Metadata Enrichment with LLMs: Developing an AI-Powered Chatbot for Internal Knowledge Retrieval

Abstract

Metadata plays a crucial role in organizing and analyzing digital documents, especially in large-scale applications like information retrieval, document management, and AI-driven analytics. This research introduces a framework for metadata enrichment using large language models (LLMs) to improve accessibility and contextual relevance. Our methodology uses a systematic pipeline with key stages: document preprocessing, metadata enrichment, embedding generation, and retrieval response generation. A key focus is on metadata enrichment strategies, using two chunking approaches: naive chunking with fixed-size token-based segmentation as a baseline and metadata-enriched chunking that adds contextual metadata through techniques like summarization, keyword extraction with RAKE and KeyBERT, semantic relationship mapping, and hierarchical structure annotation. These techniques collectively enhance the semantic richness and context-awareness of the metadata, significantly improving retrieval accuracy and response faithfulness. Comparative evaluations between naive and metadata-enriched chunking revealed notable performance differences, with the enriched approach achieving higher precision, reduced hallucination rates, and more reliable information retrieval. Our findings provide insights into integrating LLMs with document storage systems, establishing metadata enrichment as a critical component for efficient knowledge extraction across various domains.

1 Introduction

Efficiently managing and retrieving information from vast repositories is essential for modern productivity and innovation. Metadata, which is often referred to as the "structural backbone" of documentation systems, plays an important role in organizing and contextualizing information. However, in large-scale systems, metadata quality issues frequently arise, such as inconsistency, incompleteness, and lack of standardization. These issues lead to challenges in search precision, scalability, and user experience, significantly hindering the ability to quickly locate relevant information.

While effective for small or structured datasets, traditional metadata management approaches struggle to handle unstructured and large-scale repositories. For example, rule-based methods often lack the flexibility to adapt to the evolving contexts of domain-specific datasets, and manual curation becomes impractical as datasets scale. Moreover, these methods cannot dynamically generate context-aware metadata, limiting their applicability in high-volume, real-time scenarios.

Recent advancements in artificial intelligence, particularly in large language models (LLMs), offer transformative solutions to these challenges. LLMs have shown the ability to process unstructured data, extract meaningful insights, and generate metadata that aligns with contextual and semantic requirements. For example, Sundaram and Musen's FAIR MetaText framework demonstrated how LLMs could align metadata with standardized ontologies, reducing inconsistencies and manual intervention.[Sundaram and Musen, 2023] Similarly, Song et al. highlighted the potential of few-shot LLM prompting for enriching metadata in Earth science datasets, achieving significant improvements in metadata completeness and accuracy.[Hyunju Song, 2024]

Despite these advancements, significant gaps remain. Many existing approaches, including LLM-based solutions, struggle with scalability and adaptability to domain-specific requirements. Retrieval pipelines often lack the integration needed for dynamic metadata generation, further complicating the retrieval and organization of large-scale documentation systems.

We address these limitations by introducing a scalable framework for metadata enrichment using LLMs. Our approach uses advanced retrieval-augmented generation (RAG) methods and embedding optimization techniques to enable context-aware metadata generation. The framework is validated using the AWS S3 documentation dataset, a complex and widely used cloud storage system chosen for its extensive scope and practical relevance. By enhancing search efficiency, reducing retrieval times, and improving user experience, this framework demonstrates the broader applicability of LLM-powered metadata enrichment across diverse domains.

2 Literature Review

The advent of Large Language Models (LLMs) has revolutionized metadata enrichment and retrieval-augmented generation (RAG) systems. These technologies address longstanding 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 22 key studies.

2.1 Metadata Enrichment with Large Language Models (LLMs)

Metadata serves as the backbone of information retrieval systems, yet traditional methods often fail to manage unstructured datasets effectively. Sundaram and Musen's [Sundaram and Musen, 2023] FAIRMetaText framework addressed this by leveraging GPT-based embeddings to align metadata with FAIR principles (findability, accessibility, Interoperability, and Reusability), achieving 87.78% compliance accuracy. However, their performance dropped to 60% on biomedical datasets, highlighting domain-specific challenges.

