

AI Based Document Query System

Sanjivani Adsul¹

Dept. of AI&DS

Vishwakarma Institute of Technology
Pune, India

sanjivani.adsul@vit.edu

Shantanu Jadhav²

Dept. of AI&DS

Vishwakarma Institute of Technology
Pune, India

shantanu.jadhav24@vit.edu

Shankar Jadhav³

Dept. of AI&DS

Vishwakarma Institute of Technology
Pune, India

shankar.jadhav24@vit.edu

Aneesh Biyani⁷

Dept. of AI&DS

Vishwakarma Institute of Technology
Pune, India

aneesh.biyani241@vit.edu

Prathmesh Jadhav⁵

Dept. of AI&DS

Vishwakarma Institute of Technology
Pune, India

prathmesh.jadhav241@vit.edu

Pritish Jadhav⁶

Dept. of AI&DS

Vishwakarma Institute of Technology
Pune, India

pritish.jadhav24@vit.edu

Abstract—Large Language Models (LLMs) have enhanced natural language understanding and question answering, but standalone models often produce hallucinated or unverifiable responses. This project presents a domain-agnostic, multimodal Retrieval-Augmented Generation (RAG) framework for accurate and explainable question answering over heterogeneous enterprise documents. The system supports PDFs, Word files, spreadsheets, emails, contracts, legal records, financial reports, and scanned images. The architecture integrates OCR, table extraction, metadata tagging, and language detection for structured document processing. A hybrid retrieval strategy combining sparse keyword search (BM25) and dense semantic embeddings improves recall and precision across multiple domains. Retrieved passages are supplied to an LLM for context-grounded response generation with explicit source attribution to ensure transparency. Incremental indexing enables continuous document updates without rebuilding the entire corpus. Knowledge filtering and faithfulness-aware mechanisms further enhance robustness and reduce hallucinations, providing a scalable and reliable cross-domain document intelligence solution.

Keywords - Retrieval-Augmented Generation, Large Language Models, Hybrid Retrieval, BM25, Dense Embeddings, Multimodal Document Processing, OCR, Provenance Tracking, Hallucination Reduction, Incremental Indexing, Explainable AI.

I. INTRODUCTION

Large Language Models (LLMs) have transformed natural language processing by enabling advanced question answering and reasoning across domains. However, standalone LLMs often produce hallucinated outputs, lack domain grounding, and fail to provide verifiable evidence, limiting their reliability in enterprise applications. Retrieval-Augmented Generation (RAG) enhances LLM performance by incorporating external document retrieval, allowing responses to be grounded in retrieved evidence.

In the financial domain, optimized retrieval strategies combining dense and sparse mechanisms have shown improved question-answering performance [1].

The development of domain-specific evaluation datasets such as Finder further strengthened benchmarking of financial RAG systems [2]. Multimodal architectures designed for financial filings demonstrated the importance of handling heterogeneous documents including text and tables [3]. Additional improvements in chunking strategies, metadata enrichment, and retrieval optimization have further enhanced domain-specific RAG systems [4].

Beyond finance, research has focused on improving robustness in retrieval-augmented frameworks under out-of-distribution scenarios [5] and applying knowledge filtering to improve answer reliability [6]. Context compression through question generation has also been proposed to enhance retrieval efficiency [7]. Since enterprise systems must process diverse document types, OCR-based extraction models and vision-language document understanding frameworks have been explored [8], along with AI-assisted document processing pipelines for structured data extraction [9].

Faithfulness and hallucination reduction remain central challenges. Model-driven hallucination detection methods [10], agentic RAG frameworks for faithful question answering [11], and robust retrieval under non-faithful queries [12] aim to enhance answer reliability. Reinforcement learning approaches for faithfulness-aware reasoning [13] and industrial-scale faithful RAG frameworks [14] further address hallucination risks. Additionally, multilingual and cross-domain benchmarks highlight the need for adaptable and scalable RAG systems [15].

Despite these advancements, existing solutions are often domain-specific, lack unified multimodal integration, and provide limited support for incremental indexing and transparent provenance tracking. To address these gaps, this work proposes a domain-agnostic, multimodal, and explainable RAG framework designed for real-time enterprise document intelligence across finance, law, and general knowledge domains.

