

# A Systematic Framework for Enterprise Knowledge Retrieval: Leveraging LLM-Generated Metadata to Enhance RAG Systems

Pranav Pushkar Mishra, Kranti Prakash Yeole, Ramyashree Keshavamurthy, Mokshit Bharat Surana, Fatemeh Sarayloo  
 University of Illinois Chicago  
 {pmishr23, kyeole2, rkesh, msura4, fsaraylo}@uic.edu

**Abstract**—In enterprise settings, efficiently retrieving relevant information from large and complex knowledge bases is essential for operational productivity and informed decision-making. This research presents a systematic framework for metadata enrichment using large language models (LLMs) to enhance document retrieval in Retrieval-Augmented Generation (RAG) systems. Our approach employs a comprehensive, structured pipeline that dynamically generates meaningful metadata for document segments, substantially improving their semantic representations and retrieval accuracy. Through extensive experiments, we compare three chunking strategies—semantic, recursive, and naive—and evaluate their effectiveness when combined with advanced embedding techniques. The results demonstrate that metadata-enriched approaches consistently outperform content-only baselines, with recursive chunking paired with TF-IDF weighted embeddings yielding an 82.5% precision rate compared to 73.3% for semantic content-only approaches. The naive chunking strategy with prefix-fusion achieved the highest Hit Rate@10 of 0.925. Our evaluation employs cross-encoder reranking for ground truth generation, enabling rigorous assessment via Hit Rate and Metadata Consistency metrics. These findings confirm that metadata enrichment enhances vector clustering quality while reducing retrieval latency, making it a key optimization for RAG systems across knowledge domains. This work offers practical insights for deploying high-performance, scalable document retrieval solutions in enterprise settings, demonstrating that metadata enrichment is a powerful approach for enhancing RAG effectiveness.

## I. INTRODUCTION

Efficiently retrieving and leveraging information from large-scale knowledge repositories is vital for modern organizational productivity and innovation. Retrieval-Augmented Generation (RAG) systems have emerged as a powerful paradigm for enhancing Large Language Models (LLMs) by integrating external knowledge sources [21]. This approach addresses inherent limitations of LLMs, such as knowledge cut-off dates, lack of domain-specificity, and hallucinations [22].

While effective for structured datasets, traditional retrieval methods often struggle with the complexity, scale, and dynamic nature of enterprise knowledge bases. Manual curation becomes impractical as repositories expand, and these methods are prone to overlooking relevant information in lengthy contexts, resulting in the "Lost in the Middle" phenomenon.

Recent advancements in LLMs offer transformative solutions. LLMs can process unstructured data, extract meaningful insights, and generate structured metadata aligned with

semantic and contextual needs. For example, Sundaram and Musen's FAIRMetaText framework demonstrated how LLMs could align metadata with ontologies, reducing manual effort [7]. Similarly, Song et al. highlighted the potential of few-shot prompting for enriching metadata in Earth science datasets, significantly improving metadata completeness and accuracy [10].

Despite these developments, building effective RAG pipelines remains challenging, requiring management of large data volumes, optimization of retrieval strategies, and continual performance tuning.

This paper presents a comprehensive, systematic framework for metadata enrichment using LLMs to optimize RAG systems in enterprise settings. Our structured pipeline dynamically generates meaningful metadata for document segments, enhancing semantic understanding and retrieval accuracy. We compare various chunking strategies—semantic, recursive, and naive—and evaluate their effectiveness with advanced embedding techniques, including TF-IDF weighting and prefix-fusion, using the Snowflake Arctic-Embed model.

Beyond methodology, we conduct an extensive experimental analysis to measure the impact of metadata enrichment on retrieval accuracy, clustering quality, and latency. Our results show that metadata-enriched retrieval outperforms content-only approaches, achieving higher precision and hit rates while reducing latency. This framework establishes metadata enrichment as a scalable, practical optimization for RAG systems across diverse domains, providing a foundational guideline for high-performance deployment in enterprise knowledge management and technical document retrieval.

The rest of this paper is organized as follows: in Section II, we review related work in RAG, metadata enrichment, and embedding strategies. Section III details our systematic framework, covering document processing, metadata generation, embedding, retrieval architecture, and evaluation methodology. Experimental results are presented and analyzed in Section IV. Finally, Section V concludes the paper with key insights and directions for future research.

