Conceptual Framework for Document Intelligence & Routing System

Saumya Gupta
gupta.sg003@gmail.com
Department of CSE, AKTU, India

Nivedika Sharma
nivedika9719sharma@gmail.com
Department of CSE, AKTU, India

Pankhi Gupta
pankhigupta440@gmail.com
Department of CSE, AKTU, India

Sanmita Biswas
sanmitabiswas22@gmail.com
Department of CSE, AKTU, India

Dr. Jyoti Agarwal
jyoti.agarwal@srmscet.edu
Department of CSE, AKTU, India

ABSTRACT

Organizations increasingly struggle to handle the growing volume of unstructured and semi-structured documents. The use of isolated tools for individual tasks often results in operational inefficiencies, duplicated information, and reduced confidence in AI outputs, particularly due to hallucination-related inconsistencies. To overcome these issues, this paper introduces the *Unified Document Intelligence System (UDIS)*, a conceptual framework designed to consolidate the entire document-processing workflow into one integrated architecture. UDIS brings together five core elements: (1) a multi-document OCR pipeline for reliable data intake; (2) a semantic redundancy detection mechanism to manage overlapping or repetitive content; (3) an abstractive summarization engine capable of synthesizing information across multiple documents; (4) a traceability module that links generated summaries back to their source segments to ensure trustworthiness; and (5) a fine-grained role-based access control. This work outlines the UDIS architecture in full and explains the rationale behind its unified design through relevant literature.

Key terms: multiple documents, summarization, traceability, role-based access control, data redundancy.

I. INTRODUCTION

The proliferation of digital documents represents both a critical asset and a significant challenge for modern organizations. In sectors such as law, finance, and research, teams must synthesize information from dozens or hundreds of disparate reports, contracts, and transcripts. This "information overload" creates a significant bottleneck, impeding decision-making and operational efficiency.

This fragmentation introduces severe risks and inefficiencies. An analyst's workflow may involve using a standalone Optical Character Recognition (OCR) tool to digitize a document, a separate AI model to summarize it, a cloud drive to store it, and a different platform to manage permissions. This disjointed process not only creates information redundancy but also presents a critical security challenge, as granular, role-based access control (RBAC) and versioning are difficult to enforce across multiple, non-integrated systems.

Furthermore, the recent advent of powerful generative AI, while promising, has introduced a profound "crisis of trust." Abstractive summarization models are prone to "hallucinations"—inventing facts, figures, or clauses not present in the source text. When an AI-generated summary cannot be programmatically verified against its source, it becomes an unusable "black box." This lack of traceability creates an unacceptable level of risk for enterprise adoption, where a single, factually incorrect summary can have severe financial or legal consequences. Current solutions fail to address this multi-faceted problem holistically. They are

typically rigid data extraction pipelines, simple storage systems, or isolated "black box" AI models. To fill this critical gap, we propose the Unified Document Intelligence System (UDIS) framework. UDIS is a novel conceptual architecture designed to function as a single, cohesive source of truth for an organization's documents. Its primary contribution is the tight integration of five components often treated in isolation: (1) a multi-document OCR pipeline for robust ingestion; (2) semantic redundancy detection to reduce clutter; (3) a multi-document, abstractive summarization engine for synthesis; (4) a "source-to-summary" traceability module to mitigate hallucination; and (5) a granular, role-based access control (RBAC) and versioning layer to ensure security.

This paper presents the complete conceptual architecture of the UDIS framework. Section II reviews the existing research across OCR, multi-document summarization, hallucination mitigation, redundancy detection, and document clustering, and identifies the gaps that motivate the proposed framework. Section III defines the core problem and introduces the Unified Document Intelligence System (UDIS), detailing its integrated architecture and workflow. Section IV discusses the future scope and potential extensions of the framework toward real-world deployment. Finally, Section V concludes the paper by summarizing the key contributions and outlining opportunities for further research.

II. RELATED WORK

Our framework builds upon and integrates research from several distinct computer science domains. This review is structured thematically to align with the core components of our proposed framework.

