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

Kranti Yeole[1] (kyeole2@uic.edu), Ramyashree Keshavamurthy[2] (rkesh@uic.edu), Vanessa Beck (vbeck4@uic.edu), Michael Moreno (mmoren42@uic.edu), Michael Uretzky (muretz2@uic.edu) and Fatemeh Sarayloo (fsaraylo@uic.edu)

University of Illinois, Chicago IL – 60607, USA

Abstract – Metadata plays a crucial role in organizing, accessing, and analyzing digital documents, particularly in large-scale applications such as information retrieval, document management, and AI-driven analytics. However, the quality and completeness of metadata often vary, posing challenges for effective utilization. This research introduces a novel framework for **Metadata Enrichment** using **Large Language Models (LLMs)** to enhance the accessibility and contextual relevance of digital documents. The study outlines an end-to-end methodology, including document chunking strategies, embedding generation, and the application of LLMs to extract, infer, and enrich metadata. The framework was evaluated using a diverse dataset, incorporating naive and metadata-enriched ***embeddings*** stored in vector databases, enabling comparison across retrieval and augmentation tasks. Our findings highlight the significant improvements in metadata accuracy, relevance, and utility, demonstrating the transformative potential of LLMs in automated metadata enrichment. This paper contributes to the field of AI-powered document management by offering actionable insights into integrating LLMs with document storage systems for efficient knowledge extraction and retrieval. The implications span various domains, from enterprise search systems to large-scale archival solutions, positioning LLM-driven metadata enrichment as a cornerstone for future advancements.

Keywords – Metadata Enrichment, Large Language Models, Document Chunking, Information Retrieval, Vector Databases, Embedding Generation

1 Introduction

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

Traditional metadata management approaches, while effective for small or structured datasets, struggle to handle unstructured and large-scale repositories. For instance, 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 are not equipped to generate context-aware metadata dynamically, limiting their applicability in high-volume, real-time scenarios.

Recent advancements in artificial intelligence, particularly 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.[1] 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.[2]

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.

This research addresses these limitations by introducing a scalable framework for metadata enrichment using LLMs. Our approach employs 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

Metadata management has evolved as a cornerstone for optimizing the accessibility and usability of documentation systems. While traditional methods have relied on manual curation and rule-based algorithms, these approaches often fall short in handling unstructured data, addressing scalability, and generating dynamic metadata. Recent advancements in LLMs and retrieval-augmented generation (RAG) frameworks have paved the way for addressing these challenges.

2.1 Metadata Enrichment with LLMs

Metadata serves as the foundation for information retrieval systems, yet its quality is often compromised by inconsistencies and incompleteness. Sundaram and Musen (2024) introduced FAIRMetaText, an NLP-driven framework for metadata alignment based on FAIR principles.[1] Their study demonstrated how LLMs could improve metadata quality and reduce manual curation by aligning descriptions with standardized ontologies. However, they identified scalability and domain-specific adaptability as significant challenges.[1]

Similarly, Song et al. (2024) leveraged few-shot LLM prompting and hierarchical taxonomy traversal to enrich missing metadata fields in Earth science datasets.[2] Their approach achieved an F1 score of 0.928, showcasing the potential of LLMs for contextual metadata generation. However, their methodology was limited to specific datasets and lacked generalizability to diverse documentation repositories.[2]

2.2 Advancements in Retrieval-Augmented Generation

RAG frameworks bridge the gap between retrieval and generation by combining structured retrieval with LLM capabilities. Gao et al. (2023) identified three key RAG paradigms—Naive, Modular, and Advanced—highlighting their potential for improving retrieval precision while addressing challenges like hallucinations and outdated information.[3] However, they noted that RAG systems often struggled to scale in noisy, real-world datasets.[3]

Chen et al. (2023) proposed the Retrieval-Augmented Generation Benchmark (RGB) to evaluate RAG systems' noise robustness and contextual integration.[4] Their findings indicated that RAG systems outperform standalone LLMs in complex retrieval scenarios. Nonetheless, handling dynamic datasets, such as AWS S3 documentation, remains a persistent challenge .[4]