## II. RELATED WORK

The advent of Large Language Models (LLMs) has revolutionized metadata enrichment and retrieval-augmented

generation (RAG) systems. These technologies address long-standing challenges in metadata management, including inaccuracies, inconsistencies, and scalability issues. LLMs offer transformative capabilities for automating metadata structuring, refining search processes, and enhancing document retrieval. Despite these advancements, persistent challenges such as retrieval bias, hallucination in generated metadata, domain adaptability, and real-time validation continue to hinder their broader application. This review explores recent progress in metadata enrichment, RAG systems, semantic embeddings, and LLM-driven search architectures, synthesizing insights from 25 key studies.

#### A. Advanced RAG Systems and Techniques

RAG systems vary in complexity from Naive RAG, which simply indexes and retrieves document chunks, to Advanced RAG, incorporating techniques like query rewriting and re-ranking, to Modular RAG with configurable pipeline components [5]. Recent innovations include Agentic RAG with dynamic adaptation, GraphRAG utilizing structured knowledge representations [23], and Multimodal RAG integrating non-textual data formats. Our approach builds upon these foundations by specifically focusing on metadata enrichment to optimize retrieval precision.

#### B. Metadata Enrichment with Large Language Models (LLMs)

Metadata serves as the backbone of information retrieval systems, yet traditional methods often struggle with unstructured datasets. Sundaram and Musen's FAIRMetaText framework [7] aligned metadata with FAIR principles using GPT-based embeddings, though performance varied across domains. Song et al. [10] applied taxonomy-guided techniques for metadata completion in Earth sciences, while Mombaerts et al. [11] introduced Meta Knowledge Summaries, though both approaches had limitations in generalizability and dynamic adaptation. Saad-Falcon et al. [14] proposed ARES for automated RAG evaluation, offering scalability for enterprise applications.

#### C. Advancements in Retrieval-Augmented Generation (RAG)

RAG systems mitigate LLM hallucinations by integrating external knowledge bases. Gao et al. [5] classified RAG paradigms, highlighting metadata-aware retrieval potential, while Chen et al. [6] introduced evaluation benchmarks revealing limitations in LLMs' rejection and integration capabilities. Lewis et al. [1] demonstrated improved performance in knowledge-intensive tasks by combining parametric and non-parametric memory, and Shuster et al. [2] showed hallucination reduction in conversational AI through retrieval augmentation.

#### D. Embedding Optimization for Semantic Search

Karpukhin et al. [3] demonstrated dense vector embeddings' superiority over sparse methods for retrieval accuracy. Recent innovations include Harris et al.'s [12] text enrichment techniques, Cuconasu et al.'s [16] controlled randomness approach challenging traditional retrieval assumptions,

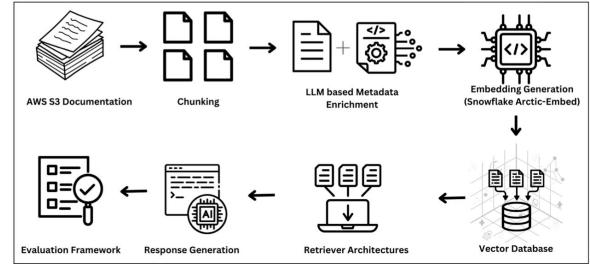


Fig. 1: The proposed system architecture for metadata enrichment and retrieval optimization .

and Anantha et al.'s [15] context tuning that outperformed GPT-4-based retrieval.

#### E. Advancements in Search System Architectures

Wang et al. [8] unified search tasks under autoregressive text generation, while others developed specialized applications like Thottempudi and Borra's [13] virtual assistant and Shao et al.'s [17] ITER-RETEGEN for multi-hop reasoning. Wang et al.'s [18] Self-Knowledge Guided Retrieval enables LLMs to recognize knowledge gaps adaptively.

Despite significant advances in metadata enrichment, chunking strategies, and retrieval techniques, several critical gaps remain unaddressed. Existing studies often lack a systematic evaluation of the interplay between document chunking methods and embedding strategies, leaving unclear how these factors jointly influence retrieval effectiveness. Moreover, there is limited rigorous analysis of metadata composition and its direct impact on retrieval metrics across diverse enterprise datasets. Addressing these gaps is essential for developing scalable, high-performance RAG systems that can adapt to complex, real-world knowledge repositories. In the following section, we introduce a systematic framework designed to fill these gaps and provide practical guidance for optimizing retrieval pipelines at scale.