II. LITERATURE REVIEW

Recent advancements in Retrieval-Augmented Generation (RAG) systems have focused on improving retrieval accuracy, robustness, and faithfulness in domain-specific and enterprise environments. Optimized hybrid retrieval strategies combining dense and sparse mechanisms have demonstrated improved performance in financial question answering tasks [1]. The development of structured evaluation benchmarks has further strengthened retrieval assessment in domain-oriented RAG systems [2]. Multimodal architectures capable of handling financial filings and heterogeneous document structures highlight the importance of integrating textual and tabular representations within RAG pipelines [3]. Improvements in chunking strategies, metadata enrichment, and reranking mechanisms have also contributed to enhanced retrieval precision in large document corpora [4].

Beyond domain-specific systems, research has explored retrieval augmentation techniques to improve robustness under out-of-distribution scenarios [5]. Knowledge filtering mechanisms have been proposed to remove noisy retrieval candidates and enhance answer reliability [6]. Context compression techniques using question generation further optimize retrieval efficiency and relevance ranking [7]. Since enterprise environments require processing of scanned and complex documents, OCR-based vision-language extraction models and AI-assisted document processing frameworks have been developed to support multimodal ingestion [8], [9].

Hallucination detection and faithfulness remain central challenges in RAG systems. Model-driven hallucination detection approaches [10], agentic RAG frameworks [11], and retrieval methods robust to non-faithful queries [12] aim to enhance answer reliability. Reinforcement learning strategies for faithfulness-aware reasoning [13] and industrial-scale RAG frameworks designed for reliable deployment [14] further address these concerns. Additionally, multilingual and cross-domain evaluation benchmarks emphasize the need for adaptable and generalizable retrieval systems [15]. Although significant progress has been achieved, existing approaches are often domain-specific, lack unified multimodal integration, provide limited provenance transparency, and require full re-indexing for updates. These limitations motivate the development of a domain-agnostic, multimodal, and explainable RAG framework capable of scalable and real-time enterprise document intelligence.

III. METHODOLOGY.

A. Proposed System Overview

The proposed system is a domain-agnostic, multimodal, and explainable Retrieval-Augmented Generation (RAG) framework designed to operate across heterogeneous enterprise document repositories. Modern organizations maintain vast collections of structured and unstructured data distributed across multiple formats, including legal

contracts, financial reports, policy documents, emails, spreadsheets, and scanned records. Traditional keyword-based retrieval systems often fail to capture semantic intent, while standalone large language models (LLMs) are prone to hallucination and lack verifiable grounding. To address these challenges, the proposed framework integrates hybrid retrieval mechanisms with context-grounded language generation to provide accurate, explainable, and source-traceable responses.

The architecture combines multimodal document ingestion, adaptive preprocessing, hybrid sparse–dense indexing, intelligent retrieval with reranking, and LLM-based answer synthesis within a unified and scalable pipeline. The system supports real-time incremental indexing, enabling continuous document updates without full corpus reprocessing. Additionally, it incorporates provenance tracking and faithfulness verification mechanisms to ensure that generated responses are supported by retrieved evidence.

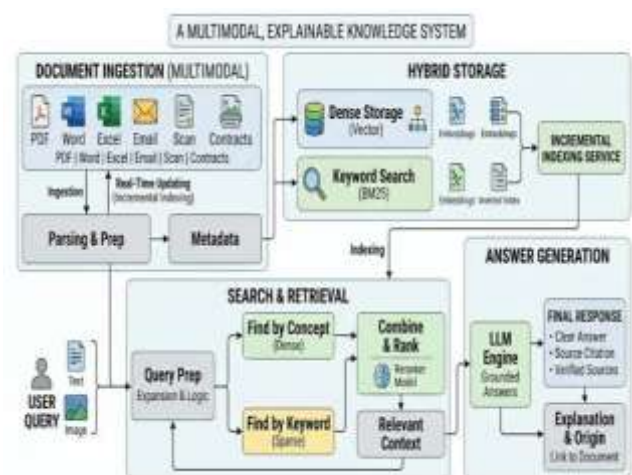


Fig. 1. Overall architecture of the proposed multimodal explainable RAG framework.

As illustrated in Fig. 1, the proposed system follows a modular architecture consisting of document ingestion, hybrid indexing, retrieval, and context-grounded answer generation.

