REVIEW ARTICLE

Intelligent Document Processing: From OCR to Multimodal Foundation Models-Architectures, Evidence, and Enterprise Integration

Atul Kumar[1,*]

[1] Department of Computer Science, R.G.M. Government College, Joginder Nagar 175015, India

Abstract

Intelligent document processing (IDP) integrates text recognition, document geometry, visual features, semantic interpretation, and information access to convert heterogeneous documents into machine-actionable information. This review examines the evolution from rule-based optical character recognition (OCR) and convolutional neural network (CNN)-recurrent sequence recognizers to layout-aware document transformers, OCR-free document models, vision-language foundation models, and visual retrieval systems. Rather than ranking architectures using metrics obtained from incompatible datasets and protocols, the review distinguishes four complementary capabilities: recognition, extraction, reasoning, and retrieval. It compares representative systems in terms of their input representations, learning objectives, strengths, limitations, and operational requirements. A structured narrative-review protocol is reported, including search sources, keywords, coverage period, and eligibility criteria. Quantitative claims are reported only when they can be traced to primary sources and compatible evaluation protocols. The review also examines industrial evidence with explicit source-quality qualifications and presents a transparent, assumption-based total-cost example. Finally, it discusses low-resource scripts, including recent reviews of handwritten and low-resource historical material, and identifies research priorities in reproducible benchmarking, evidence-grounded reasoning, privacy, and human oversight.

Keywords: intelligent document processing, optical character recognition, document layout analysis, document transformers, vision-language models, multimodal AI, visual retrieval, human-in-the-loop.

1 Introduction

Organizations process large volumes of invoices, claims, contracts, forms, receipts, reports, and archival records. Much of the information in these documents is unstructured or semi-structured, making it difficult to use directly in computational workflows. Traditional OCR converts document images into text, but transcription alone does not

Submitted: 13 October 2026
Revised: 04 November 2026
Accepted: 10 November 2026
Published: 27 November 2026

Vol. 2, **No.** 4, 2026.
10.62762/JIAP.2026.875404

***Corresponding author:**
✉ Atul Kumar
atulkmr02@gmail.com

Citation
Kumar, A. (2026). Intelligent Document Processing: From OCR to Multimodal Foundation Models—Architectures, Evidence, and Enterprise Integration. *ICCK Journal of Image Analysis and Processing*, 2(4), 277–294.

© 2026 by the Author. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

necessarily preserve page structure, reading order, table relationships, or the semantic roles of individual fields. Enterprise document workflows therefore require more than character recognition: they may need region detection, layout reconstruction, field extraction, evidence retrieval, and reasoning over information distributed across pages. Intelligent document processing (IDP) brings computer vision, OCR, document layout analysis, natural language processing, retrieval, and multimodal models together in a configurable pipeline. This review examines the architectural changes behind that evolution and assesses how far the available evidence supports claims about each change.

The transition from perceptual OCR to document intelligence is driven by the limitations of sequential recognition pipelines. Preprocessing, segmentation, recognition, and post-processing can work well on clean machine-printed pages, but performance may deteriorate with multi-column layouts, skew, bleed-through, handwriting, non-standard typography, and degraded images. More importantly, a plain text transcript does not fully represent the spatial and semantic relationships on a page. IDP systems therefore combine multiple forms of evidence, including text tokens, two-dimensional coordinates, visual regions, tables, and linguistic context.

1.1 Review Questions and Contributions

This review addresses three questions: (1) How have document processing architectures evolved from heuristic OCR to multimodal foundation models? (2) How do their representations, learning objectives, and inference pipelines differ, and what are the implications for recognition, extraction, reasoning, and retrieval? (3) What evidence, deployment constraints, governance requirements, and research gaps should guide the selection of IDP systems? The contributions are a capability-oriented taxonomy; a qualitative architectural comparison grounded in primary model papers; a transparent narrative-review protocol; a separation of task-specific evidence from incompatible cross-study metrics; and an enterprise analysis that makes assumptions and evidence limitations explicit.

2 Literature Review

The literature on document processing has progressed from conventional OCR to learned sequence recognition and multimodal document understanding. Smith [1] describes the architecture and development of the Tesseract engine. Deep sequence recognizers reduced reliance on hand-engineered character segmentation; Shi et al. [2], for example, introduced a convolutional recurrent neural network (CRNN) architecture trained with a sequence objective for image-based text recognition. Such methods improve line-level transcription but do not, by themselves, solve page-level layout interpretation or semantic field association.

Image restoration and preprocessing remain relevant when source pages are degraded. Super-resolution, denoising, binarization, and geometric correction can improve the visual input available to a recognizer, although gains depend on the degradation type and the downstream model. These operations should be evaluated on the target document collection rather than assumed to improve every OCR pipeline. In practice, preprocessing may also remove faint strokes or punctuation, so image enhancement must be validated against character- and word-level ground truth.

Layout-aware document models address a different problem from line recognition. LayoutLM combines textual tokens with two-dimensional position information [3]; LayoutLMv2 adds visual features and multimodal pretraining [4]; and LayoutLMv3 uses unified text/image masking and word-patch alignment [5]. BROS emphasizes text and layout for key information extraction [6]. DocFormer integrates textual, visual, and spatial features through multimodal self-attention and shared spatial embeddings [7]. These methods are not interchangeable: their input dependencies, pretraining objectives, and evaluated tasks differ.

OCR-free approaches attempt to avoid a separately executed OCR stage. Donut maps document images to structured outputs using an encoder-decoder transformer [8]; Nougat specializes in converting scientific documents into markup and was published at ICLR 2024 [9]; and Pix2Struct uses screenshot parsing as a pretraining task for visual-language understanding [10]. These models have different target domains and should not be treated as a single benchmark family. OCR-free inference can reduce pipeline coupling, but generated sequences still require validation for omissions, normalization errors, and unsupported content.

Document question answering and multimodal foundation models extend the task from predefined extraction to answering questions about document images. DocVQA provides a task-specific

benchmark for visual question answering on document images [11]. Qwen2-VL introduces dynamic-resolution image processing and multimodal positional representations [12]. GPT-4o is a proprietary general-purpose multimodal model whose system card [13] documents overall capabilities and safety evaluations rather than document-extraction accuracy. Its document performance must therefore be established separately and reported with the exact model/API version, date, prompt, image resolution, tool access, and evaluation set. Results from general multimodal benchmarks should not be treated as direct evidence of document extraction accuracy. Visual retrieval is a complementary capability. ColPali encodes document-page images into multi-vector representations and uses late interaction for page retrieval; the authors introduce ViDoRe as a benchmark for visually rich page-level retrieval [14]. Retrieval identifies potentially relevant evidence; it does not itself guarantee correct transcription, extraction, or reasoning, and its evaluation is discussed in Section 4.6. Commercial platforms from Google Cloud [15], Microsoft Azure [16], and Amazon Web Services [17] provide managed document-processing services and APIs. Their pricing documentation is useful for understanding the available processors, billable service units, and pricing structure, but provider documentation is not independent evidence of comparative model accuracy. Customer stories can illustrate workflow integration and reported operational outcomes, but should be labeled as vendor- or customer-reported evidence unless independently replicated. This review therefore distinguishes peer-reviewed model evidence, official technical documentation, and industry case reports.

Low-resource scripts and handwritten material remain challenging because public datasets, pretrained recognizers, and standardized evaluation protocols are unevenly distributed across languages and writing systems. Two review-level studies are relevant to this gap. Memon et al. [18] systematically review handwritten OCR research, summarizing the machine-learning techniques and databases used for different languages. More recently, Balabayev et al. [19] synthesize indirect evidence on AI methods for low-resource Chagatai handwritten manuscripts, where historical records remain difficult to access through modern digital systems. Both studies indicate where data and methods are scarce, but they do not provide a common benchmark; their conclusions should therefore not be compared directly with character error rate (CER) and word error rate (WER) results for full-page document understanding.

