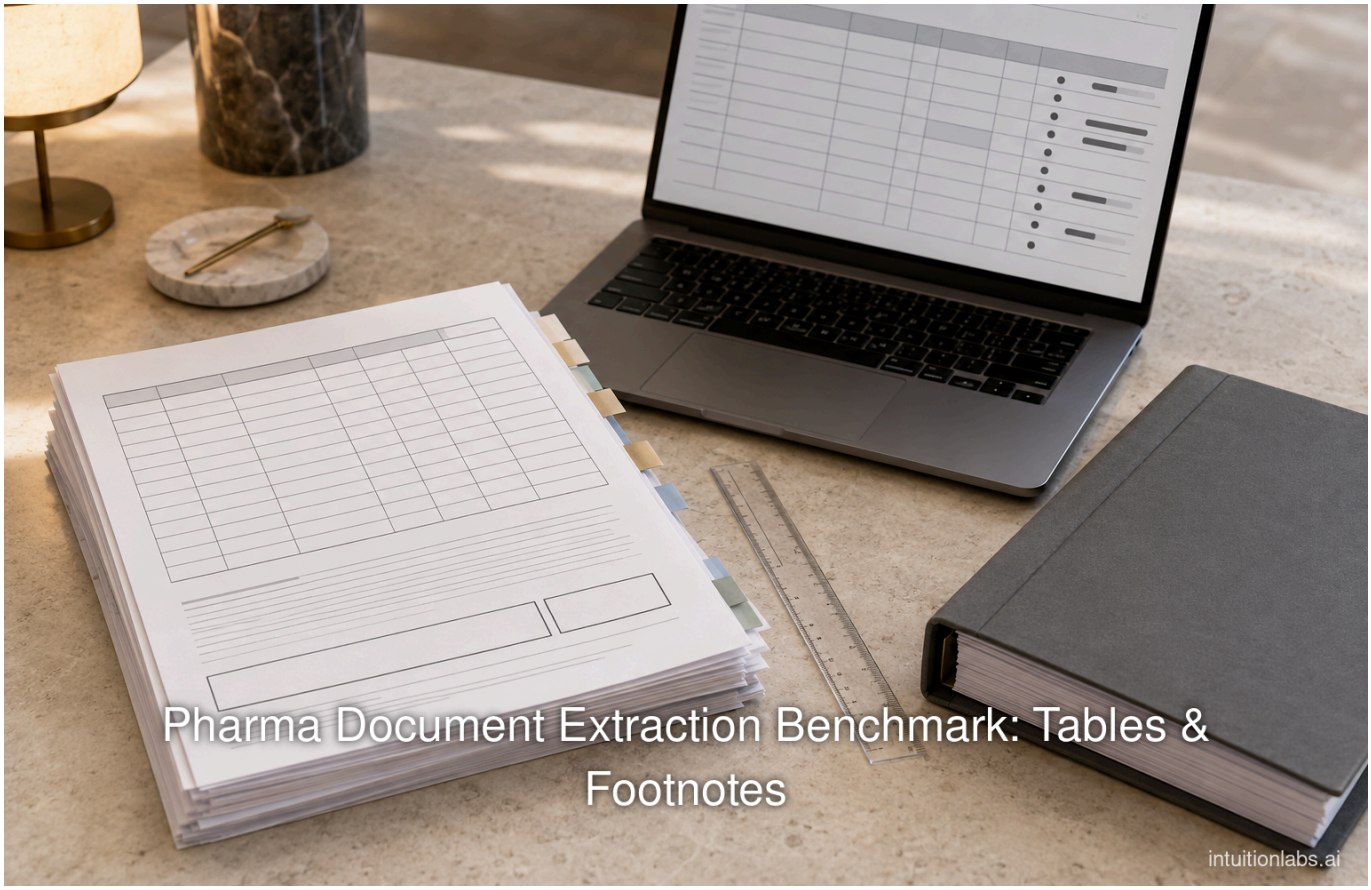
Pharma Document Extraction Benchmark: Tables & Footnotes

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · pharma document extraction benchmark · table extraction · footnote extraction · document ai · ocr accuracy · clinical trial data extraction · llm data extraction · unit normalization · life sciences ai tools



Executive Summary

This report examines what is and is not measurable, as of September 2026, when evaluating artificial intelligence (AI) tools for extracting structured data, tables, footnotes, and individual fields, from pharmaceutical documents such as clinical trial publications, FDA Structured Product Labels, and regulatory submissions. No single, independently audited benchmark tests table extraction, footnote handling, missing-value treatment, and unit normalization together on a shared corpus of real pharma documents. Instead, the evidentiary base is a patchwork: general-purpose layout datasets, LLM table-reasoning benchmarks, vendor-run comparisons, and domain-specific clinical and regulatory extraction studies, each using its own metric and test set.

