FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 19, October, 2026, pp 388 – 402
DOI: https://doi.org/10.33003/fjs-2026-1019-5746

Efficient and Interpretable Multimodal Document Understanding: A Survey of Layout Transformers, Vision–Language Models, Multi-Agent Frameworks, and Conditional Computation

***Goni Mahmud Mustapha, Nurudeen Mahmud Ibrahim, Aliyu Sulieman Muhammed and Austin Olom Ogar**

Department of Computer Science, Faculty of Computing, Nile University of Nigeria, Abuja-FCT, Nigeria.

*Corresponding authors' email: gonimahmudmustapha@gmail.com

Received: 17 Aug. 2026
Accepted: 14 Sep. 2026
Published: 22 Oct. 2026

ABSTRACT

Multimodal document understanding—the joint interpretation of textual content, visual appearance, and spatial layout in documents such as forms, receipts, contracts, scientific papers, and long reports—has advanced rapidly, moving from layout-aware transformers to large vision–language models (LVLMs) and, most recently, to retrieval-augmented and multi-agent frameworks. This progress, however, has been driven overwhelmingly by accuracy, and it has come at a steep and largely unexamined cost in computation, latency, and interpretability. Dense-by-default attention, exhaustive multi-agent pipelines, and opaque internal state have produced systems that are accurate on benchmarks yet difficult to deploy at scale and difficult to trust in regulated domains. This survey organises the field around that tension. We first establish a taxonomy that separates architectural paradigms (layout transformers, OCR-free generative models, LVLMs, and structural/graph reasoning) from efficiency and orchestration mechanisms (retrieval augmentation, multi-agent collaboration, conditional computation, and cost-aware routing). We then review each paradigm in depth, tracing how the field's persistent problems—architectural fragmentation across siloed encoders, weak explicit structural reasoning, uniform-density execution in multi-agent systems, and an interpretability deficit—have shaped successive generations of models. We give particular attention to an emerging synthesis in which conditional computation, long established for tokens and experts in mixture-of-experts models, is transferred to the orchestration of document agents, decoupling a system's capability from its per-query cost. We consolidate the benchmark and evaluation landscape, including the strict-versus-fuzzy metric problem on long-document question answering, and we survey interpretability and Green-AI considerations that are increasingly treated as first-class objectives rather than afterthoughts. We conclude with open challenges, arguing that the next decade of document intelligence will be defined less by raw accuracy than by the co-optimisation of accuracy, efficiency, and interpretability. Throughout, we cite recent surveys and primary work through 2025 and situate the field's trajectory for researchers and practitioners.

Keywords: Multimodal Document Understanding, Document AI, Layout Transformers, Vision–Language Models, Multi-Agent Systems, Conditional Computation, Dynamic Routing, Retrieval-Augmented Generation

INTRODUCTION

The digitisation of records across finance, healthcare, law, government, and science has made automated document understanding one of the most economically consequential applications of artificial intelligence. Unlike plain text, real-world documents are heterogeneous artefacts whose meaning emerges from the interaction of three modalities: the linguistic content of the text, the visual appearance of the page (font weight, colour, logos, stamps, handwriting), and the spatial layout that organises content into tables, columns, headers, and nested hierarchies. Interpreting a document therefore requires a system to reason jointly over what is written, how it looks, and where it sits on the page. Recent surveys of multimodal large language models and of visually rich document understanding document an explosive growth in capability across exactly these axes (Yin et al., 2024; Ding et al., 2025; Fu et al., 2025).

Yet the trajectory of the field has been shaped almost entirely by a single objective—benchmark accuracy—and the pursuit of that objective has produced a second, less-examined trend: a steady increase in computational cost, inference latency, and model opacity. Each new generation of document models has added parameters, modalities, and reasoning stages. Layout-aware transformers grew from hundreds of millions of parameters to multimodal encoders with quadratic attention cost on long inputs. Large vision–language models pushed accuracy higher still, at the price of dense activation across billions of parameters for every input. Most recently, multi-agent frameworks have achieved state-of-the-art results on long-document question answering by orchestrating several specialised models per query, multiplying cost accordingly. The environmental and financial footprint of this trend has become a first-order concern, echoing the broader Green-AI argument that the compute cost of a system should be reported and minimised as a matter of scientific practice, not treated as an externality (Schwartz et al., 2020).

This survey is organised around the resulting tension. We argue that multimodal document understanding is best understood not as a single ladder of increasing accuracy but as a search through a three-way trade-off between accuracy, efficiency, and interpretability—a trilemma in which historically, improving one axis has degraded the others.

Fine-tuning a general model for a specialist domain raises accuracy but erodes generality and inflates cost; dense execution maximises accuracy but makes real-time deployment infeasible; and the opaque internal state of high-accuracy models blocks the auditability that regulated domains require. The central intellectual development of the last two years, in our reading, is the beginning of a genuine escape from this trilemma through conditional computation and dynamic orchestration: mechanisms that activate only the parts of a system a given input actually requires.

The contributions of this survey are fourfold. First, we present a taxonomy that cleanly separates architectural paradigms from efficiency and orchestration mechanisms, allowing the two axes to be discussed independently and then recombined. Second, we provide an in-depth, comparative review of each paradigm—layout transformers, OCR-free generative models, vision–language models, retrieval-augmented and multi-agent systems, conditional computation and routing, and explicit structural reasoning—tracing how the field's four persistent gaps have driven its evolution. Third, we consolidate the benchmark, dataset, and evaluation landscape, including the strict-versus-fuzzy metric problem that distorts reported accuracy on long-document question answering. Fourth, we survey interpretability and efficiency/Green-AI as first-class objectives, review applications, offer a cross-paradigm synthesis, and close with open challenges and future directions.

Scope, Method, and Relation to Prior Surveys

This survey concentrates on multimodal document understanding—systems that jointly reason over text, vision, and layout to extract, classify, locate, or answer questions about documents—and on the efficiency and interpretability of those systems. We include layout-aware transformers, OCR-free generative models, vision–language models adapted to documents, retrieval-augmented and multi-agent pipelines, and the conditional-computation and routing methods that govern their cost. We deliberately foreground the efficiency and interpretability dimensions that broader surveys treat only in passing, because we regard them as the axes along which the field's next advances will be measured. We draw both on foundational works that defined each paradigm and on recent primary and survey literature through 2025, and we organise the material analytically—by paradigm and mechanism—rather than chronologically, so that trade-offs can be compared directly.

Several excellent recent surveys complement this one and clarify its distinctive contribution. General surveys of multimodal large language models chart the capabilities of the underlying backbones without a document focus (Yin et al., 2024). A survey of MLLM-based visually rich document understanding maps the document-specific model landscape and its emerging trends (Ding et al., 2025), and a comprehensive review of text-rich image understanding catalogues the models that read dense-text images (Fu et al., 2025). A survey of multimodal retrieval-augmented generation for document understanding consolidates the retrieval branch specifically (Abootorabi et al., 2025), and surveys of mixture-of-experts inference optimisation consolidate the conditional-computation branch (Liu et al., 2024). What distinguishes the present survey is its organising thesis: rather than cataloguing models by capability, we structure the entire field around the accuracy–efficiency–interpretability trilemma and around the convergence of conditional computation with multi-agent document orchestration, which we argue is the field's most consequential recent development and which no prior survey treats as its central axis.

Background and Preliminaries

The Multimodal Document Understanding Problem

Formally, a document can be described as a set of pages, each containing a sequence of textual tokens, each token associated with a two-dimensional bounding box that locates it on the page, together with the raw pixels of the page image. A document-understanding system maps this multimodal input, optionally conditioned on a natural-language query, to a structured or free-form output—an extracted field, a class label, a bounding region, or an answer. What distinguishes the problem from ordinary natural-language processing is that the correct output frequently depends on non-local, structurally mediated relationships: determining whether a date is an invoice date or a due date requires reasoning over its spatial proximity to a label, its alignment within a column, and its position relative to a section header. This is why purely textual models, however large, systematically underperform on visually rich documents.

The modern treatment of this problem rests on three architectural pillars inherited from the deep-learning revolution: the transformer's self-attention mechanism, which models dependencies among tokens (Vaswani et al., 2017); large-scale masked pre-training, which endows encoders with transferable representations (Devlin et al., 2019); and the vision transformer, which extends the same machinery to image patches (Dosovitskiy et al., 2020). The first document-specific synthesis of these ideas, LayoutLM, added two-dimensional positional embeddings to a BERT backbone so that token coordinates could inform representation, and established the template that most subsequent layout-aware models refined (Xu et al., 2020).

Core Tasks

The field is conventionally organised around a small number of canonical tasks, each with associated benchmarks. Key information extraction and entity linking, exemplified by the FUNSD form-understanding dataset, require a model to identify and connect semantic entities such as questions and answers on a form (Jaume et al., 2019). Receipt understanding, exemplified by SROIE, requires extraction of a fixed schema—company, date, address, total—from scanned receipts (Huang et al., 2019). Document visual question answering, standardised by the DocVQA benchmark, requires free-form answers grounded jointly in text and image (Matthew et al., 2021). Document image classification, standardised by RVL-CDIP, requires assigning a document to one of a set of categories (Harley et al., 2015). Document layout analysis and segmentation, exemplified by PubLayNet, require identifying the geometric regions—paragraphs, figures, tables—that compose a page (Zhong et al., 2019). These tasks differ in whether the dominant evidence is textual, visual, or structural, a distinction that becomes central when we discuss modality-conditional efficiency.

