International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

OPTIMIZING PDF INGESTION FOR LARGE LANGUAGE MODELS IN RAG ARCHITECTURES

Rishab Bansal[1], Binita Mukesh Shah[2], Abhijit Chanda[3], Vrushali Parate[4]

[1]Independent Researcher, USA

Email: connect.rishabbansal@gmail.com / ORCID: 0009-0002-5348-0872

[2]IEEE Senior Member, Independent Researcher, USA

Email: binitashah6492@gmail.com / ORCID ID: 0009-0009-1555-9134

[3]Senior Manager, Tredence Inc., Fremont, CA, USA

Email: abhijitju252024@gmail.com / ORCID: 0009-0002-6976-3922

[4]Independent Researcher, Farmington Hills, MI, USA

Email: vrushali30109@gmail.com / ORCID: 0009-0007-5031-3581

Abstract

The Portable Document Format (PDF) is widely used for enterprise information communication and archival, but its emphasis on visual fidelity presents major barriers for ingestion into Large Language Model (LLM)-based systems. High-quality data ingestion is critical for Retrieval-Augmented Generation (RAG) systems, which increasingly rely on unstructured organizational knowledge. Complex PDFs, featuring tables, figures, headers, footers, and intricate layouts, often suffer from context loss and semantic degradation during extraction, impairing RAG performance. This paper presents a survey of existing research on parsing such documents for LLM vectorization. It identifies a gap between the capabilities of current parsing techniques, often evaluated on simplified benchmarks, and the needs of real-world enterprise documents. Key challenges highlighted include layout interpretation, contextualization of tables and images, OCR noise reduction, and preservation of semantic relationships. The paper categorizes existing approaches into pipeline-based methods, holistic Vision-Language Models (VLMs), hybrid systems, and graph-based representations. Analysis of reported performance reveals persistent gaps between model accuracy and human-level understanding, especially in complex reasoning tasks, and highlights limitations in current benchmarks. Based on this review, the paper offers practical recommendations for engineers, emphasizing semantic chunking, layout-aware tool selection, multimodal strategies, and metadata enrichment. Future directions include improving multimodal model robustness, establishing realistic benchmarks, enhancing explainability, and ensuring semantic fidelity, the accurate capture and representation of a document's intended meaning and structure, in PDF ingestion pipelines for RAG systems.

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Keywords: Large Language Models, LLM, RAG, PDF Ingestion, PDF, Retrieval Augmentation

1. Introduction

1.1 Processing Challenges of PDFs

The PDF format is used as a standard for document sharing and preservation in most of the industries. Its fundamental feature is formatting uniformity across platforms LLMs and operating systems, assuring visual fidelity for readers. However, this design concept, which prioritizes visual layout preservation through complicated structures, embedded visuals, and structured text, hinders automated information extraction and analysis tools. Companies store financial statements, legal contracts, research papers, technological documentation, and corporate policies in PDFs. Given the extent of modern digital archives, manually extracting and processing this material is time-consuming, error-prone, and impractical. Thus, powerful document intelligence systems that automatically extract and analyze PDF information with high accuracy and contextual awareness are needed. The PDF format encodes a visual presentation rather than a semantic structure, making it difficult for machines to understand the content's logical flow and relationships within it.

1.2 LLMs and RAG for Enterprise Knowledge

Large Language Models (LLMs) have brought a major paradigm change in natural language processing (NLP) and understanding. Having been trained on large text datasets, these models show an ability to understand and produce human-like texts. Their possible use to record intelligence, especially PDF information extraction, presents intriguing paths to get above conventional constraints. Retrieval-Augmented Generation (RAG) has become an important architecture for improving LLMs by integrating external, current information sources [1]. This method allows models to offer context-specific responses without the need for constant retraining [1]; hence, it addresses frequent LLM problems such as hallucinations. Developing RAG systems that operate on corporate document collections, such as company rules, standard operating procedures, technical manuals, and reports, primarily stored in PDF format—is a common and essential use case for machine learning (ML) developers in enterprises. These systems seek to use natural language searches to retrieve the knowledge found inside these documents.

1.3 The PDF Ingestion Constraint

The quality and accuracy of the information base an enterprise RAG system receives essentially determine its general dependability and efficiency [1]. Reading documents is necessary to build this knowledge base; we must parse the content, chunk it into manageable chunks, and produce vector embeddings suitable for semantic search. But the RAG pipeline suffers a significant slowdown in the process of extracting data from complicated PDFs [2]. Standard parsing encounters issues when dealing with PDFs that have non-trivial structures, such as multiple

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

columns, headers, footers, footnotes, tables with merged cells, embedded images, diagrams, and various text formatting.

Improper ingestion techniques might result in loss of the document's structural and relational information, and its visual presentation is discarded [3]. Text from nearby columns, for instance, can be blended wrongly, or the link between a figure and its caption might be disrupted. The process also produces a "missing semantic layer." Layout signals (such as headings signaling hierarchy or proximity suggesting relatedness) lose their relevance, and the LLM is left with sequences of text devoid of essential context. This deterioration of input quality immediately reduces the RAG system's capacity to create accurate, contextually appropriate responses and retrieve pertinent information. The basic contradiction is in the design philosophy of the PDF, which gives visual preservation for human consumption top priority and is essentially incompatible with the LLM's demand for semantically structured, machine-readable content. There is a need for advanced parsing methods that can interpret layout and structure to preserve meaning, rather than relying solely on basic text scraping.