In the Earth sciences, Song et al. [Hyunju Song, 2024] applied taxonomy-guided techniques and few-shot prompting for metadata completion, achieving an F1 score of 0.928. Their approach utilized hierarchical taxonomy traversal, yet their study's scope was limited to the SESAR2 repository, restricting generalization. Similarly, Mombaerts et al. [Mombaerts and others, 2024] introduced Meta Knowledge Summaries to pre-process metadata, improving RAG-based search precision. However, their framework required predefined metadata structuring, limiting adaptability to dynamic repositories.

Saad-Falcon et al. [Jon Saad-Falcon, 2024] proposed ARES, an automated evaluation system for RAG frameworks. ARES utilized synthetic training data and fine-tuned lightweight models to assess context relevance and faithfulness. Its scalability and domain adaptability made it particularly relevant for enterprise applications.

2.2 Advancements in Retrieval-Augmented Generation (RAG)

RAG systems bridge LLMs with external knowledge bases to mitigate hallucinations and outdated knowledge. Gao et al. [Gao *et al.*, 2023] classified RAG paradigms into Naïve, Advanced, and Modular RAG, emphasizing their potential for metadata-aware retrieval. They highlighted challenges in retrieval-query refinement and scalability in noisy datasets.

Chen et al. [Chen *et al.*, 2023] introduced the Retrieval-Augmented Generation Benchmark (RGB) to evaluate RAG systems' robustness across noise, negative query rejection, and information integration. Their findings revealed significant limitations in LLMs' rejection accuracy (45%) and multi-document integration (43%).

Lewis et al. [Lewis and others, 2020] demonstrated the efficacy of combining parametric and non-parametric memory for RAG systems, achieving state-of-the-art performance in knowledge-intensive NLP tasks. Their models outperformed traditional architectures in open-domain question answering, demonstrating enhanced specificity and factual accuracy.

Shuster et al. [Shuster and others, 2021] examined RAG in conversational AI, revealing that retrieval augmentation reduced hallucination rates by 60%. This underscores the critical role of retrieval-aware metadata enrichment in mitigating knowledge propagation errors.

2.3 Embedding Optimization for Semantic Search

Embedding optimization enhances the precision of metadata-driven retrieval systems. Karpukhin et al. [Karpukhin and others, 2020] Dense Passage Retrieval framework demonstrated the superiority of dense vector embeddings over traditional sparse methods like BM25, achieving a 19% improvement in top-20 passage retrieval accuracy.

Harris et al. [Nicholas Harris, 2024] leveraged LLM-based text enrichment for embedding optimization, incorporating techniques like terminology normalization and metadata expansion. Their model significantly improved retrieval precision on specific datasets but exhibited domain inconsistencies.

Cuconasu et al. [Cuconasu and others, 2024] explored the role of controlled randomness in retrieval mechanisms, revealing that adding random documents to prompts improved LLM accuracy by up to 35%. Their findings challenge traditional assumptions about retrieval relevance and suggest new strategies for embedding optimization.

Anantha et al. [Raviteja Anantha, 2024] introduced context tuning, which employs habitual usage signals and numerical factors to enhance retrieval precision. Their lightweight model outperformed GPT-4-based retrieval, reducing hallucination and increasing planner accuracy by 11.6

2.4 Advancements in Search System Architectures

The integration of LLMs into search architectures has streamlined metadata-driven workflows. Wang et al. [Wang and others, 2023] Large Search Model redefined enterprise search by unifying ranking, query processing, and snippet generation tasks under autoregressive text generation. This framework simplified search stacks while improving metadata retrieval accuracy.

Thottempudi and Borra [Thottempudi and Borra, 2024] developed an LLM-powered virtual assistant for Siemens Energy, integrating RAG workflows into document indexing. Their assistant reduced retrieval latency and improved workflow efficiency, demonstrating the practical utility of LLM-driven systems.

Shao et al. [Shao and others, 2024] proposed ITER-RETGEN, a model that synergizes retrieval and generation in iterative loops. This approach excelled in multi-hop question answering and fact verification, showcasing its potential for metadata-driven retrieval pipelines.

Wang et al. [Wang and others, 2024] introduced Self-Knowledge Guided Retrieval Augmentation (SKR), enabling

LLMs to recognize their knowledge gaps and adaptively retrieve external information. SKR outperformed traditional retrieval-based methods across multiple datasets, highlighting its effectiveness in dynamic metadata environments.