2.1 Advancements in Optical Character Recognition (OCR)

Foundational work by Chen and Bloomberg [1] introduced a system for summarizing imaged documents without performing OCR. Their image-based method detects word-bounding boxes and groups them into “equivalence classes” based on shape similarity, mimicking term identification in text documents; summary sentences are then selected using a statistical classifier.

El Harraj and Raissouni [2] propose a pre-processing pipeline to improve OCR accuracy for camera-captured documents affected by poor lighting or distortion. Their technique sequentially applies CLAHE, optimized grayscale conversion, Unsharp Masking, and Otsu’s binarization, demonstrating substantial improvements in character recognition accuracy.

2.2 Multi-Document Summarization (MDS)

Pallagani et al. [3] introduce the novel problem of summarizing "Planning-Like (PL) Tasks". Their work focuses on synthesizing multiple action sequences (e.g., different recipes, travel routes, or workflows) that share a common goal into a single, coherent summary. The authors argue that this form of summarization must preserve the "executability and logical flow" of actions. Their user study confirms that while

abstractive LLMs are the preferred method, the risk of hallucination remains a "significant challenge".

2.3 The Challenge of Hallucination and Traceability

Metropolitansky and Larson [4] present VeriTrail, a novel method to address "closed-domain hallucination" by providing traceability. Their work is particularly notable for handling "multiple generative steps" (MGS) , where errors in intermediate stages can propagate to the final output. The VeriTrail system models the generative process as a directed acyclic graph (DAG) and performs a reverse traversal to enable both provenance (tracing faithful content) and error localization (identifying where hallucinations were introduced) [3] .

From a Human-Computer Interaction (HCI) perspective, Kambhamettu et al. [5] introduce "Traceable Text," an interaction primitive designed to help users critically examine AI-generated summaries and the source texts they are derived from. Their system provides explicit, phrase-level links that connect passages in the summary to the source text that informed them. A usability study conducted by the authors showed that this traceable interaction helped readers answer questions about source content more quickly and markedly improved their correctness when summaries contained hallucinations [3].

In their application-oriented survey, Li et al. [6] propose a useful taxonomy that distinguishes between knowledge-based hallucinations (content inconsistent with real-world facts) and logic-based hallucinations (flaws in the reasoning process). The authors identify Retrieval-Augmented Generation (RAG) as a primary mitigation strategy for knowledge-based hallucinations, as it introduces external knowledge sources to reduce factual errors. They also posit that RAG and reasoning-enhancement techniques are complementary, and their integration into "Agentic Systems" represents a unified framework for addressing composite hallucination problems .

Belem et al. [7] provide a critical investigation into *how* large language models (LLMs) hallucinate specifically within the context of Multi-Document Summarization (MDS) . By creating two novel benchmarks, they observe that LLM-generated summaries can contain a high percentage of hallucinated content (up to 75%). Their human evaluation reveals that the majority of these errors are not simple factual inventions but rather stem from the model failing to follow instructions (e.g., summarizing information that is not shared across documents or is unrelated to the topic) or producing overly generic, non-informative insights. The authors also note that hallucinations are more likely to occur toward the end of a generated summary.

2.4 Semantic Redundancy and Duplicate Detection

Wu et al. [8] present a practical study on file deduplication for storage devices, implementing three schemes—filename-based, filesize-based, and MD5 hash-based—on a QNAP NAS system. The filename-based and file size-based schemes

are the most intuitive and fastest, both exhibiting a time complexity of O(N \log N). However, these approaches can lead to inaccurate deduplication. To improve accuracy, the MD5 hash value-based approach is introduced. This scheme, which is an extension of the file size approach, has a higher worst-case time complexity of O(N^2) due to the hash calculation, resulting in higher CPU and memory usage. The central contribution is the finding that a partial content hashing based file deduplication offers the best trade-off between computation cost and deduplication accuracy. This is a compromised way that brings a faster response with few sacrifices in accuracy. The evaluation was conducted on a QNAP NAS device.