On table structure specifically, IBM Research's DocLayNet dataset, 80,863 human-annotated pages across 11 layout classes (raw.githubusercontent.com), shows trained object-detection models trailing human inter-annotator agreement by roughly 10 percentage points (ar5iv.labs.arxiv.org). On pure table-reasoning tasks, the TableBench and "Table Meets LLM" academic benchmarks found that even GPT-4-class models score well below human

performance, with the best LLM configuration in one study reaching only 65.43% overall accuracy on structural table-understanding tasks (arxiv.org). Reducto's vendor-run RD-TableBench, by contrast, reported a 90.2% average table-similarity score for its own extraction model, the top result among seven providers it tested (reducto.ai), a figure this report treats as a vendor claim rather than an independently verified result.

Footnotes remain a documented weak point: DocLayNet's own annotators agreed on footnote boundaries at only 83 to 91 mean Average Precision (ar5iv.labs.arxiv.org), and the 2024-2025 OmniDocBench benchmark separately flags footnotes as a source of reading-order error (arxiv.org). On clinical trial documents specifically, a 2026 study comparing OpenAI, Anthropic, Google, and Meta models on 67 trial documents found mean field extraction accuracy of 93.7% to 98.9% (www.ncbi.nlm.nih.gov), though simple binary fields reached near-100% agreement while complex categorical fields performed far worse (www.ncbi.nlm.nih.gov). On FDA labels, a 2018 annotation project (SPL-ADR-200db) recorded human inter-annotator F-scores of 74.9% to 85.5% (pmc.ncbi.nlm.nih.gov), and a 2021 DistilBERT classifier reached an F1 of 0.9607 on label text classification (pmc.ncbi.nlm.nih.gov). No dedicated benchmark for unit normalization in pharma extraction was found; the closest analogue, a materials-science LLM extraction study, reported close to 90% precision and recall (www.ncbi.nlm.nih.gov).

On pricing, as of September 2026, per-page costs for enterprise document-extraction APIs span roughly $0.0006 to $0.070 per page depending on the provider and the complexity of extraction requested (Amazon Textract, Google Document AI) (aws.amazon.com) (aws.amazon.com), while newer OCR-focused models like Mistral's OCR 4 price at $2 to $4 per 1,000 pages (mistral.ai). The global AI-in-life-science market is projected to grow from $21.58 billion in 2026 to $69.34 billion by 2031 (www.marketsandmarkets.com), a backdrop against which this fragmented benchmark landscape is likely to consolidate only gradually. Readers evaluating any extraction tool for regulated pharma use should treat every accuracy figure as conditional on its metric, test set, and vendor-versus-independent status.

Introduction and Background

Pharmaceutical organizations generate enormous volumes of unstructured and semi-structured documents: clinical study reports, drug labels, package inserts, batch records, and regulatory submissions built around the electronic Common Technical Document (eCTD) format that the U.S. Food and Drug Administration (FDA) requires for applications, amendments, supplements, and reports submitted to its Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER) (www.fda.gov). In parallel, the European Medicines Agency (EMA) is rolling out ISO Identification of Medicinal Products (IDMP) standards across four master-data domains, substance, product, organisation, and referential (SPOR), specifically to make regulatory data interoperable across systems and institutions (www.ema.europa.eu) (www.ema.europa.eu). Both requirements push the industry toward machine-readable, structured data rather than free-text PDFs, which is exactly the gap that AI-based document extraction tools are built to close.

"Document extraction benchmark," as used in this report, refers to a published, reproducible test in which a tool or model's output on a document (a table, a form, a footnote, a labeled field) is scored against a human-annotated ground truth using a defined metric, not a vendor's unverified accuracy percentage. This distinction matters because most publicly available accuracy claims are either academic benchmarks built on general (often scientific-article) documents, or vendor-run benchmarks on vendor-selected test sets; genuinely pharma-specific, third-party-audited extraction benchmarks remain rare as of September 2026.

