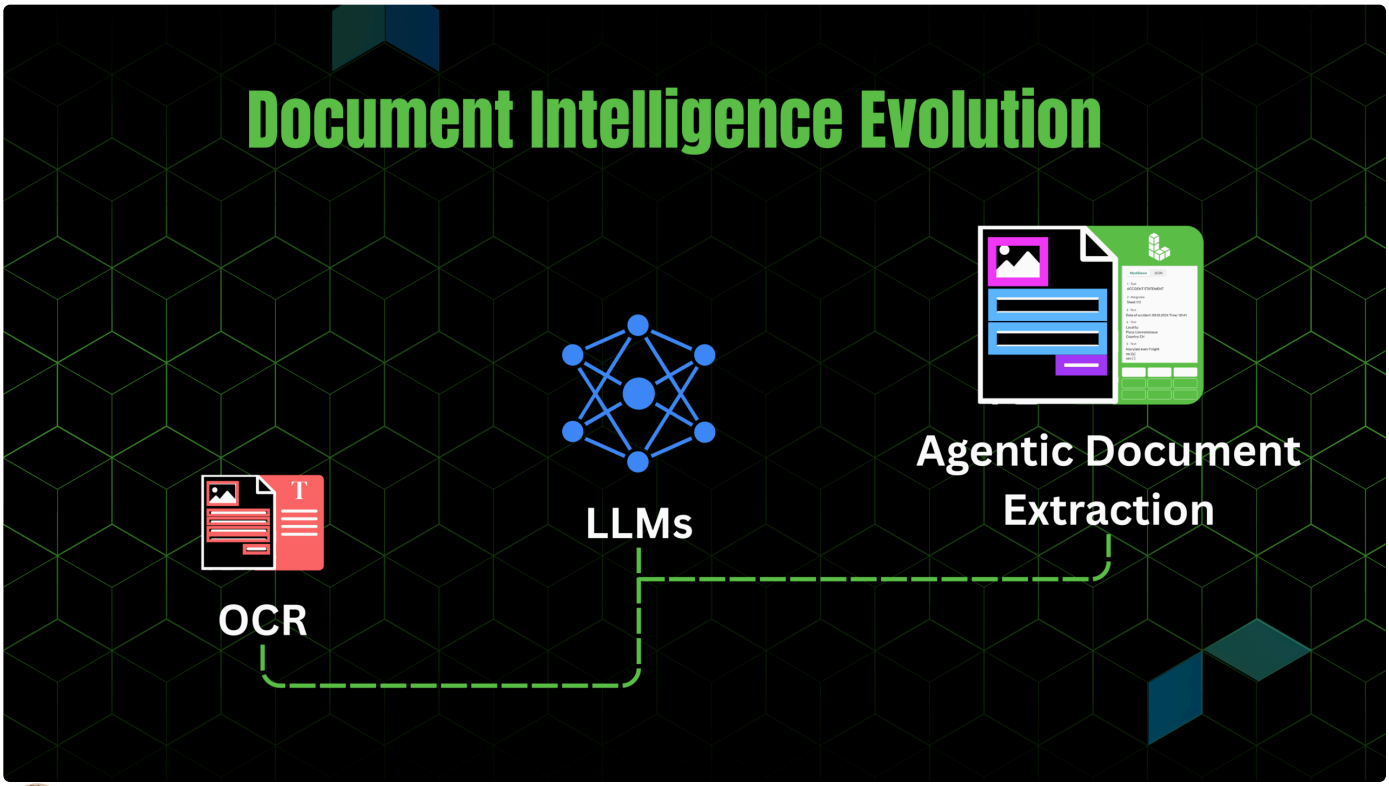
AI

OCR to Agentic Document Extraction: A look into the Evolution of Document Intelligence



Ankit Khare
September 24, 2025

Agentic Document Extraction (ADE) pioneers a new paradigm shift by introducing a truly agentic document understanding system that is **visual ai-first** and built on **data-centric practices**. Accuracy, scale, speed, cost of ownership, and developer-friendliness for downstream integration together enable pragmatic and reliable document processing for high-stakes industries like finance and healthcare. ADE takes on this challenge with its underlying agentic system powered by specialized vision/ML models.