This evidence-type distinction is worth making precise, because it is the foundation on which modality-conditional efficiency rests. A question such as 'what is the total amount due?' is answered from text and its spatial context; a question such as 'what colour is the warning label?' is answered from the image; and a question such as 'does the chart on page 4 support the claim in the abstract?' is genuinely cross-modal, requiring both. Long-document benchmarks increasingly annotate the evidence type or source of each question—text, table, chart, image, or layout—precisely because the evidence

type predicts both the difficulty of a question and the machinery required to answer it. The key observation is that a large fraction of real document questions are single-modality, so a system that treats every question as fully cross-modal wastes computation on the majority that are not.

The Accuracy–Efficiency–Interpretability Trilemma

The organising thesis of this survey is that these tasks have been pursued under an implicit trilemma, illustrated in Figure 1. Improving accuracy has typically meant more parameters, denser attention, or more reasoning stages—each of which degrades efficiency. Improving efficiency through compression or aggressive sparsity has historically cost accuracy, or has failed to preserve domain-specific knowledge under adaptation. And the highest-accuracy models tend to be the least interpretable, because their decisions are distributed across billions of densely activated parameters with no exposed reasoning trace. Regulated domains—clinical documentation, legal discovery, financial compliance—cannot accept accuracy alone; they require predictions that are transparent, traceable, and auditable. Much of the most interesting recent work can be read as an attempt to bend this trilemma, most often by making computation conditional on the input rather than uniform across all inputs.

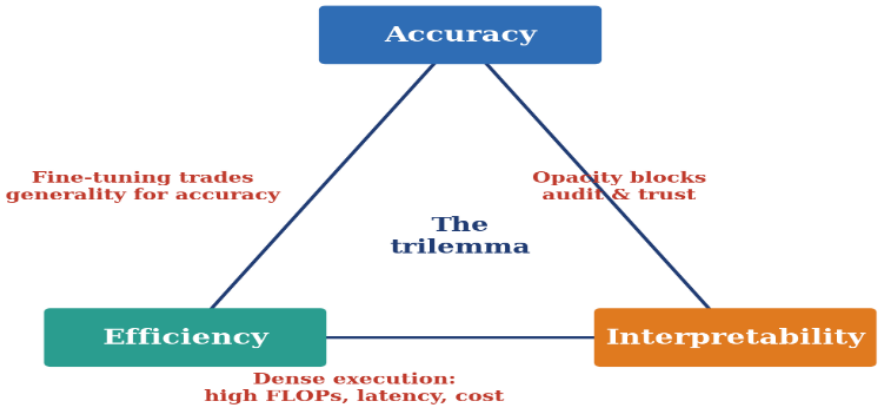


Figure 1: The accuracy–efficiency–interpretability trilemma. Historically, gains on one axis have degraded the others; conditional computation and explicit reasoning offer partial escapes

Modality Fusion Strategies

Because the document-understanding problem is intrinsically multimodal, how and when the modalities are combined is a defining design choice, and the field's history can be read as a progression through increasingly deep fusion. In late fusion, the earliest and simplest strategy, text, vision, and layout are encoded by independent networks and their outputs are concatenated or combined only at the final decision stage. Late fusion is modular and easy to engineer, but it severs the fine-grained cross-modal interactions on which document understanding depends: a late-fusion model cannot let the word 'signature' guide attention to a handwritten glyph, because by the time the two representations meet, both have already been reduced to fixed summaries. This is the architectural fragmentation that constrains many otherwise capable systems.

Early fusion combines the modalities at the input stage—for example, by adding layout and visual embeddings to token embeddings before the first transformer layer—so that all subsequent computation is jointly informed by every modality. LayoutLM's two-dimensional positional embeddings are a canonical early-fusion mechanism (Xu et al., 2020). Deep fusion goes further, interleaving cross-modal attention throughout the network so that the modalities condition one another at every layer; symmetric cross-modal attention, in which text queries attend to visual keys and visual patches simultaneously query text tokens, is the fullest expression of this idea and produces representations in which vision contextualises text and text disambiguates vision. The general trend across the field is toward deeper fusion, because depth of interaction correlates with the ability to resolve the non-local, structurally mediated relationships that make documents difficult.

The OCR Pipeline and Its Discontents

A further design axis concerns whether a system depends on optical character recognition at all. Layout-aware transformers presuppose an upstream OCR engine that converts page pixels into tokens and bounding boxes; the quality of that engine sets an upper bound on the whole system, and any recognition error—a misread digit, a merged column—propagates into every downstream prediction and cannot be corrected later. This cascade is especially damaging on degraded scans, unusual fonts, and handwriting, precisely the inputs common in archival, clinical, and legal settings. The OCR-free paradigm responds by removing the OCR stage entirely and reading pixels directly, trading the cascade for a loss of explicit coordinates. The tension between these positions—OCR-grounded models that are auditable but brittle to recognition error, and OCR-free models that are robust but hard to ground—runs through the remainder of the survey and is one of the field's genuinely unresolved trade-offs.

Taxonomy of Multimodal Document Understanding

We propose a taxonomy (Figure 2) that separates two orthogonal questions that the literature often conflates: how is a document represented and reasoned over (the architectural paradigm), and how is the cost of that reasoning managed (the efficiency and orchestration mechanism). The first axis comprises four families. Layout-aware transformers fuse text, vision, and layout through positional embeddings and masked pre-training. OCR-free generative models bypass explicit text extraction and read the page image directly. Vision–language models adapt general-purpose multimodal LLMs to documents through instruction tuning. Structural and graph reasoning approaches model the document as an explicit graph of related elements. The second axis also

comprises four families. Retrieval augmentation narrows the input to relevant evidence before generation. Multi-agent frameworks decompose the task across specialised collaborating models. Conditional computation activates only a subset of a model's parameters per input. Cost-aware routing selects, per query, the cheapest pathway or model sufficient to answer it.

The value of separating these axes is that the most promising recent systems combine a paradigm from the first axis with a mechanism from the second—for example, a vision–language backbone wrapped in a multi-agent framework governed by a cost-aware router. Figure 3 places the paradigms on a historical timeline, and the remainder of the survey works through the taxonomy branch by branch before recombining them.

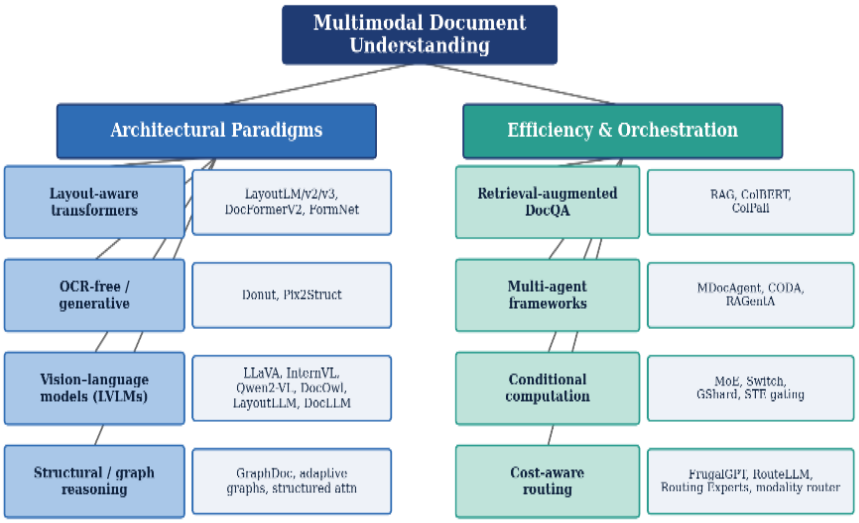


Figure 2: A taxonomy of multimodal document understanding, organised by architectural paradigm and by efficiency/orchestration mechanism

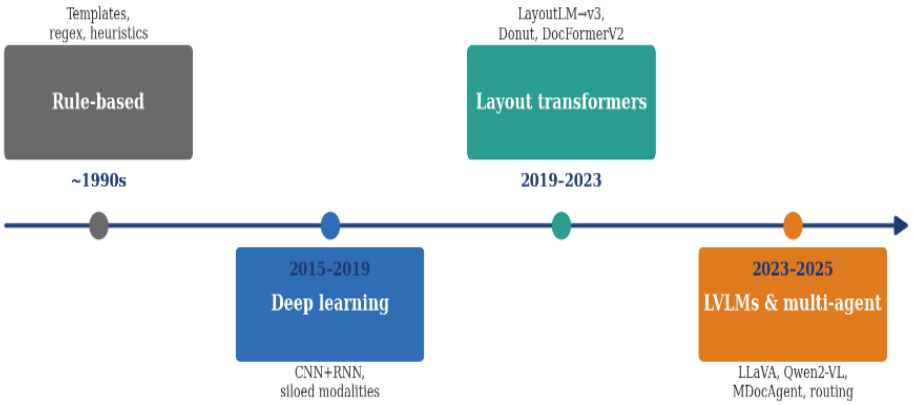


Figure 3: Four eras in the evolution of Document AI, from rule-based systems to LVLM-based multi-agent frameworks

Layout-Aware Document Transformers

Layout-aware transformers were the first architectural family to treat spatial position as a signal on par with token identity, and they remain the strongest models on single-page, structure-heavy extraction tasks. The lineage begins with LayoutLM, which augmented BERT with two-dimensional positional embeddings derived from OCR bounding boxes, allowing the model to learn that spatially adjacent tokens are often semantically related (Xu et al., 2020). LayoutLMv2 deepened the integration by bringing the page image into pre-training and introducing a spatial-aware self-attention mechanism, so that visual and textual streams were fused during representation learning rather than merely concatenated (Xu et al., 2021). LayoutLMv3 completed this trajectory with a unified text–image masked pre-training objective that removed the dependence on a separate convolutional backbone and set a durable standard on FUNSD, DocVQA, and RVL-CDIP (Huang et al., 2022).

