

Benchmarking PDF Parsers on Table Extraction with LLM-based Semantic Evaluation

Pius Horn¹[0009-0004-1911-1138] and Janis Keuper^{1,2}[0000-0002-1327-1243]

¹ Institute for Machine Learning and Analytics (IMLA), Offenburg University, Offenburg, Germany pius.horn@hs-offenburg.de

² University of Mannheim, Mannheim, Germany

Abstract. Reliably extracting tables from PDFs is essential for large-scale scientific data mining and knowledge base construction, yet existing evaluation approaches rely on rule-based metrics that fail to capture semantic equivalence of table content. We present a benchmarking framework based on synthetically generated PDFs with precise LaTeX ground truth, using tables sourced from arXiv to ensure realistic complexity and diversity. As our central methodological contribution, we apply LLM-as-a-judge for semantic table evaluation, integrated into a matching pipeline that accommodates inconsistencies in parser outputs. Through a human validation study comprising over 1,500 quality judgments on extracted table pairs, we show that LLM-based evaluation achieves substantially higher correlation with human judgment (Pearson $r=0.93$) compared to Tree Edit Distance-based Similarity (TEDS, $r=0.68$) and Grid Table Similarity (GriTS, $r=0.70$). Evaluating 21 contemporary PDF parsers across 100 synthetic documents containing 451 tables reveals significant performance disparities. Our results offer practical guidance for selecting parsers for tabular data extraction and establish a reproducible, scalable evaluation methodology for this critical task.

Keywords: PDF Document Parsing · Table Extraction · LLM-based Evaluation · OCR Benchmarking.

1 Introduction

Much of the structured knowledge in scientific publications, financial reports, and technical documents is organized in tables. As document parsing becomes central to language model pretraining, retrieval-augmented generation, and scientific data mining [43,31], the ability to accurately and reliably extract tabular data from PDFs has become increasingly important.

The landscape of PDF document parsing has evolved rapidly, with approaches ranging from rule-based extraction tools and specialized OCR models to end-to-end vision-language models [33,1]. Existing benchmarks evaluate table extraction at scales from cropped table images [46,36] to full document-level assessments [28,27], and the accompanying metrics have progressed from cell adjacency relations [7] to tree-based [46] and grid-based [37] comparison (see Section 2.2).

Yet all of these approaches rely on structural matching and surface-level string comparison, unable to assess whether the actual information conveyed by a table has been correctly preserved. Consequently, a parser that produces a structurally different but semantically equivalent representation may be penalized unfairly, while one that preserves structure but corrupts cell content may receive an inflated score.

The LLM-as-a-judge paradigm [44] offers a promising solution, having demonstrated effectiveness for evaluating complex outputs where traditional metrics fall short [14]. For table assessment, where content correctness and structural fidelity must be jointly evaluated, LLM-based evaluation can capture semantic nuances that surface-level similarity metrics miss.

We combine this evaluation approach with a benchmarking framework that embeds real arXiv tables into synthetic PDFs, providing exact LaTeX ground truth without manual annotation. Together, these contributions address both the metric and benchmark gaps:

- We pioneer LLM-as-a-judge for semantic table evaluation, demonstrating substantially higher agreement with human judgment than rule-based metrics.
- We provide 1,554 human ratings on 518 table pairs, enabling meta-evaluation of existing and future table extraction metrics against human judgment.
- We introduce a benchmarking framework that embeds real tables from arXiv into synthetic PDFs, combining realistic table diversity with exact LaTeX ground truth, and develop an LLM-based matching pipeline that reliably aligns each parsed table to its ground truth despite variations in parser output formats.
- We establish a public leaderboard evaluating 21 contemporary document parsers across 100 synthetic pages containing 451 tables, revealing significant performance disparities and providing practical guidance for practitioners.

2 Related Work

2.1 PDF Parsing Benchmarks

Existing benchmarks for table extraction fall into two broad categories: *table recognition datasets* that operate on cropped table images, and *document-level benchmarks* that evaluate table extraction in the context of full pages.

Table recognition datasets have driven progress in table structure recognition from isolated images. Large-scale datasets such as PubTabNet [46] (568K images from PubMed Central), FinTabNet [45] (113K tables from financial reports), TableBank [15] (417K tables via weak supervision), PubTables-1M [36] (nearly one million scientific tables), and SynthTabNet [26] (600K synthetic tables with controlled structure and style variation) provide extensive training and evaluation resources, while SciTSR [4] contributes 15K tables with structure labels derived from LaTeX sources. These datasets also gave rise to the dominant evaluation metrics: PubTabNet introduced TEDS, and GriTS [37] later proposed

grid-level evaluation as an alternative. The ICDAR 2021 competition [11] complemented these efforts by targeting table image to LaTeX conversion, a task recently advanced by reinforcement learning over multimodal language models [20]. While instrumental for advancing table recognition, these datasets provide cropped table images rather than full documents, making them unsuitable for benchmarking end-to-end PDF parsing pipelines where tables must first be detected within a page of mixed content.