3 Review Methodology and Evidence Scope

This paper adopts a structured narrative-review design rather than claiming to be a systematic review or meta-analysis. Targeted literature searches and reference-chain checks were used to identify representative primary studies and official technical sources published between 2006 and 1 November 2026. Search locations included Google Scholar, Crossref, PubMed/PMC, IEEE Xplore, ACM Digital Library, SpringerLink, ACL Anthology, arXiv, and official publisher or project repositories. These sources cover OCR, computer vision, document understanding, multimodal learning, and deployed document-processing systems. Because a complete record-level screening log was not maintained, the review does not claim an exhaustive search compliant with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, nor does it report record counts.

Search terms were combined using Boolean operators and model names, including: "optical character recognition" OR OCR; "handwritten text recognition"; "document layout analysis"; "document understanding"; "visually rich document"; "LayoutLM" OR "DocFormer" OR "Donut" OR "Nougat" OR "Pix2Struct"; "vision-language model" AND document; "document visual question answering" OR DocVQA; "visual document retrieval" OR ColPali OR ViDoRe; "low-resource script" AND handwriting; ("low-resource" OR historical) AND handwritten AND (manuscript OR script); and "intelligent document processing" AND (enterprise OR cost OR deployment). Backward reference checking and publisher/project pages were used to verify bibliographic details and model scope.

Sources were eligible when they described an identifiable OCR, layout-analysis, document-understanding, question-answering, retrieval, or enterprise deployment method and provided enough information to identify the task, architecture, dataset, evaluation measure, or operational context. Peer-reviewed primary papers were prioritized for model mechanisms and benchmark results; official model and provider documentation was used for version-specific capabilities and service characteristics; customer stories and industry reports were retained only as

Table 1. Classification of evidence sources used in the comparative analysis of intelligent document processing.

Evidence class	Typical source	Interpretation
A – Primary benchmark	Original peer-reviewed model paper	Direct experimental evidence under the paper's stated protocol.
B – Independent evaluation	Later peer-reviewed benchmark or replication	Useful for external validation and cross-study comparison.
C – Technical documentation	Official provider/model documentation	Useful for capabilities, deployment, and service characteristics.
D – Customer/industry case	Vendor case study or industry report	Useful for operational context; should not be treated as independently reproduced benchmark evidence.
E – Indicative assumption	Analytical assumption or derived estimate	Suitable for scenario analysis, not as a measured model result.

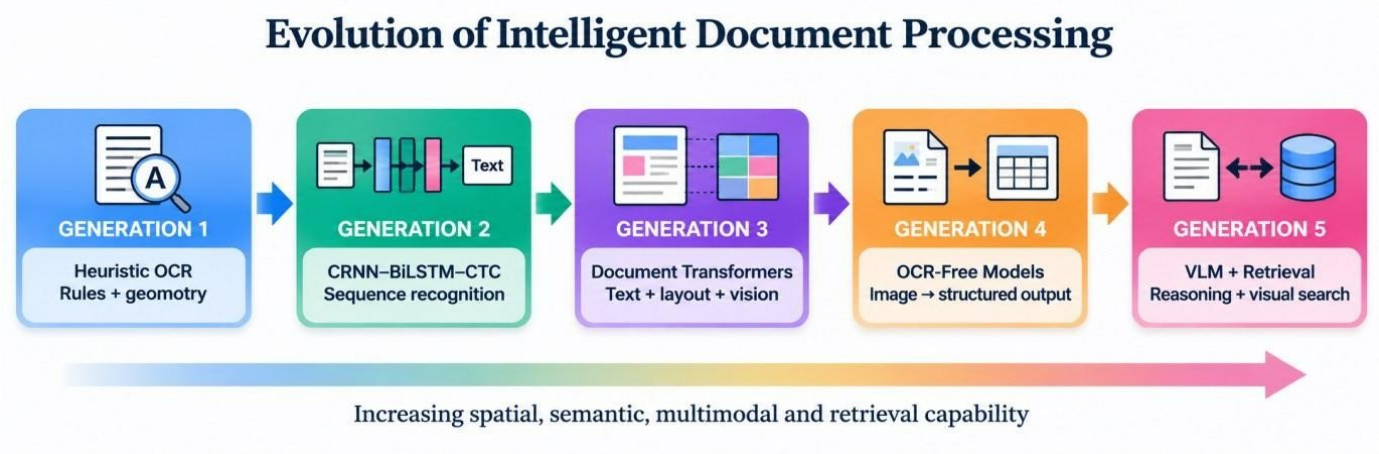


Figure 1. Architectural evolution of document intelligence.

contextual evidence. Duplicate records, sources without traceable provenance, unsupported numerical claims, and material unrelated to document processing were excluded from the synthesis. Selection was relevance-based and narrative, not a PRISMA screening exercise. The absence of a complete record-level screening log is acknowledged as a limitation.

3.1 Evidence Classification

Evidence sources are classified by provenance and evidential strength, as summarized in Table 1.

3.2 Analytical Dimensions

Architectures are compared across input representation, recognition dependence, spatial modeling, output type, task suitability, computational characteristics, and failure risks. Evidence is organized by capability rather than by a single model ranking. Reported scores are retained only when their task, dataset, metric, and source can be identified; scores from different datasets or protocols are not treated as directly comparable.

4 Architectural Evolution of Document Intelligence

Figure 1 illustrates the evolution of IDP from conventional OCR systems to modern multimodal and retrieval-augmented architectures. The progression shows how document-processing systems have moved from character recognition toward integrated visual, spatial, semantic, and reasoning capabilities.

4.1 Generation 1: Rule-Based Glyph Recognition and Heuristic Geometry

Rule-based systems locate text regions using mathematical morphology, contour processing, run-length smoothing, connected components, and horizontal or vertical projection profiles. Their advantages include interpretability, relatively low computational requirements, and the ability to encode known page structures. Their limitations include sensitivity to skew, illumination, bleed-through, irregular spacing, and layout variation. Rules designed for one document family may not transfer reliably to another.

4.2 Generation 2: CNN-BiLSTM-CTC Sequence Recognition

CRNN-style recognizers combine convolutional feature extraction with bidirectional long short-term memory (BiLSTM) sequence modeling and Connectionist Temporal Classification (CTC) [2, 20]. The CNN converts a text-line image into a feature sequence; recurrent layers model contextual dependencies; and CTC sums over valid alignments, avoiding the need to segment every character during training. These models are suited to continuous text-line recognition, including cursive writing, but generally depend on a preceding stage that identifies text lines and do not inherently reconstruct page-level reading order or key-value relations.

4.3 Generation 3: Visually Rich Document Transformers

LayoutLM-family models and DocFormer jointly represent text, page coordinates, and visual features. LayoutLM uses token and two-dimensional position embeddings with a pretrained language model; LayoutLMv2 adds stronger visual-text alignment; LayoutLMv3 uses unified text/image masking and a word-patch alignment objective. DocFormer instead emphasizes multimodal self-attention and shared spatial embeddings across text and visual modalities. These design choices are related but not identical. LayoutLM variants are particularly suited to token-level and form-field tasks when OCR tokens and bounding boxes are available; DocFormer provides a more tightly integrated multimodal representation. Both families can inherit errors from OCR and region detection, and their performance depends on the quality of the input text and coordinates.

4.4 Generation 4: End-to-End OCR-Free Document Understanding

Donut, Nougat, and Pix2Struct learn image-to-sequence or image-to-structured-output mappings without requiring a conventional external OCR stage at inference. Donut targets general visual document understanding and structured parsing; Nougat focuses on scientific-document conversion; Pix2Struct is pretrained through screenshot parsing and applies to a broader set of visual-language tasks. OCR-free models can reduce pipeline coupling and cascading OCR errors, but generated sequences may omit, normalize, or invent content when the visual evidence is ambiguous. Their suitability should be tested for the target language, document type, and output schema.