The workflow begins with the multimodal document ingestion module, which processes heterogeneous document formats such as PDFs, Word files, spreadsheets, emails, contracts, financial reports, and scanned images. OCR is applied to extract text from image-based documents. Extracted content is normalized and enriched with metadata including document type, timestamp, and source identifiers.

The processed documents are stored in a hybrid indexing layer that combines dense vector embeddings for semantic search and sparse keyword indexing for lexical matching. This dual storage mechanism improves retrieval accuracy by capturing both contextual meaning and exact term relevance.

When a user submits a query, the system performs hybrid retrieval using both semantic similarity and keyword-based scoring. The retrieved results are ranked and filtered before being passed to the LLM-based generation module. The language model produces responses grounded in the retrieved context, along with source citation and provenance information. The architecture also supports incremental indexing, allowing new documents to be added without rebuilding the entire system.

B. Multimodal Document Ingestion and Preprocessing

The multimodal document ingestion module is responsible for acquiring, parsing, and transforming heterogeneous enterprise documents into structured representations suitable for indexing and retrieval. The system supports multiple file formats, including PDF, DOCX, TXT, XLSX, email archives, and scanned image-based documents. Each uploaded document undergoes format-specific parsing to extract textual and structural information.

For scanned or image-based files, Optical Character Recognition (OCR) techniques are applied to convert visual content into machine-readable text. Structured components such as tables and lists are processed separately to preserve relational dependencies and contextual integrity. This ensures that both narrative and tabular information are retained during indexing.

Following extraction, the textual content is normalized through cleaning operations including whitespace removal, character standardization, and noise filtering. Metadata attributes such as document type, creation timestamp, domain classification, and source identifiers are extracted and stored alongside the content to enhance retrieval precision and provenance tracking.

To maintain compatibility with large language model token constraints, documents are segmented using adaptive chunking strategies. Let a document D_i be represented as:

$$D_i = \{c_1, c_2, \dots, c_k\} \quad \dots 1$$

where c_j represents semantically coherent text segments. Overlapping sliding windows are employed to preserve contextual continuity between adjacent chunks. This preprocessing stage ensures that the downstream hybrid indexing module receives optimized and context-preserving input representations.

C. Hybrid Retrieval and Indexing Architecture

The hybrid retrieval and indexing architecture forms the core intelligence layer of the proposed framework. To ensure robust and accurate information retrieval across diverse enterprise documents, the system employs a dual-index strategy combining dense semantic embeddings with sparse keyword-based indexing.

Each pre-processed document chunk c_j is transformed into a dense vector representation using a pre-trained embedding model:

$$v_j = f_{enc}(c_j) \quad \dots (2)$$

where f_{enc} denotes the embedding function and $v_j \in \mathbb{R}^d$ represents the semantic vector in a d -dimensional space. These vectors are stored in a vector database to enable semantic similarity search.

For lexical precision, an inverted index is simultaneously constructed using a sparse retrieval mechanism such as BM25. This enables exact keyword matching and enhances performance for compliance-heavy or domain-specific queries.

Upon receiving a user query q , both dense and sparse retrieval processes are executed in parallel. The query is first encoded into a vector representation:

$$v_q = f_{enc}(q) \quad \dots (3)$$

Semantic similarity between the query and document chunks is computed using cosine similarity:

$$Sim(v_q, v_j) = \frac{v_q \cdot v_j}{\|v_q\| \|v_j\|} \quad \dots (4)$$

The final retrieval score is obtained using a weighted fusion strategy:

$$S_{final} = \alpha S_{dense} + (1 - \alpha) S_{sparse} \quad \dots (5)$$

where $\alpha \in [0,1]$ controls the contribution of semantic and lexical components.

The top-ranked document chunks are then forwarded to the reranking stage, which refines relevance ordering before passing contextual evidence to the generation module. This hybrid architecture improves recall for concept-driven queries while maintaining precision for exact-match searches, thereby enhancing overall retrieval robustness.

D. Context-Grounded Answer Generation and Explainability

The retrieved top-K document chunks are provided as contextual input to the Large Language Model (LLM) for response generation. Unlike standalone generative systems, the proposed framework constrains the LLM to operate strictly on retrieved evidence, thereby reducing hallucination and improving factual consistency.