1.4 Scope and Structure

The purpose of this study is to give engineers who are responsible for developing RAG systems over PDF documents a thorough understanding of the difficulties, current approaches, assessment procedures, and useful factors for enhancing the ingesting of complicated PDFs. It explores the problems that lead to semantic gaps and context. The analysis evaluates existing technological solutions, such as modular pipelines and end-to-end multimodal models, examines performance metrics and their limitations, and provides practical recommendations. The following is the format of the following sections: The gap in existing research and practice on PDF parsing for RAG is covered in Section 2. The unique problems that arise while processing complicated PDFs are described in depth in Section 3. Section 4 examines current extraction methods and strategies. Performance considerations, measurements, and evaluation benchmarks are covered in Section 5. Future research directions and useful advice for engineers are outlined in Section 6. The list of references used in the paper is given in Section 7. In the end, solving this upstream issue of robust PDF intake is important to the success of many corporate RAG applications. Although considerable study emphasizes the optimization of retrieval algorithms or the fine-tuning of LLMs, defective data intake fundamentally constrains the system's overall capability, rendering breakthroughs in parsing an essential facilitator [1].

2. The Gap in Current Research and Practice

2.1 Limitations of Traditional Text Extraction

Conventional techniques, which usually depend on libraries like PyPDF2, usually break down when extracting text from PDFs with intricate layouts. These simpler methods frequently produce confused or partial text output when processing documents with several figures, headers, and footers; financial reports with complex tables; or scientific publications with two-column formats. The fundamental problem is that these techniques mostly concentrate on character stream extraction without sufficiently analyzing the spatial structure of the document

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[3]. As a result, the PDF's intrinsic visual and structural information, which is essential to comprehending its content, is severely lost [3]. Essential components like table data may be distorted or left out completely, or text blocks may be concatenated in an illogical manner (reading across columns rather than down them, for example). The output, which frequently poorly represents the original document's meaning, creates a weak basis for further LLM processing and vectorization.

2.2 The Semantic Gap

The idea of the "missing semantic layer" highlights a fundamental drawback of conventional PDF extraction. PDFs use formatting indications and visual layout in addition to text strings to communicate meaning. The geographical proximity of items frequently suggests relationships, bold or italicized text highlights important terms, bullet points indicate lists, and headings and subheadings show hierarchy. Visual representations of structured data are provided by tables and charts.

An LLM's capacity to thoroughly understand the content is severely hampered by this loss of semantic consistency within extracted text chunks. For example, the LLM might not be able to identify the beginning of a new part or the subject matter it covers if the visual prominence of a heading is diminished. If we extract table data without maintaining its row and column structure, we also lose the links between data points. Semantic chunking is one strategy that tries to recover semantics that were already visually evident in the original PDF, but it also divides text based on meaning rather than fixed size after extraction. The innate layout understanding capabilities of LLMs themselves have been the subject of recent study, which frequently reveals that pre-training exposure to structured material, like code, contributes to these abilities [4]. But it's frequently insufficient to rely just on the LLM to deduce structure from poorly retrieved text. Extraction techniques that actively maintain or rebuild the document's structural and relational information are necessary to close this semantic gap.

3. Issues in Complex PDF Ingestion for LLMs

Particularly for RAG systems, ingesting complicated PDFs into a format fit for LLM processing presents difficulties because of the visual character of the format and the variety of document formats used in practice. These problems cause the semantic gap and context loss.

3.1 Layout Complexity and Text Flow

Extracting text from PDFs in a logically coherent sequence that reflects human reading can be challenging since PDF content arrangement depends on exact coordinates rather than a relational structure, therefore producing programmatic extraction that might leap illogically across document sections. Multi-column layouts, which are common in many document types, can sometimes confuse basic parsers. This confusion leads the parsers to read horizontally across columns and merge text into a jumbled stream, disrupting logical flow, mixing phrases, and corrupting context. As a result, such confusion can lead to erroneous data and meaningless RAG system outputs, potentially causing failures in tasks such as Named Entity Recognition

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(NER) or topic classification [15]. Moreover, the inadvertent inclusion of recurring components like page numbers, headings, and footers might generate noise, compromise the semantic integrity of paragraphs, and unnecessarily raise the volume of text for LLMs to analyze, perhaps reducing the relevance of content chunks.

