Retrieval-Augmented Generation with Metadata-Aware Chunk Retrieval

Tayfun Jamalbayli
Middle East Technical University
Ankara, Turkiye
jamalbayli.tayfun@metu.edu.tr

***Abstract*—Retrieval-Augmented Generation (RAG) has emerged as a powerful framework to enhance large language models (LLMs) by incorporating external knowledge through retriever-generator pipelines. This study investigates the performance of various chunk retrieval strategies within RAG systems, focusing particularly on the influence of metadata and retrieval modality. A synthetically generated benchmark dataset—constructed from academic paper titles—was used to evaluate eight different retrieval configurations, including keyword-based, vector-based, and hybrid approaches. Experimental results show that keyword retrieval on titles achieves the highest accuracy (91.12%), while hybrid methods combining keyword and vector search provide more adaptable performance for realistic scholarly queries. Notably, the Keyword Search (BM25) on "Title" and Vector search on "Abstract" strategy offers a balance between lexical precision and semantic flexibility.**

Keywords—Retrieval-Augmented Generation (RAG), Large Language Models (LLM), Vector Search; Keyword Search, Metadata, Hybrid Retrieval; Academic Search, Chunk Retrieval; Benchmark Dataset; Semantic Retrieval

I. INTRODUCTION

Retrieval-Augmented Generation (RAG) is a machine learning framework that enhances the capabilities of large language models (LLMs) by integrating external knowledge sources during the generation process. Instead of relying solely on a model's internal parameters, RAG retrieves relevant information from a database or corpus and incorporates it into the response generation. This approach improves factual accuracy, reduces hallucination, and allows models to handle up-to-date or domain-specific content more effectively.s

Chen et al. [1] provide a comprehensive taxonomy of RAG approaches, dividing them into Naive RAG, Advanced RAG, and Modular RAG paradigms. They formally decompose RAG into three primary modules: retriever, augmenter, and generator. Naive RAG methods adopt simple dense or sparse retrieval followed by direct generation, while Advanced RAG incorporates sophisticated augmentation techniques such as retrieval-augmented prompt design, retriever fine-tuning, and multi-hop retrieval. Modular RAG emphasizes system architectures where retriever and generator components are trained jointly or interactively refined. Chen et al. also describes the mathematical formulation of RAG, define standard evaluation metrics, and outline the influence of retrieval recall and generation faithfulness on overall system performance [1].

Lewis et al. [2] focus on empirical evaluations of RAG systems, systematically analyzing how different retrieval methods, retriever models, passage selection strategies, and prompt augmentation mechanisms affect downstream performance. Their work benchmarks dense retrieval (e.g., DPR), sparse retrieval (e.g., BM25), and hybrid retrieval approaches across multiple knowledge-intensive tasks. They also introduce strategies for query classification to selectively trigger retrieval modules, optimizing both retrieval accuracy and computational efficiency. Lewis et al. highlight the impact of retrieval size (top-k selection) and retrieval diversity on generation quality, and demonstrate that retriever fine-tuning using contrastive loss can significantly enhance end-to-end task performance in RAG pipelines [2].

Expanding beyond text-based retrieval, Ma et al. [3] propose VisRAG, a vision-centric RAG architecture for document visual question answering (DocVQA) tasks. Unlike traditional RAG systems that operate solely on text, VisRAG directly embeds document images and queries into a shared latent space using a vision-language model. This approach preserves the structural and spatial information of documents, which is critical for answering questions dependent on visual layouts. The retrieval module retrieves relevant document regions based on query-image similarity, and the generation module conditions on both retrieved visual features and the input query to generate answers. Experimental results on benchmarks such as DocVQA and WebSRC show that VisRAG improves both retrieval precision and generation accuracy compared to standard OCR-based methods [3].

Together, these studies illustrate the technical evolution of RAG, from basic text retrieval and augmentation pipelines to more advanced architectures involving fine-tuned retrievers, query-aware retrieval strategies, and multimodal retrieval-augmented systems. Key challenges identified include optimizing retrieval quality, developing joint retriever-generator training frameworks, handling multi-hop and open-domain retrieval scenarios, and adapting RAG methods to non-textual modalities.