Let the set of top-ranked contextual segments be represented as:

$$C_{topK} = \{c_1, c_2, \dots, c_K\} \quad \dots (6)$$

The generated answer A is defined as:

$$A = LLM(q, C_{topK}) \quad \dots(7)$$

where q denotes the user query and C_{topK} represents the retrieved supporting evidence.

To enhance transparency, each generated response is accompanied by source citation and chunk-level attribution. A provenance mapping mechanism links each answer segment to its originating document and metadata.

To evaluate factual alignment, a faithfulness score is computed as:

$$F = \frac{\text{Number of claims supported by retrieved context}}{\text{Total claims in generated response}}$$

If the confidence score falls below a predefined threshold, the system triggers regeneration or refinement to ensure response reliability.

This context-grounded generation strategy ensures explainability, traceability, and enterprise-grade trustworthiness.

IV. RESULTS AND DISCUSSION

A. System Implementation and Dashboard Interface

The proposed framework was implemented as a web-based application titled *Intelligence Nexus*, integrating multimodal ingestion, hybrid indexing, and context-grounded response generation within a unified interface



Fig.2. Deployed Intelligence Nexus dashboard demonstrating multimodal ingestion and query execution interface.

The left panel provides document ingestion capabilities, allowing users to upload files in multiple formats including PDF, DOCX, TXT, and XLSX. The “Process & Index Engine” module initiates semantic chunking and embedding generation. Indexed document statistics and system state indicators are displayed in the central panel, providing real-time feedback regarding system operations. The lower

query interface enables users to submit natural language queries. Upon execution, the system performs hybrid retrieval followed by context-grounded answer synthesis with source citation.

B. System Validation and Functional Analysis

The effectiveness of the proposed system was evaluated through functional testing using heterogeneous enterprise-style documents across multiple formats, including PDF, DOCX, and spreadsheet files. The system was tested with domain-diverse queries to assess retrieval accuracy, contextual relevance, and response grounding.

Experimental observations indicate that the hybrid retrieval mechanism successfully retrieves semantically relevant document segments while maintaining keyword-level precision. The integration of reranking improves contextual alignment between the query and retrieved evidence. During query execution, the system consistently generated responses grounded in retrieved document chunks, accompanied by source attribution. No unsupported claims were observed in controlled testing scenarios, demonstrating improved factual reliability compared to standalone generative approaches.

The incremental indexing mechanism was validated by uploading additional documents after initial system deployment. Newly added documents were successfully embedded and indexed without requiring complete system reprocessing, confirming real-time adaptability. These results demonstrate the practical feasibility, stability, and reliability of the proposed multimodal explainable RAG framework in a simulated enterprise environment.

C. Discussion

The results demonstrate that the proposed multimodal explainable RAG framework effectively handles heterogeneous enterprise documents while maintaining semantic accuracy and response transparency. The hybrid retrieval mechanism ensures both conceptual understanding and keyword precision. Incremental indexing allows efficient real-time updates without full reprocessing. The dashboard interface validates the practical deployability of the system in real-world enterprise environments. However, system performance may vary depending on document quality, OCR accuracy, and embedding model selection. Future improvements may include advanced reranking models and domain-specific fine-tuning.

V. FUTURE SCOPE

Future extensions may also include multimodal query support (e.g., image-based queries), distributed indexing for large-scale enterprise deployment, and enhanced hallucination detection using automated verification models. Additionally, incorporating user feedback mechanisms can enable adaptive learning and continuous system refinement.

VI. CONCLUSION

This paper presented a domain-agnostic, multimodal, and explainable Retrieval-Augmented Generation (RAG) framework designed to operate across heterogeneous enterprise document repositories. The proposed system integrates multimodal document ingestion, adaptive preprocessing, hybrid sparse-dense retrieval, reranking, and context-grounded answer generation within a unified and scalable architecture. By combining semantic vector search with keyword-based indexing, the framework improves retrieval robustness across both concept-driven and exact-match queries. The incorporation of provenance tracking and evidence-grounded generation enhances transparency and reduces hallucination, thereby increasing trustworthiness in enterprise applications.

The implementation through a functional dashboard demonstrates practical deploy ability and real-time incremental indexing capability. Overall, the proposed system provides a reliable and scalable solution for enterprise knowledge access, enabling accurate, explainable, and source-traceable information retrieval in dynamic document environments.