3.2 Table Extraction Challenges

Although tables are rich sources of ordered information, they provide major extraction challenges. Given the sophisticated multi-level headers, complex nested structures, merged cells, and cells including multi-line content found in real-world tables, automated programs certainly find it difficult to precisely determine the limits of cells and grasp the interactions among rows, columns, and headers [2]. Financial or scientific tables are especially susceptible since accurate interpretation depends on proper column-row alignment; once this structure is disrupted, the LLM cannot effectively deduce conclusions. While useful for some analytical chores, converting tables into normalized formats like CSV or JSON usually ignores the surrounding textual context and the internal relational information the table's layout communicates. Usually, the approach removes important context even while extracting data from table cells. LLMs struggle with decontextualized data; they may not understand the table's meaning or how its values relate to each other or the document. Simple extraction may also misunderstand cell special characters or abbreviations. Multiple page tables add still another degree of complexity and demand tools to correctly identify continuity headers and sew the table portions together exactly [2]. Ignoring these standards results in either broken or partial table data.

3.3 Image and Vector Graphics Handling

PDFs are naturally multimodal, often including photos, graphs, charts, logos, scanned signatures, and other vector graphics that deliver important information. Conventional text-only extraction algorithms lose this information for downstream LLMs by treating visual components as isolated information silos, so neglecting important visually displayed data as trends in charts, links in diagrams, or validation by signatures [3, 5]. Extracting an image file alone is insufficient; real knowledge calls for connecting the visual element to its surrounding context inside the document—including associated labels, subtitles, or textual references [3]. Beyond simple extraction, a more difficult task is deciphering the substance and relevance of the visual item [3]. Respected multimodal models (such as VLMs) meant to process both visual and textual input [6] enable the system to "see" and understand the chart, therefore enabling answers to questions like "What trend does the Q3 sales chart show?" [6].

3.4 OCR Noise and Its Impact on RAG

Optical Character Recognition (OCR) is a required step for scanned PDFs or PDFs with image-based text; nonetheless, it is sometimes lacking and results in noise. Factors that affect OCR accuracy include poor scan quality (such as low resolution, skew, background noise, and watermarks), difficult document types (like historical documents and handwritten text), complex fonts or layouts, and the inclusion of specialized mathematical or scientific symbols.

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Studies on OCR noise classify it by its downstream impact, separating "semantic noise," which alters the meaning of the text by misreading digits or replacing important words, from "formatting noise," which influences structure or readability without changing core content (e.g., improper spacing, false line breaks, minor character replacement). As RAG systems are known to be sensitive to such input noise, these OCR mistakes spread through the RAG pipeline, affecting the quality of the knowledge base [1]. While even formatting problems can disturb chunking, embedding generation, and retrieval, potentially leading the system to retrieve irrelevant content or miss the correct information, semantic noise can lead to factually erroneous responses or the retrieval of wrong information. Therefore, establishing high-quality knowledge bases for RAG systems depends much on accurate OCR predictions, which are somewhat difficult [1].

Th1s 1s example 0f bad OCR exttract w1th n01se.

Revenue growth of 1O% was rec0rded.

Market share increased t0 15%.

Example 1. Errors found in OCR Parsing

Example 1 shows issues that can be generated because of wrong OCR parsing:

- 1 is read instead of I
- 0 (zero) is read instead of O
- Misspellings and character confusion

3.5 Loss of Semantic Relationships and Hierarchy

The general semantic structure of the document is often lost during intake, transcending features like tables or columns. Subtle structural cues indicating hierarchy and semantic organization inside a page are font size, bolding, indentation, and space; these cues are lost when extraction flattens the content into plain text, therefore making it more difficult for LLMs to detect the flow of information. Although tables of contents (ToCs) in structured publications provide a high-level structural map, unsophisticated ingestion techniques may treat this important data as either simple sequential text or ignore it completely, so losing important hierarchical context and affecting RAG system processing and retrieval. Naive chunking, particularly when using fixed-size windows, can disrupt semantic continuity by arbitrarily separating paragraphs or words, which may lead to LLMs lacking the necessary preceding or succeeding context for complete understanding. Such behavior results in less accurate or incomplete responses and fails to address the problem that semantic chunking aims to solve by segmenting content along logical boundaries. Moreover, the existence of internal cross-references in complicated documents presents a difficulty for simple linear chunking to adequately depict; this impairs the RAG system's ability to follow links or synthesize

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

knowledge from various, potentially distant parts of the document without a comprehensive understanding of its content organization.

4. Existing Techniques for PDF Content Extraction

Dealing with the difficulties of sophisticated PDF intake has resulted in the creation of several approaches, generally classifiable into pipeline-based methods, end-to-end multimodal models, hybrid tactics, and graph-based representations.