Parallel lines of work refined the fusion mechanism and the structural objectives. DocFormerV2 introduced aggressive spatial–semantic alignment objectives that improved dense form parsing, targeting precisely the cases where key–value relationships are geometrically intricate (Appalaraju et al., 2023). FormNet added structural encoding beyond sequential modelling, using graph-based components to capture the relationships that pure self-attention represents only weakly (Lee et al., 2022). mmLayout explored multi-grained fusion, aggregating information at several structural scales to better capture both fine-grained token relations and coarse layout regions (Wang et al., 2022). Across these models, the recurring insight is that layout is not an auxiliary feature to be appended to text but a first-class semantic signal that, when properly modelled, materially improves extraction accuracy.

Two limitations, however, constrain the entire family and motivate the paradigms that follow. First, these models rely on implicit self-attention to discover structural relationships; they are not given, and do not construct, an explicit representation of document structure, so relationships that require multi-hop reasoning over distant elements—linking a total at the bottom of a page to a subtotal at the top—are captured fragilely and degrade under layout variability or noisy scans. Second, self-attention's cost grows quadratically with sequence length, which makes these models expensive or infeasible on long, multi-page documents. Both limitations become acute exactly where modern applications are heading: long, complex, visually rich documents. The first limitation motivates the explicit structural-reasoning work; the second motivates the retrieval and routing work.

Pre-training Objectives

The strength of the layout-transformer family derives less from architectural novelty than from the pre-training objectives that teach a model to align modalities before it ever sees a downstream task. Masked language modelling, inherited from BERT, teaches the model to predict masked tokens from surrounding context (Devlin et al., 2019). Layout-aware variants add masked-image modelling, in which masked image patches are reconstructed, and, most importantly, cross-modal alignment objectives such as text–image matching and word–patch alignment, which explicitly reward the model for associating a token with the region of the page it occupies. LayoutLMv3's unified text–image masking objective was influential precisely because it treated both modalities symmetrically under a single masking scheme, removing the asymmetry of earlier designs that leaned on a heavyweight convolutional vision backbone (Huang et al., 2022). DocFormerV2's spatial–semantic alignment objectives pushed this further for dense forms (Appalaraju et al., 2023). The lesson generalises: the quality of a document encoder is determined largely by whether its pre-training forces genuine cross-modal grounding rather than allowing the modalities to be learned in parallel and merged superficially.

The Long-Document Scaling Wall

The quadratic cost of self-attention is not merely an inconvenience; it is a wall that the single-encoder paradigm cannot climb. A model whose attention cost grows with the square of the token count becomes prohibitively expensive on documents of tens or hundreds of pages, both in memory and in latency, and this is exactly the regime into which applications are moving. The field has responded along two axes. One response keeps the single-encoder architecture but sparsifies or linearises attention to reduce its asymptotic cost; these methods help but do not change the fundamental fact that a single model must still ingest the entire document. The other response, which has proven more influential in practice, abandons the assumption that the whole document must be processed at once: retrieval augmentation selects a small, relevant subset of the document before any expensive model runs, and multi-agent and routing approaches distribute and gate the work. The scaling wall is thus one of the principal forces pushing the field from monolithic encoders toward orchestrated, conditional systems.

OCR-Free and Generative Document Models

A distinct paradigm sidesteps the optical-character-recognition (OCR) pipeline entirely. Traditional layout transformers depend on an upstream OCR system to produce tokens and boxes, and any error in that system cascades into every downstream prediction. OCR-free models instead take the page image as their sole input and generate the output text autoregressively. Donut, the canonical example, is an OCR-free encoder–decoder that reads the document image and emits structured output directly, eliminating the OCR error cascade and performing competitively on document classification and information extraction (Kim et al., 2022). Pix2Struct generalised this idea by pre-training a pixel-to-text model on the task of parsing masked web-page screenshots into simplified HTML, yielding a flexible backbone that transfers to a wide range of visually situated language tasks (Lee et al., 2023).

The trade-off of the OCR-free paradigm is the mirror image of the layout-transformer family's strength. By discarding explicit coordinates, these models lose the ability to ground an extracted entity to a specific page region, which is precisely the capability that regulated, auditable extraction depends upon. They are strong where holistic reading suffices and weaker where fine-grained spatial grounding is required. This tension—between the robustness of end-to-end pixel reading and the auditability of explicit grounding—recurs throughout the field and is not fully resolved by any single model; it reappears in the vision–language models of the next section, which inherit the generative, image-reading stance of the OCR-free paradigm while attempting to recover grounding through scale and instruction tuning.

The pre-training strategy of the OCR-free paradigm is itself instructive. Rather than aligning tokens to boxes, these models learn to read by mapping images to a linearised textual rendering of their content—Pix2Struct, for instance, pre-trains on the task of parsing masked web-page screenshots into simplified HTML, which teaches a single model both to recognise text and to recover the structural markup implicit in a page's appearance (Lee et al., 2023). This objective is appealing because the supervision is abundant and self-labelling: any rendered web page is simultaneously an image and its own structured description. The limitation is that the recovered structure is textual and approximate rather than a set of precise coordinates, so downstream grounding remains coarser than in coordinate-based models. OCR-free models thus occupy a distinctive position in the accuracy–interpretability trade-off: robust and simple to deploy, but harder to audit at the level of individual page regions, which again explains why hybrid strategies that re-inject explicit structure—pursued by the document-adapted LVLMs—have become the dominant direction.

Large Vision–Language Models for Documents

The most visible recent shift in the field has been the adaptation of general-purpose large vision–language models (LVLMs) to documents. These models pair a vision encoder with a large language-model decoder, typically bridged by a projection module, and are trained on broad image–text corpora before being specialised. The contrastive vision–language pre-training of CLIP established that image and text could be aligned in a shared space at scale, providing the encoders on which many later systems are built (Radford et al., 2021). LLaVA demonstrated that visual instruction tuning—fine-tuning an LLM on image-grounded instructions—produces strong open-ended visual reasoning with a remarkably simple architecture (Liu et al., 2023). InternVL scaled the vision foundation model and its alignment to the language model, closing much of the gap to proprietary systems on general multimodal benchmarks (Chen et al., 2024), and Qwen2-VL introduced dynamic-resolution processing and improved cross-modal attention

that are particularly relevant to the high-resolution, text-dense images that documents present (Wang et al., 2024).

A dedicated sub-line adapts these general LVLMs specifically to documents. mPLUG-DocOwl 1.5 introduced unified structure learning for OCR-free document understanding, teaching the model to recover layout structure from the image without an OCR front end (Hu et al., 2024). LayoutLLM injected explicit layout instructions into an LLM, reintroducing the spatial signal that pure image reading tends to lose, and reporting gains on document reasoning tasks (Luo et al., 2024). TextMonkey scaled input resolution and introduced shifted-window attention to cope with the dense text of document images (Liu et al., 2024). DocLLM took a lightweight route, extending a generative language model with disentangled spatial attention over bounding boxes rather than a heavy vision encoder, achieving strong results at lower cost (Wang et al., 2024). Across these systems, the dominant strategy is to raise absolute accuracy through scale and instruction tuning; comprehensive reviews of text-rich image understanding and of visually rich document understanding document both the rapid gains and the accompanying escalation in inference cost (Ding et al., 2025; Fu et al., 2025).

The salient limitation of the LVLM paradigm, for the purposes of this survey, is that it addresses the accuracy axis of the trilemma while largely ignoring the efficiency axis. LVLMs are typically invoked in full for every input: every query pays for the full vision encoder, the full projector, and dense decoding over billions of parameters, regardless of whether the question requires visual reasoning at all. This is efficient neither in compute nor in cost, and it is the specific inefficiency that the orchestration and conditional-computation paradigms are designed to remove.

Anatomy of a Document LVLM

Understanding why LVLMs are expensive requires a look at their common anatomy. A document image is first divided into patches and encoded by a vision transformer, frequently a CLIP- or SigLIP-derived encoder, into a sequence of visual tokens (Radford et al., 2021). Because documents are text-dense and reward high resolution, the number of visual tokens can be large—often larger than the textual prompt—and this token count dominates the cost of the subsequent decoder. A projection module, either a lightweight multilayer perceptron as in LLaVA or a query-based resampler such as a Q-Former, maps the visual tokens into the language model's embedding space (Liu et al., 2023). The language-model decoder then attends jointly over the projected visual tokens and the textual prompt to generate an answer. Each of these stages scales with input size, and the decoder's cost in particular grows with the combined visual-plus-text token count. Innovations such as Qwen2-VL's dynamic resolution and TextMonkey's shifted-window attention are, in large part, attempts to control this token explosion for the high-resolution images that documents demand (Wang et al., 2024; Liu et al., 2024). The architectural point for this survey is that the cost is incurred uniformly: absent an explicit gating mechanism, a text-only question over a document still pays the full price of encoding and attending to a large field of visual tokens.