4.5 Generation 5: Multimodal Foundation Models and Visual Retrieval

Vision-language foundation models (VLMs) accept document images and support flexible question answering, summarization, and extraction through prompts or task-specific adaptation. Their breadth is useful when document requirements are not known in advance, but output reliability, reproducibility, latency, privacy, and inference cost require explicit evaluation. ColPali serves a different role: it creates page-level visual embeddings for retrieval and uses late interaction to rank pages against a query. It should be evaluated with retrieval metrics such as Recall@K, mean reciprocal rank (MRR), or normalized discounted cumulative gain (nDCG), not with extraction F1 or question-answering average normalized Levenshtein similarity (ANLS). A retrieval model can supply relevant pages to a downstream reasoning model, but retrieval and reasoning remain separate components.

4.6 Capability-Oriented Taxonomy

Document understanding encompasses distinct tasks that should not be conflated. Recognition converts visible characters or text lines into transcription and is evaluated with CER/WER. Information extraction identifies predefined entities, key-value pairs, tables, or structured fields and is evaluated with field-level precision, recall, F1, normalized edit distance, or exact match. Document question answering interprets document evidence to answer a query; evaluation should consider answer correctness, evidence support, source localization, and consistency, rather than relying on OCR accuracy alone. Document retrieval ranks documents or page regions by relevance and is evaluated with Recall@K, MRR, and nDCG. These tasks are complementary: a workflow may use recognition to produce text, extraction to populate fields, retrieval to locate supporting pages, and question answering to synthesize evidence. Each component should be assessed separately, and the complete workflow should also be tested for error propagation.

4.7 Comparative Architectural Synthesis

Architectural selection involves trade-offs rather than a universal ranking. Rule-based OCR and conventional OCR engines have low computational demands and can be economical at high throughput, but depend on image quality, language support, and document-specific rules. CNN-BiLSTM-CTC recognizers offer learned line-level transcription and

Table 2. Critical comparison of IDP architectural families.

Approach	Strengths / suitable tasks	Limitations / dependencies	Compute and scalability	Operational trade-off
Rule-based / conventional OCR	Low-cost transcription for clean, stable print; interpretable rules	Sensitive to skew, noise, fonts, layout variation; limited semantics	Low-to-moderate compute; scales efficiently with CPU batch processing	Best for stable templates and simple text; requires rules and exception handling
CNN-BiLSTM-CTC	Learned line recognition; useful for continuous text and handwriting	Usually needs line segmentation; limited page-level structure	Moderate training/inference; throughput depends on line count and image size	Useful where transcription dominates and labeled line data exist
Layout-aware transformers	Combines OCR tokens, coordinates and visual context for field extraction	OCR/box errors propagate; token limits and domain adaptation matter	Higher memory/compute than classical OCR; batching and GPU often useful	Strong for forms and semi-structured documents; integration includes OCR and schema mapping
OCR-free encoder-decoder	Direct image-to-text/ structured output; removes separate OCR dependency	Autoregressive latency; output may omit or hallucinate; validation needed	Moderate-to-high GPU demand; decoding cost scales with output length	Simplifies pipeline but requires constrained decoding, schema checks and fallback review
Vision-language foundation models	Flexible document QA, summarization and zero/few-shot extraction	Version drift, prompt sensitivity, privacy, hallucination and cost concerns	Potentially high inference cost/latency; rate limits and image resolution constrain throughput	Good for variable tasks; use strict grounding, output validation and human review
Visual retrieval (e.g., ColPali)	Retrieves visually relevant pages without exhaustive OCR transcription	Retrieval is not extraction or reasoning; index size and updates matter	Multi-vector indexing/storage and late-interaction compute can be substantial	Useful for vision retrieval-augmented generation (Vision RAG); evaluate retrieval separately and retain page-level evidence

moderate inference requirements, yet normally need text-line segmentation and do not model document semantics. Layout-aware transformers (LayoutLM variants and DocFormer) combine text, geometry, and visual features for structured extraction; they require OCR tokens and bounding boxes in common configurations, and their memory and latency rise with token count and image resolution. OCR-free encoder-decoder models (Donut, Nougat, Pix2Struct) avoid a separately deployed OCR engine and can simplify integration, but autoregressive decoding can be slower and generated outputs require schema and evidence validation. General vision-language models can address diverse question-answering and extraction prompts with little task-specific engineering, but inference cost, latency, context/image limits, version drift, and hallucination risk must be controlled. Visual retrieval models such as ColPali optimize page discovery, not field extraction or answer generation; their multi-vector indexes can increase storage and retrieval computation, although they can avoid complex OCR-based indexing. Scalability therefore depends on workload volume, page resolution, batching, model size, index size, hardware utilization, concurrency, and human-review capacity. A fair comparison requires a matched corpus and protocol, not scores collected from unrelated benchmarks. Table 2 summarizes these trade-offs.

5 Mathematical Foundations

5.1 Connectionist Temporal Classification

Let $X = (x_1, x_2, \ldots, x_T)$ denote the feature sequence of length T extracted from a text-line image, and let the label alphabet be extended with the blank symbol ε. Following [20], the probability of a target sequence l is obtained by summing the probabilities of all valid alignment paths π that collapse to l, as given in Eq. (1):

$$P(l \mid X) = \sum_{\pi \in \mathcal{B}^{-1}(l)} \prod_{t=1}^{T} P(\pi_t \mid X), \tag{1}$$

where $\mathcal{B}$ is the CTC collapse operation, which first merges consecutive repeated labels and then removes blank symbols, and $\mathcal{B}^{-1}(l)$ denotes the set of all alignment paths that collapse to the target sequence l. A dataset-level objective can be expressed as Eq. (2):

$$\mathcal{L}_{\mathrm{CTC}} = -\sum_{i=1}^{N} \log P(l_i \mid X_i), \tag{2}$$

where N denotes the number of training examples, X_i represents the input feature sequence for example i, and l_i denotes its corresponding target transcription.

5.2 Spatially Aware Attention in Document Transformers

For a document token i with hidden state $\mathbf{h}_i$ and normalized bounding-box coordinates $[x_{i0}, y_{i0}, x_{i1}, y_{i1}]$, attention between tokens i and j can incorporate conventional token similarity together with relative spatial information. A generalized spatially biased attention logit for head m is:

$$\begin{aligned} \alpha_{ij}^{(m)} = {} & \frac{(\mathbf{h}_i W_Q^{(m)})(\mathbf{h}_j W_K^{(m)})^{\top}}{\sqrt{d_k}} + b_{1\mathrm{D}}^{(m)}(j-i) \\ & + b_x^{(m)}(\Delta x_{ij}) + b_y^{(m)}(\Delta y_{ij}), \end{aligned} \tag{3}$$

where d_k is the key dimension, Δx_{ij} and Δy_{ij} are the relative horizontal and vertical offsets between the two boxes, and $b(\cdot)$ are learnable (typically bucketed) bias functions. Equation (3) is used here to illustrate the role of spatial bias; it should not be interpreted as the exact attention equation of every LayoutLM-family or DocFormer implementation.

5.3 Autoregressive OCR-Free Modeling

For an image I, an encoder produces a visual representation Z_V. An autoregressive decoder models the output sequence $Y = (y_1, \ldots, y_M)$ of length M as in Eq. (4):

$$P(Y \mid I) = \prod_{k=1}^{M} P(y_k \mid y_{<k}, Z_V), \tag{4}$$

where $y_{<k}$ denotes the tokens generated before position k. The formulation is attractive because the explicit OCR stage can be removed. Its limitation is that generation is probabilistic and may produce outputs that are not fully supported by the visual evidence.