2.3 Embedding Optimization for Semantic Search

Embedding optimization is critical for improving semantic search and metadata-driven retrieval systems. Harris et al. (2024) demonstrated how LLM-based text enrichment significantly enhances embedding quality, resulting in improved search precision for domain-specific datasets.[6] However, they emphasized the high computational costs associated with embedding generation, which limits their applicability in large-scale applications.[6]

2.4 Integration of LLMs in Search Systems

LLMs have redefined traditional search stacks by integrating tasks like metadata generation, document retrieval, and summarization under a unified framework. Wang et al. (2024) proposed the concept of a "large search model," simplifying search architectures for complex queries.[7][8] While promising, their study underscored the scalability challenges in handling high query loads and real-time applications.[7][8]

2.5 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

2.6 Demonstrating the Need for This Research

This study introduces a scalable, LLM-powered framework for metadata enrichment that addresses the limitations of existing approaches. By integrating RAG methods, embedding optimization, and metadata-driven retrieval workflows, this research provides a replicable methodology for improving 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.

3 Methodology

This section outlines the research methodology employed in this Project, detailing the systematic approach used to analyze Different Models, The RAG framework integrates **retrieval mechanisms** with **large language models (LLMs)** to augment their generative capabilities. Instead of solely relying on pre-trained knowledge in the LLM, RAG retrieves relevant information from external datasets or databases and uses it to provide informed and accurate outputs.

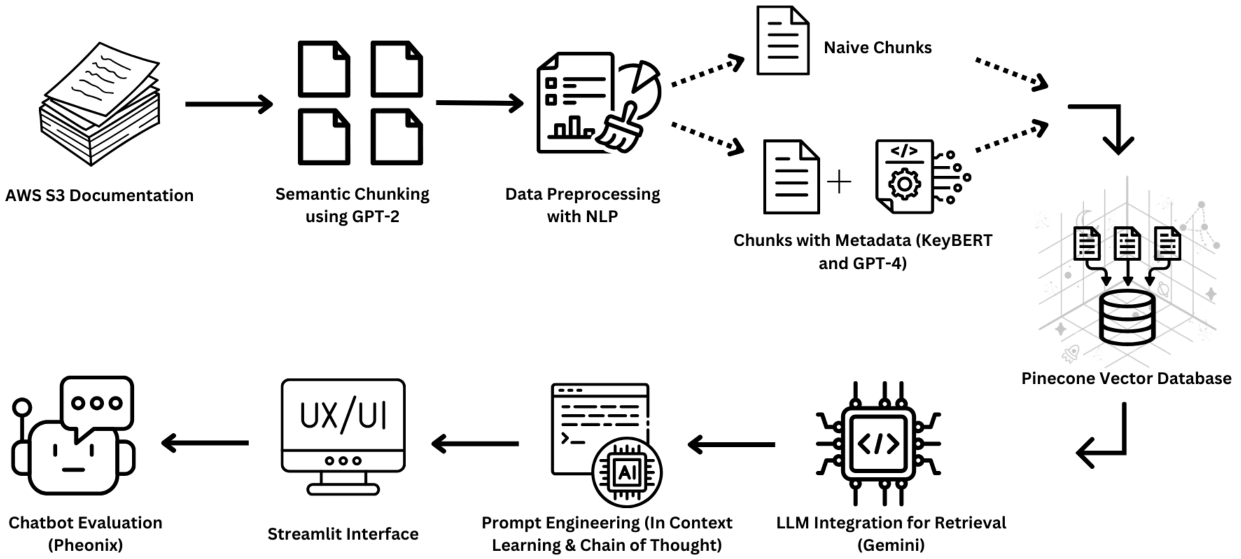

AWS S3 Documentation
Semantic Chunking using GPT-2
Data Preprocessing with NLP
Naive Chunks
Chunks with Metadata (KeyBERT and GPT-4)
Pinecone Vector Database
LLM Integration for Retrieval (Gemini)
Prompt Engineering (In Context Learning & Chain of Thought)
Streamlit Interface
UX/UI
Chatbot Evaluation (Pheonix)
</>