2.5 Document Clustering and Knowledge Graph Representation

Sampaio and Maxcici's [9] work on unsupervised document clustering using multimodal embeddings addresses a critical gap in document management: the need to automatically categorize heterogenous document collections across both category and template levels without manual labelling. Their systematic evaluation of eight pe trained encoders (text-only, layout aware, vision-only, and vision-language models) combined with classical clustering algorithms (k-means, DBSCAN, HDBSCAN k-NN, BIRCH) across five diverse corpora reveals that document clustering performance is fundamentally modality-dependent. Their key finding is that vision-based features excel in template discovery while textual embeddings show greater robustness to document degradation and noise. The work demonstrates that multimodal fusion strategies achieve the best trade-off for automatic document categorization and clustering without manual labelling, providing practical foundation for intelligent document organization systems.

Gretarsson et al. [10] present TopicNets, a system that visualizes document clusters and topics as interactive knowledge graphs where documents and topics are nodes connected through discovered semantic relationships. By enabling real-time topic refinement and interactive exploration of document relationships, TopicNets demonstrates how graph-based visualization transforms document clustering into accessible, navigable knowledge networks that support organizational discovery, compliance auditing, and human-in-the-loop exploration of document landscapes.

2.6 Synthesis and The Research Gap

The reviewed literature demonstrates significant progress in the OCR [1,2,11], multiple document summarization [3,15] hallucination mitigation and traceability [4,5,6,7], redundancy detection [8,12], and document clustering [9], yet these advancements remain largely disconnected. OCR [1,2,11] studies improve text extraction, while summarization [3,15] research strengthens coherence but continues to face hallucination issues. Traceability methods [4,5,6,7], including provenance graphs and phrase-level linking, help verify outputs but function as standalone verification layers. Likewise, work on deduplication and clustering [8,9,12] offers useful techniques for managing redundancy and organizing

document sets, but these remain independent utilities with limited interaction across ingestion, summarization, or security workflows. Overall, the field lacks a holistic, end-to-end framework that unifies these capabilities. As a result, organizations are left with fragmented toolchains that compromise trust, efficiency, and operational reliability. The Unified Document Intelligence System (UDIS) addresses this gap by integrating these components into a single conceptual architecture designed to support accurate, secure, and trustworthy large-scale document management.

III. PROPOSED SOLUTION

A. PROBLEM DEFINITION

Modern organizations increasingly rely on large volumes of unstructured and semi-structured documents originating from diverse sources such as scanned reports, legal contracts, research articles, and financial records. Existing document-processing ecosystems are highly fragmented, requiring separate tools for OCR, summarization, redundancy detection, access control, and verification of generative outputs. This fragmentation results in critical challenges such as:

- Redundant and inconsistent information across repositories due to lack of semantic duplicate detection.
- Low trust in generative AI outputs, especially abstractive summaries that may contain hallucinated or unverifiable information.
- Absence of unified traceability to link summary content back to its source, making verification difficult.
- Weak or inconsistent access control, as RBAC and versioning mechanisms are scattered across different tools.
- Operational inefficiencies due to manual switching between multiple, non-integrated systems.

Therefore, the core problem addressed in this research is the absence of a unified, secure, and traceable document-intelligence architecture that integrates OCR, semantic redundancy detection, multi-document summarization, hallucination mitigation, and granular RBAC into a single system.

The Unified Document Intelligence System (UDIS) is proposed as a conceptual framework to solve this problem holistically.

B. PROPOSED FRAMEWORK

The Unified Document Intelligence System (UDIS) integrates the complete lifecycle of enterprise documents into a single, cohesive architecture. The workflow begins with the ingestion of diverse document types such as scanned records, legal agreements, and financial files. These documents are digitized through a robust OCR pipeline that ensures accurate text extraction even from noisy or low-quality inputs. To maintain information clarity and reduce storage inefficiencies, UDIS incorporates semantic redundancy detection, which identifies and removes duplicate or near-duplicate content using advanced hashing and embedding-based similarity techniques.