### III. METHODOLOGY

This section outlines the detailed methodology, which involves a systematic pipeline for transforming raw technical documentation into optimized retrieval components. The primary research question focuses on examining how various document chunking strategies, metadata generation techniques, and embedding approaches impact retrieval accuracy and efficiency. The entire process, from data ingestion to evaluation, is illustrated in Figure 2 and described in detail below.

#### A. Document Processing and Chunking Strategies

We employ the following three distinct chunking approaches, as illustrated in the first part of Fig. 2:

1) *Naive Chunking*: As a baseline, we implemented fixed-size token-based segmentation. This approach creates uniform chunks without considering semantic boundaries or document structure.

2) *Recursive Chunking*: Recursive chunking implements a hierarchical splitting algorithm that preserves the structure of the document while maintaining size restrictions. The process begins with coarse-grained delimiters (e.g., paragraph breaks) before proceeding to finer splits (e.g., sentence boundaries) only when necessary. This approach balances contextual integrity with size consistency, avoiding splits within semantic units when possible.

3) *Semantic Chunking*: Semantic chunking leverages sentence-transformer embeddings to group textual content based on inherent semantic relationships, enabling the preservation of topical coherence within document segments. The methodology identifies natural breakpoints by detecting significant similarity drops between consecutive sentence vectors in the embedding space, indicating potential shifts in thematic focus. To optimize chunk coherence, the algorithm merges smaller fragmentary sections with their most semantically similar neighbors, followed by the computation of quantitative coherence metrics to validate the semantic integrity of each resulting chunk.

### B. Metadata Enrichment Pipeline

A central component of this study is the systematic generation of metadata to enhance retrieval performance, as illustrated in the second part of Fig. 2. Our LLM-based metadata enrichment pipeline creates structured, semantic annotations for each document segment, which support more precise and context-aware retrieval.

For each chunk, the system generates three categories of metadata that capture different aspects of the document content. Content metadata includes content type classification spanning procedural, conceptual, reference, warning, and example categories, along with extracted keywords, entities, and code example detection. Technical metadata identifies primary and secondary categories, mentioned services, and technical tools referenced within the chunk. Semantic metadata provides concise summaries of chunk content, identifies user intents such as how-to guidance, debugging assistance, comparison information, or reference material, and generates potential user questions that the chunk addresses.

The metadata generation system processes chunked documents using a specialized LLM prompt engineered to extract structured metadata consistently. The system employs batched processing for efficiency and maintains a consistent schema across different chunking strategies. Generated metadata is stored alongside the original content in a structured format suitable for the embedding phase.

### C. Embedding Techniques

We evaluated the following distinct embedding approaches to represent document chunks in vector space:

1) *Snowflake Arctic-Embed Model*: The Snowflake Arctic-Embed model is considered as the primary embedding framework. This model was selected for its superior performance on technical documentation and its ability to effectively capture domain-specific terminology. The arctic-embed-m variant offers an optimal balance between retrieval

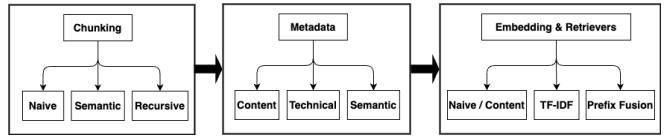


Fig. 2: RAG enhancement pipeline showing the progression from document chunking strategies (left), through metadata enrichment (center), to embedding generation and retrieval methods (right).

performance and computational efficiency, supporting the long contexts required for technical documentation [24].

2) *Embedding Approaches*: We systematically compared three embedding strategies that leverage the Arctic-Embed model in different configurations. Naive embeddings generate vector representations using only the raw chunk content without metadata integration, serving as our baseline approach. TF-IDF weighted embeddings combine content embeddings weighted at 70% with TF-IDF vectors derived from metadata weighted at 30%, creating a hybrid representation that incorporates both semantic content and metadata-derived statistical features. Prefix-fusion embeddings inject structured metadata directly into document text as formatted prefixes before the embedding process, allowing the model to jointly encode content and metadata in a unified representation.