Today's Intelligent Document Processing (IDP) is largely built on top of **LLMs** that rely on tools like OCR to extract meaning from documents. But LLMs fail to traverse the **structural complexities** of documents commonly seen in financial services, healthcare, and legal industries.

In this blog post, we discuss the **evolution of IDP**—from early OCR, to deep learning–based methods, to the rise of modern LLM-based systems—and introduce ADE as the next paradigm.

Agenda

1. **The First Wave — OCR and Its Limits**
How pattern recognition enabled the first era of digitization—and why OCR fell short.
2. **The Second Wave — Statistical, Rule-Based, and Early Deep Learning Methods**
From templates to NLP pipelines to early machine learning models: incremental progress, but still brittle.
3. **The Third Wave — LLMs and the New Challenges for IDP**
The promise of semantic reasoning and multimodality—and the risks of hallucinations, inconsistency, and lack of traceability.
4. **The Fourth Wave — Agentic Document Extraction (ADE)**
A vision-first, agentic, and data-centric paradigm that unifies accuracy, auditability, and developer-friendliness.

The First Wave — The Era of Optical Character Recognition (OCR)

OCR: A Foundational Breakthrough for Digitization

Before intelligent document processing, there was OCR: the first great leap that turned paper into pixels. For its time, it was revolutionary: businesses could finally convert decades of paper archives into machine-readable text. Invoices, receipts, contracts, and lab reports became searchable, indexable, and storable in digital systems.

OCR worked through a deterministic pipeline: image acquisition, enhancement, character segmentation, and pattern recognition. For clean, structured text, it achieved impressive accuracy and jumpstarted digital workflows across industries.

Where OCR Fell Short

But OCR recognized characters; it didn't understand documents. Its design focused on character-level pattern matching created critical blind spots:

- **Poor input quality**: Handwritten text, stylized fonts, or low-resolution scans produced garbled outputs.
- **Loss of structure**: OCR flattened rich layouts into a linear stream of text, discarding context. For example, it could read *"Hemoglobin"* and *"12.5 g/dL"* but failed to connect them as a single test result

in a lab report. In invoices or financial statements, tables and multi-column layouts were stripped of their relationships.