Strategies for Document Adaptation

Adapting a general LVLM to documents follows three broad strategies, which are not mutually exclusive. The first is instruction tuning on document-specific data, exemplified by LLaVA-style visual instruction tuning applied to document corpora, which teaches the model the conventions of document questions and answers (Liu et al., 2023). The second is structure injection, in which layout information is fed back into a model that would otherwise read only pixels: LayoutLLM supplies explicit layout instructions (Luo et al., 2024), DocLLM adds disentangled spatial attention over bounding boxes without a heavy vision encoder (Wang et al., 2024), and mPLUG-DocOwl 1.5 learns document structure through a unified structure-learning objective (Hu et al., 2024). The third is resolution and tokenisation engineering, in which the visual front end is redesigned to preserve the fine detail that dense text requires while keeping the token count manageable (Wang et al., 2024; Liu et al., 2024). Each strategy raises accuracy; none, on its own, reduces the per-query cost of invoking the full model, which is why the efficiency gains surveyed are increasingly layered on top of, rather than baked into, the LVLM backbone.

It is worth noting the bifurcation of the LVLM landscape into proprietary and open-weight systems, because it shapes what efficiency research is even possible. The most capable general models are frequently proprietary and accessible only through metered application-programming interfaces, which turns every inference into a direct monetary cost and makes token-level efficiency an immediate commercial concern rather than an abstraction; it also means that the internal-state interpretability methods discussed cannot be applied, because the weights are inaccessible. Open-weight models such as the LLaVA, InternVL, and Qwen2-VL families (Liu etl., 2023; Chen et al., 2024; Wang et al., 2024) permit both fine-grained interpretability analysis and architectural modification—expert routing, gating, structure injection—that closed models do not. This division has a practical consequence for the efficiency programme: system-level routing, which treats models as opaque callable services, applies to both proprietary and open models and is therefore the most broadly deployable efficiency lever, whereas model-internal conditional computation is available only where weights are open. Much of the appeal of agent-level routing for documents is precisely that it works over proprietary APIs, where the bulk of production DocQA is served.

Retrieval-Augmented and Multi-Agent Document Understanding

As documents have grown longer, two complementary strategies have emerged to keep the input to a model tractable: retrieval augmentation and multi-agent decomposition. Retrieval-augmented generation (RAG) prepends a retrieval step that selects the most relevant passages before generation, allowing a bounded-context model to answer questions over arbitrarily long documents (Lewis et al., 2020). The retrieval component itself has become a locus of innovation: ColBERT introduced efficient late-interaction passage search that scales to long documents while preserving fine-grained relevance (Khattab & Zaharia, 2020), and ColPali extended retrieval into the visual domain, indexing document pages directly as images with a vision–language model and thereby avoiding a lossy OCR-then-retrieve pipeline (Faysse et al., 2025). A recent survey specifically synthesises the fast-growing area of multimodal retrieval-augmented generation for document understanding, signalling that retrieval has become a core, not peripheral, component of document intelligence (Abootorabi et al., 2025).

Multi-agent frameworks take a different but composable approach: they decompose the task across several specialised models that collaborate on each query. MDocAgent is the canonical document-specific example, orchestrating five agents—a general agent, a critical-evidence agent, a text specialist, an image specialist, and a summariser—that perform retrieval-augmented generation and exchange

evidence to answer long-document questions, achieving state-of-the-art accuracy on long-document benchmarks (Han et al., 2025). CODA extended multi-agent thinking with a dual-brain, decoupled reinforcement-learning design for computer-use agents (Sun et al., 2025), and RAGentA demonstrated a training-free multi-agent RAG pipeline in which successive agents retrieve, filter, and compose answers with in-line citations (Fabian et al., 2025). A parallel line studies iterative routing within multi-agent systems, in which a controller repeatedly directs a query to the most appropriate specialist rather than running all specialists at once (Chen et al., 2025).

The accuracy of these systems is not in doubt; their inefficiency is. The defining problem of the multi-agent paradigm, which we term uniform-density execution, is that the orchestration layer is typically static: every agent runs on every query regardless of the query's actual modality or evidence requirement. A purely textual question over a document still triggers high-resolution image encoding and the full image-grounding agent; a purely visual question still runs the heavy text pipeline. Because each agent invocation is an expensive call to a large model, latency and token cost scale linearly with the number of agents and with document volume, making these systems difficult to deploy at scale despite their accuracy. This is the single largest efficiency gap in current document intelligence, and it is precisely the gap that conditional computation and cost-aware routing—developed in a different research community—are positioned to close.

Retrieval Mechanics for Long Documents

The effectiveness of retrieval augmentation hinges on granularity and modality. Text-only pipelines typically chunk a document into passages, embed each passage, and retrieve the top-k passages most similar to the query; the choice of chunk size trades recall against precision, and fixed-size chunking can sever evidence that spans a boundary. Late-interaction retrieval, introduced by ColBERT, mitigates the coarseness of single-vector retrieval by representing a passage as a set of token-level vectors and scoring query–passage relevance through fine-grained maximum-similarity interactions, which improves recall on the long, heterogeneous passages that documents produce (Khattab & Zaharia, 2020). The more recent and document-specific advance is visual retrieval: ColPali indexes each page directly as an image using a vision–language model, so that charts, tables, and layout—information that an OCR-then-embed pipeline would flatten or lose—remain available to the retriever (Faysse et al., 2025). A dedicated survey of multimodal retrieval-augmented generation documents the rapid maturation of these ideas and argues that retrieval has become a core architectural component rather than a preprocessing convenience (Abootorabi et al., 2025). Retrieval is, in efficiency terms, the cheapest lever available: it reduces the input a downstream model must process before that model is ever invoked.

Multi-Agent Coordination Patterns

Multi-agent document systems differ in how their agents are organised and how control flows among them. In role-specialised pipelines such as MDocAgent, distinct agents own distinct responsibilities—general reasoning, critical-evidence selection, textual grounding, visual grounding, and summarisation—and evidence flows between them in a largely fixed order (Han et al., 2025). In verify-and-refine pipelines such as RAGentA, successive agents retrieve, filter, and compose, with later agents scrutinising the outputs of earlier ones and attaching citations (Fabian et al., 2025). In controller-directed systems, a coordinating agent decides which specialist to consult next, potentially iterating; this iterative-routing pattern brings the multi-agent paradigm closest to the conditional-computation ideas discussed, because the controller is, in effect, making per-step activation decisions (Chen et al., 2025). CODA's dual-brain design, separating high-level planning from low-level action with decoupled reinforcement learning, illustrates a further axis of variation in how planning and execution are divided (Sun et al., 2025). What unites the role-specialised and verify-and-refine patterns, and distinguishes them from controller-directed ones, is that the set of agents invoked is fixed in advance; it is this fixity that produces uniform-density execution.

A Cost Model for Uniform-Density Execution

The cost of uniform-density execution can be stated simply. If a pipeline comprises a set of agents and each agent's invocation carries some latency and token cost, then a static pipeline pays the sum of all agents' costs on every query. If, however, a large fraction of queries require only a subset of the agents—a text-centric question needs no image agent, a chart-reading question needs no heavy text-grounding agent—then the static pipeline wastes, on those queries, the cost of the unneeded agents. Because the wasted cost is incurred per query and multiplied across a corpus, it dominates the operating expense of large-scale deployment. The remedy is to make the set of invoked agents conditional on the query: a lightweight classifier predicts which agents are necessary and gates the rest, defaulting to the full set only when the prediction is uncertain. Provided the classifier is accurate and fails safe, this removes the wasted cost while preserving accuracy, because the agents that do run are unchanged. Reported instances of this design achieve substantial latency and token reductions—on the order of fifteen per cent latency and forty per cent tokens in one document-routing system—at statistically indistinguishable accuracy, which demonstrates that the uniform-density overhead is real and recoverable.

Conditional Computation and Cost-Aware Orchestration

The idea that a model should activate only the parts of itself that a given input requires—conditional computation—predates the current document literature by nearly a decade, and it provides the theoretical foundation for escaping uniform-density execution. Sparsely gated mixture-of-experts (MoE) models introduced a learned gating network that routes each token to a small subset of expert sub-networks, decoupling a model's total capacity from the compute expended on any single input (Shazeer et al., 2017). Switch Transformers simplified and stabilised this routing to scale models to trillions of parameters while keeping per-token compute nearly constant (Fedus et al., 2022), and GShard demonstrated that such conditional routing could be sharded efficiently across accelerators for very large models (Lepikhin et al., 2021). A persistent technical challenge across this line is that the routing decision is discrete and therefore not directly differentiable; the straight-through estimator, which passes gradients through a hard gate as if it were the identity, is the standard remedy and underlies most trainable gating mechanisms (Bengio et al., 2013).

Whereas MoE routes among experts inside a single model, a second line routes among whole models or pathways at the system level, which maps more directly onto document agents. FrugalGPT showed that a cascade of language models of increasing cost, invoked in sequence and halted as soon as

an adequate answer is produced, can match the quality of the most expensive model at a fraction of the cost (Chen et al., 2023). RouteLLM trained a lightweight router that sends each query directly to the cheapest model able to answer it well, reporting large cost reductions at minimal quality loss (Ong et al., 2024). Within multimodal models specifically, Routing Experts learns to route dynamic experts inside an MLLM, adapting the classic MoE idea to the vision–language setting (Yang et al., 2024). A recent survey of inference-optimisation techniques for MoE models consolidates the design space—expert design, compression, dynamic routing, and system-level scheduling—and notes an important caveat: naive conditional routing does not automatically yield wall-clock speedups on modern hardware, so realised efficiency depends heavily on implementation (Liu et al., 2024).