A related IntuitionLabs report, "AI OCR Models: PDF-to-Structured-Text Comparison," previously compared a broad set of OCR and document-AI models on their capability to convert PDFs into structured text (intuitionlabs.ai), including the observation that classical OCR engines such as Tesseract do not inherently preserve complex layout structures like tables or forms (intuitionlabs.ai). This report does not restate that comparison. Instead, it focuses narrowly on what is separately and independently measurable: table-structure extraction accuracy, footnote handling, missing-value treatment, and field-level accuracy on clinical and regulatory document types, anchored to the specific academic benchmarks and vendor disclosures available as of September 2026.

Methodology: Benchmark Datasets, Documents, and Metrics

This report synthesizes evidence rather than running a new benchmark from scratch. Every figure below was traced to a specific published dataset paper, an open-source benchmark repository, or a vendor's own current pricing or accuracy-disclosure page, fetched directly during research for this report. Three categories of source material were used, and each is labeled accordingly throughout.

- **Academic layout and table-structure benchmarks**: general-purpose document datasets such as **DocLayNet** (IBM Research, 80,863 human-annotated pages across 11 layout classes including Table and Footnote, published at KDD 2022) (raw.githubusercontent.com), **PubLayNet** (over 360,000 document images built automatically from more than one million PubMed Central PDF/XML article pairs) (arxiv.org), and **PubTables-1M** (947,642 tables extracted from scientific articles, 52.7% of which are structurally complex) (arxiv.org) (arxiv.org) and **FUNSD** (199 fully annotated real-world forms used for spatial layout and form-understanding evaluation) (guillaumejaume.github.io) (guillaumejaume.github.io).
- **Table- and field-reasoning benchmarks for large language models (LLMs)**, such as **TableBench** and the "Table Meets LLM" study, which test whether general-purpose LLMs (GPT-4, GPT-3.5, GPT-4o) can correctly answer questions about or reconstruct the structure of tabular data, scored against human performance on the same tasks (arxiv.org).
- **Domain-specific extraction studies**, run against real clinical trial publications, regulatory drug labels, or case reports, including **ExaCT** (randomized controlled trial characteristic extraction) (pmc.ncbi.nlm.nih.gov), **SPL-ADR-200db** (200 FDA Structured Product Labels annotated for adverse drug reactions) (pmc.ncbi.nlm.nih.gov), and a 2026 multi-model comparison of large language models extracting parameters from 67 clinical trial documents.

Assumptions and limits of this method should be stated plainly. First, no benchmark identified during this research tests extraction tools against a shared, purpose-built corpus of pharma-specific documents (clinical study reports, SmPC labels, batch records) using a single standardized metric; the closest analogues are FDA-label-based datasets like SPL-ADR-200db and case-report datasets like CaseReportBench, described in the sections below. Second, vendor-published benchmarks (Reducto's RD-TableBench, Mistral's OCR 4 comparison against OlmOCRBench) disclose their methodology and, in Reducto's case, open-source the scoring code and dataset (github.com) (raw.githubusercontent.com), but the vendor still selects which competing tools to include and how tests are run, so these figures are labeled as vendor claims throughout, not independent audits. Third, benchmark scores computed with different metrics, exact cell match, hierarchical alignment similarity, or micro-F1, are not directly comparable to one another; this report keeps each metric attached to its source rather than normalizing across benchmarks.

Table Structure Extraction: Benchmark Accuracy Across Models and Tools

Table extraction, correctly identifying rows, columns, merged cells, and header hierarchy, then reading the text inside each cell, remains one of the hardest sub-tasks in document AI. *Table 1* below summarizes the major published benchmarks used to measure it.