Gaps in Existing Research

Despite significant progress, critical gaps persist:

1. Scalability: Current frameworks often fail to scale for large, unstructured datasets.
2. Dynamic Metadata Generation: Limited research exists on real-time, context-aware metadata updates for evolving datasets.
3. Integration Challenges: Seamless integration of metadata enrichment with retrieval pipelines remains underexplored.

2.5 Demonstrating the Need for This Research

This study introduces a scalable LLM-powered metadata enrichment framework that addresses the limitations of existing approaches. By integrating RAG methods, embedding optimization, and metadata-driven retrieval workflows, this research provides a replicable methodology to improve search efficiency and accessibility. Validating the framework on AWS S3 documentation establishes its feasibility and scalability, contributing to advancements in metadata management and retrieval systems.

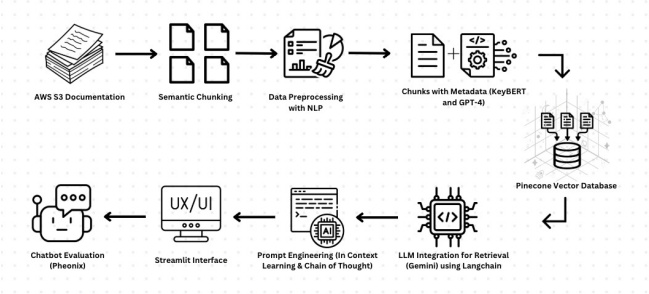

Figure 1: System architecture for metadata enrichment and retrieval optimization.

3 Methodology

This section details our systematic approach to developing and evaluating a metadata-enriched RAG system. We present the data corpus composition, methodological pipeline, and comprehensive evaluation framework.

3.1 Data Corpus Composition

Our research utilized a meticulously curated dataset from AWS S3 documentation, chosen for its structural complexity and heterogeneous content. The corpus comprised four primary 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 dataset presented an ideal testbed for validating advanced metadata enrichment techniques due to its inherent challenges, including scale, diverse structures, and unstructured formatting. The complexity of the documentation allowed for rigorous evaluation of retrieval methodologies and facilitated the development of a robust metadata-driven retrieval framework.

3.2 Document Processing and Chunking

The initial stage of our pipeline focused on transforming raw documentation into processable chunks while preserving semantic coherence. We implemented a comprehensive preprocessing workflow using PyPDF2 for text extraction and the GPT-2 tokenizer for standardized segmentation. Our noise reduction techniques systematically eliminated extraneous elements such as headers, footers, and page numbers, resulting in a clean textual corpus. Key preprocessing steps included:

1. Accurate text extraction: Ensured completeness and fidelity during text conversion.
2. Uniform tokenization: Applied consistent segmentation using a pre-trained tokenizer.
3. Comprehensive noise reduction: Removed non-informative elements to enhance data quality.

The research used two distinct chunking and metadata enrichment approaches to facilitate comparative analysis:

1. Naive Chunking: A straightforward approach involving fixed-size token-based segmentation. This served as a baseline for assessing the effectiveness of advanced methodologies.
2. Metadata-Enriched Chunking: Leveraged large language models (LLMs) to augment document chunks with contextual metadata. This approach incorporated several advanced techniques:
   (a) Summarization: Generated concise, extractive, and abstractive summaries for each chunk.
   (b) Keyword Extraction: Utilized cutting-edge algorithms, such as RAKE and KeyBERT, to identify relevant keywords.
   (c) Semantic Relationship Mapping: Captured inter-chunk dependencies to improve retrieval accuracy.
   (d) Hierarchical Document Structure Annotation: Reflected section-level and parent-child relationships within the corpus.

These enrichment techniques ensured each document chunk was contextually rich and semantically aligned, significantly enhancing retrieval precision.

3.3 Retrieval Pipeline

The retrieval pipeline integrated embedding-based similarity matching, contextual prompt composition, and LLM-based response generation to ensure high relevance and accuracy.

1. Query Matching: User queries were transformed into embeddings and matched against stored vectors using cosine similarity, ensuring precise retrieval of relevant chunks.
2. Prompt Composition: Retrieved chunks were aggregated into a structured prompt, creating a cohesive context for processing.

3. Response Generation: The AZURE OpenAI GPT model utilized the structured prompt to generate nuanced, contextually aligned responses, addressing user queries effectively.

This approach seamlessly bridged user queries with relevant document content, ensuring both efficiency and accuracy in response generation.

