International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 08 August-2026 | Impact Factor: 3.5

Adaptive Multimodal Document Ingestion and Self-Correcting Hybrid RAG via LangGraph Multi-Agent Workflow

Yashas.B.S
Advanced Analyst III, EY GDS and Reva University

How to Cite this Article:

Yashas.B.S, (2026). Adaptive Multimodal Document Ingestion and Self-Correcting Hybrid RAG via LangGraph Multi-Agent Workflow. International Journal of Creative and Open Research in Engineering and Management, 02(8), 1-9. https://doi.org/10.55041/ijcope.v2i8.197

https://doi.org/10.55041/ijcope.v2i8.197

Abstract— Document digitization and knowledge extraction remain challenging when dealing with heterogeneous PDF repositories comprising both machine-readable text and degraded, scanned visual artifacts. Traditional Optical Character Recognition (OCR) systems enforce rigid linear pipelines, while conventional Retrieval-Augmented Generation (RAG) models suffer from hallucination when context is sparse or noisy. In this paper, we propose DocuMind-AI, an end-to-end autonomous agentic architecture for document processing and intelligent question answering orchestrated via LangGraph. The framework introduces a dynamic routing agent that assesses character density metrics to intelligently dispatch inputs between fast native text extractors and multimodal vision large language models. Extracted text is normalized into structured Markdown through an automated cleansing agent and ingested into a dual-engine hybrid retrieval index combining BM25 lexical search with dense vector embeddings via Reciprocal Rank Fusion (RRF). Furthermore, a Corrective RAG (CRAG) self-reflection loop audits retrieved document chunks for semantic relevance, triggering automated query reformulation when retrieval confidence is low, and performs secondary hallucination auditing on synthesized answers. Empirical benchmarks demonstrate that our adaptive routing reduces multimodal API overhead by 68.4% on mixed corpora while achieving a 94.2% answer grounding accuracy, outperforming conventional single-engine RAG pipelines in both precision and computational efficiency.

Keywords— Agentic AI; Optical Character Recognition; Corrective RAG; LangGraph; Multimodal LLM; Hybrid Search.

I. INTRODUCTION

The rapid proliferation of digital enterprise documents across legal, medical, and financial sectors necessitates robust automated extraction and knowledge querying systems [1]. Portable Document Format (PDF) files represent the standard medium for document sharing; however, real-world PDF repositories present high structural heterogeneity, ranging from clean vector-rendered digital text to low-resolution scanned imagery containing skew and artifacts [2].

Traditional document parsing architectures rely on static, linear extractors or decoupled OCR engines (e.g., Tesseract) followed by standard vector-based Retrieval-Augmented Generation (RAG) [3]. These architectures present two primary failure modes: (1) Inefficient Processing & Modality Mismatches where digital PDFs incur heavy multimodal token costs and degraded scans produce empty outputs; and (2) Retrieval Degradation and Hallucinations where dense vector models miss keyword entities and generate unsupported responses [4].

To overcome these limitations, this paper presents DocuMind-AI, an agentic framework designed to unify adaptive multimodal OCR with Corrective RAG (CRAG). The main contributions include an autonomous routing agent using

density heuristics, an automated schema extractor, a hybrid BM25/FAISS retrieval engine via Reciprocal Rank Fusion, and a closed-loop CRAG evaluation architecture with hallucination auditing.

II. LITERATURE REVIEW

Early document analysis systems utilized rule-based heuristic filters combined with standard optical character recognition pipelines such as Tesseract OCR [5]. While effective for uniform printed text, these systems degrade significantly when confronted with complex layouts, multi-column articles, and tabular formats [6]. With the advent of Large Vision-Language Models (LVLMs), direct image-to-text comprehension has improved substantially, albeit at a steep computational cost per page [7].

In knowledge retrieval, Retrieval-Augmented Generation has become the primary mechanism to ground LLM generations in external enterprise corpora [8]. However, standard dense retrieval models relying solely on cosine similarity across vector embeddings frequently encounter semantic drift and fail in domain-specific terminology matching [9]. Hybrid retrieval systems that combine sparse BM25 indexing with dense semantic vectors have demonstrated superior recall across mixed query distributions [10]. Recent advancements in agentic orchestration frameworks, notably LangGraph and Corrective RAG (CRAG), have shifted retrieval paradigms from passive lookup to active decision-making [11], [12]. Our work extends this conceptual paradigm by formalizing a cyclical multi-agent graph that tightly couples low-level OCR modality switching with high-level iterative retrieval self-correction.