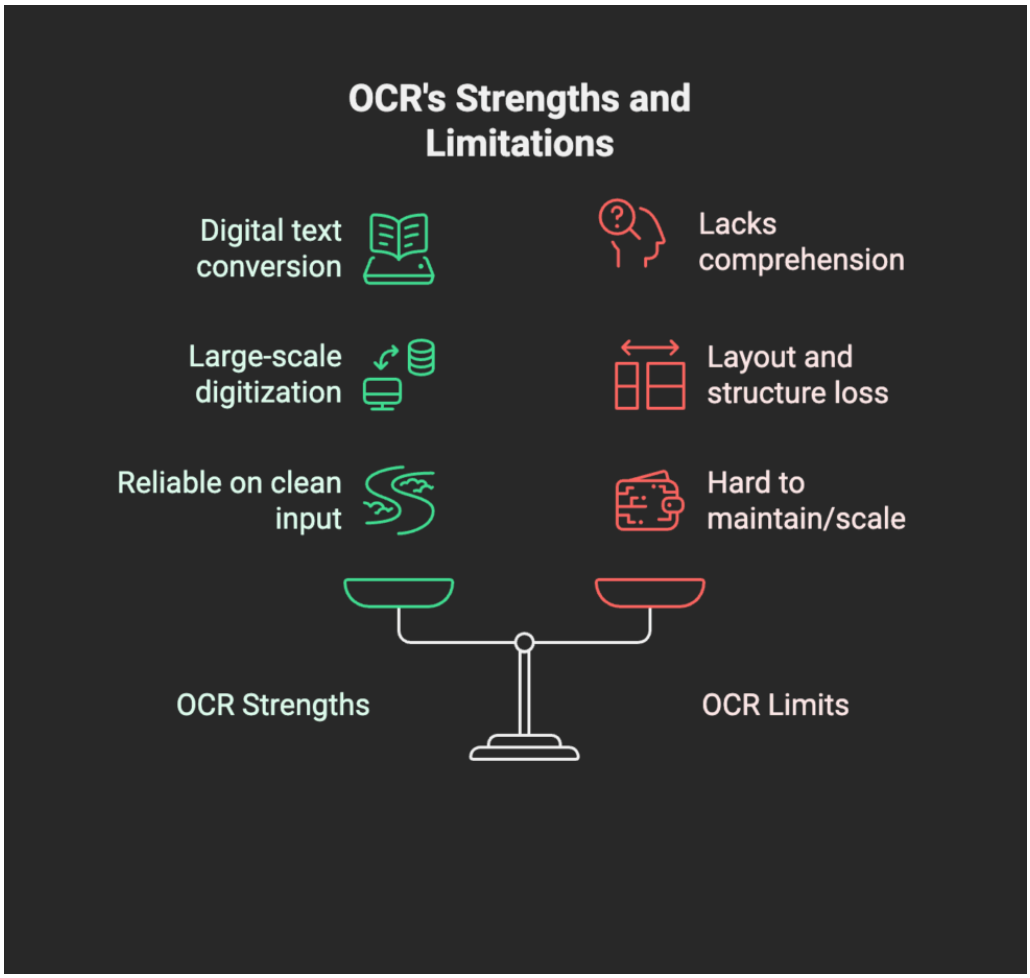
This meant businesses got digitization without comprehension, text was captured, but meaning was lost.

Templates and Rule-Based Patches

To fill the gap, template-based and rule-driven systems emerged. They layered logic on top of OCR, specifying exactly *where to look* for fields on a document. For standardized forms like W-9s, this delivered high accuracy.

But these systems were fragile. Even a minor layout change, for instance – a shifted table column, an updated header, or a new logo, could break the entire pipeline. Maintenance costs soared, and scalability collapsed.

In the end, the first wave solved digitization but left enterprises stuck with brittle, unscalable tools that separated recognition from true understanding.



The Second Wave — Statistical Methods, NLP, and Early Machine Learning

The Search for Comprehension Beyond OCR

OCR solved recognition but not understanding. Enterprises needed systems that could interpret meaning, preserve structure, and adapt across formats. This led to a wave of statistical methods, rule-

based NLP, and early machine learning approaches that tried to bridge the gap.

Statistical & Rule-Based NLP

Early natural language processing pipelines introduced tokenization, part-of-speech tagging, and statistical models such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs). These powered tasks like **named entity recognition (NER)**, enabling extraction of structured elements such as names, addresses, or dates.

However, performance depended heavily on **handcrafted features** and domain-specific lexicons. Systems worked well in narrow domains (e.g., tagging patient names in clinical notes), but scaling across industries or formats was difficult.

Classical Machine Learning

Techniques like **Support Vector Machines (SVMs)** and **decision trees** expanded capabilities into document classification and field extraction. An SVM could distinguish between an invoice, a contract, and a résumé, while classifiers could identify vendor names or payment amounts inside those documents.

Yet, this adaptability came at the cost of **extensive feature engineering**. Generalizing across diverse layouts or unseen document types remained out of reach.

Early Deep Learning Models

By the 2010s, deep learning made its way into document tasks.