This study investigates various chunk retrieval strategies within Retrieval-Augmented Generation (RAG) systems, using a synthetically generated benchmark dataset. A particular focus is placed on the role of metadata in enhancing retrieval performance. Specifically, the research seeks to answer two central questions: (1) To what extent does the inclusion of metadata improve the retrieval of the most relevant text chunks? and (2) Which retrieval method yields the best overall performance?

The scope of this research is confined to the domain of academic literature, with a use case centered on question-driven chunk retrieval from research papers. The proposed system allows users to query whether specific research topics have already been explored, thereby filtering content based on relevance and prior investigation. This targeted application provides a practical framework for evaluating retrieval effectiveness in scholarly contexts.

II. DATASET

The initial dataset presented numerous challenges due to its inconsistent and complex structure, making it unsuitable for direct use in a vector-based retrieval system. To address this, we undertook a data preprocessing phase aimed at restructuring the dataset into a clean and standardized tabular format. This format would serve as the foundation for populating a vector database used in our Retrieval-Augmented Generation (RAG) system.

To achieve this, a custom Python pipeline was developed to extract and organize the relevant information into the following columns:

- **Abstract:** A brief summary of the article content
- **Title:** The title of the article
- **Url:** A link to the article or its repository
- **ID:** A unique identifier for each article
- **Conference:** The name of the conference where the article was submitted or presented
- **Decision:** The acceptance or rejection status of the article
- **Authors:** A list of contributing authors

This structured dataset ensured consistency and facilitated further processing steps. Once the tabular dataset was prepared, we populated the vector database using embeddings derived from the DistilBERT model. For simplicity and to reduce computational overhead, only the title column was vectorized. Titles typically encapsulate the core idea of each article and were deemed sufficient for our initial retrieval experiments.

A. *Data Analysis*

To understand the structural characteristics of the dataset, we analyzed word counts in titles and abstracts, as well as the most frequent words.

As shown in Fig. 1, abstract word counts are right-skewed, with most abstracts containing between 200 and 500 words. The peak occurs near 300 words, which aligns with common academic standards. A long tail indicates some abstracts are significantly longer, possibly due to variations in journal or discipline guidelines.

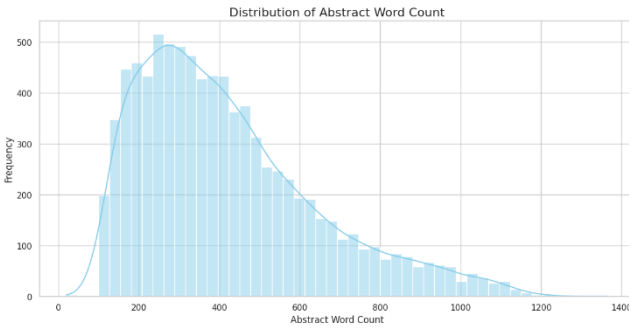

Fig. 1. Distribution of abstract word counts

Fig. 2 shows a near-normal distribution of title lengths, with most titles containing 6–9 words.

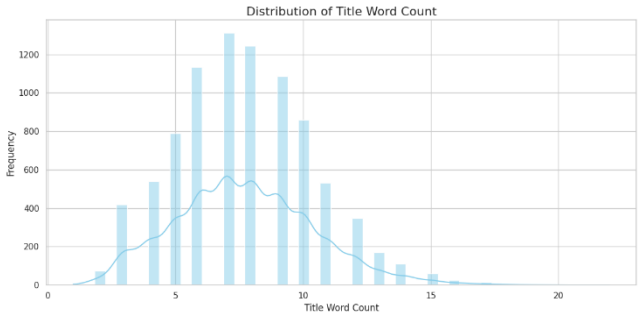

Fig. 2. Distribution of title word counts

In Fig. 3, the most frequent abstract words are common stopwords such as “the,” “of,” and “to.” While expected, these findings highlight the need for preprocessing (e.g., stopword removal) before performing advanced textual analysis.