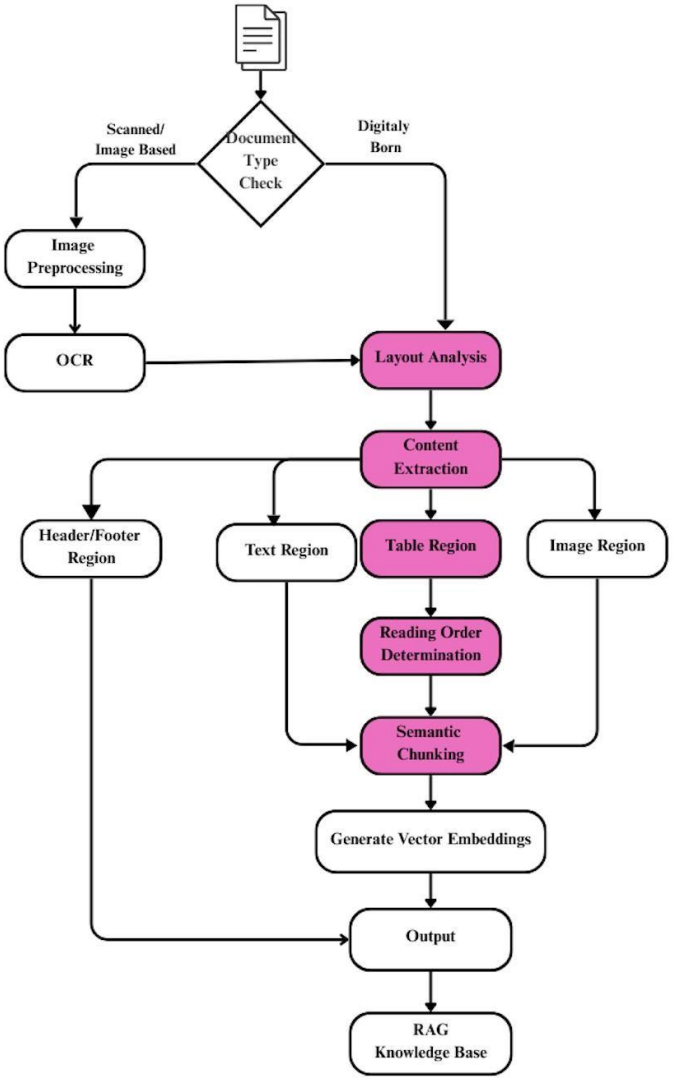

Figure 1. Flowchart of a typical pipeline-based PDF ingestion process for RAG systems. Stages prone to common challenges are pink-filled.

4.1 Methodologies Based on Pipelines

This conventional paradigm breaks out the difficult choreography of document content extraction into a set of separate, sequential modules, each addressing a different sub-task [2].

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

As per shown in Figure 1, usually, the phases consist of

- Analysis of Layout: Along with their spatial coordinates (bounding boxes), this first phase finds the structural parts of a document page—text blocks, paragraphs, headings, graphics, tables, and mathematical expressions—along with, ideally, their logical reading order[5]. Correct layout analysis is fundamental for later content extraction [5]. Deep learning techniques, among other models, have been developed for this work [2].

- OCR engines translate visual character representation into machine-readable text for scanned documents or text integrated as images [1]. The process entails examining patterns and forms in character development [5]. Common open-source engines like Tesseract are extensively utilized, together with commercial cloud services with sophisticated OCR features. Given that mistakes spread downstream, this stage's precision is crucial [1].

- Reading Order Estimation: Algorithms utilize layout information to ascertain the appropriate sequence for reading text blocks, which is especially vital for multi-column or intricate non-linear layouts [2].

- Extraction of Specific Elements: Many pipelines include specialized modules for element extraction, such as table recognition, which identifies table boundaries and structures, extracts data, and may convert it to formats like LaTeX or CSV, thereby addressing the specific issues associated with different types of content. Mathematical expression recognition handles complex symbols and layouts, detects and converts mathematical formulas into standardized formats like LaTeX or MathML, and uses tools like Nougat especially effective for academic literature. Chart recognition involves identifying various types of charts and extracting the underlying data or structural relationships, often converting visual data into data tables or JSON. Form processing then is focused on identifying and extracting data from radio buttons, checkboxes, and form fields.

Many tools combine these parts into a seamless process. Marker offers an optional LLM-enabled version that improves the handling of multi-page tables and inline math, along with open-source models designed to convert papers into structured formats such as Markdown, JSON, or HTML [2]. MinerU splits the page using layout detection first, then applies task-specific models to various areas, and lastly generates material in Markdown with a decided reading order [1]. Libraries such as PyMuPDF and pdfplumber offer strong tools for extracting text, photos, and occasionally basic table structures. Another well-liked open-source tool with thorough parsing powers, including table structure inference, is Unstructured.io. Pipeline methods have mostly advantaged in that they allow one to use specific, expert models for every sub-task and maybe attain outstanding efficiency by parallelization [2]. They can, however, suffer from error propagation, in which case early-stage (e.g., layout analysis or OCR) mistakes severely affect all later stages.

4.2 End-to-End Vision-Language Models/Multimodal Models

Based on Transformer architecture, a more contemporary and fast-developing method uses

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

single, strong models that handle document inputs holistically, integrating visual, textual, and layout information simultaneously [2]. Often directly from document images or pages, these models seek to do end-to-end document understanding chores, including content extraction or question answering. Their extensive pre-training allows them to generalize to many and even invisible document types without requiring fine-tuning specific data, offering a major potential advantage [2].

These models have evolved greatly over time.