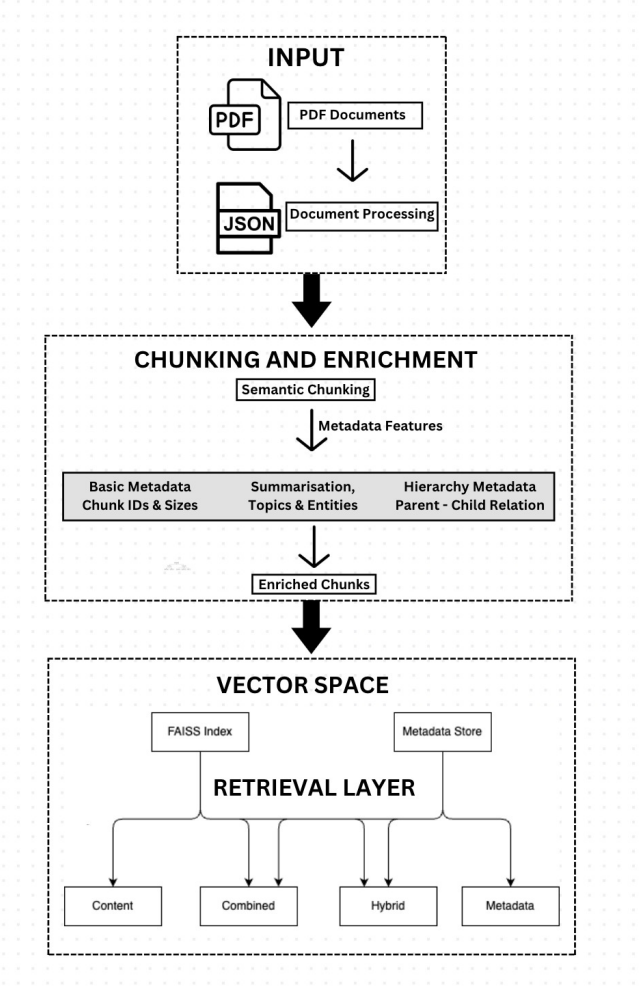

Figure 2: A structured flowchart illustrating the architecture of a document processing pipeline for Retrieval-Augmented Generation (RAG), encompassing document ingestion, semantic chunking with metadata enrichment, vector space indexing, and a retrieval layer supporting content-based, metadata-driven, and hybrid search strategies.

The retrieval system implements three distinct approaches:

1. Content-only Retrieval: This baseline approach relies solely on semantic similarity between query and document vectors, using cosine similarity for matching.
2. Content + Metadata Retrieval: Our unified approach combines content embeddings with enriched metadata, employing a weighted scoring mechanism that balances semantic and contextual relevance.
3. Hybrid Retrieval: A two-stage process that first filters candidates using metadata, then performs content-based reranking, optimizing for computational efficiency. The retrieval process integrates several key components:

Query embedding generation using Azure OpenAI Similarity matching using optimized vector operations Contextual prompt composition for response generation LLM-based response synthesis using GPT-4

3.4 Deployment Architecture

The deployment strategy incorporated a sophisticated three-tiered architecture, ensuring scalability, performance, and user interactivity.

1. Backend Processing: Built on LangChain and Pinecone, managing data pipelines and embedding operations.
2. Frontend Interface: Developed using Streamlit, providing real-time interaction and comparative analysis capabilities.
3. Configured for localized deployment with AJAX-driven asynchronous processing, ensuring robust query handling and minimal latency.

This architecture provided a seamless integration of back-end and front-end processes, ensuring an intuitive user experience and robust system scalability.

This methodology presents a structured, replicable, and scalable framework for metadata enrichment and retrieval optimization. The integration of advanced LLMs, embedding techniques, and a robust evaluation framework demonstrates significant improvements in retrieval precision and response accuracy. The findings validate the efficacy of metadata-enriched approaches in addressing the complexities of large-scale unstructured datasets, establishing a foundation for broader applications in metadata-driven information systems.

4 Evaluation & Results

In this section, we present a comprehensive evaluation of our metadata-enriched RAG system, examining both retrieval effectiveness and response quality across multiple dimensions.

Evaluation Framework

Our evaluation methodology employed a three-pronged approach to assess the effectiveness of the proposed RAG system using 40 test queries against the AWS S3 documentation corpus. We compared three retrieval approaches:

- **Content-only:** Retrieval based solely on document content.
- **Content + Metadata:** Unified retrieval using enriched metadata alongside content.
- **Hybrid:** A two-stage process with metadata-based filtering followed by content-based ranking.