The synthesis that we regard as the most consequential recent development is the transfer of this system-level routing principle from model selection to agent orchestration in document understanding. If a lightweight classifier can decide, per query, which model to invoke, it can equally decide which agents to invoke within a multi-agent document pipeline. A modality router that classifies a query as text-centric, vision-centric, or cross-modal, and conditionally gates the activation of the corresponding specialist agents—defaulting to the full pipeline only when uncertain—applies exactly the conditional-computation principle to the uniform-density problem. Crucially, because such a router changes only which agents run and not how they reason, it can preserve accuracy while removing the cost of redundant agent calls, and because it can be implemented as a zero-shot prompt-based classifier, it can be layered onto an existing pipeline without retraining. This convergence of the efficiency community's routing methods with the document community's multi-agent architectures is, in our assessment, the clearest current path through the efficiency corner of the trilemma.

The Mechanics and Pitfalls of Routing

Whatever the level at which routing acts—experts, models, or agents—it shares a common set of mechanics and a common set of pitfalls. A router is a function from input to a selection over a set of computational units. When that selection is learned, it faces three recurring difficulties. The first is differentiability: a hard selection is a discrete decision that blocks gradient flow, which is why the straight-through estimator and its relatives, which pass gradients through the hard decision as if it were soft, are ubiquitous in trainable routers (Bengio et al., 2013). The second is load balancing: naive routing tends to collapse onto a few favoured units, leaving others under-trained, and MoE systems therefore add auxiliary losses that encourage balanced utilization (Shazeer et al., 2017; Fedus et al., 2022). The third, emphasised in the inference-optimisation literature, is the gap between theoretical and realised efficiency: a routing decision that reduces floating-point operations does not necessarily reduce wall-clock latency or energy, because modern accelerators are optimised for dense, regular computation and can be starved by sparse, irregular activation (Liu et al., 2024). This last point is a crucial caveat for the whole efficiency programme: routing must be evaluated by measured latency, token, and energy costs, not by parameter or FLOP accounting alone.

From Experts to Agents

The move from expert-level to agent-level routing changes the economics of these pitfalls in a way that favours the agentic setting. When routing selects among agents that are themselves calls to large models, each avoided call saves a large, discrete, and easily measured quantity of cost—a full model invocation—rather than a fraction of a dense matrix multiplication. The realised-efficiency gap that plagues fine-grained expert routing is therefore far less severe at the agent level: skipping an agent removes a genuine network round-trip and a genuine block of tokens, which translates directly into lower latency and lower billing. Moreover, agent-level routers can be zero-shot: because the decision is coarse—which of a handful of agents to run—a capable language model can make it from the query alone, with no training and no labelled routing data, which sidesteps the differentiability and load-balancing difficulties entirely. The trade-off is that a zero-shot router's accuracy is bounded by the underlying model's judgement and must be made fail-safe, defaulting to the full pipeline when uncertain so that the cost of a wrong routing decision is bounded by the small fraction of confidently mis-routed queries. The empirical evidence from document-routing systems—high routing F1, mis-routing costs below the measurement noise on accuracy, and large, statistically significant efficiency gains—suggests that this trade-off is favourable in practice, and that agent-level routing is where conditional computation most cleanly pays off for document understanding.

A Taxonomy of Routing Granularity

Routing decisions can be made at several granularities, and the appropriate choice depends on where the redundancy in a workload lies. At the coarsest level, query-level routing decides, once per query, which model or which agents to invoke; this is the granularity of most current document routers and of model-selection systems such as FrugalGPT and RouteLLM (Chen et al., 2023; Ong et al., 2024). At a finer level, modality-level routing decides which modality pathways—textual, visual, or both—are required, which is natural for documents because the modality demanded by a query is often predictable from the query alone. Finer still, page-level and region-level routing decide, within a long document, which pages or regions warrant expensive processing; because long documents frequently concentrate the answer to a query in a small fraction of their pages, this granularity promises the largest savings on exactly the inputs that are otherwise most expensive. At the finest level lies token- and expert-level routing, the classic mixture-of-experts regime, which operates inside a single model on every forward pass (Shazeer et al., 2017; Fedus et al., 2022).

These granularities are composable: a system might route at the query level to decide whether to engage a visual pathway at all, at the page level to decide which pages to encode, and at the expert level within the model that finally runs. Coarser routing is cheaper to decide and easier to make zero-shot but leaves finer redundancy unexploited; finer routing captures more savings but is harder to decide reliably and is more exposed to the realised-efficiency gap. The design frontier, largely unexplored for documents, is the principled combination of granularities so that each level of redundancy is removed by the routing decision best suited to it, with a consistent fail-safe policy across levels.

Structural Reasoning: Implicit versus Explicit

A cross-cutting theme that distinguishes document understanding from generic image–text tasks is the central role of structure. Documents are not flat sequences; they are hierarchically and relationally organised, and much of their meaning lives in relationships—row–column dependencies in tables, key–value links in forms, logical associations between checkboxes and clauses, and section hierarchies. Two

paradigms for capturing this structure coexist. The implicit paradigm, embodied by the layout transformers and the LVLMs, hopes that self-attention will discover structural relationships from data given enough scale and appropriate pre-training. The explicit paradigm instead represents the document as a graph whose nodes are elements and whose edges are relationships, and reasons over that graph directly. Concretely, a document graph is typically initialised by connecting spatially proximate elements—via a k-nearest-neighbour rule over bounding-box coordinates—to provide a spatial skeleton, after which message passing propagates information along edges so that each node's representation is informed by its structural neighbours. The critical design question is whether the edges are fixed or learned.

A static graph fixes its edges once, from spatial heuristics, and reasons over that fixed topology; an adaptive graph instead re-weights its edges at every layer through a learnable compatibility function that scores each pair of nodes on both their learned representations and a spatial bias derived from their relative positions, and then propagates features along the re-weighted edges by multi-head attention message passing. The practical effect of this refinement is that the model can discover logical links that the initial spatial skeleton missed—connecting a total at the bottom of a page to a subtotal at the top—and can suppress spurious edges introduced by mere proximity. This combines the generative flexibility of self-attention with the structural inductive bias of graph neural networks, and it is the mechanism by which explicit reasoning recovers relationships that implicit attention leaves fragile.

Explicit graph approaches have consistently shown value on layout-heavy tasks, but early instantiations suffered from a brittleness: a graph constructed once, from spatial heuristics, cannot recover if its initial edges are wrong. The more recent direction is adaptive graph reasoning, in which edges are refined dynamically across layers by a learnable compatibility function, combining the generative flexibility of transformers with the structural rigour of graph neural networks and allowing the model to learn logical links even between geometrically distant elements. Complementary evidence for the importance of structure comes from analyses showing that structured attention—biasing or organising attention according to document structure—measurably improves multimodal LLMs on document understanding, indicating that even within the implicit paradigm, injecting structural priors helps (Wang et al., 2025). The broader lesson, supported by studies of information redundancy and evaluation bias in document extraction benchmarks (Ratnamogan et al., 2023), is that explicit structural signals capture regularities that implicit attention recovers only partially, and that the two paradigms are complementary rather than competing.

Benchmarks, Datasets, and Evaluation

Progress in the field is mediated by its benchmarks, and the benchmark landscape has shifted decisively toward longer, more visually complex documents. The single-page era was defined by FUNSD for form understanding (Jaume et al., 2019), SROIE for receipts (Huang et al., 2019), DocVQA for visual question answering (Matthew et al., 2021), RVL-CDIP for classification (Harley et al., 2015), and PubLayNet for layout segmentation (Zhong et al., 2019). These remain valuable for measuring fine-grained extraction and grounding, and Table 3 summarises them alongside the newer long-document benchmarks. The long-document era is defined by benchmarks that stress cross-page evidence integration and retrieval. DUDE broadened document VQA to multi-page, multi-domain documents with structured layouts (Van Landeghem et al., 2023). MMLongBench-Doc benchmarks long-context document understanding over lengthy PDFs—averaging tens of pages—with a large fraction of questions requiring evidence from multiple pages and from heterogeneous sources including text, tables, charts, and layout (Ma et al., 2024). LongDocURL integrates understanding, reasoning, and locating over thousands of question–answer pairs spanning hundreds of long documents and dozens of sub-tasks, and has become a common testbed for long-document systems (Deng et al., 2025).

Evaluation methodology has not kept pace with this shift, and a specific measurement problem deserves emphasis. Long-document question answering is frequently scored with a strict exact-match metric, often adjudicated by an LLM judge, which penalises answers that are semantically correct but lexically divergent—returning a currency-formatted total where the gold answer is an unformatted number, for instance. Strict exact-match therefore systematically understates true accuracy, and the resulting low absolute scores can be mistaken for model failure when they are in fact metric artefacts. The standard mitigation is to report a fuzzy metric such as the Average Normalized Levenshtein Similarity (ANLS) alongside the strict metric, and analyses of redundancy and bias in extraction benchmarks reinforce the case for fuzzy scoring as a default (Ratnamogan et al., 2023). Beyond accuracy, a complete evaluation of a modern system must report efficiency—inference latency, API or token cost per document, and throughput—and, increasingly, interpretability. Table 4 organises the efficiency and orchestration mechanisms surveyed against the axis of the trilemma each primarily targets.