Table 1: Table-structure and layout-extraction benchmarks referenced in this report

Benchmark	Scale	Document domain	Metric	Headline result (as reported)
DocLayNet (IBM Research, KDD 2022)	80,863 annotated pages, 11 layout classes (raw.githubusercontent.com)	Mixed: financial reports, scientific articles, patents, manuals, laws, tenders	Mean Average Precision (mAP) vs. human inter-annotator agreement	Trained object-detection models trail human agreement by roughly 10 percentage points overall
PubTables-1M	947,642 tables, 52.7% structurally complex (arxiv.org)	Scientific articles	Content accuracy (AccCont): exact cell-text match across the whole table	Defines full-table exact-match as the strictest correctness bar for structure recognition
RD-TableBench (Reducto, Nov. 2024, vendor-run, open methodology)	1,000 PhD-labeler-annotated complex tables (scanned, handwritten, merged cells, multilingual) (github.com)	Diverse real-world documents	Table similarity via hierarchical Needleman-Wunsch alignment (raw.githubusercontent.com)	Reducto's own model reported an average table similarity of 90.2%, the top score among seven providers it tested (reducto.ai), with Azure and AWS outperforming most newer entrants and GPT-4o following behind (reducto.ai)
TableBench	Multi-task table question-answering suite	General tabular data	Automated accuracy metric vs. human baseline	GPT-4, the strongest model tested, still achieved only "a modest score compared to humans" (arxiv.org)
"Table Meets LLM" benchmark	Seven structural table-understanding tasks	General tabular data	Overall task accuracy	Best LLM configuration (HTML-serialized tables with role prompting) reached 65.43% overall accuracy; removing in-context examples cut accuracy by 30.38 percentage points

The pattern across independent, non-vendor benchmarks is consistent: general-purpose LLMs (GPT-4, GPT-4o) perform well below both specialized table-structure models and human baselines on pure structure-recognition and structural-reasoning tasks, while vendor-run benchmarks on curated real-world tables report considerably higher similarity scores for tools purpose-built for extraction (reducto.ai). IBM Research's own DocLayNet-adjacent evaluation found that a specialized object-detection model (YOLOv5x6) can exceed human inter-annotator agreement on some individual layout classes even while trailing on the aggregate benchmark, illustrating that structure detection and content-accuracy are related but distinct measurement problems. Because AccCont on PubTables-1M requires every cell in a table to match exactly, the metric is intentionally unforgiving of partial errors, which is why Reducto's RD-TableBench instead uses a partial-credit alignment score (raw.githubusercontent.com), and the two numbers should never be quoted interchangeably.

Footnotes, Missing Values, and Field-Level Completeness

Footnotes are a persistent failure point in document extraction because they sit outside a table's grid but semantically qualify specific cells or rows, an association most extraction pipelines do not explicitly model. DocLayNet treats "Footnote" as one of its eleven core layout classes, alongside Caption, Table, and Section-header, annotated across the same 80,863-page, multi-annotator corpus used for its table and layout figures. Human annotators agreed with each other on Footnote boundaries at a mean Average Precision of 83 to 91 across the dataset's document categories, an agreement band that is useful as a ceiling: no automated system evaluated against DocLayNet-style ground truth should be expected to exceed roughly this range, since the annotators themselves did not agree perfectly. A separate 2024 to 2025 document-parsing benchmark, OmniDocBench, explicitly flags footnotes (along with figure and table captions) as a source of reading-order ambiguity, since their placement on the physical page is often inconsistent relative to the content they annotate.

Missing and absent values pose a different but related challenge: an extraction system must distinguish "this field was not reported in the source document" from "this field exists but the model failed to find it," and scoring rules must not reward a model for silently omitting a hard field. Recent schema-guided extraction research addresses this directly. One 2026 biomedical PDF-extraction pipeline instructs its underlying model explicitly not to infer or guess, and to set a field to null whenever the source information is ambiguous or absent (arxiv.org). The open-source ExtractBench framework enforces this at the scoring level: an omitted key is scored as an explicit null, every scalar field counts in both the numerator and denominator of the accuracy calculation, and a correct null on a genuinely blank field is credited as a correct answer rather than ignored (github.com). A complementary metric, field completeness rate (FCR), defined in a 2026 clinical-research LLM extraction study as the proportion of required fields that are non-empty and valid, gives a way to separately track whether a pipeline is even attempting to populate a field, independent of whether the value it produces is correct (pmc.ncbi.nlm.nih.gov); the same study reported an overall extraction accuracy near 89.5% and recall near 85.3% for its LLM-based medication-extraction task (pmc.ncbi.nlm.nih.gov).