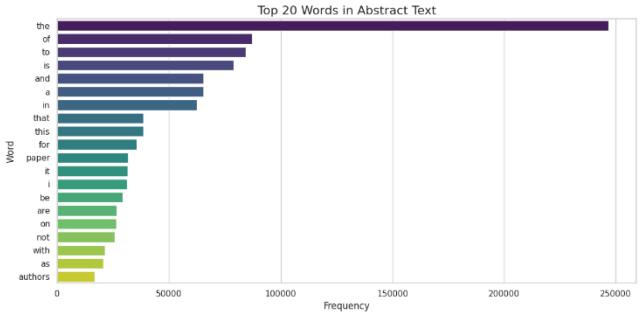

Fig. 3. Distribution of title word counts

Fig. 4 shows the 20 most frequent words used in abstract texts. The most common words—such as “the,” “of,” “to,” and “is”—are mostly stopwords, which are typical in longer, descriptive writing. This suggests that abstracts are written in full sentences and aim to clearly explain the research, relying heavily on grammatical structure rather than technical keywords.

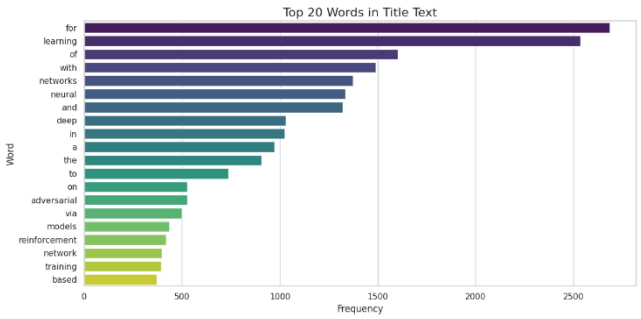

Fig. 4. Distribution of title word counts

presents the 20 most frequent words in title texts. Unlike abstracts, the top words here—such as “learning,” “neural,” “networks,” “deep,” and “reinforcement”—are more technical and field-specific. This indicates that titles are more focused on highlighting key topics and concepts, particularly within areas like machine learning and artificial intelligence. Titles tend to use fewer stopwords and instead prioritize clarity and relevance to help readers quickly grasp the subject of the work.

B. *Benchmark Dataset*

To evaluate the effectiveness of the RAG system, we created a custom benchmark dataset derived from the cleaned

tabular data. This benchmark was designed to test the system's ability to retrieve relevant documents based on user queries. We used GPT-4o-mini, a lightweight and efficient version of GPT-4, to generate synthetic question-article ID pairs. In each case, a question was formulated using only the title of an article, without access to the abstract or full content. This constraint simulates a real-world scenario where a user may have only partial information.

Due to time and resource limitations, we generated 260 question-ID pairs. Although the dataset size is modest, it serves its intended purpose of providing a controlled environment to compare the performance of different retrieval strategies. We believe that this benchmark is sufficiently representative for evaluating and contrasting multiple methods within the same framework.

III. MODELING

To evaluate chunk retrieval performance in RAG systems, we compare a range of retrieval strategies that vary by search modality (keyword-based vs. vector-based) and by metadata type (title, abstract, or both). The methods are categorized into two groups: Preliminary and Hybrid Retrieval Methods. The full implementation used in this study is publicly available on GitHub, and all training and evaluation logs are tracked via Weights & Biases.

These initial baselines rely solely on keyword-based matching using different metadata fields:

- **KW-Title:** Keyword search using only the title column.
- **KW-Abstract:** Keyword search using only the abstract column.
- **KW-Both:** Keyword search using a combination of title and abstract columns.

These methods serve as baseline indicators for how traditional lexical search performs when limited to specific metadata fields. The second group incorporates vector-based semantic search, which leverages text embeddings to capture deeper contextual similarity, either alone or in hybrid configurations:

- **Vec-Title:** Vector search using only the title column.
- **Vec-Abstract:** Vector search using only the abstract column.

To investigate the synergistic effect of combining search modalities across different metadata types, these methods are also evaluated:

- **KW-Title + Vec-Abstract:** Keyword search on the title column combined with vector search on the abstract column.
- **KW-Abstract + Vec-Title:** Keyword search on the abstract column combined with vector search on the title column.
- **KW-Title + Vec-Title:** Keyword search and vector search on the title column.

Each method is tested against a synthetically generated benchmark dataset designed to mimic academic queries.

