and comprehensive context recall are critical. Semantic Search continues to offer strong performance but falls slightly behind Hybrid Search in overall effectiveness. Future research could focus on continuously optimizing chunking and search techniques to improve retrieval accuracy and answer generation. One potential direction is to explore more sophisticated chunking strategies that could dynamically adapt based on the complexity and context of the input data, offering even greater precision and relevance in the retrieved information.

Moreover, integrating RAG systems with long-context LLMs offers the potential to enhance the scalability of these systems. This could allow them to process even larger datasets and handle more complex queries without losing accuracy, as demonstrated by recent developments in the field (Yu et al., 2023). This scalability is crucial in terms of real-world implications for applications in E-Learning environments, where learners increasingly interact with diverse and growing bodies of information. Enhanced LLM-driven systems would furthermore enable learners to receive precise answers and retrieve the exact information they need, exactly when they need it, ensuring efficiency in self-study.

Additionally, exploring new datasets, more diverse learning materials and experimenting with alternative chunking strategies could provide further insights into how these systems can be optimized for various educational scenarios. These developments would not only improve the accuracy and scalability of RAG systems but would also make them invaluable in fields where high precision and contextual understanding are essential, such as STEM (Science, Technology, Engineering, and Mathematics).

REFERENCES

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv Preprint, arXiv:2303.12712.

Finardi, P., Avila, L., Castaldoni, R., Gengo, P., Larcher, C., Piau, M., & Costa, P. (2024). *The chronicles of RAG: The retriever, the chunk, and the generator*. arXiv Preprint, arXiv:2401.07883.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey*. arXiv Preprint, arXiv:2302.07842.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 671–684.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2022). *Large language models struggle to learn long-tail knowledge*. arXiv Preprint, arXiv:2211.08411.

LangChain. (2023). *Text splitter*. LangChain Documentation. Retrieved from https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/

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

Pinecone. (2023). *Chunking strategies for LLM applications*. Pinecone Documentation.

Stöckl, A., & Ibrovic, E. (2024). *Document Segmentation for Topic Modelling with Sentence Embeddings*, in *Proceedings of the International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*. pp. 1–5.

Yu, T., Xu, A., & Akkiraju, R. (2024). *In defense of RAG in the era of long-context language models*. arXiv Preprint, arXiv:2409.01666.

Zhao, S., Sun, Y. et al. (2024). *Retrieval-augmented generation (RAG) and beyond: A comprehensive survey on how to make your LLMs use external data more wisely*. arXiv Preprint, arXiv:2409.14924.