- **Early Layout-Aware Models:** Series LayoutLM uses 2D position embeddings to show the spatial positioning of tokens on the document page. LayoutLM was a pioneering approach that clearly included layout information in the pre-training process along with text [7]. For jobs including form and receipt interpretation, this simultaneous modeling of text and layout proved helpful. Later iterations included visual signals like font styles or colors, hence, improving performance [7]. At the time of its publication, LayoutLM produced state-of-the-art results on benchmarks like FUND, SROIE (receipts), and RVL-CDIP (classification) [7].

- **Document Vision Transformers (DiT)** directly implemented the Vision Transformers (ViT) architecture on document pictures, extending its success in computer vision [7]. Using huge-scale unlabeled document image datasets (like IIT-CDIP), DiT uses self-supervised pre-training procedures (like BEiT) [7]. The process lets the model learn rich visual images pertinent to document structure and content free from human labels [7]. Pre-trained DiT models provide a strong foundation for many downstream document AI tasks by significantly improving document picture categorization, layout analysis, and table/text identification [8]. Related projects, like DocTr, also utilize transformers for tasks like geometric unwarping and lighting corrections in document images [9].

- General **Multimodal Models** and LLM Integration: The newest trend is integrating sophisticated document understanding capabilities either directly with or into Multimodal Large Language Models (MLLMs) or Large Language Models (LLMs) [10]. Using layout-aware pre-training and fine-tuning activities, LayoutLLM, for instance, employs instruction tuning specifically designed to improve an LLM's comprehension and use of document layout information, so introducing concepts like layout chain-of-thought (Note: two distinct papers share the name LayoutLLM, one focusing on instruction tuning generally, the other concentrating on layout instruction tuning) [10, 11]. Strong performance in document content extraction tasks has also been shown by potent general-purpose MLLMs like OpenAI's GPT-4o 5 and Qwen2-VL [2], which can directly parse interleaved text and images straight from PDF files. These models gain greatly from extensive pre-training, including text, graphics, and maybe layout ideas obtained from online data or code [4].

Using typically VLM-like foundations, commercial cloud solutions, including Azure Document Intelligence (previously Form Recognizer), Google Document AI, and AWS Tesseract, provide powerful document analysis tools. They offer APIs for OCR, layout analysis, table extraction, key-value pair extraction, and classification with the ability for

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

custom model training. For managing different document types, these end-to-end models shine in understanding both visual layouts and textual content, perhaps providing a better mix between accuracy and performance than complicated pipelines [2]. Large VLMs can be computationally costly, nevertheless, for training and running.

4.3 Graph-Based Representation

This technique emphasizes representing the acquired data as a knowledge graph instead of as linear text chunks. Documents are used to find important entities (like people, companies, subjects, and particular policy provisions) and their interactions. LLMs themselves can efficiently extract entities and relationships from unstructured or semi-structured text within the PDFs [12]. The resulting graph structure explicitly captures connections between various sections of a text, links between related documents, or sophisticated multi-hop connections between entities—connections that could be lost in simple chunking. For RAG systems, this ordered form has possible benefits. The system can query the graph to retrieve entities, their properties, and associated entities or context, thereby possibly spanning several documents, instead of retrieving sometimes fragmented text portions based just on vector similarity. More complicated reasoning and multi-hop question answering can benefit from this, as the response requires synthesizing knowledge from several facets of the knowledge base. This shift towards graph representations provides a structured knowledge backbone for LLM reasoning, hence addressing the inherent constraints of just sequential text in capturing hierarchical linkages and long-range dependencies.

This illustrates continual progress in techniques as it moves from modular, specialized components to more integrated, multimodal systems that are capable of better comprehension. Models that use pre-training on large datasets and can jointly process text, layout, and visual information clearly show direction. For practical, large-scale deployment in business environments, however, hybrid or intelligently routed techniques remain rather important despite ongoing trade-offs between accuracy, speed, cost, and expertise. The needs of the application, including the kinds of papers being handled, the necessary degree of accuracy, and the available computer resources, greatly influence the technique chosen.

Technique	Description	Advantages	Disadvantages	Use Cases
Pipeline-Based Methods	Utilize multiple specialized components to handle different aspects of document parsing, such as OCR, layout analysis, and entity extraction.	Modular and flexible and be optimized for specific tasks	Complexity in integration and potential for fragmented results	Legal document processing, form extraction

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

End-to-End Vision-Language Models	Use single models that integrate visual, textual, and layout information simultaneously to understand documents holistically.	Holistic processing Generalizes well to diverse document types	High computational cost May act as a "black box"	Content extraction, question answering
Hybrid Systems	Combine multiple techniques intelligently routed or hybridized to balance performance and accuracy.	Balanced performance Optimizes resource usage; Leverages strengths of different methods.	Complexity in design. Routing logic can be complex to design and maintain.	Business applications, large-scale deployments
Graph-Based Representations	Represent document data as knowledge graphs to capture entities and their relationships, providing structured knowledge for LLM reasoning.	Captures connections Effective for multi-hop reasoning	Potential for LLM hallucination during extraction Extraction complexity (defining schema, ensuring accuracy)	Multi-document analysis, complex reasoning tasks

Table 1. Summarizes all the existing techniques mentioned in the paper.