Table 1: Representative Architectural Paradigms for Multimodal Document Understanding

Paradigm	Representative Models	Key Strength	Key Limitation
Layout transformers	LayoutLM/v2/v3, DocFormerV2, FormNet, mmLayout	Strong single-page extraction; explicit coordinates	Implicit structure; quadratic cost on long docs
OCR-free generative	Donut, Pix2Struct	No OCR error cascade; end-to-end reading	Weak coordinate grounding / auditability
Vision–language models	LLaVA, InternVL, Qwen2-VL, DocOwl 1.5, LayoutLLM, DocLLM, TextMonkey	High absolute accuracy; open-ended reasoning	Dense per-query cost; invoked in full
Structural / graph reasoning	GraphDoc, adaptive graph reasoning, structured attention	Explicit multi-hop structural relations	Added complexity; initial-edge brittleness (static graphs)

Beyond individual metrics, the methodology of evaluation itself warrants scrutiny, because efficiency claims are peculiarly fragile. Inference latency and, to a lesser extent, token cost vary from run to run with network conditions, API load, and backend scheduling, so a single measurement of a fifteen-per-cent latency reduction may reflect noise as much

as signal. Rigorous practice therefore repeats timing and cost runs several times and reports the mean with a dispersion measure, and it uses paired statistical tests—McNemar's test for paired binary accuracy, the Wilcoxon signed-rank test for paired latency and token comparisons—together with effect sizes, so that the distinct claims of accuracy parity and of real efficiency gains can each be substantiated separately. Accuracy on long-document benchmarks is likewise best reported with bootstrap confidence intervals over questions rather than as a single point estimate. This methodological discipline is not pedantry: because the efficiency programme advocated in this survey stands or falls on measured cost reductions, the credibility of the whole enterprise depends on evaluations that separate signal from variance.

A complete evaluation therefore spans several distinct dimensions, each with its own metrics, and Table 2 organises them. Accuracy is captured by exact-match rates and by fuzzy similarity such as ANLS, ideally reported per question type so that strengths and weaknesses are visible rather than averaged away. Efficiency is captured by latency, token or API-unit consumption, peak memory, and throughput. Statistical rigour is captured by confidence intervals, paired significance tests, and effect sizes. Interpretability is captured by attention–ground-truth alignment, human verification time, and calibration error. A system reported on only one of these dimensions—accuracy alone, as remains common—cannot be situated in the trilemma, and the field's move toward multi-dimensional reporting is a precondition for the co-optimisation this survey advocates.

Table 2: Evaluation Dimensions and Representative Metrics for Document Understanding

Dimension	Representative metrics	What it measures
Accuracy	Exact-match / binary correctness, ANLS, per-type F1	Correctness, strict and fuzzy
Efficiency	Latency, tokens or API units per document, peak memory, throughput	Deployable cost and speed
Statistical rigour	Bootstrap 95% CI, McNemar, Wilcoxon signed-rank, effect size	Signal versus run-to-run variance
Interpretability	Attention–GT alignment, human time-to-verification, calibration error	Trustworthiness and auditability

Table 3: Representative Benchmarks for Multimodal Document Understanding

Benchmark	Task	Scale / Scope	Primary Metric
FUNSD	Form key-value extraction / linking	199 forms	F1
SROIE	Receipt information extraction	973 receipts	F1
DocVQA	Document visual QA (single-page)	~50k images	ANLS
RVL-CDIP	Document classification	400k images, 16 classes	Accuracy
PubLayNet	Layout segmentation	360k documents	IoU / mAP
DUDE	Multi-page document VQA	Multi-domain, multi-page	ANLS
MMLongBench-Doc	Long-context document QA	135 long PDFs, ~47 pages avg.	Accuracy (multi-source)
LongDocURL	Long-document understand / reason / locate	2,325 QA over 396 docs, 20 sub-tasks	Binary correctness + ANLS

Interpretability and Trustworthy Document AI

Accuracy is necessary but not sufficient for deployment in regulated domains, where a prediction must be not only correct but also explainable and auditable. Interpretability research in document understanding proceeds along two complementary lines. The first is representational: understanding what the model has internally learned. Neuron-level analyses such as MMNeuron discover domain-specific neurons inside multimodal LLMs, showing that specialised knowledge is localised in identifiable sub-populations of units and that this localisation can be measured and, in principle, exploited (Huo et al., 2024). This finding has a practical corollary for efficiency: if domain-specific knowledge is concentrated in a small, identifiable set of neurons, then activating only those neurons for a given domain is a form of conditional computation grounded in interpretability, linking the two axes of the trilemma directly. Complementary techniques probe the model's layerwise dynamics: the logit lens projects intermediate representations onto the output vocabulary to observe how a prediction forms across depth, and calibration analyses ask whether a model's confidence matches its accuracy, which matters acutely in high-stakes settings where an over-confident error is more dangerous than an acknowledged uncertainty. Attention-based analyses ask whether a model attends to the regions that actually contain the answer; a natural metric is the overlap between the model's peak attention and the ground-truth answer region, which quantifies whether the model is 'looking' in the right place. Evidence that structured attention improves document understanding (Wang et al., 2025) suggests that these internal signals are not merely diagnostic but causally related to accuracy.

The second line is human-centric: measuring whether a model's exposed reasoning actually helps a person verify its output. Purely qualitative attention visualisations, common in early work, are difficult to evaluate and easy to over-interpret. A more rigorous approach ties interpretability to an externally observable cost—for example, the time a trained human verifier needs to confirm or reject a prediction with, versus without, an exposed reasoning trace such as a graph-edge visualisation. When explicit reasoning traces measurably reduce human verification time, interpretability is demonstrated in operational terms rather than asserted. This human-cost framing connects document AI to the broader trustworthy-AI and regulatory agenda, in which auditability is a precondition for deployment in clinical, legal, and financial settings, and it argues for interpretability metrics to be reported alongside accuracy and efficiency as a matter of course.

A further advantage of the explicit, orchestrated systems surveyed is that they are, by construction, more interpretable than monolithic dense models. When a system routes a query to a subset of named agents, the routing decision is itself an explanation of how the query was handled; when a system reasons over an explicit graph of document elements, the graph edges are a legible account of which relationships

informed the answer. In this sense, the efficiency mechanisms and the interpretability mechanisms advocated here are not merely compatible but mutually reinforcing: making computation conditional and explicit tends to make it inspectable. This contrasts with the interpretability of dense LVLMs, which must be recovered after the fact through neuron and attention analysis because nothing in the architecture exposes its own reasoning. The design lesson is that interpretability is cheaper to obtain when it is built in through explicit structure and orchestration than when it is retrofitted onto an opaque model.

Efficiency and the Green-AI Imperative

The efficiency axis of the trilemma has moved from an engineering afterthought to a first-class scientific and ethical concern. The Green-AI argument holds that the computational cost of a system—and the energy and carbon it entails—should be reported and minimised as a standard part of research practice, not treated as a free externality (Schwartz et al., 2020). In document intelligence, where systems are deployed over corpora of millions of documents and where multi-agent pipelines multiply per-query cost, the stakes are substantial: a reduction in per-document token consumption translates directly into lower API expenditure, higher throughput within a fixed budget, and a smaller carbon footprint per query.

The magnitude of the concern is easy to underestimate. Inference, not training, dominates the lifetime compute of a widely deployed model, because a model is trained once but queried billions of times; contemporary estimates place the annual energy of serving a single large multimodal model at hundreds of gigawatt-hours, and the water and carbon footprints scale accordingly. In document intelligence specifically, the multi-agent multiplier is pernicious: a five-agent pipeline does not merely add a constant overhead but multiplies the per-query footprint, and it does so on every document in a corpus that may number in the millions. A forty-per-cent reduction in per-document token consumption is therefore not a marginal optimisation but, at corpus scale, a transformative reduction in operating cost and emissions, and it corresponds to processing roughly one and two-thirds as many documents within an unchanged budget. This is why the efficiency techniques below should be read not as engineering conveniences but as the primary lever by which document AI can be scaled responsibly.

The efficiency techniques surveyed here can be organised by where they act. At the model level, conditional computation through mixture-of-experts routing activates only a subset of parameters per input (Shazeer et al., 2017; Fedus et al., 2022; Lepikhin et a., 2021), and compression techniques—pruning, quantisation, and distillation—reduce the cost of the parameters that remain, as consolidated in recent inference-optimisation surveys (Liu et al., 2024). These compression techniques are worth distinguishing because they occupy a different position in the trilemma from routing. Pruning removes weights judged unimportant; quantisation reduces the numerical precision of weights and activations, often to eight or four bits, cutting memory and bandwidth; and distillation trains a smaller student model to imitate a larger teacher. Each reduces cost uniformly across inputs and each risks a small accuracy loss, in contrast to conditional routing, which reduces cost selectively and can be made accuracy-neutral. Compression and routing are complementary: a system can quantise the model that a router chooses to run, compounding a per-input selection with a per-parameter reduction. For document intelligence served over proprietary APIs, however, compression is typically inaccessible—the provider controls the weights—so system-level routing again emerges as the efficiency lever available to practitioners who do not own the model. At the input level, retrieval augmentation narrows the context that a model must process (Lewis et al., 2020; Khattab & Zaharia, 2020; Faysse et al., 2025). At the system level, cost-aware routing selects the cheapest sufficient model or pathway per query (Chen et al., 2023; Ong et al., 2024), and, as argued, the same principle applied to agent orchestration removes the uniform-density overhead of multi-agent pipelines. Table 4 maps these mechanisms to the trilemma axis each primarily improves and notes their interaction with accuracy. A recurring and important caveat, emphasised in the MoE inference literature, is that theoretical FLOP reductions do not automatically become wall-clock or energy savings on real hardware; efficiency claims must therefore be validated empirically with latency, token, and, ideally, energy measurements rather than asserted from parameter counts alone (Liu et al., 2024).

Table 4: Efficiency and Orchestration Mechanisms Mapped to the Trilemma.