5.4 Multi-Vector Late-Interaction Retrieval

For a query q represented by vectors $E_q = \{\mathbf{e}_{q,1}, \ldots, \mathbf{e}_{q,|q|}\}$ and a document page d represented by visual vectors $E_d = \{\mathbf{e}_{d,1}, \ldots, \mathbf{e}_{d,|d|}\}$, where $|q|$ and $|d|$ are the numbers of query and page vectors, a MaxSim-style late-interaction score [14, 21] can be expressed as:

$$S(q,d) = \sum_{i=1}^{|q|} \max_{j=1,\ldots,|d|} \cos(\mathbf{e}_{q,i}, \mathbf{e}_{d,j}). \tag{5}$$

Equation (5) supports semantic retrieval directly over page-level visual representations and can preserve information contained in charts, tables, diagrams, and typography that may be lost during text indexing.

6 Evidence and Evaluation

A central methodological problem in document-AI comparison is that published results are measured on different datasets, splits, annotation schemas, model sizes, input pipelines, and evaluation protocols. A field-extraction F1 score on FUNSD (Form Understanding in Noisy Scanned Documents), a DocVQA ANLS score, and a ColPali retrieval Recall@K value measure different tasks. They cannot be placed in one ranking. For this reason, this review does not tabulate cross-study scores whose experimental settings cannot be traced consistently to primary papers. Tables 3 and 4 instead record the task families, metrics, and reporting details required for a defensible comparison, without substituting unverified scores with new estimates.

6.1 Task-Specific Evidence Matrix

The task-oriented evaluation framework is summarized in Table 3, while Table 4 presents the primary sources and protocol requirements for reporting model performance.

Table 3. Task-oriented evaluation framework for intelligent document processing architectures.

Task family	Representative approaches	Typical metrics	Interpretation
Text recognition	Tesseract; CRNN; OCR engines	CER, WER	Measures transcription accuracy; sensitive to image quality, script, font, and preprocessing.
Form information extraction	LayoutLM family; BROS; DocFormer; related layout-aware models	Precision, Recall, F1	Measures extraction of entities or key-value information from structured/semi-structured pages.
Receipt/invoice extraction	Layout-aware transformers; multimodal parsers	Field-level F1 / exact match	Useful for business fields such as totals, dates, vendors, and tax information.
Document question answering	Donut; Pix2Struct; VLMs	ANLS, EM, accuracy	Measures answer quality conditioned on a document image and question.
Visual document retrieval	ColPali and related late-interaction models	Recall@K, MRR, nDCG	Measures retrieval rather than extraction or transcription.
Operational performance	All architecture families	Latency, throughput, cost/page, review rate	Must be measured under a defined hardware, batch, resolution, and deployment configuration.

Table 4. Primary sources and protocol requirements for model performance reporting.

Model family / example	Capability and primary source	Minimum reporting details
Layout-aware transformers: LayoutLMv3; DocFormer	Form/key-value extraction; original LayoutLMv3 and DocFormer papers [5, 7]	Dataset/split; model size; OCR source; token and box inputs; metric; fine-tuning protocol
OCR-free parsing: Donut; Nougat; Pix2Struct	Image-to-structured-output or image-to-text; original model papers [8–10]	Dataset/split; output schema; decoding; checkpoint; normalization; task metric
Vision-language models: Qwen2-VL; GPT-4o	Document QA, visual extraction, and reasoning; Qwen2-VL model report [12]; GPT-4o system card [13] (general capabilities and safety, so document results require a reproducible benchmark)	Exact version/date; prompt; resolution; tools/OCR access; shot setting; split; metric
Visual retrieval: ColPali	Page-level visual retrieval; ColPali paper and ViDoRe benchmark [14]	Index corpus; query set; Recall@K/MRR/nDCG; page granularity; embedding/reranking setup
Low-resource handwritten recognition (reviews)	Memon et al. [18]; Balabayev et al. [19]	Script/language coverage; dataset availability and size; writer-independent split; metric (CER/WER); leakage controls

6.2 Evaluation Measures

CER and WER measure recognition errors; exact-match rate assesses whether the normalized extracted fields are completely correct; and expected calibration error (ECE), introduced for neural-network classifiers [22], evaluates whether confidence agrees with empirical correctness and is used here to assess the confidence scores that drive review routing in IDP.

$$\mathrm{CER} = \frac{S_c + D_c + I_c}{N_c}, \tag{6}$$

where S_c, D_c, and I_c denote the numbers of character substitutions, deletions, and insertions, respectively, and N_c is the total number of characters in the reference transcription.

$$\mathrm{WER} = \frac{S_w + D_w + I_w}{N_w}, \tag{7}$$

where S_w, D_w, and I_w denote the numbers of word substitutions, deletions, and insertions, respectively, and N_w is the total number of words in the reference transcription. For K evaluated fields, the exact-match rate is:

$$\mathrm{EMR} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}[\mathrm{Norm}(\hat{y}_k) = \mathrm{Norm}(y_k)], \tag{8}$$

where K is the number of evaluated fields; $\hat{y}_k$ is the predicted value for field k; y_k is the corresponding reference value; $\mathrm{Norm}(\cdot)$ denotes the predefined normalization function; and $\mathbb{1}[\cdot]$ is an indicator function that equals 1 when the normalized values match and 0 otherwise. ECE estimates the difference

between confidence and empirical accuracy:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{B}_b|}{n} \left|\mathrm{acc}(\mathcal{B}_b) - \mathrm{conf}(\mathcal{B}_b)\right|, \qquad (9)$$

where B is the number of confidence bins; $\mathcal{B}_b$ is the set of predictions assigned to bin b; n is the total number of predictions; $\mathrm{acc}(\mathcal{B}_b)$ is the empirical accuracy in bin b; and $\mathrm{conf}(\mathcal{B}_b)$ is the mean predicted confidence in that bin.

7 Industrial Applications and Case Evidence

IDP is used in settings where large volumes of heterogeneous documents must be converted into structured information and then validated or interpreted. The examples below distinguish documented workflow use from independently measured model performance. Reported customer outcomes are not treated as universal benchmarks.

7.1 Healthcare Administration

Healthcare organizations process claims, medical forms, invoices, prescriptions, and supporting records. An IDP workflow may combine image-quality checks, OCR or handwriting recognition, layout analysis, domain-specific terminology, structured validation, and human review. Figures reported for such deployments, such as reductions in per-document processing time, extraction-error rates, or monthly staff hours, are often published without enough methodological detail to verify the document population, baseline, error definition, or independent evaluation. Such figures are therefore not reported in this review. Healthcare deployments should be assessed using field-level accuracy, clinically relevant error severity, review rate, provenance, and privacy controls; OCR accuracy alone is not an adequate measure of safety or operational value.

7.2 Financial Underwriting

Financial institutions process loan applications, income statements, bank statements, tax records, mortgage documents, and identity evidence. In addition to OCR and layout analysis, financial IDP may require field validation, cross-document consistency checks, anomaly detection, and audit trails. An Ocrolus customer story [23] describes a deployment at Neighborhood Loans, a US mortgage lender. This is a vendor-published customer case and can illustrate a reported deployment context, but its processing-time claims should not be generalized or presented as independent benchmark results. A reproducible evaluation should state the document population, fields evaluated, exact-match or field-F1 protocol, false-positive and false-negative rates, manual-review frequency, and baseline workflow.

Customer-reported performance depends on the document mix, workflow design, baseline process, integration, and criteria used to define success. Accordingly, this review retains the example only as contextual industry evidence and does not attach an unsupported percentage improvement to it.

Errors involving income, loan amounts, account numbers, tax liabilities, or repayment obligations can have substantial consequences. Confidence thresholds should therefore be field-specific, and low-confidence or high-impact outputs should be routed to qualified human reviewers. Performance reports should distinguish model errors from errors introduced by source quality, data matching, or downstream business rules.