5. Evaluations and Performance

Evaluation of performance for PDF ingestion is crucial but complex. Various benchmarks and metrics specific to the tasks mentioned above are required to measure the performance of the pipeline.

5.1 Common Benchmarks and Datasets

The research community relies on several standard datasets to benchmark progress in document AI, each focusing on specific tasks and document types:

- **Document Image Classification:** RVL-CDIP [16] is a large dataset (400k images) for classifying documents into 16 categories (letter, email, report, invoice, etc.).
- **Layout Analysis/Detection:** PubLayNet [17] contains over 360k document images

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

derived from PubMed articles, annotated for layout elements like text, title, list, figure, and table. FUNSD [18] includes scanned forms and is used for tasks including text detection.

○ **Information Extraction (Forms/Key-Value/Entities):** FUNSD [18] is also used for form understanding (extracting semantic entities). Kleister Charity (KLC) (charity reports) [21] and SciREX (scientific articles) [19] are used for information extraction, often framed as question-answering. Datasets like DocRED and its variants focus on document-level relation extraction [20].

○ **Document Visual Question Answering (DocVQA):** DocVQA [22] uses industry documents with question-answer pairs requiring visual understanding. TAT-DQA [23] focuses specifically on documents containing tables and text, demanding discrete reasoning (e.g., calculations) over financial reports. ChartQA [24] targets question answers over charts. WikiTableQuestions (WTQ) [25] involves QA over tables from Wikipedia. XFundQA and FetaQA are other multilingual/table-focused QA datasets [4].

○ **Table Detection/Structure Recognition:** ICDAR 2019 cTDaR [26] provides datasets specifically for detecting table regions (Track A) and recognizing internal table structure, covering both historical and modern documents.

○ **Receipt Understanding:** SROIE is a common benchmark for extracting information from scanned receipts, used in evaluating models like LayoutLM [27].

○ **OCR Impact on RAG:** OHRBench [1] is specifically designed to evaluate the end-to-end impact of OCR quality on RAG systems, using real-world documents from diverse domains and Q&A pairs requiring understanding, reasoning, and multi-page context.

5.2 Examples of State-of-the-Art (SotA)

On several benchmarks, pre-trained models that incorporate layout and visual information have shown notable performance gains:

● Compared to earlier text-only BERT-based models, LayoutLM demonstrated significant gains in FUNSD form comprehension (F1 score increased from 70.72 to 79.27), SROIE receipt comprehension (F1 score increased from 94.02 to 95.24), and RVL-CDIP document classification (accuracy increased from 93.07 to 94.42) [7].

● DiT made RVL-CDIP more accurate (increased from 91.11 to 92.69), improved PubLayNet layout analysis (mAP went up from 91.0 to 94.9), and enhanced ICDAR 2019 cTDaR table detection (weighted F1 rose from 94.23 to 96.55) by using self-supervised pre-training on document images [8].

● Significant conflicts between perception (such as OCR capability) and cognition (such as responding to questions based on the perceived text) were discovered in research evaluating multimodal large language models (MLLMs) like GPT-4o. GPT-4o only achieved 68.6% consistency on document understanding tasks, suggesting the possibility of unreliability [6]. Even though they performed noticeably better than baselines on the reasoning-intensive TAT-

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DQA dataset, models such as MHST still performed far worse than expert humans [23].

5.3 Evaluating Impact on Downstream RAG

Evaluation must go beyond component-level accuracy (such as OCR or layout detection) to evaluate the RAG application's overall performance, which is crucial for engineers developing RAG systems [1]. The quality of the initial document parsing and OCR has a direct, cascading effect, as the OHRBench benchmark emphasizes [1]. Understanding how different types of noise—formatting noise affects structure, while semantic noise changes meaning—affect both the retrieval stage (for example, failing to locate the correct chunk) and the generation stage (for example, producing incorrect answers based on flawed retrieved text) is crucial [1]. It has been discovered that even the most advanced OCR solutions may not be able to reliably build the high-quality knowledge bases needed for reliable RAG systems working on various real-world documents [1]. Evaluators should include various question types (testing simple lookup, reasoning, and multi-page synthesis) to assess the quality of the final response derived from the ingested PDF content [1].

5.4 Limitations of Current Evaluations

As was previously mentioned, standards frequently lack document diversity and place a strong emphasis on document types, such as academic papers or forms, which might not accurately reflect the complexity of enterprise documents (such as financial filings and legal contracts) [2]. Metrics used for evaluation may vary from study to study or may overlook important factors like reading order accuracy or semantic fidelity [5]. Furthermore, privacy, confidentiality, and proprietary concerns make it difficult to find large-scale, diverse, and representative datasets of actual enterprise documents for public research [14]. This dependence on a small number of corpora makes it more difficult to create and validate models that are resilient to the diversity found in practice [14].

It is challenging to conduct comprehensive comparisons of various end-to-end PDF ingestion systems due to this evaluation fragmentation [1]. More significantly, there aren't many benchmarks that thoroughly assess the critical relationship between the final quality and factual accuracy of the responses produced by the RAG system in a realistic environment and the quality of the parsed, chunked, and embedded data [1]. Even though SotA models show remarkable improvements on certain benchmark tasks, the ongoing performance gap with humans, particularly on complex reasoning or noisy, out-of-distribution data [14], raises the possibility that current evaluation techniques do not adequately account for the needs of obtaining trustworthy, human-level document understanding in real-world enterprise applications.