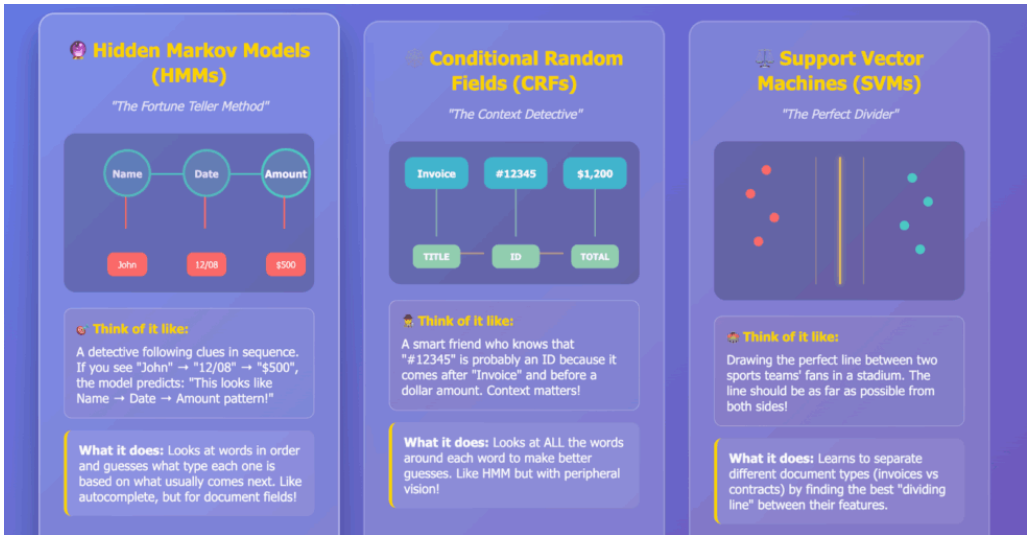
- **CNNs** supported layout analysis and handwriting recognition.
- **RNNs** improved text recognition and sequence tagging.

These advances reduced reliance on handcrafted features but introduced new challenges: the need for **large labeled datasets**, **expensive computation**, and difficulties maintaining **consistency and traceability** in production.

Progress, But Still Patchwork

These methods represented critical steps forward, moving beyond rigid templates and enabling more flexible interpretation of text. They introduced statistical reasoning, feature learning, and higher-level pattern recognition than OCR could achieve.

But they were still **piecemeal solutions**. Each advanced specific subtasks like classification or entity recognition, yet none delivered robust, end-to-end document understanding at enterprise scale. This created the opening for the **third wave: LLMs and VLLMs**, which promised unification of reasoning and multimodal comprehension.



The Third Wave — The Promise of Large Language Models (LLMs) and Vision-Language Models (VLLMs)

A Leap in Semantic Reasoning

LLMs and VLLMs arrived with the promise of comprehension—the missing piece OCR and statistical methods could never deliver. By reasoning over text and images, they brought semantic understanding much closer to human reading. These models could:

- Generalize across varied document layouts.
- Handle linguistic variations (e.g., “Hemoglobin,” “Hb,” and “HGB” as the same concept).
- Perform advanced tasks such as summarizing entire PDFs, answering questions about a document, or interpreting data within tables and mixed layouts.

For developers, this was a breakthrough: rapid deployment with simple prompting, no brittle templates, and minimal fine-tuning.

How LLMs Process Documents

When you upload a PDF or scanned document to an LLM-powered system, the workflow typically looks like this:

1. **OCR (if scanned)** – Images are converted into raw text.
2. **Segmentation** – Text is chunked into blocks (paragraphs, sections), often losing the original layout.
3. **Tokenization** – The text is transformed into tokens the LLM can process.
4. **LLM Reasoning** – The model generates answers, summaries, or extractions.

At no point does the model maintain a **native understanding of layout, spatial relationships, or bounding-box provenance**. This loss of structure is the root of many downstream challenges.