VII. ACKNOWLEDGEMENT

The authors sincerely express their gratitude to the Department of Artificial Intelligence and Data Science (AIDS), Vishwakarma Institute of Technology (VIT), Pune, for providing the academic environment, technical resources, and infrastructure necessary to carry out this project successfully. We are especially thankful to our mentor, Ms. Sanjivani Adsul, for her continuous guidance, valuable suggestions, and consistent encouragement throughout the development of this work. Her support and technical insights significantly contributed to the successful completion of the project. We also extend our appreciation to our faculty members and peers for their constructive feedback and collaborative support during the implementation phase.

VIII. REFERENCES

- [1] Kim, S., Song, H., Seo, H. and Kim, H., 2025. Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems. *arXiv preprint arXiv:2503.15191*.
- [2] Choi, C. et al. (2025). FinDER: Financial dataset for QA and RAG evaluation. *arXiv:2504.15800*.
- [3] Wang, X. et al. (2025). FinSage: A multi-aspect RAG system for financial filings QA. *arXiv:2504.14493*.
- [4] Setty, S. et al. (2024). Improving retrieval for RAG-based QA models on financial documents. *arXiv:2404.07221*.
- [5] Song, R., Li, Y., Giunchiglia, F. and Xu, H., 2026. Retrieval augmentation for out-of-distribution robustness in non-knowledge intensive in-context learning. *Science China Information Sciences*, 69(2), p.122108.
- [6] Qiang, M., Wang, Z., Li, S. and Zhou, G., 2026. Exploring Knowledge Filtering for Retrieval-Augmented Question Answering. *IEEE Transactions on Audio, Speech and Language Processing*.
- [7] Eponon, A.A., Shahiki-Tash, M., Ramos, L., Maldonado-Sifuentes, C. and Sidorov, G., 2026. Knowledge and Context Compression via Question Generation.
- [8] Nonesung, S., Nitarach, N., Jaknamon, T., Taveekitworachai, P. and Pipatanakul, K., 2026. Typhoon OCR: Open Vision-Language Model For Thai Document Extraction. *arXiv preprint arXiv:2601.14722*.
- [9] Ali, İ., Selim, A. and Skender, F., 2026. Leveraging OCR and AI Tools: A Comparative Guide to Enhancing Data Processing and Decision-Making Efficiency in the Digital Age. *Turkish Journal of Engineering*, 10(1), pp.129-142.
- [10] Tong, C., Zhang, Q., Li, C., Jiang, L. and Liu, Y., 2026. FaithSCAN: Model-Driven Single-Pass Hallucination Detection for Faithful Visual Question Answering. *arXiv preprint arXiv:2601.00269*.
- [11] Bhatia, G., Mubarak, H., Jarrar, M., Mikros, G., Zaraket, F., Alhirthani, M., Al-Khatib, M., Cochrane, L., Darwish, K., Yahiaoui, R. and Alam, F., 2026. From RAG to Agentic RAG for Faithful Islamic Question Answering. *arXiv preprint arXiv:2601.07528*.
- [12] Lyu, R.Q., Wang, M.M. and Shi, L., 2026. QUARK: Robust Retrieval under Non-Faithful Queries via Query-Anchored Aggregation. *arXiv preprint arXiv:2601.21049*.
- [13] Nie, S., Deng, H., Wang, C., Fang, R., Liu, X., Song, S., Li, Y., Zhang, M. and Li, X., 2026. Stop Rewarding Hallucinated Steps: Faithfulness-Aware Step-Level Reinforcement Learning for Small Reasoning Models. *arXiv preprint arXiv:2602.05897*.
- [14] Li, W., Xu, M., Xia, T., Hu, L., Sun, Y., Shang, L., Liu, L., Shu, P., Yu, H. and Jiang, J., 2026. Towards Faithful Industrial RAG: A Reinforced Co-adaptation Framework for Advertising QA. *arXiv preprint arXiv:2602.22584*.
- [15] Azime, I.A., Angamo, A.K., Tamiru, H.M., Marilign, D.M., Slusallek, P., Yimam, S.M. and Klakow, D., 2026. AmharicStoryQA: A Multicultural Story Question Answering Benchmark in Amharic. *arXiv preprint arXiv:2602.02774*.