6. Conclusion and Future Directions

6.1 Synopsis

Leveraging the full potential of large language models and retrieval-augmented generation systems inside businesses still depends on the ability to consume challenging PDF documents.

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Significant difficulties arise from the intrinsic tension between the LLM's demand for semantically organized input and the visual fidelity design of the PDF format. Often complex layouts, sophisticated tables, embedded images, and OCR flaws cause context loss and the absence of a required semantic layer in the obtained data. The main concerns discussed in this work include handling non-linear text flow, retaining table context, interpreting visual features, reducing OCR noise, and preserving hierarchical relationships. Current techniques span modular pipelines using specialized components to end-to-end vision-language models, hybrid methods, and new graph-based representations, providing holistic processing. Each method involves the trade-offs involve accuracy, speed, cost, and complexity. Although helpful, current evaluation standards sometimes lack the diversity and end-to-end attention needed to completely anticipate real-world RAG performance, where a gap between state-of-the-art models and human dependability exists.

6.2 Practical Takeaways for Engineers

Navigating these obstacles for developers creating RAG systems utilizing internal PDF policy papers or comparable corpora calls for a practical and knowledgeable approach:

- Use a refined tool selection strategy. Avoid adopting a universal solution. Faster and typically open-source technologies like PyMuPDF can be rather effective for basic, digitally born PDFs with simple layouts. Invest in more advanced solutions for scanned documents, especially those with complicated multi-column layouts, detailed tables, forms, or crucial graphic features. To balance cost and accuracy, think about using an intelligent routing system that examines document complexity upfront to choose the most suitable parser.

- Preserve the layout: Use output forms and extraction techniques that maintain the layout context of the document whenever at all feasible. Markdown, for instance, can more LLM-friendlyly show headings, lists, and even table structures than plain text [2]. Tools especially made for layout-preserving extraction can greatly enhance the LLM's capacity to grasp the structure and relationships within the information.

- Create tables and image strategies. Tables call for particular care. Using dedicated table extraction models inside a pipeline, storing structured table data separately (e.g., in JSON files or a database), or using LLMs to translate extracted tables into descriptive natural language summaries to retain context—all of which will help you to link this structured data back to the pertinent text chunks during retrieval [2]. Prioritize utilizing multimodal models' (VLMs/MLLMs) capability of directly processing visual content for papers where images, charts, and diagrams are essential [3].

- Evaluate and improve OCR quality. OCR quality is the priority for scanned document processes. Analyze the result of your selected OCR engine. Before OCR, apply picture pre-processing techniques (e.g., deskewing, noise reduction) for best results. Be particularly mindful that OCR mistakes immediately affect downstream RAG performance and could call for quality control or remedial actions [1].

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Use semantic chunking. Semantic chunking methods are meant to split the material along logical lines—sentences, paragraphs, or semantic themes. This technique preserves contextual integrity inside every piece, therefore giving the LLM more coherent information during retrieval.

- Experiment with multimodal models: Many complicated PDFs have inherent multimodal character; thus, actively explore end-to-end VLMs or MLLMs [3]. These models perhaps provide a stronger answer for managing complicated layouts, tables, and figures than text-only pipelines by processing visual and textual input together, therefore streamlining the whole intake process.

- User Knowledge Graphs for Complex Domains. Graph-based RAG (GraphRAG) techniques are investigated for applications containing densely connected material, many cross-references, or the necessity of multi-hop reasoning across documents—common in legal or complicated policy domains. By extracting items and relationships from LLMs into a knowledge graph, one can create a structured backbone clearly modeling linkages sometimes lost in linear chunking.

6.3 Suggestions for Future Studies

Although some improvement has been made, a major study is still required to fully address the challenges associated with complicated PDF intake for LLMs:

- Better and realistic benchmarks: Development of more varied and thorough benchmarks is vital. These should encompass a greater spectrum of real-world document categories (financial, legal, technical) and assess performance on end-to-end RAG tasks, specifically quantifying the effect of parsing quality on final answer accuracy and robustness [1].

- Further developments in MLLMs are required to reach greater integration and reasoning between visual (layout, charts, images) and textual modalities. The key is to solve the noted tensions between cognition and perception [6]. Models must consistently understand intricate tables, graphs, and the interactions among text and images [5].

- Effective and easily reachable VLMs, particularly for on-site deployment or in cost-sensitive applications, and research on building smaller, faster, but strong VLMs are crucial for enabling more general adoption [2].

- Better methods are needed for robustly handling noise, geometric distortions (skew, warp), low resolution, watermarks, and other abnormalities typical of scanned or badly digitized materials.

- Developing effective strategies for customizing big, general-purpose document AI models to company domains, terminologies, and special document layouts remains a challenging field [14].