Real-world field coverage in clinical documents is highly uneven, which is precisely why missing-value handling matters more in this domain than in general text extraction. CaseReportBench, a 138-case dense information-extraction benchmark built from real published case reports, found that some clinical categories (laboratory and imaging findings, patient history) were annotated in 98.55% of cases, 136 of 138, while a narrow clinical category,

lymph node findings, was present in only 1.45% of cases (arxiv.org) (arxiv.org). Any extraction pipeline evaluated only on high-coverage fields will look far more accurate than one evaluated honestly across the full, sparse field schema a real clinical document implies.

Clinical Trial and Regulatory Document Field Extraction

Extracting structured fields from clinical trial publications and regulatory drug labels has a longer research history than general-purpose LLM table extraction, and the reported accuracy numbers are correspondingly more mature. ExaCT, a 2010 system for automatically extracting trial characteristics (sample size, interventions, outcomes) from full-text randomized controlled trial (RCT) publications, combined a sentence-classification stage with rule-based fragment extraction; its sentence-retrieval stage achieved 88% top-5 recall and 80% top-1 precision (pmc.ncbi.nlm.nih.gov), and its downstream extraction rules reached 93% precision and 91% recall (pmc.ncbi.nlm.nih.gov). Across a 1,050-task end-to-end test set, the system produced fully or partially correct answers on 992 tasks, a 94% task-level success rate.

More recent work applies large language models to the same problem at greater scale. A 2026 study comparing OpenAI, Anthropic, Google, and Meta models on structured-parameter extraction from 67 clinical trial documents found mean extraction accuracy of 93.7% to 98.9% against an expert-curated gold standard across all tested features. That aggregate figure hides sharp variation by field type: simple binary classification fields, such as whether a trial involved non-invasive intervention, reached near-perfect inter-model agreement (100% and 99.3%, respectively), and numeric fields such as stimulation duration showed some models (OpenAI's o4-mini-high and Anthropic's Claude Sonnet 4) reaching perfect agreement with the reference value across 19 comparisons, while more complex categorical fields, such as anatomical targets, showed much lower agreement under strict string-matching rules. This is the central lesson for anyone evaluating LLM extraction on clinical documents: aggregate accuracy figures can conceal large field-by-field variance, and a tool that performs well on yes/no fields is not thereby validated for open-ended categorical or numeric fields.

On regulatory labeling specifically, a broader evidentiary base exists because FDA-published Structured Product Labels (SPLs) are public. SPL-ADR-200db, a dataset of 200 real SPLs manually annotated for adverse drug reactions (ADRs), contains 5,098 distinct annotated ADR mentions, and its own human annotators improved from a 74.9% average inter-annotator F-score on the first 100 labels to 85.5% on the second 100 as they gained experience with the annotation task, a useful reminder that human-labeled "ground truth" itself has a measurable error rate that automated systems are being compared against. A 2021 study applying a DistilBERT transformer model to FDA drug-labeling text classification reported precision of 0.9649, recall of 0.9565, and an F1 score of 0.9607 (pmc.ncbi.nlm.nih.gov), outperforming the rule-based and traditional machine-learning baselines it was compared against in that same study.

On the question of unit normalization specifically, converting equivalent expressions such as milligrams per milliliter, international units, or millimoles per liter into a single canonical form, no dedicated, published benchmark quantifying this sub-task in a pharma extraction context was located as of September 2026. The closest available quantified analogue is a 2024 Nature Communications study (the ChatExtract method) that measured LLM-based numeric and scientific data extraction from materials-science research papers, finding that the best-performing conversational LLMs, led by GPT-4, achieved precision and recall both close to 90% (www.ncbi.nlm.nih.gov). Because that figure comes from a different scientific domain, it should be read as a general indicator of LLM numeric-extraction capability, not as a pharma-specific or unit-normalization-specific accuracy claim.

Vendor Landscape and Pricing: Comparative Positioning

Enterprise document-extraction tools relevant to pharma workflows span general-purpose cloud OCR services, specialized document-AI platforms, and startups focused specifically on table and layout parsing. *Table 2* summarizes current, vendor-published pricing and accuracy disclosures as of September 2026; all figures are vendor claims unless otherwise noted, and all prices should be re-verified against the live pricing page before use in a procurement decision, since cloud vendor pricing changes frequently.