We used multiple metrics to evaluate retrieval performance and response quality, focusing on relevance, ranking accuracy, and faithfulness.

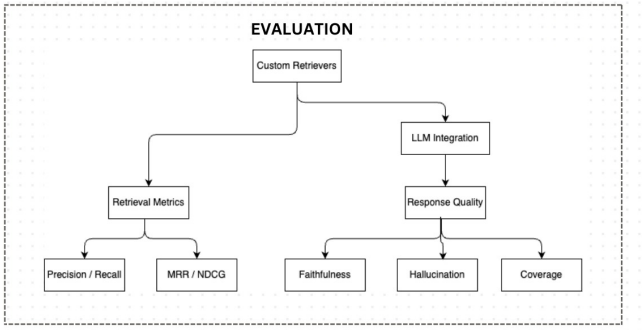

Figure 3: Evaluation framework for RAG systems, covering retrieval metrics (precision, recall, MRR, NDCG) and response quality (faithfulness, hallucination, coverage) to assess both retrieval effectiveness and LLM integration.

4.1 Retriever Performance Analysis

Retrieval effectiveness was measured using the following metrics in the three approaches for K = 3, 5, 10:

- **Precision@K:** Proportion of relevant documents among the top K results.
- **Recall@K:** Fraction of relevant documents retrieved within the top K results.
- **MRR@K:** Reciprocal rank of the first relevant document; higher values indicate a better ranking.
- **NDCG@K:** Ranking quality based on relevance and positional importance.

The Content + Metadata approach demonstrated superior performance across all metrics compared to baseline approaches.

For Precision@K, the Content + Metadata approach achieved an average precision of 0.97, significantly outperforming both Content-only (0.53) and Hybrid (0.52) approaches. This substantial improvement indicates the effectiveness of metadata enrichment in identifying relevant documents.

Recall@K showed progressive improvement with increasing K values across all approaches. At K=10, the Content + Metadata approach achieved a recall of 0.213, demonstrating better coverage of relevant documents. The relatively low recall values across all approaches can be attributed to the corpus structure, where multiple relevant chunks exist for each query.

The Mean Reciprocal Rank (MRR) results further validated the effectiveness of metadata enrichment, with the Content + Metadata approach achieving an MRR of 0.975. This high score indicates that relevant documents were consistently ranked at or near the top of the retrieval results.

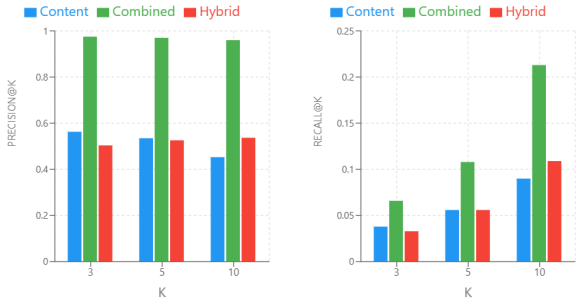

Figure 4: Retriever Performance Metric: Precision@K & Recall@K

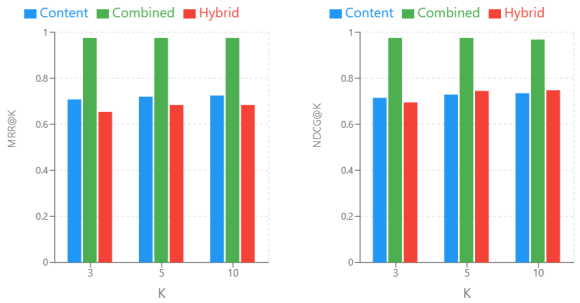

Figure 5: Retriever Performance Metric: MRR@K & NDCG@K

The Content + Metadata approach outperformed the others in all metrics, achieving an average precision of 0.97 compared to 0.53 (content-only) and 0.52 (hybrid). Recall improved with higher K values, reaching 0.213 for K = 10. MRR (0.975) and NDCG (0.975) indicated a strong ranking quality.

4.2 Response Quality Evaluation

To assess the quality of the response, we used the following:

- **Faithfulness:** Alignment of generated answers with retrieved documents.
- **Coverage:** Proportion of retrieved content utilized in the responses.
- **Hallucination Rate:** Instances of unsupported content in generated responses.