- Modern Graph Strategies: More effectively than existing approaches, ongoing research

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

on LLM-powered knowledge graph building, refining, and querying strategies promises to capture complicated relationships both inside and across texts.

The best approach for PDF intake is probably hybrid and adaptable, dynamically choosing solutions depending on document properties and job needs. The emphasis must move from simple technical accuracy of extraction to achieving true semantic authenticity, ensuring that the information passed to the LLM accurately reflects the meaning, structure, and relationships present in the original document, enabling more reliable and intelligent RAG systems.

7. **References**

1. Zhang, J., Zhang, Q., Wang, B., Ouyang, L., Wen, Z., Li, Y., ... & Zhang, W. (2024). OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation. arXiv preprint arXiv:2412.02592.
2. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., ... & He, C. (2024). Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. arXiv preprint arXiv:2412.07626.
3. Xie, X., Yan, H., Yin, L., Liu, Y., Ding, J., Liao, M., ... & Bai, X. (2024). WuKong: A Large Multimodal Model for Efficient Long PDF Reading with End-to-End Sparse Sampling. arXiv preprint arXiv:2410.05970.
4. Li, W., Duan, M., An, D., & Shao, Y. (2024). Large Language Models Understand Layout. arXiv preprint arXiv:2407.05750.
5. Zhang, Q., Huang, V. S. J., Wang, B., Zhang, J., Wang, Z., Liang, H., ... & Zhang, W. (2024). Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. arXiv preprint arXiv:2410.21169.
6. Shao, Z., Luo, C., Zhu, Z., Xing, H., Yu, Z., Zheng, Q., & Bu, J. (2024). Is Cognition consistent with Perception? Assessing and Mitigating Multimodal Knowledge Conflicts in Document Understanding. arXiv preprint arXiv:2411.07722.
7. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020, August). Layoutlm: Pre-training of text and layout for document image understanding. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 1192-1200).
8. Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., & Wei, F. (2022, October). Dit: Self-supervised pre-training for document image transformer. In Proceedings of the 30th ACM international conference on multimedia (pp. 3530-3539).
9. Feng, H., Wang, Y., Zhou, W., Deng, J., & Li, H. (2021). Doctr: Document image transformer for geometric unwarping and illumination correction. arXiv preprint arXiv:2110.12942.
10. Fujitake, M. (2024). Layoutllm: Large language model instruction tuning for visually rich document understanding. arXiv preprint arXiv:2403.14252.

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

11. Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., & Yao, C. (2024). Layoutllm: Layout instruction tuning with large language models for document understanding. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 15630-15640).
12. Kumar, R., Ishan, K., Kumar, H., & Singla, A. (2025). LLM-Powered Knowledge Graphs for Enterprise Intelligence and Analytics. arXiv preprint arXiv:2503.07993.
13. Sun, S., An, W., Tian, F., Nan, F., Liu, Q., Liu, J., ... & Chen, P. (2024). A review of multimodal explainable artificial intelligence: Past, present and future. arXiv preprint arXiv:2412.14056.
14. Nourbakhsh, A., Shah, S., & Rose, C. (2024). Towards a new research agenda for multimodal enterprise document understanding: What are we missing?. Findings of the Association for Computational Linguistics ACL 2024, 14610-14622.
15. Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. International Journal of Computer Science and Network Security, 8(2), 339-344.
16. Harley, A. W., Ufkes, A., & Derpanis, K. G. (2015, August). Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th international conference on document analysis and recognition (ICDAR)* (pp. 991-995). IEEE.
17. Zhong, X., Tang, J., & Yepes, A. J. (2019, September). Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)* (pp. 1015-1022). IEEE.
18. Jaume, G., Ekenel, H. K., & Thiran, J. P. (2019, September). Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)* (Vol. 2, pp. 1-6). IEEE.
19. Jain, S., Van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). SciREX: A challenge dataset for document-level information extraction. *arXiv preprint arXiv:2005.00512*.
20. Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., ... & Sun, M. (2019). DocRED: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.
21. Stanisławek, T., Graliński, F., Wróblewska, A., Lipiński, D., Kaliska, A., Rosalska, P., ... & Biecek, P. (2021, September). Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition* (pp. 564-579). Cham: Springer International Publishing.
22. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., & Jawahar, C. V. (2022). Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1697-1706).
23. Zhu, F., Lei, W., Feng, F., Wang, C., Zhang, H., & Chua, T. S. (2022, October). Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM*

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025

Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

International Conference on Multimedia (pp. 4857-4866).

24. Masry, A., Long, D. X., Tan, J. Q., Joty, S., & Hoque, E. (2022). Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.
25. Pasupat, P., & Liang, P. (2015). Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*.
26. Gao, L., Huang, Y., Déjean, H., Meunier, J. L., Yan, Q., Fang, Y., ... & Lang, E. (2019, September). ICDAR 2019 competition on table detection and recognition (cTDaR). In *2019 International conference on document analysis and recognition (ICDAR)* (pp. 1510-1515). IEEE.
27. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., & Jawahar, C. V. (2019, September). Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1516-1520). IEEE.

Received: July 24 2025; Accepted: 13 September 2025; Published: 25 September 2025