7.3 Legal and Contract Analysis

Law firms, corporate legal departments, and financial institutions handle contracts, case files, correspondence, regulatory documents, and supporting evidence. In this domain, useful capabilities include clause identification, obligation extraction, entity linking, semantic retrieval, and reasoning across long documents. The economic value may arise primarily from reduced search and review effort rather than improved transcription accuracy. Retrieval results should point to exact pages or regions so that lawyers can inspect the underlying evidence.

As a banking example of contract review, JPMorgan Chase's 2016 annual report [24] describes its Contract Intelligence (COiN) platform as using unsupervised machine learning to analyze legal documents and extract data points and clauses. The report states that an initial implementation could extract 150 attributes from 12,000 annual commercial credit agreements in seconds, compared with as many as 360,000 hours per year under manual review. This is a company-reported operational estimate in a corporate annual report, not a peer-reviewed or independently reproduced benchmark; the report does not provide a controlled evaluation protocol, uncertainty interval, or independent audit of the time estimate. The figure is therefore reported only as company-attributed context, not as a measured effect size applicable to other organizations. Legal IDP evaluations should additionally measure clause-level recall, source

localization, evidence traceability, false-positive rates, and human-review requirements.

8 Economic Analysis and Deployment

8.1 Total Cost of Ownership

A realistic comparison between conventional OCR and modern IDP should include infrastructure, integration, inference, human review, error correction, and the expected cost of undetected errors. A generalized lifecycle model is:

$$\begin{aligned} \mathrm{TCO}_{\mathrm{annual}} = C_{\mathrm{cap}} + C_{\mathrm{int}} + C_{\mathrm{maint}} + C_{\mathrm{monitor}} + C_{\mathrm{storage}} \\ + C_{\mathrm{network}} + C_{\mathrm{security}} + \sum_{p=1}^{N} \Big[C_{\mathrm{inf}}(p) \\ + P_{\mathrm{rev}}(p)\, C_{\mathrm{human}} + P_{\mathrm{err}}(p)\, C_{\mathrm{pen}} \Big]. \end{aligned} \tag{10}$$

Here, N is the number of pages processed in the accounting period; C_{cap} is the annualized cost of compute hardware or fixed infrastructure; C_{int} is one-time integration and implementation cost amortized over the evaluation period; $C_{\mathrm{inf}}(p)$ is inference cost for page p, including API charges or allocated GPU/CPU time; $P_{\mathrm{rev}}(p)$ is the probability that page p is routed to human review; C_{human} is the fully loaded labor cost per reviewed page; $P_{\mathrm{err}}(p)$ is the probability of an undetected error; and C_{pen} is the expected financial, operational, regulatory, or service impact of that error. For a complete deployment comparison, annual maintenance, model monitoring, storage, networking, security controls, backup, vendor support, and retraining costs must also be included as fixed or variable terms. The model should be evaluated over the same annual page volume, document mix, service-level target, and error tolerance for each candidate solution.

For practical use, the organization should calculate each fixed term over a common accounting horizon (for example, one year) and amortize one-time implementation costs over the expected useful life of the solution. Variable terms should be estimated from a representative pilot and measured at the expected operating volume. The comparison should include at least three workload scenarios (low, expected, and peak volume) because API pricing, GPU utilization, queueing latency, and staffing needs may change nonlinearly with volume. Report both total annual cost and cost per successfully processed page; a page with an undetected critical error should not be counted as successfully processed:

$$C_{\mathrm{page}} = \frac{\mathrm{TCO}_{\mathrm{annual}}}{N_s}, \tag{11}$$

where $N_s \leq N$ is the number of pages processed without an undetected critical error.

8.1.1 *Worked Illustrative Cost Example*

Consider an organization processing 100,000 pages annually. Assume, solely for illustration, that manual processing takes 4 minutes per page and loaded labor costs INR 24 per hour. The manual baseline is $100{,}000 \times 4/60 \times 24 =$ INR 160,000 per year. Now assume an IDP service costs INR 0.03 per page, 20% of pages require review, and each reviewed page takes 1.5 minutes. Assume further that annualized integration and fixed operating costs are INR 10,000. Inference cost

Table 5. Comparative characteristics of enterprise IDP deployment models.

Deployment	Advantages	Limitations	Best fit
Managed cloud API	Fast integration; elastic scaling; managed infrastructure	Variable recurring cost; data residency and external processing considerations; vendor dependency	High-volume workloads where rapid deployment is more important than infrastructure control.
Hybrid / private gateway	Balances scalability and sensitive-data controls	More integration and networking complexity	Enterprises with regulated or sensitive document classes.
Self-hosted	Data control; predictable model serving; customization	GPU CapEx; maintenance; model lifecycle management	Stable high-volume workloads and organizations with strong ML infrastructure.
Edge / batch	Low connectivity dependence; privacy; potentially low marginal inference cost	Resource constraints; model optimization required	Scanners, field operations, archives, and disconnected environments.

is $100{,}000 \times 0.03 =$ INR 3,000; review labor is $20{,}000 \times 1.5/60 \times 24 =$ INR 12,000; fixed cost is INR 10,000. The illustrative annual IDP cost is therefore INR 25,000, or INR 0.25 per page ($N_s = N$, since error penalties are not modeled), compared with INR 1.60 per page for the manual baseline. The illustrative difference is INR 135,000 per year before error penalties, taxes, financing, and other unmodeled costs. Under these assumptions, 80% of pages are processed without human review; this is a scenario automation rate, not a published or guaranteed target. Actual values must be replaced with local wages, measured review time, supplier quotations, integration costs, and observed error rates. This example is an analytical scenario, not empirical evidence or a vendor price quotation.

8.2 Cloud, Hybrid, and Self-Hosted Deployment

The principal enterprise deployment options are compared in Table 5.

8.3 Pricing and Break-Even Analysis

Commercial prices vary by provider, processor, region, document type, volume, service tier, custom-model requirements, training, storage, and retention. Any price cited in a paper should specify the provider, exact processor or API, unit, currency, region, tier, date accessed, and whether discounts or volume commitments apply. The current review does not quote static vendor prices because they change frequently. The cost comparison should additionally account for integration, exception handling, human review, maintenance, security, monitoring, retraining, and the consequences of undetected errors.

To compare alternatives, calculate annualized total cost for each candidate using the same page volume, document mix, service-level target, security requirements, and acceptable error rate. A managed API typically combines per-page charges with integration, storage, and review labor. A self-hosted option adds annualized GPU/CPU hardware, utilization, electricity, cooling, operations staffing, patching, monitoring, and model lifecycle costs, while potentially lowering marginal inference cost at high utilization. A hybrid option may add gateway, networking, and duplicated operations costs but reduce data-transfer or residency risk. Break-even volume is the page volume at which the annualized costs of two alternatives are equal. It should be computed from local cost inputs; a universal threshold is not defensible.

The appropriate economic choice depends on document volume, document complexity, required service levels, review capacity, sensitivity of the data, and the consequences of incorrect extraction.

9 Failure Modes, Governance, Privacy, and Human Oversight

Deploying IDP in enterprise settings introduces risks that extend beyond document recognition. Errors may stem from visual degradation, misinterpretation of spatial information, domain shift, generative behavior, lack of confidence calibration, and inappropriate incorporation into downstream systems. Reliable IDP therefore requires a combination of model-level evaluation, governance measures, security, and suitably designed human oversight.

9.1 Principal Failure Modes

Several failure modes are particularly relevant to document-processing systems.