### D. Retriever Architectures

We implement three distinct retriever architectures, each corresponding to a specific embedding methodology as shown in the third part of Fig. 2. The content retriever utilizes naive embeddings for baseline similarity matching between queries and documents. The TF-IDF retriever combines content embeddings with metadata-derived TF-IDF vectors using a 70:30 weighting scheme, enabling retrieval that considers both semantic content and metadata-derived features. The prefix-fusion retriever leverages embeddings generated with metadata prefixes and features automatic intent detection for query enhancement, allowing for more sophisticated query-document matching.

Each retriever architecture is instantiated across all three chunking strategies (semantic, recursive, and naive), yielding a  $3 \times 3$  experimental matrix comprising nine distinct retriever configurations. This design enables systematic ablation analysis to isolate the effects of both chunking granularity and embedding methodology on retrieval performance.

### E. Ground Truth Generation and Evaluation Framework

A critical component of our methodology involves establishing high-quality ground truth relevance judgments for retrieving document chunks. To achieve this, we employed a cross-encoder reranking approach, which provides fine-grained relevance assessments by jointly processing query and candidate chunk pairs.

Specifically, for each user query, we initially retrieve the top 50 candidate chunks using each retriever configuration under evaluation. These candidates are then scored by a

cross-encoder model—*BAAI/bge-reranker-base* [25]—which captures detailed interaction patterns between queries and chunks, surpassing the capabilities of bi-encoder models in relevance discrimination. The scores are then normalized to a common 0-1 range to enable a fair comparison across different retrievers. Subsequently, relevance rankings are generated based on the normalized scores, with higher scores indicating greater relevance. This process yields a robust set of ground truth labels, which serve as benchmarks for evaluating the accuracy of each retrieval configuration.

Employing such a cross-encoder based reranking method ensures that our ground truth is aligned with human judgment and reflects a realistic standard for relevance, thereby supporting rigorous, reproducible performance comparisons. This approach mitigates biases associated with raw similarity metrics and provides a solid foundation for quantitatively assessing the impact of metadata enrichment and chunking strategies on retrieval effectiveness.

This systematic framework provide a structured and reproducible approach for evaluating the impact of metadata enrichment within RAG pipelines. In the following section, we present a comparative analysis of different chunking strategies and embedding techniques, enabling empirical identification of the most effective configurations for retrieving technical documentation.

#### IV. EXPERIMENTAL RESULTS

In this section, we present a comprehensive analysis of the experimental outcomes obtained from evaluating our proposed systems. We begin by discussing the dataset used, including its composition and inherent challenges, followed by a detailed description of the experimental setup encompassing data processing, chunking strategies, embedding techniques, and retrieval architectures. Subsequently, we systematically compare the performance of different components—such as semantic, recursive, and naive chunking methods—across various evaluation metrics. This structured comparison enables us to identify the strengths and limitations of each approach, providing valuable insights into the optimal configurations for enhancing retrieval accuracy and efficiency in enterprise knowledge management contexts.

##### A. Data Corpus Composition and Experiment setup

Our study employed a carefully curated dataset derived from AWS S3 documentation (<https://docs.aws.amazon.com/s3/>), selected for its intricate structure and diverse content types. The corpus consists of four main components: the S3 User Guide (2,499 pages), the API Reference (3,013 pages), the S3 Glacier Developer Guide (558 pages), and the S3 on Outposts documentation (217 pages). This comprehensive dataset offers a realistic and challenging environment for assessing the effectiveness of metadata-enriched retrieval strategies in practical, real-world technical contexts.

