PDF Extraction Complexities:
Layout and Structure

- Inconsistent Layout: Documents as PDFs often have inconsistent layout, which means that the content is positioned in a nondeterministic manner, making it hard to programmatically extract text, tables, or images in a way that retains the original meaning or structure.
- Lack of Structural Metadata: PDFs frequently lack explicit structural metadata, such as headings, paragraphs, or the semantics of tables. This absence complicates the identification of different content types and the understanding of their hierarchy and interrelationships.
- Complex Layouts and Formats: Documents with multi-column layouts, footnotes, sidebars, or tables pose additional challenges. Typically, extraction tools may strip away context during the process, diminishing the usefulness of the data retrieved.

Content Quality and Types

- Quality Issues: When dealing with scanned PDFs, the quality of the scan can significantly affect the ability to extract text. Poor quality scans may result in OCR errors, missing characters, or incorrect text, requiring manual correction or sophisticated error-handling algorithms.
- Multimodal Content: PDFs can contain a mix of text, images, graphics, and sometimes even multimedia elements. Extracting data from these varied content types requires different approaches and technologies, complicating the extraction process.

Security and Encryption

- Security Features & Encrypted Content: Some PDFs contain embedded or encrypted content, which can require additional steps to access and decode before extraction is possible. PDFs can have security settings that restrict copying, printing, or editing of the document. These features can prevent extraction tools from accessing the content unless appropriate permissions are provided or bypassed.

⚠️ **Issue**	📄 **Applies To**	🧠 **Description**
No semantic structure	Native + Scanned	PDFs don’t store headings, paragraphs, or tables semantically—just positioned text/graphics.
Absolute positioning	Native	Content is stored as X-Y coordinates; there's no flow or DOM-like structure.
Encoding inconsistencies	Native	Fonts may use custom encodings or missing glyph mappings (e.g. ligatures, symbol fonts).
Multi-column layouts	Native + Scanned	Difficult to infer logical reading order across columns.
Embedded images	Native + Scanned	Text inside images needs OCR; images mixed with text confuse structure.

Vector graphics or charts	Native	Not easily parsed—these are drawings, not data.
Tables without borders	Native + Scanned	Tables might be represented as just spaced text (no lines/cells).
Hyphenated/line-broken text	Native + Scanned	Lines split across rows (e.g. "informa-\ntion") must be intelligently rejoined.
Rotated/skewed text	Native + Scanned	Text at angles (e.g. vertical headers, rotated page scans) requires special handling.
Page headers and footers	Native + Scanned	Repeating noise across pages; must be removed to avoid duplication.
Scanned documents	Scanned only	Must first convert image to text using OCR (adds noise, can misread).
Low resolution or noisy scans	Scanned	OCR accuracy drops if the image is blurry, compressed, or skewed.
Math equations	Native + Scanned	Often laid out as images or special fonts—hard to parse or understand.
Forms (checkboxes, fields)	Native + Scanned	Form fields are visual; mapping input areas to semantic labels is non-trivial.
Language issues	Both	Mixed languages, RTL scripts (e.g., Arabic), or special symbols are challenging.
Watermarks and annotations	Both	Appear in content layer or separate annotation layer; hard to distinguish.
Page reordering	Native + Scanned	Some PDFs store pages out of logical order (e.g., booklet layouts).
Password protection	Native	Prevents extraction unless decrypted.

Feature	Native PDF	Scanned PDF
Embedded text	✅ Yes	❌ No (must OCR)
Accurate fonts/layout	⚠️ Sometimes	❌ No
Tables (structured)	⚠️ Unreliable	❌ No (OCR might work)

Images/Graphics	✅ Yes	✅ Yes
OCR required	❌ No	✅ Yes
Multi-language support	⚠️ Complex	⚠️ OCR must support

Problem Area	Best Tools / Models
OCR of scanned PDFs	PaddleOCR, Tesseract, LLM Whisperer
Table detection	Docling, LayoutLMv3, TableNet
Complex layouts	Docling, Unstructured.io, VisionParse
Form parsing	DeepForm, LayoutLM
Math equation parsing	Nougat
Image-text extraction	VisionParse, LLMs w/ vision