Mechanism	Level	Primary axis improved	Effect on accuracy	Representative work
Mixture-of-experts routing	Model	Efficiency (compute)	Preserved (conditional capacity)	Switch, GShard (Fedus et al., 2022; Lepikhin et a., 2021)
Compression (prune/quantise/distill)	Model	Efficiency (memory, compute)	Small loss possible	MoE inference survey (Liu et al., 2024)
Retrieval augmentation	Input	Efficiency (context) + accuracy	Improved on long docs	RAG, ColBERT, ColPali (Lewis et al., 2020; Khattab & Zaharia, 2020; Faysse et al., 2025)
Model-level routing / cascades	System	Efficiency (cost)	Preserved	FrugalGPT, RouteLLM (Chen et al., 2023; ong et al., 2024)
Agent-level modality routing	System	Efficiency (cost, latency)	Preserved (fail-safe fallback)	Modality-conditional orchestration (Han et al., 2025; Chen et al., 2025)
Explicit structural reasoning	Model	Accuracy + interpretability	Improved on structure	Adaptive graphs, structured attn (Wang et al., 2025)

Applications and Deployment Domains

The paradigms surveyed above are not academic exercises; they are deployed across sectors whose requirements differ sharply, and those differences illuminate why the trilemma matters. In finance, document intelligence underpins invoice and receipt processing, statement reconciliation, and know-your-customer verification, where volume is enormous and the marginal cost of each API call is a direct operating expense; here the efficiency axis dominates, and modality-conditional routing that removes redundant agent calls translates immediately into lower cost and higher throughput within a fixed budget. In healthcare, clinical forms, laboratory reports, and scanned records must be extracted with high fidelity and, critically, with an auditable trail, because a prediction that cannot be verified cannot be acted upon; here the interpretability axis dominates, and the human-cost framing—reducing the time a clinician or coder needs to confirm a prediction—is the operative benefit. In legal services, contract analysis and discovery require parsing long, structurally intricate documents across jurisdictions, stressing both the long-document scaling wall and explicit structural reasoning. In science and government, extracting evidence from figures, tables, and prose across millions of papers and records rewards retrieval augmentation and long-document benchmarks such as LongDocURL and MMLongBench-Doc (Deng et al., 2025; Ma et al, 2024).

A domain that deserves particular attention, because it is under-served by the current literature, is document intelligence for low-resource languages and historical archives. The overwhelming majority of models and benchmarks target printed English, yet an enormous share of the world's consequential documents—civil records, land titles, medical registries, and cultural heritage—exists in other languages, in handwriting, and on degraded media. These inputs stress precisely the weakest points of current systems: the OCR cascade fails on unfamiliar scripts and noisy scans, layout conventions differ across administrative traditions, and instruction-tuned LVLMs inherit the language biases of their pre-training. The efficiency techniques surveyed here are doubly relevant in this setting, because low-resource deployment often occurs on constrained infrastructure where dense execution is simply infeasible; a system that can route away redundant computation may be the difference between a deployable tool and none at all. Extending both the architectural backbones and the routing logic to these settings is therefore not only a matter of fairness but a test of whether the field's methods generalise beyond their comfortable evaluation regime.

Across these domains, three regulatory and operational pressures recur. The first is auditability: sectors governed by frameworks that require traceable, explainable decisions cannot deploy black-box models regardless of accuracy, which elevates interpretability from a research nicety to a deployment precondition. The second is cost at scale: because these systems run over very large corpora, per-query inefficiency compounds into substantial expense and carbon, making the efficiency techniques economically decisive rather than merely elegant. The third is robustness to real-world inputs—degraded scans, handwriting, and low-resource languages—where the OCR cascade is most damaging and where OCR-free and structurally aware approaches have the most to offer. The practical upshot is that no single paradigm is universally best; the right system is the one whose position in the trilemma matches the dominant constraint of the deployment domain.

Cross-Paradigm Synthesis and Discussion

Read together, the paradigms of this survey are complementary rather than competing, and the strongest emerging systems are compositions across the two taxonomy axes. A capable modern document system might use a vision–language backbone for open-ended reasoning, a layout-aware or structurally aware component for fine-grained extraction, retrieval to bound the input on long documents, a multi-agent decomposition to divide the labour, and a cost-aware router to gate the agents so that only the necessary ones run per query. Each component addresses a different corner of the trilemma: the backbone and structural reasoning serve accuracy, retrieval and routing serve efficiency, and exposed reasoning traces serve interpretability. The central argument of this survey is that these axes should be co-designed rather than layered opportunistically, because the interactions between them are where both the risks and the opportunities lie.

Two interactions deserve emphasis. The first is that efficiency mechanisms can be made accuracy-neutral by construction. A router that changes only which agents run, not how they reason, and that fails safe to the full pipeline when uncertain, cannot degrade accuracy beyond the small, measurable cost of its own mis-routing; this is qualitatively different from compression methods that trade accuracy for efficiency, and it is why agent-level routing is such an attractive lever. The second is that structure and interpretability reinforce one another: an explicit graph of document elements is simultaneously a mechanism for better structural reasoning and a ready-made reasoning trace that a human can inspect, so the same architectural choice can strengthen two corners of the trilemma at once. These interactions suggest that the trilemma, while real, is not immutable: the historical pattern in which improving one axis degraded the others reflected particular design choices, and newer designs can improve two or three axes together. The task for the field is to find and generalise such designs, and the convergence of conditional computation with multi-agent document architectures is, we have argued, the clearest current example.

At the same time, the discussion must be tempered by measurement discipline. Many reported efficiency and accuracy gains rest on partial evaluations—small samples, single runs, strict metrics that understate accuracy, or FLOP accounting that overstates realised savings. The credibility of the synthesis we advocate depends on evaluating it honestly: with repeated runs and confidence intervals for the noisy quantities, with fuzzy metrics reported alongside strict ones, and with measured latency, token, and energy costs rather than theoretical proxies. The field's most important methodological need is not another architecture but a shared commitment to reporting that makes cross-paradigm comparison trustworthy.

Open Challenges and Future Directions

Several challenges will, in our assessment, define the next phase of the field. The first is evaluation at credible scale with honest metrics. Long-document benchmarks such as LongDocURL and MMLongBench-Doc have raised the difficulty of the task, but reported results are still often based on partial samples, single runs, and strict exact-match scoring that understates true accuracy. The community would benefit from pre-registered evaluation protocols with repeated runs, bootstrap confidence intervals, paired significance testing, and fuzzy metrics reported alongside strict ones, so that efficiency claims in particular—which are highly sensitive to run-to-run variance—can be trusted.

The second is finer-grained and more general orchestration. Current routing operates largely at the level of whole models or whole queries; page-level and region-level routing, which decide agent activation per page or per region within a long document, promise further savings on documents in which only a small subset of pages contains the relevant evidence. More broadly, the modality-conditional routing principle is not specific to documents: any multi-agent system in which running every agent on every input is wasteful—medical decision support, autonomous perception, scientific-discovery agents—may benefit from the same conditional-computation treatment, and establishing whether these routing policies transfer is an open empirical question.

The third is multilingual, low-resource, and noisy-input robustness. The bulk of the literature evaluates printed English documents; extending both the architectural backbones and the routing logic to low-resource languages, handwritten text, and degraded scans is essential for document intelligence to serve emerging markets and historical archives. The fourth is the maturation of interpretability from diagnostic to operational: tying internal signals to auditable, machine-verifiable evidence chains that satisfy regulatory requirements, and doing so at inference time rather than only in post-hoc analysis. The fifth is the principled unification of the two taxonomy axes—jointly designing the architectural paradigm and the efficiency mechanism rather than bolting one onto the other—so that accuracy, efficiency, and interpretability are co-optimised from the outset. The convergence of conditional computation with multi-agent document architectures, surveyed, is an early and promising instance of exactly this unification.

A sixth challenge concerns the realised-efficiency gap that recurs throughout the conditional-computation literature. Because sparse and conditional computation does not automatically translate theoretical savings into wall-clock or energy savings on hardware optimised for dense computation (Liu et al., 2024), the community needs standard, hardware-grounded efficiency reporting—latency, tokens, and, where possible, measured energy—so that claimed savings are credible and comparable across systems. Closely related is a seventh challenge: the co-design of routers and the units they route over. Most current routing treats the downstream agents or experts as fixed and optimises the router around them; jointly optimising the router and the agents, so that the agents are designed to be cleanly separable and the router is calibrated to their boundaries, is largely unexplored and may yield both better accuracy and larger savings.

An eighth, more speculative direction is the generalisation of document-derived orchestration principles to other agentic domains. The modality-conditional routing that removes uniform-density overhead in document pipelines is not intrinsically about documents; it is about any multi-agent system in which the cost of running every agent on every input is wasteful. Establishing whether the routing policies, calibration methods, and fail-safe designs developed for documents transfer to medical decision support, autonomous perception, or scientific-discovery agents would both broaden their impact and test the depth of the underlying principle. Finally, the field would benefit from living, community-maintained benchmarks that jointly score accuracy, efficiency, and interpretability, so that progress on the trilemma as a whole—rather than on any single axis—can be tracked over time.

CONCLUSION

This survey has traced multimodal document understanding from layout-aware transformers, through OCR-free generative models and large vision–language models, to retrieval-augmented and multi-agent frameworks, and it has argued that the field's dominant trajectory—ever-higher accuracy at ever-higher cost and opacity—is best understood as movement within a three-way trade-off between accuracy, efficiency, and interpretability. Organising the literature along two orthogonal axes—architectural paradigm and efficiency/orchestration mechanism—clarifies both what each family of models contributes and what it leaves unresolved. The layout transformers gave the field explicit spatial grounding but scale poorly; the LVLMs raised absolute accuracy but ignored per-query cost; the multi-agent frameworks reached state-of-the-art accuracy but at the price of uniform-density execution.