We implemented our evaluation framework on technical documentation datasets processed through three distinct chunking strategies: recursive (max\_tokens=512, over-

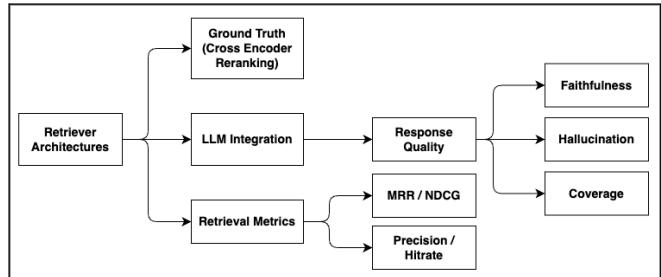


Fig. 3: Evaluation framework for RAG systems, covering retrieval metrics and response quality to assess both retrieval effectiveness and LLM integration.

lap=128), naive (fixed\_length=1024), and semantic (adaptive with max\_tokens=1024). For each strategy, we generated embeddings using the Snowflake Arctic-Embed model (dimensions=1536) with three configurations: content-only, TF-IDF weighted (70% content, 30% metadata), and prefix-fusion approaches.

Retrieval experiments were conducted using exact vector search with cosine similarity across varying k values. For ground truth generation, we employed a cross-encoder methodology using Snowflake Arctic-Embed to establish relevance judgments with threshold  $\tau=0.8$ . Statistical significance was assessed across all retrieval runs ( $p<0.05$ ).

Performance evaluation employed standard information retrieval metrics: Hit Rate@10, Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), and precision. We maintained consistent query processing procedures and embedding dimensionality across all experimental configurations to isolate the effects of metadata enrichment and chunking strategies.

Metadata generation utilized GPT-4o (gpt-4o-2024-05-13) with controlled temperature (0.5) and structured output formats for consistency.

##### B. Evaluation Metrics

We established a rigorous evaluation methodology as shown in Fig. 3 to assess the impact of metadata enrichment on RAG performance. The evaluation utilized a test set of diverse technical queries executed against our documentation corpus. Our framework evaluates both retrieval effectiveness through standard information retrieval metrics and response quality through semantic and factual accuracy measurements.

For the retriever evaluation, we used standard information retrieval metrics:

- **Hit Rate@k:** Proportion of queries with at least one highly relevant document (above 95th percentile threshold based on ground truth reranker scores) in the top-k results
- **Metadata Consistency:** Consistency of document categories within top-k results, measuring the retriever's tendency to return semantically coherent chunks
- **Precision@k:** Proportion of relevant documents among top-k results

- **MRR (Mean Reciprocal Rank):** Average reciprocal rank of first relevant document
- **NDCG@k:** Ranking Quality Based on Relevance and Positional Importance

### C. Embedding and Chunking Analysis

We first analyzed the characteristics of embeddings generated using different techniques and chunking methods.

The TF-IDF weighted embedding technique consistently demonstrated superior clustering characteristics across all chunking methods, as evidenced by the lowest average nearest neighbor distances (0.833-0.839). This indicates that TF-IDF embeddings create more cohesive semantic clusters, which typically translates to improved retrieval performance. Conversely, prefix-fusion embeddings exhibited the highest average distances, suggesting greater separation between vectors, which may enhance discriminative power but potentially reduce semantic connections between related content. Semantic chunking produced significantly more chunks (5,706) compared to recursive chunking (4,099), representing a 39% increase. While this results in a larger index size, the finer granularity provided by semantic chunking enables more precise retrieval targeting.

### D. Retriever Performance Analysis

Based on our evaluation metrics, retrieval performance demonstrates complex interactions between chunking strategies and embedding techniques. The Hit Rate@10 results (Table I) reveal that naive chunking with TF-IDF embeddings achieves the highest performance (0.925), closely followed by naive chunking with prefix-fusion (0.900). This contradicts conventional assumptions about semantic chunking superiority.

For MRR, which evaluates ranking quality (Table II), recursive chunking with content-only embeddings unexpectedly outperforms (0.713) both metadata-enriched approaches in the semantic configuration. Naive chunking with prefix-fusion demonstrates strong performance (0.750), suggesting that simpler chunking strategies can excel when appropriately augmented.

NDCG scores (Table III), which assess overall ranking quality with position weighting, favor recursive chunking with TF-IDF embeddings (0.807) and naive chunking with prefix-fusion (0.813). The precision metrics (Table IV) further confirm this pattern, with recursive chunking and TF-IDF embeddings achieving the highest precision (0.825).

These results demonstrate that no single configuration universally outperforms across all metrics, indicating that retrieval system design requires careful consideration of evaluation priorities. The consistent underperformance of certain content-only configurations confirms the value of metadata enrichment, though the specific implementation must be tailored to prioritized retrieval objectives.