Fig. 1. RAG Framework: This diagram represents a pipeline for developing an AI-powered chatbot leveraging AWS S3 documentation

3.1 Data Corpus Composition

The research utilized a meticulously curated dataset from AWS S3 documentation, deliberately chosen for its structural complexity and heterogeneous content. The corpus comprised four distinct components: the **S3 User Guide** (2,499 pages), providing comprehensive operational insights; the **API Reference** (3,013 pages), detailing technical specifications; the **S3 Glacier Developer Guide** (558 pages), focusing on long-term storage mechanisms; and the **S3 on Outposts** documentation (217 pages), exploring hybrid infrastructure integration.

This dataset presented an ideal testbed for validating advanced metadata enrichment techniques due to its inherent challenges, including large 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 Methodological Pipeline

The proposed pipeline comprises four key stages: document preprocessing, metadata enrichment, embedding generation, and retrieval-response generation. Each stage was designed to address specific challenges associated with unstructured data management while ensuring scalability and replicability.

3.2.1 Document Preprocessing

The initial stage involved comprehensive preprocessing to transform the raw dataset into a machine-readable format. Using PyPDF2, text was systematically extracted from PDF documents and converted into .txt files. The GPT-2 tokenizer was employed to segment the text uniformly, ensuring standardized chunking across the dataset.

To maximize the signal-to-noise ratio, sophisticated noise reduction techniques were implemented. These techniques systematically eliminated extraneous elements, such as headers, footers, and page numbers, resulting in a clean textual corpus optimized for subsequent metadata generation.
Key preprocessing steps included:

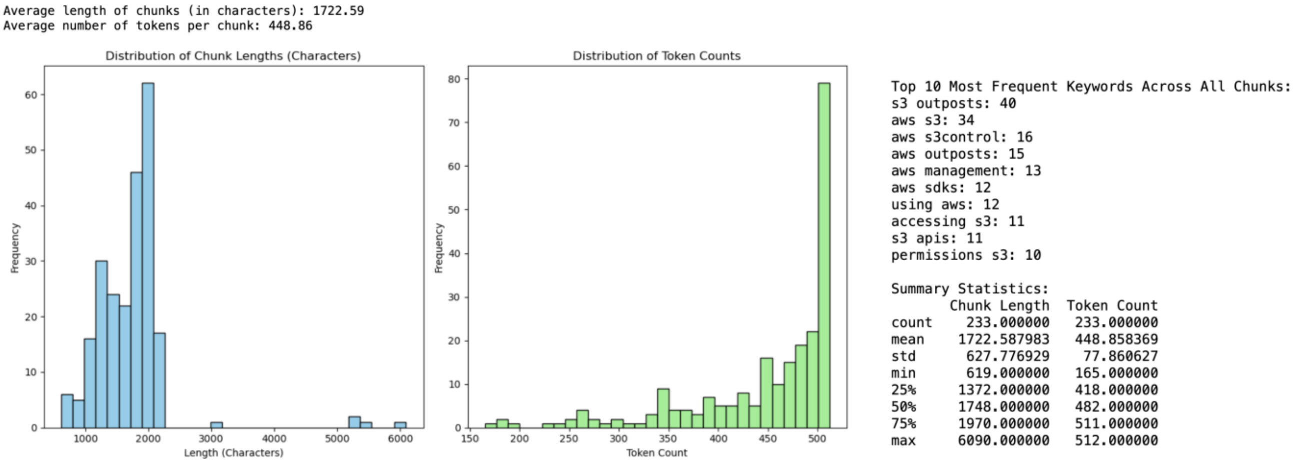
- Accurate text extraction: Ensured completeness and fidelity during text conversion.
- Uniform tokenization: Applied consistent segmentation using a pre-trained tokenizer.
- Comprehensive noise reduction: Removed non-informative elements to enhance data quality.