The Production Gap: Inconsistency at Scale

Despite their flexibility, LLMs introduced new risks that blocked reliable enterprise adoption:

- **Probabilistic outputs** – Different runs may produce different answers for the same document.
- **Lack of auditability** – No inherent way to trace extracted values back to their source.

- **Manual overhead** – Developers had to constantly tune prompts and clean outputs to make them usable.

Hallucinations and Factual Inconsistency

One of the most dangerous failure modes is hallucination—outputs that look correct but are subtly wrong.

- **In-context hallucinations**: Contradict the source (e.g., misquoting a metric from a table or altering a financial figure).
- **Extrinsic hallucinations**: Introduce entirely new, unverifiable information (e.g., guessing an invoice number).

Unlike OCR errors, which are often obvious and consistent (e.g., garbled text), LLM errors are **plausible and hidden**, making them far harder to detect at scale in high-stakes industries.

The Challenge of Structured Output and Traceability

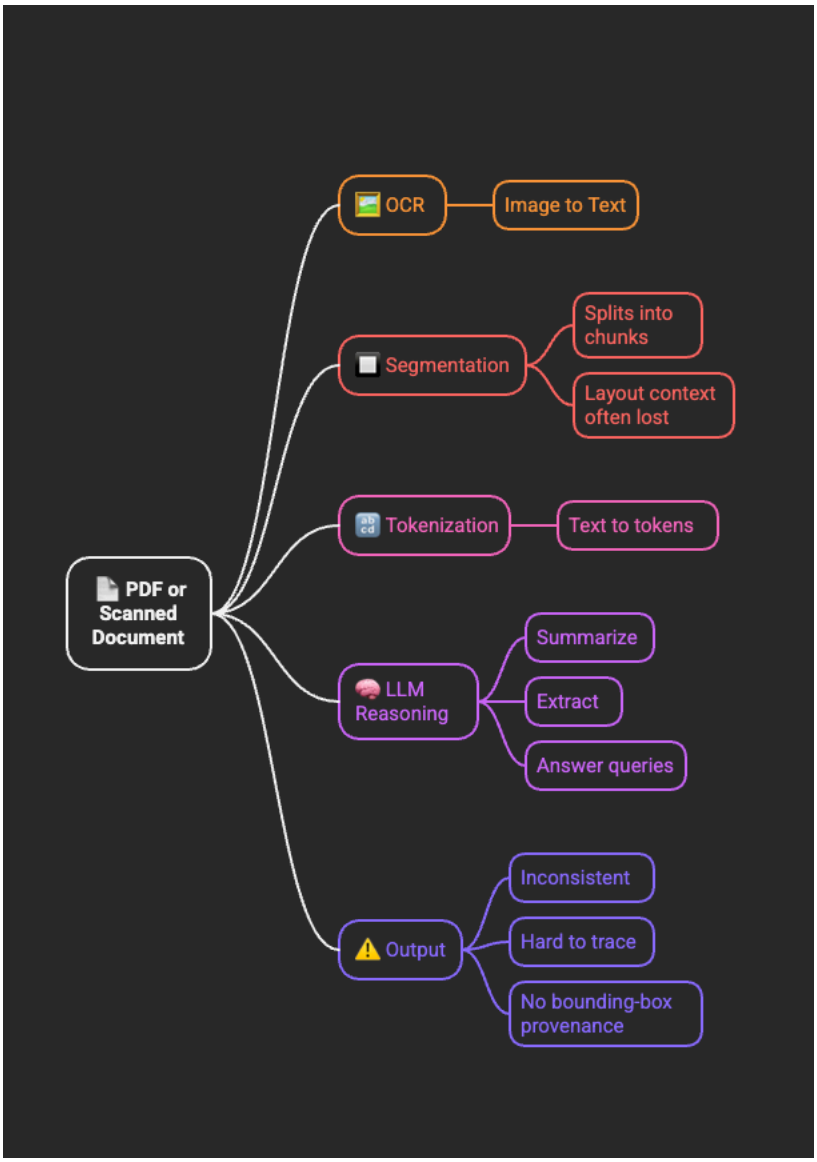
Enterprises require structured, schema-aligned data. LLMs, however, generate **free-form text** that often:

- Fails to match schema requirements.
- Produces inconsistent formats across runs.
- Requires extensive post-processing to normalize.