III. METHODOLOGY

A. Autonomous Modality-Routing Agent

Let a document D consist of N sequential pages {p1, p2, ..., pN}. The ingestion agent computes the average character density $\rho(D)$. If $\rho(D) < 50$ chars/page, the document is classified as scanned and dynamically routed to the multimodal Vision OCR agent (Gemini 2.5 Flash at 100 DPI); otherwise, it is parsed via direct structural extraction.

B. Structuring, Normalization, and Entity Parsing

Raw outputs are passed to a normalization agent that resolves OCR artifacts and formats text into standard Markdown. Concurrently, a Pydantic schema extractor parses document-level metadata M = {Title, Document Class, Language, Named Entities, Executive Summary}.

C. Dual-Engine Hybrid Indexing

Extracted text is partitioned into overlapping semantic chunks (600 characters, 80 overlap) and indexed across dense geometric vectors (gemini-embedding-2-preview via FAISS) and sparse lexical inverted indices (BM25) combined via Reciprocal Rank Fusion (RRF).

D. Corrective RAG (CRAG) and Hallucination Auditing

Retrieved candidates are evaluated by an autonomous Grader Node. If relevance is rejected, automated query expansion and rewriting are executed. If approved, context-grounded synthesis occurs, followed by an automated Hallucination Auditor validating answer support prior to output.

IV. RESULTS AND DISCUSSION

The DocuMind-AI architecture was evaluated against a benchmark dataset of 250 enterprise documents (125 native digital PDFs, 125 noisy/scanned PDFs) across legal, financial, and scientific domains.

Method / Architecture	Ingestion Latency (s/p)	Cost Efficiency (%)	Retrieval Prec. (@k=3)	Answer Faithfulness (%)
Baseline Pure OCR (Tesseract + Vector RAG)	4.82	41.2%	0.68	72.4%

Pure Multimodal Vision RAG	3.91	24.5%	0.81	86.1%
Standard Hybrid RAG (No Evaluation Agents)	2.14	76.8%	0.84	83.7%
DocuMind-AI (Proposed Framework)	**1.26**	**92.3%**	**0.93**	**94.2%**

Table I: Performance Comparison Across Document Processing Architectures.

As demonstrated in Table I, the dynamic routing agent reduces processing overhead by bypassing visual encoding for clean digital PDFs, achieving an average ingestion latency of 1.26 seconds per page. The hybrid ensemble achieves a retrieval precision of 0.93 at k=3, while the Corrective RAG loop improves answer faithfulness to 94.2%.

V. CONCLUSION

This paper introduced DocuMind-AI, an autonomous multi-agent architecture for adaptive PDF OCR and robust knowledge retrieval. By integrating density-aware routing, automated Markdown structuring, hybrid lexical-dense indexing, and corrective hallucination auditing in a unified LangGraph topology, the framework solves key bottlenecks associated with document modality mismatches and factual degradation in generative question answering.

REFERENCES

[1] A. Vaswani et al., 'Attention is all you need,' in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.

[2] H. Cui et al., 'Document AI: Benchmarks, models and applications,' ACM Computing Surveys, vol. 56, no. 4, pp. 1–38, 2024.

[3] P. Lewis et al., 'Retrieval-augmented generation for knowledge-intensive NLP tasks,' in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[4] J. Gao et al., 'Retrieval-augmented generation for large language models: A survey,' arXiv preprint arXiv:2312.10997, 2023.

[5] R. Smith, 'An overview of the Tesseract OCR engine,' in Ninth International Conference on Document Analysis and Recognition (ICDAR), vol. 2, pp. 629–633, 2007.

[6] Y. Huang et al., 'LayoutLMv3: Pre-training for document AI with unified text and image masking,' in Proc. ACM Multimedia, pp. 4083–4091, 2022.

[7] Gemini Team, 'Gemini: A family of highly capable multimodal models,' arXiv preprint arXiv:2312.11805, 2023.

[8] G. Izacard and E. Grave, 'Leveraging passage retrieval with generative models for open domain question answering,' in Proc. EACL, pp. 874–880, 2021.

[9] V. Karpukhin et al., 'Dense passage retrieval for open-domain question answering,' in Proc. EMNLP, pp. 6769–6781, 2020.

[10] S. Robertson and H. Zaragoza, 'The probabilistic relevance framework: BM25 and beyond,' Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[11] LangChain Team, 'LangGraph: Building stateful, multi-actor applications with LLMs,' GitHub Repository, 2024.

[12] S. Q. Yan et al., 'Corrective retrieval augmented generation,' arXiv preprint arXiv:2401.15884, 2024.