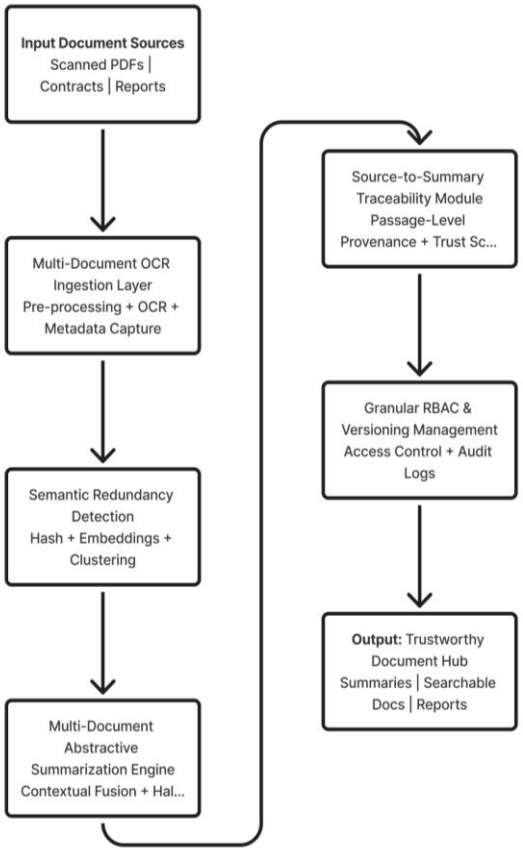

Figure1: Unified Document Intelligence System Framework

Then a multi-document abstractive summarization engine synthesizes key information into concise summaries. To address the widely acknowledged challenge of hallucination in generative AI, UDIS integrates a source-to-summary traceability mechanism, ensuring that each segment of the summary can be directly mapped back to the original document passages. This embedded verification capability enables high trust and transparency in critical enterprise usage scenarios.

The final stage involves secure sharing and collaboration using granular Role-Based Access Control (RBAC) and versioning. This guarantees that document visibility aligns with organizational hierarchies and compliance policies. By unifying OCR, summarization, redundancy management, traceability, and access governance into one system, UDIS effectively eliminates workflow fragmentation, enhances security, and enables reliable AI-driven decision support in document-intensive environments.

IV. FUTURE SCOPE

The Unified Document Intelligence System (UDIS) addresses a critical gap in current enterprise workflows by integrating OCR ingestion, semantic redundancy detection, multi-document summarization, traceability, and role-based access control into a single cohesive framework. Existing solutions treat these components separately, leading to fragmented workflows, inefficiencies, and a lack of trust in AI-generated outputs. UDIS resolves these challenges by offering a unified architecture that enhances accuracy, reduces redundancy, and ensures verifiable, secure, and transparent document processing. Although the framework is conceptual at this stage, it lays a strong foundation for developing future document-intelligence systems that are trustworthy, scalable, and suitable for real-world deployment. By emphasizing explainability, security, and integration, UDIS moves toward enabling next-generation AI systems capable of delivering reliable and enterprise-grade document management.

V. CONCLUSION

In summary, this paper presented the Unified Document Intelligence System (UDIS), a consolidated framework designed to address

the fragmentation that currently characterizes document-processing workflows. By integrating robust OCR, semantic redundancy detection, multi-document abstractive summarization, source-to-summary traceability, and role-based access control into a single architecture, UDIS provides a coherent approach to managing large-scale, heterogeneous document collections. The framework not only bridges gaps identified across existing literature but also outlines a path toward more trustworthy, efficient, and verifiable AI-driven document intelligence systems. Future extensions may operationalize UDIS in real-world enterprise settings, further validating its potential to support reliable and transparent document-centric AI solutions.

REFERENCES

[1] D. S. Bloomberg and F. R. Chen, "Document image summarization without OCR," in *Proc. 3rd IEEE Int. Conf. Image Process.*, Lausanne, Switzerland, Sep. 19, 1996, pp. —. doi: 10.1109/ICIP.1996.560744.

[2] A. E. Harraj and N. Raissouni, "OCR accuracy improvement on document images through a novel pre-processing approach," *Signal & Image Processing: An International Journal (SIPIJ)*, vol. 6, no. 4, pp. 1–13, Aug. 2015, doi: 10.5121/sipij.2015.6401.