IV. RESULTS

This section presents the retrieval accuracy results of all evaluated methods on the benchmark dataset, which consists of 260 question-document pairs. Each method retrieved only the top-ranked result per query (top-1 retrieval). The accuracy metric reflects the percentage of cases where the retrieved document matched the correct document ID associated with the query.

A. *Preliminary Methods*

Three keyword-based retrieval strategies were tested as baselines. These rely solely on direct lexical matching between the query and specific metadata fields:

TABLE I. ACCURACY RESULTS OF PRELIMINARY METHODS

Method	Accuracy (%)
KW-Title	91.12
KW-Abstract	46.72
KW-Both	67.18

These methods serve as reference points for evaluating the performance of hybrid retrieval strategies.

B. *Hybrid Methods*

Vector-based retrieval strategies, and combinations of keyword and vector search across metadata types, were evaluated as follows:

TABLE II. ACCURACY RESULTS OF HYBRID METHODS

Method	Accuracy (%)
Vec-Title	27.80
Vec-Abstract	0
KW-Title + Vec-Abstract	79.54
KW-Abstract + Vec-Title	58.30
KW-Title + Vec-Title	71.81

V. DISCUSSION

The results clearly show that the title field plays a crucial role in top-1 chunk retrieval when queries are drawn directly from titles. The KW-Title method achieved 91.12% accuracy, demonstrating that keyword matching on titles—where the words in the query and the document are almost identical—is highly effective for this benchmark.

In contrast, methods relying only on abstracts performed much worse. The KW-Abstract approach reached 46.72% accuracy, and Vec-Abstract failed to retrieve any correct results. Abstracts contain longer, more detailed descriptions and specialized terms. While these make abstracts richer sources of information, they also reduce the number of direct word matches a keyword search can use, and they broaden the

semantic space in a way that pure vector search struggles to handle in a strict top-1 setting.

When we combine keyword search on titles with vector search on abstracts—our KW-Title + Vec-Abstract hybrid—we strike a balance between precise matching and deeper understanding. This two-step process first uses title keywords to select likely candidates and then applies abstract embeddings to refine the ranking. The result is 79.54% accuracy, which, while lower than title-only search on our synthetic dataset, offers greater flexibility. It can catch relevant papers even when user queries do not exactly match the title wording.

For real-world academic search—where researchers ask, "Has anyone already studied topic X?"—this hybrid approach is particularly valuable. It keeps the strength of exact title matches but also adapts to varied phrasing by using the abstract's broader context. Therefore, although KW-Title is best for title-based benchmarks, KW-Title + Vec-Abstract provides a more reliable and adaptable solution for practical question-driven retrieval in scholarly environments.

VI. Conclusion

This study systematically evaluated multiple chunk retrieval strategies within Retrieval-Augmented Generation (RAG) systems using a synthetically generated academic benchmark dataset. The findings highlight the critical role of metadata—particularly titles—in retrieval performance. Keyword search on titles (KW-Title) demonstrated the highest accuracy (91.12%) due to its direct lexical overlap with query content. However, hybrid approaches combining keyword and vector search modalities, especially KW-Title + Vec-Abstract (79.54%), offered a compelling trade-off between lexical precision and semantic flexibility.

While abstract-based retrieval alone proved ineffective—especially for vector search—its integration with title-based keyword methods enhanced adaptability, making hybrid strategies better suited for real-world academic applications where queries may not directly match document titles. These results underscore the importance of tailored retrieval configurations in RAG pipelines and suggest that hybrid approaches leveraging both lexical and semantic cues can significantly improve relevance and robustness in scholarly information retrieval tasks.

VII. References

[1] Z. Chen, Y. Jiang, M. Bendersky, and M. Najork, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2302.14045, 2023.

[2] M. Lewis, P. Rajpurkar, X. Liang, D. Chen, and Y. Tay, "Searching for Best Practices in Retrieval-Augmented Generation," arXiv preprint arXiv:2306.01607, 2023.

[3] H. Ma, Y. Wang, X. Guo, J. Wu, X. Yang, X. Li, D. Cai, and X. He, "VisRAG: Vision-Centric Retrieval-Augmented Generation for Document Visual Question Answering," arXiv preprint arXiv:2402.06620, 2024.