Even more critically, LLMs lack **visual grounding**. They cannot link extracted values back to their precise page location or bounding box. For finance, healthcare, and legal use cases where compliance and auditability are non-negotiable, this gap is unacceptable.

Closing the Wave

LLMs and VLLMs extended document intelligence with semantic reasoning and multimodality. But without consistency, structure, or traceability, they fell short as production-ready solutions for enterprise IDP. The next step required combining deep understanding with verifiable, layout-aware grounding, ushering in the Fourth Wave: **Agentic Document Extraction (ADE).**



The Fourth Wave — Agentic Document Extraction (ADE): The Next-Generation Paradigm

ADE: A Fundamentally Different Approach

Agentic Document Extraction (ADE) represents a new paradigm for document intelligence. Rather than treating a document as a flat stream of text or relying on brittle templates, ADE treats the document as a **visual representation of information**. This is why we call ADE, **Visual AI-first**.

ADE rests on three pillars:

- **Visual AI-first** — Documents are parsed as visual representations, preserving context, layout, and relationships.
- **Agentic AI** — The system can plan, decide, and act to extract high-confidence data by orchestrating parsing logic, specialized vision/ML models, and an LLM that sequences steps, calls tools, and verifies outputs.
- **Data-Centric AI** — Specialized vision/ML models are trained on high-quality datasets curated with systematic data-centric practices. Model misses in production trigger a structured loop of review, annotation, and dataset updates—ensuring predictable, reliable improvement over time.

LandingAI's years of R&D in Visual AI/Computer Vision, combined with Andrew Ng's supervision in building agentic AI frameworks and establishing a data-centric regime, culminated in the creation of ADE.

Agentic Document Extraction API

ADE is offered as an API that extracts, enriches, and processes complex documents without layout-specific training, turning unstructured data into actionable insights in seconds.

It delivers three core capabilities:

- **Structured JSON output** — A clean hierarchical structure across different file types like PDFs and images.
- **Schema-driven field extraction** — Define fields (e.g., date, vendor name, invoice items) and ADE parses into that schema zero-shot, eliminating inconsistency.
- **Visual grounding** — Every piece of information is tied back to precise coordinates in the original document, creating an auditable trail for compliance and trust.

Schema-First Extraction: Eliminating Ambiguity and Ensuring Consistency

Field extraction begins with a **schema**. You define the exact fields that matter to your workflow—e.g., patient_name, hemoglobin_value, or invoice_total—and ADE parses every incoming document into that precise, structured format.

- Eliminates the **output inconsistency** common with unconstrained LLM outputs.
- Enforces **type and format consistency** across diverse templates.
- Drastically reduces the need for **manual cleanup**.

Developers can apply schemas **zero-shot**, enabling reliable extraction across heterogeneous documents without custom training.

Layout-Aware Parsing: Preserving Context and Relationships

ADE's layout-aware parsing overcomes OCR's fundamental failure to understand structure. The system recognizes **sections, headers, tables, and form fields**, ensuring values are extracted in their proper context.

Example: A number like *"12.5 g/dL"* is not just text; ADE understands it as a **hemoglobin value** inside a results table, associated with its label and position.

This layout-aware parsing enables accurate extraction from:

- Complex tables
- Multi-column layouts
- Visual elements

—without brittle, template-based rules. By preserving relationships, ADE captures the **true meaning** behind document data.