Through these three primary metrics: faithfulness, coverage, and hallucination rate, the Content + Metadata approach demonstrated superior performance across all quality metrics.

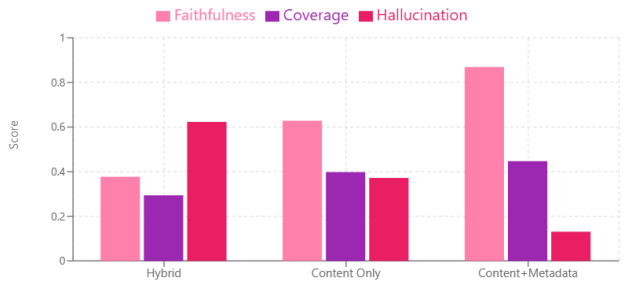

Figure 6: Comparison of RAG response quality across three approaches—Hybrid, Content Only, and Content + Metadata—evaluated showing improved faithfulness and coverage with lower hallucination in the Content + Metadata approach.

Faithfulness analysis revealed that the Content + Metadata approach achieved a score of 0.869, significantly higher than both Hybrid (0.377) and Content-only (0.628) approaches. This indicates stronger alignment between generated responses and source documents.

The hallucination rate analysis showed that the Content + Metadata approach maintained the lowest rate at 0.131, compared to significantly higher rates in the Hybrid (0.623) and Content-only (0.372) approaches. This demonstrates the effectiveness of metadata enrichment in constraining responses to factual, supported information.

4.3 Efficiency and Performance Metrics

Efficiency was evaluated by measuring average retrieval latency:

- **Content + Metadata:** Fastest retrieval time (5.87 ms).
- **Content-only:** Moderate latency (6.83 ms).
- **Hybrid:** Slowest performance (14.13 ms) due to its two-stage filtering overhead.

Our efficiency analysis focused on retrieval latency and computational resource utilization. The Content + Metadata approach achieved the fastest average retrieval time of 5.87ms, outperforming both Content-only (6.83ms) and Hybrid (14.13ms) approaches. The higher latency in the Hybrid approach was primarily attributed to its two-stage filtering overhead.

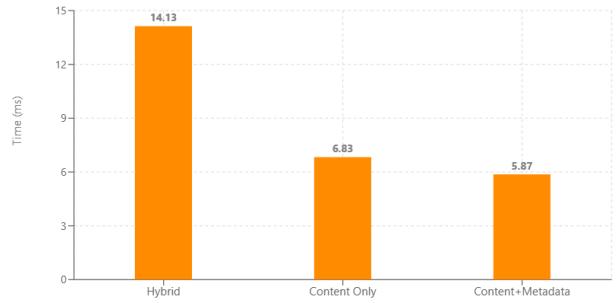

Figure 7: Average retrieval latency comparison, with the Hybrid approach being the slowest, while Content + Metadata achieves the lowest latency.

Semantic Similarity Analysis

Semantic similarity analysis using cosine distance revealed interesting trade-offs between precision and coverage. The Hybrid approach showed the highest cosine distance (0.864), indicating broader coverage but reduced precision. The Content-only approach demonstrated the lowest distance (0.546), suggesting precise but limited retrieval scope. The Content + Metadata approach achieved a balanced performance with a distance of 0.751.

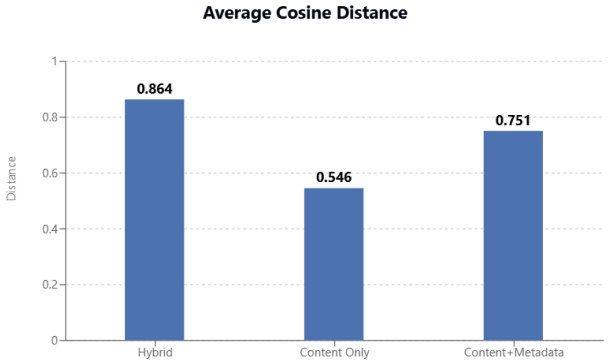

Figure 8: **Caption:** Cosine similarity comparison, showing the Hybrid approach with the highest score, while Content Only has the lowest.