Table 2: Vendor-published pricing and accuracy claims for document/table extraction tools (as of September 2026)

Tool	Vendor	Base OCR / text pricing	Table or structured-extraction pricing	Accuracy claim (source)
Amazon Textract	AWS	$0.0015/page for the first 1M pages/month, $0.0006/page above that (aws.amazon.com)	Tables: $0.015/page; Forms + Tables + Queries: $0.070/page (first 1M pages) (aws.amazon.com) (aws.amazon.com)	No public accuracy percentage disclosed on the pricing page; see AWS Bedrock Data Automation note below
Google Document AI	Google Cloud	$1.50 per 1,000 pages up to 5 million pages/month, $0.60 per 1,000 above that, for Enterprise Document OCR (cloud.google.com)	Layout Parser: $10.00 per 1,000 pages; Form Parser: $30.00 per 1,000 pages below 1M pages (cloud.google.com)	No public benchmark accuracy percentage disclosed on the pricing page
Mistral OCR 4	Mistral AI	$4 per 1,000 pages via standard API, $2 per 1,000 pages via Batch API (mistral.ai)	Same document-parsing pipeline handles tables and layout	Vendor reports the top overall score among tested models on the public OlmOCRBench (85.20), while itself cautioning that OlmOCRBench and OmniDocBench "have known limitations in how they score certain outputs" (mistral.ai) (mistral.ai)
Unstructured.io	Unstructured Technologies	$0.015 per page after the first 10,000 free pages per month (unstructured.io)	Same per-page rate covers structured partitioning	No public third-party accuracy percentage located
LlamaParse / LlamaCloud	LlamaIndex	Credit-based: 1,000 credits = $1.25 (www.llamaindex.ai); per-page cost varies with parsing mode selected	Same credit system	No public third-party accuracy percentage located
Reducto	Reducto AI	Not publicly disclosed in the sources reviewed for this report	Not publicly disclosed in the sources reviewed for this report	Vendor's own RD-TableBench reports 90.2% average table similarity, the top score among 7 tested providers
ABBYY	ABBYY	Not publicly disclosed in the sources reviewed for this report	Not publicly disclosed in the sources reviewed for this report	Vendor states its pre-trained extraction models achieve "over 90% straight-through processing

Tool	Vendor	Base OCR / text pricing	Table or structured-extraction pricing	Accuracy claim (source)
				(STP) right out of the gate" (www.abbyy.com)

Two points of methodological honesty apply to this table. First, several vendors (Reducto, ABBYY) publish accuracy claims but not standard public per-page pricing in the pages reviewed for this report; enterprise pricing for these tools is typically negotiated directly and was not fabricated here to fill the cell. Second, AWS's Amazon Bedrock Data Automation added a blueprint-instruction-optimization feature in December 2025 that returns evaluation results, including exact-match rates and F1 scores, measured against a customer's own ground truth (aws.amazon.com), but AWS's own announcement discloses no fleet-wide accuracy percentage for the feature, so no such number is asserted here.

Market context helps explain why this vendor landscape is expanding quickly: the research firm MarketsandMarkets projects the global AI in life-science market will grow from $21.58 billion in 2026 to $69.34 billion by 2031, a compound annual growth rate (CAGR) of 26.3% (www.marketsandmarkets.com), a category that includes but is not limited to document extraction and automation tooling. Life-sciences organizations comparing tools should treat this table as a snapshot rather than a static reference, given both the pace of vendor price changes and the rate at which new benchmark results are published.

Data Analysis and Evidence

Table 3 consolidates the field-level and document-level accuracy figures gathered across independent (non-vendor) studies discussed in the sections above, to make cross-study comparison easier while preserving each study's own metric definition.