### E. Response Quality Evaluation

We conducted supplementary quality assessment of RAG-generated responses across three retrieval configurations:

content-only baseline, TF-IDF weighted metadata enrichment, and prefix-fusion approaches. These evaluations served primarily as validation checks rather than primary experimental outcomes. The TF-IDF metadata-enriched approach demonstrated consistently superior performance across all quality dimensions compared to both the content-only baseline and prefix-fusion alternatives. Specifically, we observed substantial improvements in response faithfulness and corresponding reductions in hallucination rates with the TF-IDF approach. Coverage metrics similarly favored this configuration. These indicators, while directionally informative, should be interpreted with appropriate caution as they are susceptible to various confounding factors including reference selection methodology, language model parameterization, and prompt design considerations that remain outside the scope of our primary experimental controls.

### F. Key Findings and Insights

Our comprehensive evaluation revealed several important insights:

- First, the combination of TF-IDF embeddings with recursive chunking consistently delivered superior retrieval performance across key metrics, achieving a precision of 82.5% and NDCG of 0.807. The naive chunking strategy with prefix-fusion demonstrated remarkable effectiveness for hit rate metrics, reaching 0.900, while content-based approaches with recursive chunking yielded the highest MRR at 0.713. These performance differentials across metadata-enriched retrievers empirically validate the substantial impact of strategic chunking and embedding selection on RAG system effectiveness.
- Second, the analysis of retriever configurations revealed that naive chunking, when combined with prefix-fusion techniques, achieves the highest Hit Rate@10 (0.925), significantly outperforming semantic approaches. This challenges conventional assumptions about semantic boundary preservation and suggests that simpler chunking strategies may excel in specific retrieval scenarios when augmented with appropriate metadata enrichment.
- Third, recursive chunking demonstrated consistent performance across all embedding techniques, with particularly strong results for precision metrics (78.3%-82.5%). This stability across configurations indicates that recursive approaches provide robust document representations that maintain context integrity regardless of the underlying embedding methodology, offering a reliable foundation for production RAG systems where performance consistency is critical.

These results empirically validate that metadata enrichment frameworks must be carefully tailored to specific retrieval objectives. The findings establish quantitative benchmarks for implementing high-performance RAG systems, highlighting the non-trivial interactions between chunking strategies and embedding techniques in determining retrieval effectiveness across standardized information retrieval metrics.

TABLE I: Hit Rate (@10)

Retriever	Semantic	Naive	Recursive
Content	0.788	0.713	0.875
Prefix-Fusion	0.775	0.900	0.875
TF-IDF	0.775	0.925	0.825

TABLE II: MRR (@10)

Retriever	Semantic	Naive	Recursive
Content	0.669	0.538	0.713
Prefix-Fusion	0.534	0.750	0.673
TF-IDF	0.573	0.570	0.686

TABLE III: NDCG Across Configurations (@10)

Retriever	Semantic	Naive	Recursive
Content	0.730	0.669	0.782
Prefix-Fusion	0.699	0.813	0.800
TF-IDF	0.670	0.717	0.807

TABLE IV: Precision (@10)

Retriever	Semantic	Naive	Recursive
Content	0.733	0.640	0.783
Prefix-Fusion	0.745	0.798	0.785
TF-IDF	0.722	0.702	0.825

## V. CONCLUSION

This study demonstrates that integrating LLM-based metadata enrichment into retrieval pipelines significantly enhances the performance of RAG systems in technical documentation contexts. Our systematic evaluation of different chunking strategies—semantic, recursive, and naive—paired with advanced embedding techniques such as TF-IDF weighting and prefix-fusion, reveals nuanced interactions influencing retrieval metrics. Notably, recursive chunking combined with TF-IDF embeddings consistently delivers superior precision (up to 82.5%) and ranking quality, while naive chunking with prefix-fusion achieves the highest Hit Rate@10 (0.925). These findings challenge conventional assumptions about semantic chunking’s supremacy and highlight the importance of tailored configurations aligned with specific retrieval objectives.

Our framework employs cross-encoder reranking to establish high-quality relevance ground truth, enabling rigorous, reproducible evaluations. The empirical insights gained affirm that metadata enrichment enhances vector clustering, reduces retrieval latency, and improves overall system robustness—core requirements for enterprise-level knowledge bases.