Visual Grounding: The Cornerstone of Trust and Auditability

Visual grounding is one of ADE's **core differentiators**, directly addressing the lack of traceability in LLM-based solutions. Every extracted value is tied back to its **exact source location** in the original document:

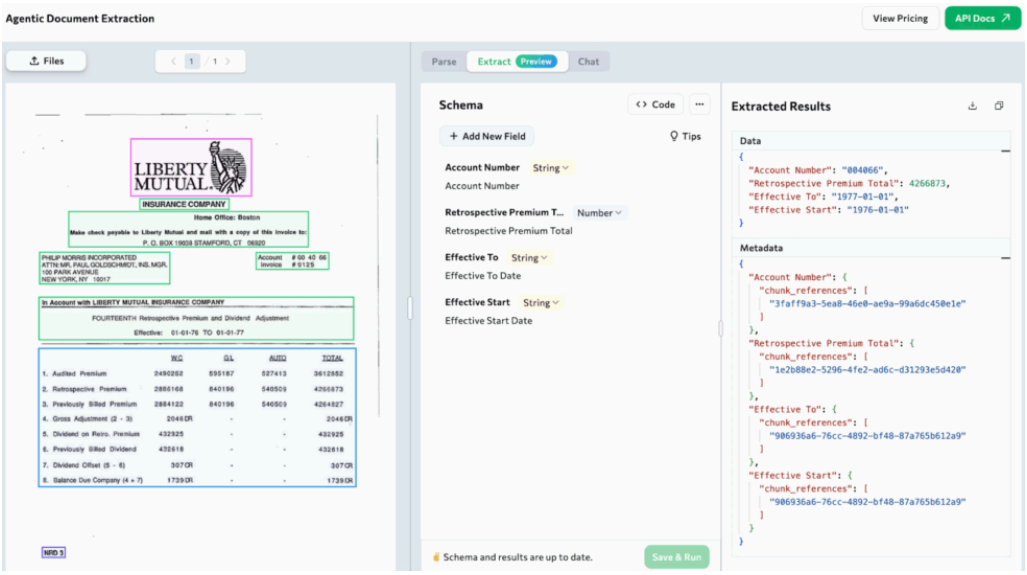
- Specific page

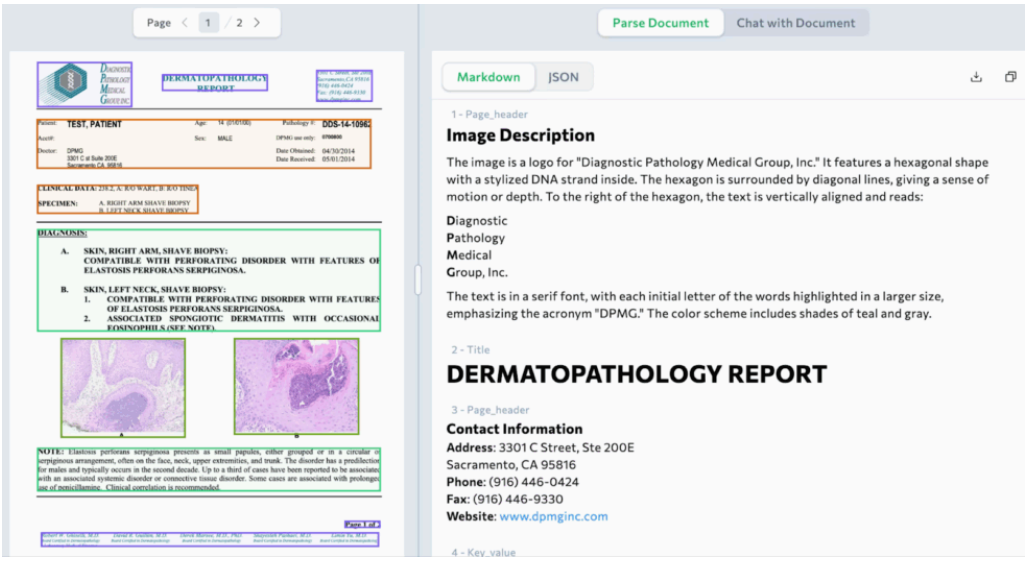
- Bounding box coordinates
- Cropped image snippet

This ensures:

- **Transparency** — users instantly verify where data came from.
- **Explainability** — every value is visually backed by evidence.
- **Trust** — hallucinations are eliminated through verifiable grounding.