[3] V. Pallagani, B. Srivastava, and N. Gupta, "PLANTS: A Novel Problem and Dataset for Summarization of Planning-Like (PL) Tasks," arXiv:2407.13597 [cs.CL], Jul. 2024, doi: 10.48550/arXiv.2407.13597.

[4] D. Metropolitansky and J. Larson, "VeriTrail: Closed-Domain Hallucination Detection with Traceability," arXiv:2505.21786 [cs.CL], May 2025, doi: 10.48550/arXiv.2505.21786.

[5] H. Kambhamettu, J. Flores, and A. Head, "Traceable Text: Deepening Reading of AI-Generated Summaries with Phrase-Level Provenance Links," arXiv:2409.13099 [cs.HC], Sep. 2024, doi: 10.48550/arXiv.2409.13099.

[6] Y. Li, X. Fu, G. Verma, P. Buitelaar, and M. Liu, "Mitigating Hallucination in Large Language Models (LLMs): An Application-Oriented Survey on RAG, Reasoning, and Agentic Systems," arXiv:2510.24476 [cs.CL], Oct. 2025, doi: 10.48550/arXiv.2510.24476.

[7] C. G. Belém, P. Pezeshkpour, H. Iso, S. Maekawa, N. Bhutani, and E. Hruschka, "From Single to Multi: How LLMs Hallucinate in Multi-Document Summarization," *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, NM, USA, Apr. 2025, pp. 5276–5309, doi: 10.18653/v1/2025.findings-naacl.293.

[8] Y.-T. Wu, M.-C. Yu, J.-S. Leu, E.-C. Lee, and T. Song, "Design and implementation of various file deduplication schemes on storage devices," in *Proc. 2015 11th Int. Conf. Heterogeneous Networking for Quality, Reliability, Security and Robustness (QSHINE)*, Taipei, Taiwan, Aug. 19–20, 2015, doi: *(DOI not provided in your data)*, ISBN: 978-1-6319-0063-1.

[9] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth, "TopicNets: Visual Analysis of Large Text Corpora with Topic

Modeling," *ACM Trans. Intell. Syst. Technol. (TIST)*, vol. 3, no. 2, Art. no. 23, pp. 1–26, 2012, doi: 10.1145/2089094.2089099.

[10] P. R. Sampaio and H. Maxcici, "Unsupervised Document and Template Clustering using Multimodal Embeddings," arXiv:2506.12116 [cs.CL], Jun. 2025, doi: 10.48550/arXiv.2506.12116.

[11] Madhavi, H., Cherian, J., Khamkar, Y., & Bhagat, D. (2025, May 4). Low-resource language processing: An OCR-driven summarization and translation pipeline. *Dr. Vishwanath Karad MIT World Peace University*. DOI: https://doi.org/10.48550/arXiv.2505.1117

[12] Ho, P.-T., & Kim, S.-R. (2014). Fingerprint-based near-duplicate document detection with applications to SNS spam detection. *International Journal of Distributed Sensor Networks*, *2014*, Article ID 612970. DOI: http://dx.doi.org/10.1155/2014/612970

[13] Sunitha, C., Jaya, A., & Ganesh, A. (2019). Automatic summarization of Malayalam documents using clause identification method. *International Journal of Electrical and Computer Engineering (IJECE)*, *9*(6), 4929–4938. DOI: 10.11591/ijece.v9i6.pp4929-4938

[14] Fan, J., & Huang, T. (2012). A fusion of algorithms in near duplicate document detection. In L. Cao et al. (Eds.), *PAKDD 2011 Workshops: Advances in Knowledge Discovery and Data Mining* (pp. 234–242). Springer. DOI: 10.1007/978-3-642-28320-8_20

[15] Godbole, A. , George, J., & Shandilya, S. (2024). Leveraging Long-Context Large Language Models for Multi-Document Understanding and Summarization in Enterprise Applications. DOI: 10.48550/arXiv.2409.18454