Future work should explore scalability to larger datasets, dynamic metadata generation for evolving content, and integration within real-time retrieval systems. Overall, this research provides a practical, adaptable blueprint for deploying high-performance, metadata-optimized RAG architectures, advancing enterprise knowledge retrieval in complex, domain-specific settings.

## ACKNOWLEDGMENT

This project has been completed under the supervision of Fatemeh Sarayloo. The first four authors were pursuing their master’s programs while working on this project. We appreciate the support of the University of Illinois Business College.

## REFERENCES

- [1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Facebook AI Research, 2020.
- [2] K. Shuster et al., "Retrieval Augmentation Reduces Hallucination in Conversation," Facebook AI Research, 2021.
- [3] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," Facebook AI Research, 2020.
- [4] G. Izacard et al., "Atlas: Few-shot Learning with Retrieval Augmented Language Models," Meta AI Research, 2022.
- [5] Y. Gao, Y. Xiong, X. Gao, and others, "Retrieval-Augmented Generation for Large Language Models: A Survey," Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, 2023.
- [6] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," Chinese Academy of Sciences, 2023.
- [7] S. S. Sundaram and M. A. Musen, "Making Metadata More FAIR Using Large Language Models," Stanford University, 2023.
- [8] L. Wang et al., "Large Search Model: Redefining Search Stack in the Era of LLMs," Microsoft Corporation, 2023.
- [9] H. Husain et al., "CodeSearchNet Challenge: Evaluating the State of Semantic Code Search," GitHub, 2019.
- [10] H. Song, S. Bethard, and A. K. Thomer, "Metadata Enhancement Using Large Language Models," University of Arizona, 2024.
- [11] L. Mombaerts et al., "Meta Knowledge for Retrieval-Augmented Large Language Models," Amazon Web Services, 2024.
- [12] N. Harris, A. Butani, and S. Hashmy, "Enhancing Embedding Performance through Large Language Model-based Text Enrichment and Rewriting," Arizona State University, 2024.
- [13] S. G. Thottempudi and S. Borra, "Leveraging Large Language Models to Enhance an Intelligent Agent with Multifaceted Capabilities," SRH University Berlin, 2024.

- [14] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," Stanford University, 2024.
- [15] R. Anantha, T. Bethi, D. Vodianik, and S. Chappidi, "Context Tuning for Retrieval-Augmented Generation," Apple, 2024.
- [16] F. Cuconasu et al., "The Power of Noise: Redefining Retrieval for RAG Systems," Sapienza University of Rome, 2024.
- [17] Z. Shao et al., "Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy," Tsinghua University, 2024.
- [18] Y. Wang et al., "Self-Knowledge Guided Retrieval Augmentation for Large Language Models," Tsinghua University, 2024.
- [19] H.-T. Chen, F. Xu, S. A. Arora, and E. Choi, "Understanding Retrieval Augmentation for Long-Form Question Answering," University of Texas at Austin, 2024.
- [20] 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:2312.10997, 2024.
- [21] B. E. Perron, L. Goldkind, Z. Qi, and B. G. Victor, "Human Services Organizations and the Responsible Integration of AI: Considering Ethics and Contextualizing Risk(s)," *Journal of Technology in Human Services*, vol. 43, no. 1, pp. 20-33, 2025.
- [22] S. Liu, A. B. McCoy, and A. Wright, "Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines," *Journal of the American Medical Informatics Association*, vol. 32, no. 4, pp. 605-615, Jan. 2025.
- [23] A. Ramachandran, "Advancing Retrieval-Augmented Generation (RAG) Innovations, Challenges, and the Future of AI Reasoning," Feb. 2025.
- [24] Snowflake, "Snowflake Arctic Embed M V2.0: Multilingual Embedding Model," Dataloop, 2025. [Online]. Available: [https://dataloop.ai/library/model/snowflake\\_snowflake-arctic-embed-m-v20/](https://dataloop.ai/library/model/snowflake_snowflake-arctic-embed-m-v20/)
- [25] Beijing Academy of Artificial Intelligence, "bge-reranker-v2-m3," Hugging Face, 2025. [Online]. Available: <https://huggingface.co/BAAI/bge-reranker-v2-m3>