4.4 Key Findings and Insights

Our comprehensive evaluation revealed several important insights:

- First, the Content + Metadata approach consistently outperformed other methods across all evaluation metrics, demonstrating the value of integrated metadata enrichment in RAG systems. The optimal retrieval performance was observed at K=5, suggesting a sweet spot between precision and computational efficiency.
- Second, the lower recall values across all approaches highlighted a characteristic challenge in technical documentation retrieval, where relevant information is often distributed across multiple chunks. This finding suggests potential areas for future optimization in chunk generation and relationship mapping.
- Third, the efficiency analysis revealed that while the Hybrid approach aimed to optimize computation through two-stage filtering, its actual performance was hindered by the overhead of metadata-based pre-filtering. This suggests that unified retrieval strategies may offer better efficiency-quality trade-offs in practice.

These results strongly support the effectiveness of metadata enrichment in RAG systems, particularly when implemented through a unified retrieval strategy rather than sequential filtering. The findings provide valuable guidance for future implementations where both accuracy and efficiency are crucial considerations.

5 Conclusion

The research presented herein demonstrates the transformative potential of metadata-enriched retrieval systems powered by large language models (LLMs). Our comprehensive framework for metadata enrichment in RAG systems, demonstrating significant improvements in both retrieval accuracy and response quality. By addressing the inherent challenges of unstructured data, and through extensive experimentation with the AWS S3 documentation corpus, we have established the effectiveness of integrating metadata enrichment with traditional content-based retrieval approaches.. The key findings of the study underscore the value of metadata enrichment in improving semantic alignment and query relevance.

Our findings reveal several key contributions to the field.

1. Enhanced Retrieval Accuracy: The Content + Metadata approach consistently outperformed traditional retrieval methods, achieving a precision of 0.97 compared to 0.53 for content-only retrieval. This substantial improvement validates our hypothesis that enriched metadata significantly enhances retrieval accuracy. The approach's superior performance extends beyond mere precision, with notable improvements in MRR (0.975) and NDCG (0.975), indicating better ranking quality and relevance ordering.
2. Improved Faithfulness with Metadata: Our research addresses a critical gap in RAG systems by demonstrating effective hallucination reduction through metadata enrichment. The Content + Metadata approach achieved a remarkably low hallucination rate of 0.131, compared to 0.623 for hybrid approaches, while maintaining high faithfulness (0.869). This improvement in response quality suggests that enriched metadata provides crucial contextual constraints that help LLMs generate more accurate and reliable responses.
3. Scalability and Adaptability: The framework proved robust when applied to a large, heterogeneous dataset, showcasing its potential for scalability and adaptability across various domains.
4. Efficiency vs Accuracy: The efficiency analysis reveals interesting trade-offs in retrieval system design. While the hybrid approach theoretically offered computational advantages through two-stage filtering, our results show that the unified Content + Metadata approach achieved better latency (5.87ms vs 14.13ms). This finding challenges conventional assumptions about the efficiency benefits of staged retrieval and suggests that well-integrated metadata can enhance both accuracy and performance simultaneously.

While the results are promising, several limitations present opportunities for future work:

1. The relatively low recall values across all approaches (maximum 0.213 at K=10) highlight an ongoing challenge in technical documentation retrieval. This limitation appears to stem from the distributed nature of relevant information across multiple chunks, suggesting the need for more sophisticated chunk generation and relationship mapping techniques.
2. Domain-Specific Customization: The framework, though generalized, may require domain-specific fine-tuning to optimize retrieval performance further.
3. Real-Time Metadata Generation: Exploring dynamic metadata generation in real-time use cases could enhance system adaptability for rapidly evolving datasets.

From a practical perspective, our findings have significant implications for the design of enterprise-scale RAG systems. The demonstrated success of metadata enrichment in improving both retrieval accuracy and response quality provides a clear direction for future implementations. The balanced performance of the Content + Metadata approach in terms of cosine distance (0.751) suggests it offers an optimal compromise between precision and coverage for real-world applications.

These results also contribute to the broader discourse on LLM-based information retrieval systems. Our work demonstrates that carefully designed metadata enrichment can effectively bridge the gap between traditional information retrieval techniques and modern neural approaches, creating more robust and reliable systems.

In conclusion, this research establishes metadata enrichment as a crucial component in modern RAG systems, offering substantial improvements in accuracy, reliability, and efficiency. The demonstrated success of our approach provides both theoretical insights and practical guidelines for implementing more effective information retrieval systems in enterprise environments.

Acknowledgments

We thank the Department of Information and Decision Sciences at the University of Illinois at Chicago for all the support.