3.2.2 Metadata Enrichment Strategies

The research employed 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:
    - **Summarization**: Generated concise, extractive, and abstractive summaries for each chunk.
    - **Keyword Extraction**: Utilized cutting-edge algorithms, such as **RAKE** and **KeyBERT**, to identify relevant keywords.
    - **Semantic Relationship Mapping**: Captured inter-chunk dependencies to improve retrieval accuracy.
    - **Hierarchical Document Structure Annotation**: Reflected section-level and parent-child relationships within the corpus.

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



```
Average length of chunks (in characters): 1722.59
Average number of tokens per chunk: 448.86
```

```
Top 10 Most Frequent Keywords Across All Chunks:
s3 outposts: 40
aws s3: 34
aws s3control: 16
aws outposts: 15
aws management: 13
aws sdks: 12
using aws: 12
accessing s3: 11
s3 apis: 11
permissions s3: 10

Summary Statistics:
       Chunk Length  Token Count
count    233.000000   233.000000
mean    1722.587983   448.858369
std      627.776929    77.860627
min      619.000000   165.000000
25%     1372.000000   418.000000
50%     1748.000000   482.000000
75%     1970.000000   511.000000
max     6090.000000   512.000000
```

Fig. 2. Avg. Chunk size and Token count distribution

3.2.3 Metadata Enrichment Strategies

High-dimensional vector representations were generated for all document chunks to enable semantic search. Azure OpenAI Embeddings were employed to compute these vectors, which were then stored in Pinecone, a high-performance vector database.

To facilitate a comparative evaluation, two parallel vector databases were created:

- Naive Embeddings Database: Containing vector representations of chunks without metadata.
- Metadata-Enriched Embeddings Database: Storing vectors generated from enriched chunks.

This dual-database strategy enabled a comprehensive assessment of the impact of metadata enrichment on retrieval performance.

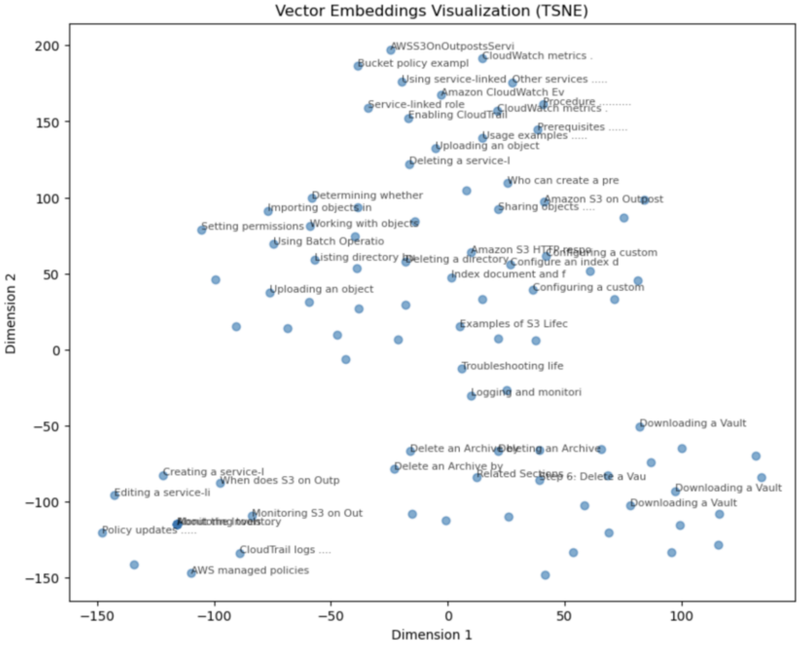


Fig. 3. Visualization of Chunks Embeddings in Pinecone.