Table 3: Field-level extraction accuracy across independent studies on clinical and regulatory documents

Study / system	Document type	Metric	Reported score	Year
ExaCT	Full-text RCT publications	Precision / Recall (extraction rules)	93% / 91%	2010
ExaCT	Full-text RCT publications	End-to-end task success	94% (992/1,050 tasks)	2010
SPL-ADR-200db	FDA Structured Product Labels	Inter-annotator F-score	74.9% (first 100 labels) to 85.5% (second 100)	2018
DistilBERT FDA label classifier	FDA drug-labeling text	Precision / Recall / F1	0.9649 / 0.9565 / 0.9607	2021
Multi-model clinical trial extraction study	67 clinical trial documents	Mean field extraction accuracy	93.7% to 98.9%	2026
ChatExtract	Materials-science research papers	Precision / Recall (numeric data)	Both close to 90%	2024

The overall pattern in Table 3 is that extraction accuracy on real clinical and regulatory text clusters in a wide band, roughly 75% to 99%, and the position within that band depends heavily on field type (binary versus categorical versus free numeric), document homogeneity (a narrow SPL schema versus heterogeneous case reports), and how strictly the scoring counts a near-miss. The multi-model 2026 clinical trial study is the most directly relevant to current LLM-based pharma workflows because it tests contemporary frontier models (OpenAI, Anthropic, Google, Meta) rather than earlier rule-based or fine-tuned classifier systems, and its finding that simple fields reach near-100% agreement while complex categorical fields fall much lower is consistent with the layout-benchmark finding that specialized structure-detection models can approach or exceed human agreement on well-defined classes while general table-reasoning tasks remain far from human-level. Two connected data points from the layout literature reinforce this: DocLayNet's own trained models trail human inter-annotator agreement by about 10 percentage points in aggregate, and its Footnote class specifically shows a human agreement band of 83 to 91 mAP, both of which set a realistic ceiling against which any single vendor's higher-sounding accuracy claim should be read.

Where a discrepancy exists between sources, this report has preserved it rather than resolved it in a vendor's favor. Reducto's vendor-run RD-TableBench places its own tool ahead of Azure, AWS Textract, GPT-4o, Google Cloud Document AI, Unstructured, and Chunkr on table similarity, while Mistral's own comparison places its OCR 4 model at the top of a different public leaderboard, OlmOCRBench, and explicitly cautions that both OlmOCRBench and OmniDocBench have scoring limitations. Neither claim has been independently re-verified by a third party in the sources reviewed for this report, and both are reported here as vendor claims, not as settled fact.

Implications and Future Directions

Three practical implications follow from the evidence above. First, buyers evaluating document-extraction tools for pharma use cases should demand the underlying benchmark methodology, not just a headline accuracy number: a 90.2% table-similarity score on RD-TableBench and a "90% straight-through processing" claim from a different vendor (www.abbyy.com) are not comparable without knowing the test documents, the metric, and who selected the competing tools. Second, footnote handling and missing-value treatment deserve explicit evaluation criteria in any procurement process, since general accuracy figures typically describe body-table extraction and can silently exclude footnote-to-cell association and null-field correctness, exactly the areas where DocLayNet and ExtractBench (github.com) show measurement is still maturing. Third, given that no dedicated pharma-specific table-structure or unit-normalization benchmark was located as of September 2026, organizations validating a tool for regulatory or clinical use should expect to build or commission a domain-specific validation set, using the general-purpose metrics described in this report (AccCont-style exact match, alignment-based similarity, field completeness rate) as a starting framework rather than assuming a general benchmark score transfers directly to pharma documents.

Regulatory momentum toward structured, interoperable data, illustrated by FDA's eCTD requirement (www.fda.gov) and EMA's phased ISO IDMP rollout across substance, product, organisation, and referential master data (www.ema.europa.eu), will likely keep increasing the volume of documents that life-sciences organizations need to convert from PDF into validated structured fields, independent of which specific extraction vendor or model architecture eventually dominates. Life-sciences organizations navigating this landscape, particularly where the target data model overlaps with Veeva-based commercial and regulatory systems, sometimes engage outside advisors to translate general-purpose benchmark evidence into a validation plan appropriate for a specific regulated

workflow. IntuitionLabs, a life-sciences-focused technology consultancy that describes itself as an "emerging Silicon Valley firm focused on Veeva CRM consulting, custom software development, and big data solutions for pharmaceutical companies" (intuitionlabs.ai) and positions its work as helping to "transform your operations with our cutting-edge AI solutions designed specifically for pharmaceutical and life science organizations" (intuitionlabs.ai), is one example of this advisory category;

Looking forward, the LLM-versus-specialized-model gap documented in Table 1, where general LLMs trail both human baselines and purpose-built structure-recognition models on pure table-structure tasks, is a reasonable area to watch for rapid change, since multimodal model releases (including the Mistral OCR 4 update referenced above) are iterating quickly enough that a benchmark snapshot from late 2024 or 2025 may already understate current-generation performance by the time this report is read.