- **Autoregressive hallucination:** A generative model produces content that is not supported by the visual evidence. This is especially critical for monetary values, dates, identifiers, medical data, and legal clauses; in high-risk applications, generated content should therefore be grounded in identifiable source passages.
- **Reading-order inversion:** Page content is serialized into a one-dimensional sequence in the wrong order. This is common in multi-column newspapers, forms, magazines, and complex reports, where a simple top-to-bottom sequence does not match the intended reading order.
- **Spatial-association and bounding-box errors:** A recognized value may be associated with the wrong label or bounding box because of perspective distortion, folds, shadows, camera capture, or inaccurate region detection. The association can be wrong even when the text itself is recognized correctly.
- **Loss of table structure:** Individual cell contents may be recognized correctly while relationships among rows, columns, merged cells, or headers and data are reconstructed incorrectly. This can yield semantically incorrect structured data despite apparently high OCR accuracy.
- **Script and domain drift:** Scripts, fonts, terminology, layouts, or templates that differ substantially from the training distribution may reduce recognition or extraction quality. Such

distribution shift is especially common for historical documents and low-resource scripts.

- **Confidence miscalibration:** A model can be highly confident in an incorrect prediction. Raw model confidence should therefore not be taken as a faithful measure of correctness, and calibration should be assessed before confidence thresholds are used to automate business decisions.
- **Semantic extraction errors:** A model may recognize the words in a document correctly but fail to determine their semantic function. For example, a numerical value may be assigned to the wrong field, or a contract clause may be interpreted without its surrounding context.
- **Error propagation:** Errors in image preprocessing, OCR, layout detection, or reading-order reconstruction can carry over into later extraction and reasoning stages. A representation error may therefore result in a plausible but incorrect answer from a downstream language model.

9.2 Governance Requirements

The governance needs of an IDP system are determined by its purpose, scope, data type, and effects on people and organizations. The governance structure should encompass the entire system lifecycle, including data acquisition, model development, deployment, monitoring, change, and retirement. Higher-risk applications should have governance that includes:

- **Documented risk assessment:** Identify potential technical, operational, legal, privacy, security, and business risks prior to deployment.
- **Data governance:** Set guidelines for data provenance, quality, access, retention, minimization, and use.
- **Model validation:** Validate the system on representative documents, including difficult and failure-prone documents, not only on average benchmark performance.
- **Auditability:** Record model versions, processing settings, document identifiers, confidence scores, human corrections, and key decisions made by the system.
- **Human oversight:** Specify the conditions under which automated outputs must be inspected, corrected, or rejected by competent staff.
- **Post-deployment monitoring:** Continuously monitor changes in document distribution, model performance, error rates, confidence calibration, and review requirements.
- **Change management:** Revalidate the system after significant changes to models, OCR engines, document templates, prompts, preprocessing procedures, or downstream integrations.

Governance should therefore be treated as an ongoing activity rather than a one-off compliance exercise.

9.3 Privacy and Security

Enterprise documents can contain personally identifiable information, financial and health information, confidential contracts, and other sensitive content. Privacy and security controls should therefore be part of the architecture from the outset rather than added afterwards. Key controls include:

- **Purpose limitation:** Transmit and retain only the information required for the intended purpose, and minimize sensitive content before external inference where technically and legally appropriate.
- **Encryption:** Encrypt documents, intermediate representations, embeddings, extracted fields, and model outputs in transit and at rest.
- **Role-based access control:** Restrict access to source documents, extracted information, model configurations, and audit logs to authorized roles.
- **Retention management:** Define retention periods and secure-deletion procedures for source images, OCR output, embeddings, prompts, model responses, and review records.
- **Audit logging:** Log document identifiers, processing timestamps, model and software versions, confidence data, review decisions, and material corrections without storing sensitive document content in the logs.
- **Deployment isolation:** Private-cloud or on-premises inference may be preferable for highly sensitive workloads when external APIs raise data-governance, confidentiality, or residency concerns.
- **Third-party risk management:** Before sending sensitive documents to an external IDP or VLM service, assess the provider's security measures, data-processing and retention terms, service dependencies, and contractual obligations.

Table 6. Architecture and deployment recommendations for representative enterprise document archetypes.

Document archetype	Structural complexity	Primary bottleneck	Recommended stack	Straight-through processing (STP) target	Deployment
Standardized forms	Low variance, fixed grids	Checkboxes, low-resolution scans	OCR + fine-tuned layout model	Set by pilot	On-premises / VPC
Invoices and receipts	Variable tables, dynamic geometry	Line items and tax calculations	Multimodal transformer / cloud API	Set by pilot	Hybrid cloud gateway
Legal contracts	Dense, multipage, unstructured	Semantic interpretation and clause risk	Visual retrieval + foundation VLM	Set by pilot	Private cloud / VPC
Degraded historical archives	Severe noise and script variation	Fragmentation and script drift	Image restoration + OCR-free encoder or specialized OCR	Set by pilot	Batch processing cluster

Privacy and security protections should be applied at every stage of the IDP pipeline, from document capture, preprocessing, and inference to storage and retrieval, human review, and integration with downstream systems.

9.4 Calibrated Human-in-the-Loop Processing

In a practical IDP architecture, it cannot be assumed that every page can be processed without human intervention. Instead, an uncertainty threshold τ can be set to route uncertain predictions to human inspection. Thresholds should be set per field type because the impact of an error can vary; a wrong invoice total, for instance, is usually far more consequential than a wrong telephone number. Calibration measures such as ECE (Eq. (9)) should therefore be reported alongside accuracy and F1 to show whether the confidence estimates used for routing are meaningful.

10 Strategic Enterprise Implementation Blueprint

Deploying IDP in an enterprise is not straightforward and does not depend only on choosing the right recognition or multimodal model. An implemented system should tie together document acquisition, preprocessing, visual and textual understanding, information extraction, validation, human review, security measures, and downstream business applications. IDP should therefore be considered an end-to-end operational pipeline rather than a single machine-learning component. Figure 2 shows a typical enterprise IDP architecture covering the key steps from document capture to integration with enterprise systems. Representative document archetypes and their recommended deployment strategies are summarized in Table 6.

STP percentages are not prescribed in Table 6 because no cited deployment study or defined pilot protocol supports universal values. STP should be defined as the proportion of documents completed without manual intervention under a stated quality and risk threshold. The target must be established through a representative pilot, with separate reporting of document coverage, field accuracy, exception rate, human-review rate, and error severity. Table 6 therefore treats STP as a locally determined design objective rather than a universal benchmark.

10.1 Architecture Selection Decision Framework

The resulting architecture-selection framework is presented in Table 7.

11 Open Research Challenges and Future Directions

Although intelligent document processing has advanced considerably, several technical, methodological, and deployment problems remain open. Future research should go beyond recognition accuracy toward document reasoning, evidence-based retrieval, efficient deployment, multilingual support, and reproducible evaluation, as shown in Figure 3.

11.1 Long-Document Reasoning

Contracts, case files, technical reports, medical records, and regulatory submissions often run to hundreds of pages. Processing such documents with conventional self-attention can be computationally expensive because attention complexity typically grows quadratically with sequence length. Scalable IDP systems can address this with hierarchical