3.3 Retrieval and Response Generation

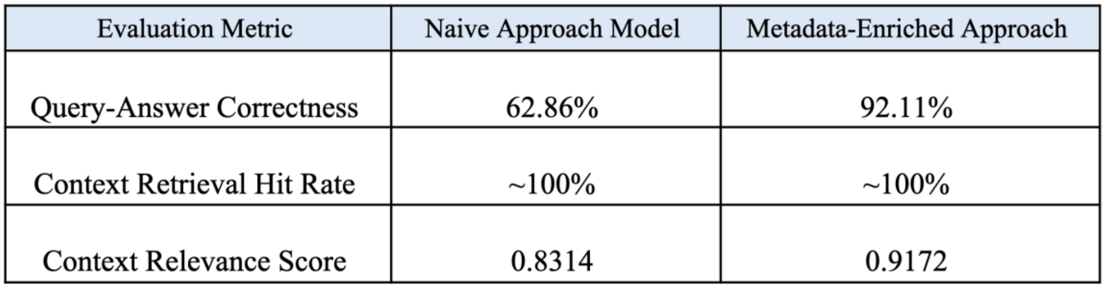
The retrieval pipeline integrated embedding-based similarity matching, contextual prompt composition, and LLM-based response generation to ensure high relevance and accuracy.

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.

3.4 Evaluation Metrics

The performance of the proposed methodology was rigorously evaluated using a combination of context relevance and query-answer correctness metrics. A comparison was conducted between the naive and metadata-enriched approaches, as detailed below:



Evaluation Metric	Naive Approach Model	Metadata-Enriched Approach
Query-Answer Correctness	62.86%	92.11%
Context Retrieval Hit Rate	~100%	~100%
Context Relevance Score	0.8314	0.9172

Fig. 4. Evaluation Metrics for Both the Models.

- **Query-Answer Correctness**: The metadata-enriched approach demonstrated a significant improvement in correctness, achieving a rate of 92.11%, compared to 62.86% for the naive method. This improvement highlights the impact of metadata enrichment in aligning retrieved information with user queries.

- **Context Retrieval Hit Rate**: Both approaches achieved near-perfect hit rates, indicating the system's ability to retrieve relevant document chunks consistently.
- **Context Relevance Score**: The enriched approach showed a marked improvement in relevance scores, underscoring the importance of metadata in improving semantic alignment.

The results address the initial problem statement by demonstrating that metadata-enriched retrieval significantly enhances both the precision and relevance of retrieved information. These findings validate the hypothesis that context-aware metadata generation improves the effectiveness of document retrieval systems.

- **Scalability and Robustness**: The ability of the framework to handle a large, heterogeneous dataset demonstrates its scalability and robustness, making it suitable for diverse domains beyond AWS S3 documentation.

The findings align with the conclusions of previous studies in the literature:

- **Retrieval-Augmented Generation (RAG)**: Consistent with Gao et al. (2023), the study demonstrates that metadata-enriched embeddings outperform naive approaches in terms of precision and relevance.[3]
- **Semantic Search with LLMs**: The results corroborate the findings of Zhu et al. (2023), which emphasized the transformative role of LLMs in improving semantic search capabilities.[5]

Differences observed include the integration of hierarchical metadata and real-time adaptability in this study, which were not explicitly explored in earlier works.

An unexpected finding was the relatively high hit rate (~100%) achieved by both naive and enriched approaches. While this indicates robust embedding generation, it suggests that the naive approach was sufficient for basic relevance retrieval but lacked the precision required for detailed contextual responses.

Additionally, the study revealed that enriched embeddings significantly outperformed naive embeddings in queries requiring detailed and domain-specific information. This finding underscores the importance of metadata-enrichment in handling complex queries and suggests potential applications in highly specialized fields, such as legal or medical documentation.

3.5 Deployment Architecture

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

1. **Backend Processing**: Managed data pipelines, embedding storage, and API communications. Implemented using **LangChain** and **Pinecone**.
2. **Frontend Interface**: A user-friendly interface was developed using **Streamlit**, enabling real-time interaction and facilitating a comparative analysis between naive and enriched databases.
3. **Execution Environment**: Configured for localized deployment with **AJAX-driven asynchronous processing**, ensuring robust handling of multiple queries and minimizing latency.