The most promising current direction, in our reading, is the transfer of conditional computation—long established for tokens in mixture-of-experts models and for models in cost-aware cascades—to the orchestration of document agents, which decouples a system's capability from its per-query cost and offers a genuine, measurable escape from the efficiency corner of the trilemma without sacrificing accuracy. Combined with explicit structural reasoning to strengthen the accuracy corner and operational, human-cost interpretability to strengthen the trust corner, this points toward a generation of document-understanding systems that are accurate, efficient, and auditable at once. We expect the next decade of document intelligence to be defined less by incremental accuracy on saturated single-page benchmarks than by this co-optimisation of the three objectives at the scale and under the constraints that real deployment demands.

REFERENCES

Abootorabi, M. M., et al. (2025). *Scaling beyond context: A survey of multimodal retrieval-augmented generation for document understanding* (arXiv:2510.15253). arXiv. https://doi.org/10.48550/arXiv.2510.15253

Appalaraju, S., Tang, Y., Dong, Q., Sankaran, S., Zhang, Y., & Manmatha, R. (2023). DocFormerv2: Local features for document understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(2), 949–957. https://doi.org/10.1609/aaai.v38i2.27828

Bengio, Y., Léonard, N., & Courville, A. (2013). *Estimating or propagating gradients through stochastic neurons for conditional computation* (arXiv:1308.3432). arXiv. https://doi.org/10.48550/arXiv.1308.3432

Chen, L., Zaharia, M., & Zou, J. (2023). *FrugalGPT: How to use large language models while reducing cost and improving performance* (arXiv:2305.05176). arXiv. https://doi.org/10.48550/arXiv.2305.05176

Chen, Y., et al. (2025). *Talk to right specialists: Iterative routing in multi-agent systems for question answering* (arXiv:2501.07813). arXiv. https://doi.org/10.48550/arXiv.2501.07813

Chen, Z., et al. (2024). InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24115–24126. https://doi.org/10.1109/CVPR52733.2024.02283

Deng, C., et al. (2025). LayoutmLongDocURL: A comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2412.18424

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Ding, Y., Luo, S., Dai, Y., Jiang, Y., Li, Z., Sun, Q., Martin, G., Liu, W., & Peng, Y. (2025). A Survey on MLLM-based Visually Rich Document Understanding: Methods, Challenges, and Emerging Trends. *ArXiv*. https://arxiv.org/abs/2507.09861

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*. https://arxiv.org/abs/2010.11929

Fabian, D., et al. (2025). *RAGentA: Multi-agent retrieval-augmented generation for attributed question answering* (arXiv:2506.16988). arXiv. https://doi.org/10.48550/arXiv.2506.16988

Faysse, M., et al. (2025). ColPali: Efficient document retrieval with vision language models. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2407.01449

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, *23*(120), 1–39. https://doi.org/10.48550/arXiv.2101.03961

Fu, P., Guan, T., Wang, Z., Guo, Z., Duan, C., Sun, H., Chen, B., Ma, J., Jiang, Q., Zhou, K., & Luo, J. (2025). Multimodal Large Language Models for Text-rich Image Understanding: A Comprehensive Review. *ArXiv*. https://arxiv.org/abs/2502.16586

Han, S., et al. (2025). *MDocAgent: A multi-modal multi-agent framework for document understanding* (arXiv:2503.13964). arXiv. https://doi.org/10.48550/arXiv.2503.13964

Harley, A. W., Ufkes, A., & Derpanis, K. G. (2015). Evaluation of deep convolutional nets for document image classification and retrieval (RVL-CDIP). *Proceedings of the 2015 International Conference on Document Analysis and Recognition*, 991–995. https://doi.org/10.1109/ICDAR.2015.7333910

Hu, A., et al. (2024). mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 3089–3103. https://doi.org/10.18653/v1/2024.findings-emnlp.175

Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*, 4083–4091. https://doi.org/10.1145/3503161.3548112

Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., & Jawahar, C. V. (2019). ICDAR2019 competition on scanned receipt OCR and information extraction (SROIE). *Proceedings of the 2019 International Conference on Document Analysis and Recognition*, 1516–1520. https://doi.org/10.1109/ICDAR.2019.00244

Huo, J., Yan, Y., Hu, B., Yue, Y., & Hu, X. (2024). MMNeuron: Discovering neuron-level domain-specific interpretation in multimodal large language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4467–4482. https://doi.org/10.18653/v1/2024.emnlp-main.387

Jaume, G., Ekenel, H. K., & Thiran, J.-P. (2019). FUNSD: A dataset for form understanding in noisy scanned documents. *Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops*, *2*, 1–6. https://doi.org/10.1109/ICDARW.2019.10020

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48. https://doi.org/10.1145/3397271.3401075

Kim, G., et al. (2022). OCR-free document understanding transformer (Donut). *European Conference on Computer Vision*, 493–517. https://doi.org/10.1007/978-3-031-19815-1_29

Lee, C.-Y., et al. (2022). FormNet: Structural encoding beyond sequential modeling in form document information extraction. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*, 3735–3754. https://doi.org/10.18653/v1/2022.acl-long.260

Lee, K., et al. (2023). Pix2Struct: Screenshot parsing as pretraining for visual language understanding. *International Conference on Machine Learning*, 18832–18843. https://doi.org/10.48550/arXiv.2210.03347

Lepikhin, D., et al. (2021). GShard: Scaling giant models with conditional computation and automatic sharding. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2006.16668

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning (LLaVA). *Advances in Neural Information Processing Systems*, *36*, 34892–34916. https://doi.org/10.48550/arXiv.2304.08485

Liu, J., et al. (2024). *A survey on inference optimization techniques for mixture of experts models* (arXiv:2412.14219). arXiv. https://doi.org/10.48550/arXiv.2412.14219

Liu, Y., et al. (2024). *TextMonkey: An OCR-free large multimodal model for understanding document* (arXiv:2403.04473). https://doi.org/10.48550/arXiv.2403.04473

Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., & Yao, C. (2024). LayoutLLM: Layout instruction tuning with large language models for document understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15647. https://doi.org/10.1109/CVPR52733.2024.01480

Ma, Y., et al. (2024). MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, *37*. https://doi.org/10.48550/arXiv.2407.01523

Mathew, M., Karatzas, D., & Jawahar, C. V. (2021). DocVQA: A dataset for VQA on document images. *Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision*, 2199–2208. https://doi.org/10.1109/WACV48630.2021.00171

Ogar, A. O., Abah, J., Suleiman, M. A., Akande, O. N., & Muhammed, F. O. (2026). Optimising domain-specific neuron activation for efficient multimodal language understanding in cloud AI systems. *International Journal of Computer Information Systems and Industrial Management Applications*, *18*(5s), 816–831.

Ong, I., et al. (2024). *RouteLLM: Learning to route LLMs with preference data* (arXiv:2406.18665). arXiv. https://doi.org/10.48550/arXiv.2406.18665

OpenAI. (2024). *GPT-4o system card*. https://openai.com/index/gpt-4o-system-card/

Radford, A., et al. (2021). Learning transferable visual models from natural language supervision (CLIP). *International Conference on Machine Learning*, 8748–8763. https://doi.org/10.48550/arXiv.2103.00020

Ratnamogan, P., Tang, J., Lam, L., Vanhuffel, W., & Caspani, F. (2023). *Information redundancy and biases in public document information extraction benchmarks* (arXiv:2304.14936). arXiv. https://doi.org/10.48550/arXiv.2304.14936

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, *63*(12), 54–63. https://doi.org/10.1145/3381831

Shazeer, N., et al. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1701.06538

Sun, Z., et al. (2025). *CODA: Coordinating the cerebrum and cerebellum for a dual-brain computer use agent with decoupled reinforcement learning* (arXiv:2508.20096). arXiv. https://doi.org/10.48550/arXiv.2508.20096

Van Landeghem, J., et al. (2023). Document understanding dataset and evaluation (DUDE). *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17616–17627. https://doi.org/10.1109/ICCV51070.2023.01789

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Wang, D., et al. (2024). *DocLLM: A layout-aware generative language model for multimodal document understanding* (arXiv:2401.00908). https://doi.org/10.48550/arXiv.2401.00908

Wang, P., et al. (2024). *Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution* (arXiv:2409.12191). https://doi.org/10.48550/arXiv.2409.12191

Wang, W., et al. (2022). mmLayout: Multi-grained multimodal transformer for document understanding. *Proceedings of the 30th ACM International Conference on Multimedia*, 4242–4251. https://doi.org/10.1145/3503161.3548406

Wang, Y., et al. (2025). *Structured attention matters to multimodal LLMs in document understanding* (arXiv:2506.21600). arXiv. https://doi.org/10.48550/arXiv.2506.21600

Xu, Y., et al. (2021). LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, *1*, 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). LayoutLM: Pre-training of text and layout for document image understanding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1192–1200. https://doi.org/10.1145/3394486.3403172

Yang, Q., et al. (2024). *Routing experts: Learning to route dynamic experts in multi-modal large language models* (arXiv:2407.14093). arXiv. https://doi.org/10.48550/arXiv.2407.14093

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, *11*(5), Article nwae403. https://doi.org/10.1093/nsr/nwae403

Zhong, X., Tang, J., & Jimeno-Yepes, A. (2019). PubLayNet: Largest dataset ever for document layout analysis. *Proceedings of the 2019 International Conference on Document Analysis and Recognition*, 1015–1022. https://doi.org/10.1109/ICDAR.2019.00166

©2026 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.