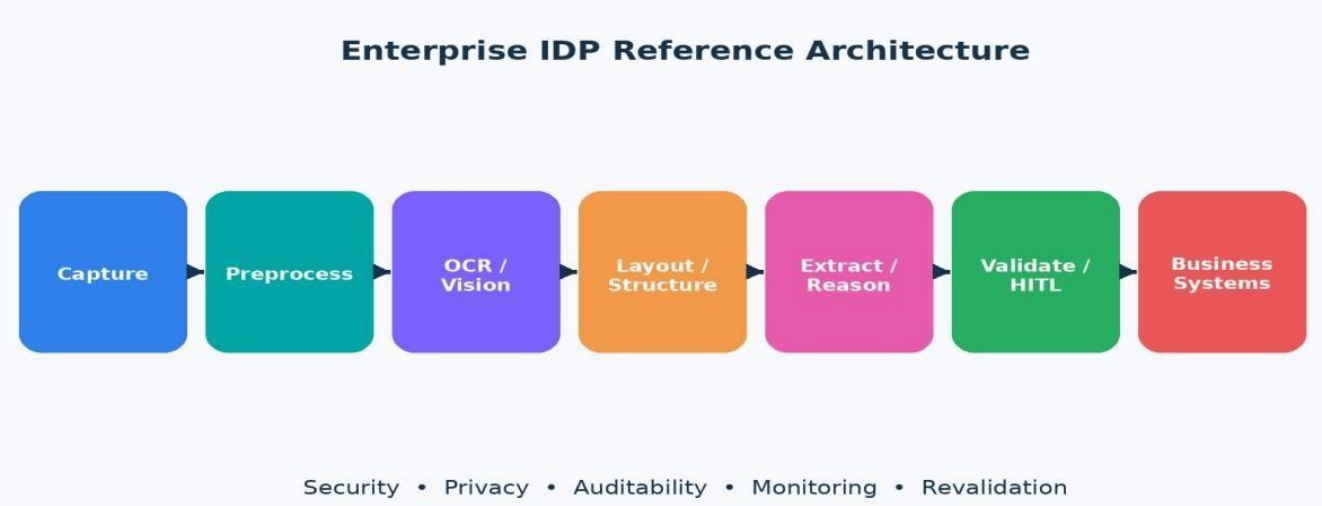


Figure 2. Enterprise implementation blueprint.

Table 7. Architecture selection framework for different enterprise document processing requirements.

Requirement	Preferred approach	Rationale
Clean, regular printed text	Conventional OCR	Efficient and predictable when document structure is simple.
High-volume text transcription	OCR + lightweight recognition	Optimizes throughput and operational cost.
Forms and semi-structured documents	Layout-aware transformer	Uses text and spatial relationships jointly.
Invoices and receipts	OCR + layout model or multimodal parser	Supports key-value and line-item extraction.
Complex visual documents	Multimodal VLM	Handles text, diagrams, tables, and broader visual context.
Long contracts and case files	Visual/text retrieval + VLM	Reduces context load and enables evidence-grounded reasoning.
Large document repositories	Hybrid text/visual retrieval	Balances searchable text with visual structure.
Sensitive documents	Private or hybrid deployment	Improves control over data movement and retention.
Low-resource scripts	Fine-tuned OCR + layout analysis	Allows domain- and script-specific adaptation.
High-risk extraction	Model + validation + human-in-the-loop review	Reduces the probability that a single model error becomes a business error.

representations, selective retrieval, memory compression, sparse or efficient attention mechanisms, and page- or section-level routing, which reduce computational overhead while preserving access to relevant document evidence.

Long-document systems should also be assessed beyond final-answer accuracy. Evidence localization, cross-page consistency, citation accuracy, missed relevant evidence, and preservation of entity and event relationships dispersed across a document should be considered. Such evaluation is crucial in legal, financial, and healthcare applications, where an answer may sound plausible yet lack supporting evidence.

11.2 Visual Retrieval and Vision RAG

Visual retrieval and Vision RAG remove the assumption that documents must be fully converted to text before search or reasoning. Visual representations can retain information that is lost in text extraction, such as charts and tables, typography, spatial relationships, handwriting, stamps, signatures, and document structure, at the page or region level.

Several issues remain open. Large-scale visual indexing can impose high storage and computation demands, and updated documents can be costly to

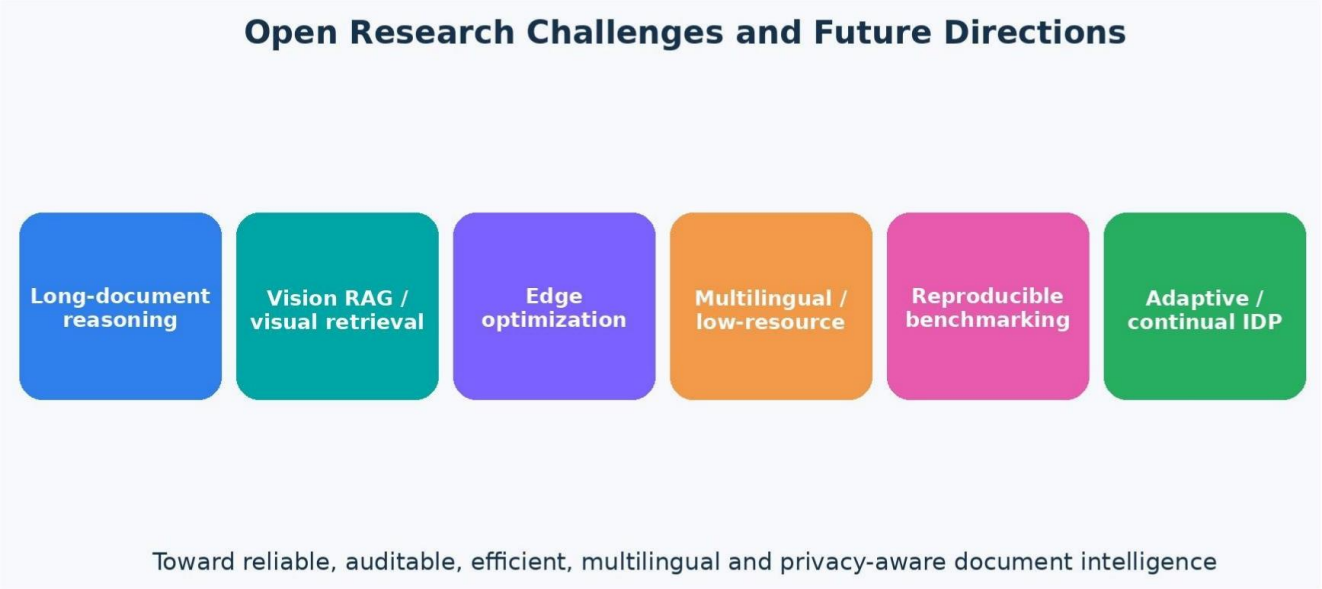


Figure 3. Research challenges and future directions.

re-index. Retrieval quality also depends on page complexity, visual similarity, query formulation, and document type. Future research should explore efficient indexing strategies, incremental updates, privacy-preserving visual retrieval, multimodal relevance metrics, and evidence-grounded generation that allows responses to be traced back to specific parts of a document.

11.3 Edge Optimization

Model size and computational load must be reduced substantially to deploy multimodal document models on scanners, mobile devices, embedded systems, or other edge platforms. Promising approaches include quantization, knowledge distillation, pruning, parameter-efficient adaptation, memory-efficient inference, and hardware-aware model design.

The main research challenge is to achieve low latency, low memory usage, and low energy consumption without unacceptable degradation in recognition or extraction quality. Evaluation should therefore consider character-level and field-level accuracy, confidence calibration, inference latency, memory usage, energy consumption, and model size together.

11.4 Multilingual and Low-Resource Scripts

Low-resource and historical scripts and region-specific document types remain underrepresented in commercial document-intelligence datasets and benchmarks. This seriously limits systems designed for multilingual public-sector records, cultural archives, newspapers, education, and historical collections.

Future work should create representative datasets with both textual and structural annotations, such as tables, figures, headings, and semantic fields, in addition to reading order and bounding boxes. Mixed-script documents, code switching, degraded print, historical typography, non-standard layouts, and culturally specific document structures also deserve attention. Evaluation protocols should report not only aggregate multilingual scores but also scores for each script, document condition, and layout type.

11.5 Reproducible Benchmarking

Future benchmarking studies should report the document collection, train/test partitions, preprocessing steps, hardware, software and model versions, inference settings, prompts, decoding method, evaluation metrics, calibration procedure, and cost assumptions. Benchmark reports should also include representative failure cases, confidence distributions, human-review rates, latency, throughput, and resource consumption. Where possible, recognition, extraction, retrieval, reasoning, and workflow-level performance should be reported as separate scores rather than reduced to a single number.