This architecture provided a seamless integration of backend and frontend 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 Conclusion

The research presented herein demonstrates the transformative potential of metadata-enriched retrieval systems powered by Large Language Models (LLMs). By addressing the inherent challenges of unstructured data, the proposed methodology significantly enhances retrieval precision, response accuracy, and user experience. The key findings of the study underscore the value of metadata enrichment in improving semantic alignment and query relevance, with notable improvements in query-answer correctness (92.11% for metadata-enriched systems versus 62.86% for naive approaches) and consistent high context retrieval hit rates (~100%).

4.1 Main Findings and Significance

1. **Enhanced Retrieval Accuracy**: The study validates that integrating contextual metadata with chunk embeddings substantially improves query relevance and response precision, making information retrieval more efficient and accurate.
2. **Scalability and Adaptability**: The framework proved robust when applied to a large, heterogeneous dataset, showcasing its potential for scalability and adaptability across various domains.
3. **Practical and Academic Implications**: The methodology bridges the gap between theoretical advancements in LLM-based retrieval and their practical implementation, paving the way for broader industry adoption and academic exploration.

4.2 Limitations and Future Research

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

1. **Domain-Specific Customization**: The framework, though generalized, may require domain-specific fine-tuning to optimize retrieval performance further.
2. **Real-Time Metadata Generation**: Exploring dynamic metadata generation in real-time use cases could enhance system adaptability for rapidly evolving datasets.
3. **Extended Dataset Testing**: Applying the methodology to more diverse datasets, including multilingual and multimodal content, would provide deeper insights into its scalability and robustness.
4. **Integration with Advanced Models**: Future research could focus on integrating emerging LLMs or hybrid RAG (Retrieval-Augmented Generation) frameworks for even greater accuracy and efficiency.

This research contributes significantly to the field of analytics and AI by showcasing the utility of LLMs in metadata enrichment and retrieval optimization. The proposed framework demonstrates how AI-driven methodologies can streamline workflows, reduce search times, and enhance user productivity in large-scale documentation systems. Its adaptability positions it as a versatile solution for industries such as customer support, compliance, and knowledge management, where efficient data retrieval is critical.

In conclusion, this study reaffirms the critical role of metadata-enriched retrieval systems in advancing the capabilities of AI-powered tools, highlighting their potential to redefine how unstructured data is accessed and utilized across domains. By building on these findings, future research can continue to innovate, addressing limitations and unlocking new possibilities for metadata-driven systems.

Acknowledgments: We would like to thank the Department of Information and Decision Sciences at University of Illinois Chicago for all the supports.

Disclosure of Interests: The authors declare no conflict of interest.

References

[1] Sundaram, S., & Musen, M. (2024). Making Metadata More FAIR Using Large Language Models. DOI: [10.4126/FRL01-006444995].

[2] Song, H., et al. (2024). Metadata Enhancement Using Large Language Models. DOI: [10.48550/arxiv.2404.12283].

[3] Gao, Y., et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. DOI: [10.48550/arxiv.2312.10997].

[4] Chen, J., et al. (2023). Benchmarking Large Language Models in Retrieval-Augmented Generation. DOI: [10.48550/arxiv.2309.01431].

[5] Mombaerts, L., et al. (2024). Meta Knowledge for Retrieval-Augmented Large Language Models. DOI: [10.48550/arxiv.2408.09017].

[6] Harris, N., et al. (2024). Enhancing Embedding Performance through Large Language Model-based Text Enrichment and Rewriting. DOI: [10.48550/arxiv.2404.12283].

[7] Wang, L., et al. (2024). Large Search Model: Redefining Search Stack in the Era of LLMs. DOI: [10.48550/arxiv.2310.14587].

[8] Thottempudi, S., & Borra, S. (2024). Leveraging Large Language Models to Enhance an Intelligent Agent with Multifaceted Capabilities. DOI: [10.20944/preprints202409.1446.v1]..