Conclusion

As of September 2026, no single, independently audited benchmark measures pharma document extraction end to end, across tables, footnotes, missing values, and unit normalization, on a shared corpus of real regulatory and clinical documents. What exists instead is a patchwork of credible but narrower measurements: general-purpose layout and table-structure datasets (DocLayNet, PubLayNet, PubTables-1M) that establish human-agreement ceilings and structure-detection baselines; LLM table-reasoning benchmarks (TableBench, "Table Meets LLM") that show a persistent gap between general-purpose models and human performance; vendor-run benchmarks (RD-TableBench, Mistral's OlmOCRBench comparison) that are methodologically disclosed but not independently verified; and domain-specific studies on clinical trial publications, FDA labels, and case reports that report accuracy figures ranging roughly from the mid-70s to the high 90s depending on field type and document homogeneity. Reading any single accuracy number from this landscape without its metric, test set, and vendor-versus-independent status attached is the most common way this kind of evidence gets misused. Organizations evaluating extraction tools for regulated pharma workflows should treat every figure in this report, and every figure a vendor presents outside it, as conditional on those three attributes.

Frequently Asked Questions (FAQs)

Is there a standardized pharma document extraction benchmark? Not as of September 2026. General-purpose layout and table-structure benchmarks such as DocLayNet (raw.githubusercontent.com) and PubTables-1M exist, and domain-specific datasets built from FDA labels (SPL-ADR-200db) or clinical trial publications (ExaCT) exist separately, but no single benchmark combines a pharma-specific document corpus with a shared, independently audited extraction metric.

How accurate is AI table extraction on pharma documents specifically? No study identified in this research measured table-structure extraction accuracy on a pharma-specific document set using a standard metric like TEDS or AccCont. The closest general evidence is Reducto's vendor-run RD-TableBench, reporting 90.2% average table similarity on complex real-world tables, and the TableBench academic benchmark, where even GPT-4 scored well below human performance on structural table reasoning.

How well does AI extract data from clinical trial PDFs? A 2026 study comparing frontier LLMs on 67 clinical trial documents reported mean extraction accuracy of 93.7% to 98.9% across tested fields, though accuracy varied sharply by field type, with simple binary fields far more reliable than complex categorical fields. Earlier rule-based systems such as ExaCT reported 93% precision and 91% recall on extraction rules applied to full-text randomized controlled trials.

How do extraction systems handle missing or unreported values? Current best practice, illustrated by a 2026 schema-guided biomedical extraction pipeline, is to explicitly instruct the model not to guess and to return a null value when information is absent (arxiv.org), combined with scoring rules, as used in the open-source ExtractBench framework, that credit a correct null rather than ignoring blank fields (github.com). A related metric, field completeness rate, separately tracks what proportion of required fields a pipeline even attempts to populate (pmc.ncbi.nlm.nih.gov).

Is there a benchmark specifically for footnote extraction? No dedicated, standalone footnote-extraction benchmark was located. Footnote handling is measured as one layout class within broader document-layout datasets, most notably DocLayNet, where human annotators themselves only agreed on footnote boundaries at 83 to 91 mAP, and within newer parsing benchmarks like OmniDocBench that flag footnote placement as a source of reading-order error.

Does unit normalization have a published accuracy benchmark in pharma extraction? No dedicated benchmark for pharma-specific unit normalization (converting between equivalent dose, concentration, or measurement units) was located as of September 2026. The nearest quantified analogue is a materials-science study reporting close to 90% precision and recall for LLM-based numeric data extraction, which is a different domain and should not be read as a pharma-specific figure.

How do life-sciences document extraction tools compare on price? As of September 2026, per-page cloud OCR pricing ranges from roughly $0.0006 to $0.0015 per page for basic text extraction (Amazon Textract) up to $0.070 per page for combined forms, tables, and query extraction on the same platform, with Google Document AI's structured parsers priced separately at $10 to $30 per 1,000 pages (cloud.google.com) and Mistral's OCR 4 priced at $2 to $4 per 1,000 pages depending on batch usage. Exact pricing changes frequently and should be reconfirmed against the current vendor page before use in a purchasing decision.