11.6 Adaptive and Continual Document Processing

Enterprise document collections change continually. Even if the underlying model is unchanged, new templates, suppliers, terminology, layouts, regulations, and scanning conditions can degrade performance.

Continual learning and adaptive document processing are therefore important research opportunities. Future systems should be able to detect distribution shifts, adapt to new document types, and update specialized components without full retraining. Such adaptation must be accompanied by controlled validation, version management, rollback features, and monitoring so that performance degradation does not go unnoticed.

12 Limitations of the Comparative Analysis

The principal limitation is evidence heterogeneity. Studies use different datasets, document populations, preprocessing, hardware, model versions, prompts, inference configurations, and evaluation protocols. Scores should not be compared directly unless the underlying conditions and metric definitions are sufficiently aligned. The review is narrative and does not claim exhaustive systematic coverage or a quantitative meta-analysis.

The evidence base also varies in methodological strength. Peer-reviewed benchmark papers provide experimental detail; technical documentation describes product capabilities; and customer case reports provide operational context but may omit dataset composition, baselines, error definitions, uncertainty estimates, or independent replication. This review labels these evidence types separately and omits numerical claims that could not be traced to a suitable source.

The scope is also time-sensitive because foundation models and commercial document services change rapidly. API versions, model checkpoints, context limits, pricing, and inference capabilities may change after publication. Any model-specific result should therefore identify the exact version and access date. Organizational factors, including data governance, regulation, staff capability, integration complexity, procurement constraints, and change management, are important but difficult to compare across studies.

The review focuses on architecture, evaluation, and operational considerations. It does not conduct a controlled head-to-head experiment. Its conclusions are therefore a synthesis of architectural patterns and evidence limitations, not a definitive ranking of models or commercial platforms.

13 Conclusion

Intelligent document processing has evolved from conventional OCR into a family of systems that combine visual perception, spatial representation, language modeling, retrieval, and multimodal reasoning. Conventional OCR remains useful for clean printed text; sequence recognizers support continuous text transcription; layout-aware transformers combine text and spatial information for structured extraction; OCR-free models generate structured outputs directly from page images; vision-language models support flexible document question answering and interpretation; and visual retrieval models locate relevant pages using visual representations. These capabilities are complementary, and no single architecture is best for every document collection or business task.

The proposed taxonomy distinguishes recognition, extraction, reasoning, and retrieval because each operation has a different output and requires different evaluation measures. Recognition should be measured with CER/WER; extraction with field-level precision, recall, F1, or exact match; reasoning with answer correctness and evidence support; and retrieval with Recall@K, MRR, or nDCG. A robust IDP system should report component-level results alongside workflow-level measures such as latency, cost per page, review rate, straight-through-processing rate, and error impact.

The contribution of this review is an evidence-aware architectural synthesis that links model design to task requirements, source quality, operational constraints, and governance. Numerical comparisons are limited to those traceable to primary sources, and the methodology section identifies the search sources, search terms, coverage period, and inclusion criteria, while acknowledging that the work is a structured narrative review rather than a PRISMA systematic review. Industrial cases are treated as contextual reports, and the cost example is explicitly identified as an assumption-based scenario.

Future research should prioritize reproducible multilingual benchmarks, low-resource and historical scripts, writer-independent handwritten recognition datasets, long-document evidence localization, privacy-preserving visual retrieval, and efficient edge inference. The two cited reviews of handwritten and low-resource material illustrate the scarcity of datasets and evaluation protocols for such scripts, and the need to distinguish script-level recognition from full-page OCR and document understanding. Reliable IDP will require not only accurate recognition, but also grounded extraction, traceable reasoning, relevant retrieval, calibrated uncertainty, and appropriate

human oversight.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The author declares no conflicts of interest.

AI Use Statement

The author declares that ChatGPT-5 was used for language editing of the manuscript. The author has carefully reviewed, revised, and verified the AI-assisted output and takes full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

Not applicable. This review does not involve human participants or animals.

References

[1] Smith, R. (2007). An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 2, pp. 629-633). IEEE. [CrossRef]

[2] Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(11), 2298-2304. [CrossRef]

[3] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). 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). [CrossRef]

[4] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., & Zhou, L. (2021, August). LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 2579-2591). [CrossRef]

[5] Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 4083-4091). [CrossRef]

[6] Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., & Park, S. (2022, June). BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 10767-10775). [CrossRef]

[7] Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., & Manmatha, R. (2021). DocFormer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 993-1003). [CrossRef]

[8] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2022, October). OCR-free document understanding transformer. In *European Conference on Computer Vision* (pp. 498-517). Cham: Springer Nature Switzerland. [CrossRef]

[9] Blecher, L., Cucurull Preixens, G., Scialom, T., & Stojnic, R. (2024, May). Nougat: Neural optical understanding for academic documents. In *International Conference on Learning Representations*.

[10] Lee, K., Joshi, M., Turc, I. R., Hu, H., Liu, F., Eisenschlos, J. M., Khandelwal, U., Shaw, P., Chang, M.-W., & Toutanova, K. (2023, July). Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning* (pp. 18893-18912). PMLR. https://proceedings.mlr.press/v202/lee23g.html

[11] Mathew, M., Karatzas, D., & Jawahar, C. V. (2021). DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 2200-2209). [CrossRef]

[12] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., & Lin, J. (2024). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*. [CrossRef]

[13] OpenAI. (2024). GPT-4o system card. *arXiv preprint arXiv:2410.21276*. [CrossRef]

[14] Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., & Colombo, P. (2025, April). ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*.

[15] Google Cloud. (n.d.). Document AI documentation and pricing. Retrieved October 10, 2026, from https://cloud.google.com/document-ai/pricing

[16] Microsoft. (n.d.). Azure AI Document Intelligence pricing. Retrieved October 10, 2026, from https://azure.microsoft.com/pricing/details/ai-document-intelligence/

[17] Amazon Web Services. (n.d.). Amazon Textract

pricing. Retrieved October 10, 2026, from https://aws.amazon.com/textract/pricing/

[18] Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). *IEEE Access, 8*, 142642-142668. [CrossRef]

[19] Balabayev, Z., Biloshchytska, S., Abdikenov, B., Zhaksylyk, T., Ayanbayev, B., & Rakishev, D. (2026). Navigating the digitization gap: An indirect evidence synthesis of AI methods for low-resource Chagatai manuscripts. *Information, 17*(7), 681. [CrossRef]

[20] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning* (pp. 369-376). [CrossRef]

[21] Khattab, O., & Zaharia, M. (2020, July). ColBERT: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval* (pp. 39-48). [CrossRef]

[22] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70 (pp. 1321-1330). https://proceedings.mlr.press/v70/guo17a.html

[23] Ocrolus. (2025). Neighborhood Loans cuts processing time by 90% with Ocrolus automation [Customer story]. Retrieved October 10, 2026, from https://www.ocrolus.com/customer-stories/neighborhood-loans-ai-lending-processing/

[24] JPMorgan Chase & Co. (2017). *Annual Report 2016* (Letter to shareholders from M. Zames, Chief Operating Officer, pp. 47-51). https://www.jpmorganchase.com/content/dam/jpmc/jpmorgan-chase-and-co/investor-relations/documents/ar2016-lettertoshareholders.pdf



Atul Kumar is an Assistant Professor in the Department of Computer Science at R.G.M. Government College, Joginder Nagar, District Mandi, Himachal Pradesh, India. He received the Ph.D. degree in Computer Science from Punjabi University, Patiala, Punjab, India, and has over 13 years of teaching experience in higher education. He has authored 23 research papers published in national and international journals and conference proceedings. His research interests include natural language processing (NLP), educational technology, pattern recognition, OCR, and newspaper layout analysis, with a particular focus on computational approaches for document analysis, language processing, and technology-enabled educational applications. (Email: atulkmr02@gmail.com)