By combining deep AI understanding with a **deterministic, auditable framework**, ADE brings enterprise-grade reliability to document intelligence.





For example, below is the JSON output for a `text` chunk. The `grounding` object indicates that the text is on the first page, and the `box` object indicates the bounding box coordinates.


```
{
      "text": "## INSURANCE COMPANY",
      "grounding": [
        {
          "box": {
            "l": 0.35,
            "t": 0.22619999999999998,
            "r": 0.565,
            "b": 0.24033749999999998
          },
          "page": 0
        }
      ],
      "chunk_type": "text",
      "chunk_id": "9475461e-0686-4b16-b503-ccec7d7f115c"
    }
```


Conclusion and the Road Ahead: Towards Agentic Document Intelligence

A NeurIPS Outstanding Paper award recipient, *Are Emergent Abilities of Large Language Models a Mirage?* (Rylan Schaeffer, Brando Miranda, Sanmi Koyejo), studied emergent properties of LLMs and concluded:

"… emergent abilities appear due to the researcher's choice of metric rather than fundamental changes in model behavior with scale. Specifically, nonlinear or discontinuous metrics produce apparent emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance."

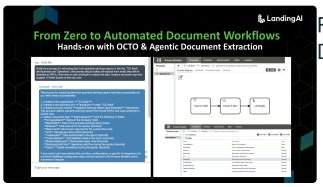
Public perception often experiences discontinuities—when many people suddenly become aware of a technology, it looks like a breakthrough. In reality, growth in AI capabilities is more **continuous and incremental** than it may appear.

That is why we expect the path of Intelligent Document Processing (IDP) to follow a similar pattern: **step by step improvements, each building on the last**. OCR digitized text and made it accessible. Template-based methods, statistical NLP, classical machine learning, and early deep learning techniques added layers of interpretation but remained brittle. LLMs and VLLMs extended this trajectory further with semantic reasoning and multimodal understanding, allowing more flexible interpretation of documents.

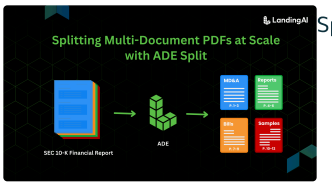
ADE now establishes the next paradigm, moving IDP toward structured, verifiable, and auditable data that enterprises can depend on. Its roots are firmly grounded in being **Visual AI-first, Agentic, and Data-Centric**. Looking ahead, the trajectory points to an **increasing spectrum of "Agentic"** in ADE's internal workflows—systems that don't just parse documents but plan, verify, and continuously improve as part of an ongoing evolution in document intelligence.

Table Of Content

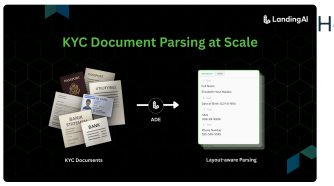
Related Posts



From Zero to Automated Document Workflows: Hands-on with OCTO & Agentic Document Extraction



Splitting Multi-Document PDFs at Scale with ADE Split



How to Automate KYC Document Parsing at Scale with ADE

LandingAI's cutting-edge software platform makes computer vision easy for a wide range of applications across all industries

Headquarters

400 Castro St., Suite 600,
Mountain View, CA 94041
support@landing.ai

Industries

Financial Services

Healthcare

Resources

Blogs

Webinars & Events

Discord

Circle Community

Youtube

Product

Agentic Document Extraction

LandingLens

Snowflake Native Apps

Company

About Us

Careers

Partners

Trust

Security & Compliance

Trust Center