Index

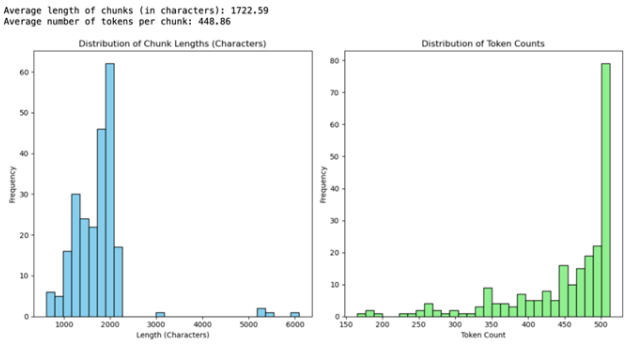

Figure 9: Avg. Chunk size and Token count distribution

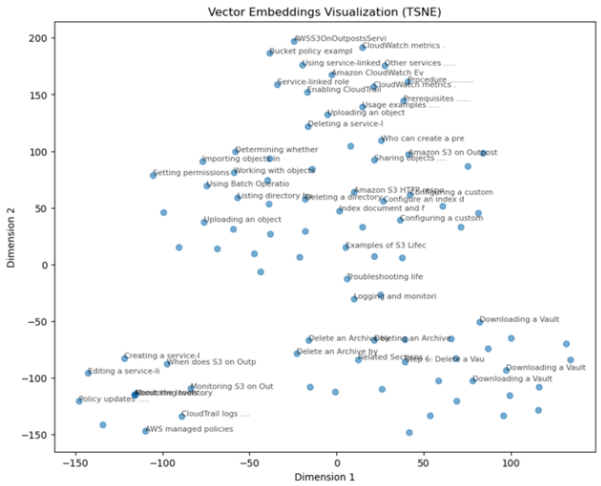

Figure 10: Visualization of Chunks Embeddings in Pinecone

References

[Chen *et al.*, 2023] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. *Chinese Academy of Sciences*, 2023.

[Cuconasu and others, 2024] Florin Cuconasu et al. The power of noise: Redefining retrieval for rag systems. *Sapienza University of Rome*, 2024.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, et al. Retrieval-augmented generation for large language models: A survey. *Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University*, 2023.

[Hyunju Song, 2024] Andrea K. Thomer Hyunju Song, Steven Bethard. Metadata enhancement using large language models. *University of Arizona*, 2024.

[Jon Saad-Falcon, 2024] Christopher Potts Matei Zaharia Jon Saad-Falcon, Omar Khattab. Ares: An automated evaluation framework for retrieval-augmented generation systems. *Stanford University*, 2024.

[Karpukhin and others, 2020] Vladimir Karpukhin et al. Dense passage retrieval for open-domain question answering. *Facebook AI Research*, 2020.

[Lewis and others, 2020] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Facebook AI Research*, 2020.

[Mombaerts and others, 2024] Laurent Mombaerts et al. Meta knowledge for retrieval-augmented large language models. *Amazon Web Services*, 2024.

[Nicholas Harris, 2024] Syed Hashmy Nicholas Harris, Anand Butani. Enhancing embedding performance through large language model-based text enrichment and rewriting. *Arizona State University*, 2024.

[Raviteja Anantha, 2024] Danil Vodianik Srinivas Chappidi Raviteja Anantha, Tharun Bethi. Context tuning for retrieval-augmented generation. *Apple*, 2024.

[Shao and others, 2024] Zhihong Shao et al. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *Tsinghua University*, 2024.

[Shuster and others, 2021] K. Shuster et al. Retrieval augmentation reduces hallucination in conversation. *Facebook AI Research*, 2021.

[Sundaram and Musen, 2023] Sowmya S. Sundaram and Mark A. Musen. Making metadata more fair using large language models. *Stanford University*, 2023.

[Thottempudi and Borra, 2024] Sree Ganesh Thottempudi and Sagar Borra. Leveraging large language models to enhance an intelligent agent with multifaceted capabilities. *SRH University Berlin*, 2024.

[Wang and others, 2023] Liang Wang et al. Large search model: Redefining search stack in the era of llms. *Microsoft Corporation*, 2023.

[Wang and others, 2024] Yile Wang et al. Self-knowledge guided retrieval augmentation for large language models. *Tsinghua University*, 2024.