Document-level benchmarks evaluate table extraction from complete pages where tables appear alongside text and figures. The OmniAI OCR Benchmark [27] evaluates overall document extraction accuracy but lacks table-specific metrics, while olmOCR-Bench [31] includes 1,020 table-specific unit tests across 1,402 PDFs but focuses on cell-level pass/fail verification rather than holistic table quality assessment. OmniDocBench [28] (1,355 pages) and READoc [19] (3,576 documents, 15 parsing systems) go further by including explicit table evaluation, yet both rely on TEDS and edit distance metrics that capture structural similarity without assessing semantic equivalence. PubTables-v2 [35] provides the first large-scale benchmark for full-page and multi-page table extraction (467K single pages with 548K tables and 9,172 multi-page documents), extending PubTables-1M from cropped images to document-level evaluation using GriTS. The ICDAR table competitions [7,6] established early document-level benchmarks on small collections using cell adjacency relations, and SCORE [16] more recently addresses evaluation methodology by proposing interpretation-agnostic metrics that handle legitimate structural ambiguity, though still operating at the structural rather than semantic level. Soric et al. [38] benchmark nine extraction methods across three document collections totaling approximately 37K pages, relying on TEDS and GriTS for evaluation.

2.2 Table Extraction Evaluation Metrics

Evaluating extracted tables against ground truth requires metrics that jointly assess structural fidelity and content correctness. Directed adjacency relations (DAR) [7], introduced in the ICDAR 2013 Table Competition as the first metric designed specifically for table structure evaluation, captured local cell neighborhoods but could not represent global table structure, motivating the tree-level and grid-level approaches that followed.

Tree Edit Distance-based Similarity (TEDS). TEDS [46], introduced alongside PubTabNet, has become the de facto standard for table recognition evaluation. Both predicted and ground-truth tables are represented as HTML trees whose leaf nodes (`<td>`) carry colspan, rowspan, and character-level tokenized content. The tree edit distance is computed with a unit cost for structural mismatches and normalized Levenshtein distance for cell content, yielding $TEDS = 1 - d / \max(|T_{pred}|, |T_{gt}|)$, where d is the edit distance and $|T|$ the number of nodes in each tree. Since both structure and content are compared at the character level, the score is sensitive to markup choices (e.g., `<thead>` vs. `<tbody>`, or `<th>` vs. `<td>`) and surface-level string differences.

Grid Table Similarity (GriTS). GriTS [37] addresses the HTML sensitivity of TEDS by operating directly on the table’s 2D grid representation, using factored row and column alignment via dynamic programming. It defines separate metrics for structural topology (GriTS_{Top}, via intersection-over-union on relative span grids) and content (GriTS_{Con}, via longest common subsequence similarity), each yielding precision, recall, and F-score. By avoiding the tree representation, GriTS treats rows and columns symmetrically and is robust to markup variations, though content comparison remains string-based.

SCORE. Although SCORE [16] presents itself as a “semantic evaluation framework” addressing the format rigidity of TEDS and GriTS, it normalizes tables into format-agnostic cell tuples and evaluates *index accuracy* by checking whether cells occupy correct grid positions and *content accuracy* via edit distance on cell text, with tolerance for small row/column offsets. This avoids penalizing markup differences across output formats, though the structural tolerance may also mask genuine errors, and the underlying cell comparison remains string-based, leaving true semantic equivalence (such as notational variants or equivalent value formats) unaddressed.

Text-based metrics such as Levenshtein edit distance [13] or BLEU [29] operate at a strictly lower level of granularity: while the metrics above preserve cell-level structure, text-based approaches flatten the table into a token sequence, making scores dependent on serialization order and unable to distinguish structural from content errors.

While LLM-based evaluation has recently shown promise for formula extraction from PDFs [10], substantially outperforming text-based, tree-based, and image-based metrics in correlation with human judgment, no comparable study exists for table extraction.

Our work addresses both gaps: on the benchmark side, we use synthetic PDFs whose LaTeX source serves as exact ground truth, eliminating the need for manual annotation. On the metric side, we apply LLM-as-a-judge for semantic table evaluation. Together with a broad comparison of 21 contemporary parsers, this yields a reproducible evaluation framework that we validate against human judgment in Section 4.

3 Methodology

Our benchmarking methodology rests on two key design decisions: (1) using real tables extracted from arXiv to ensure realistic diversity, and (2) embedding them into synthetically generated PDFs to obtain exact ground truth without manual annotation. This section describes the resulting benchmark construction and the matching pipeline that aligns parser outputs to ground truth tables.

3.1 Benchmark Dataset: Synthetic PDFs with Ground Truth

We collect LaTeX table sources from arXiv papers published in December 2025 to avoid overlap with established datasets [46,36] that may already be part

of parser training data. All top-level `tabular/tabular*` environments are extracted, cleaned of non-content commands (citations, cross-references), and compiled standalone to verify validity and record rendered dimensions; invalid tables are discarded. Each valid table is classified by structural complexity using an LLM-based classifier: *simple* (regular grid), *moderate* (limited cell merging), or *complex* (multi-dimensional merging, nested structures).

Each benchmark page is generated by sampling a random layout configuration (document class, font family, page margins, font size, line spacing, and single- or two-column layout) and iteratively appending content blocks, either filler text or tables from the extracted pool. The document is recompiled with `pdflatex` after each addition; blocks that trigger overflow or typesetting warnings are discarded, and the process terminates when no further content fits. Tables are pre-filtered by their recorded dimensions against the remaining page space and scaled to column width via `adjustbox` when moderately oversized. To ensure deterministic positioning, tables are placed as non-floating centered blocks, avoiding the unpredictable reordering of LaTeX float environments that would complicate ground truth alignment. Figure 1 gives an overview of the pipeline.

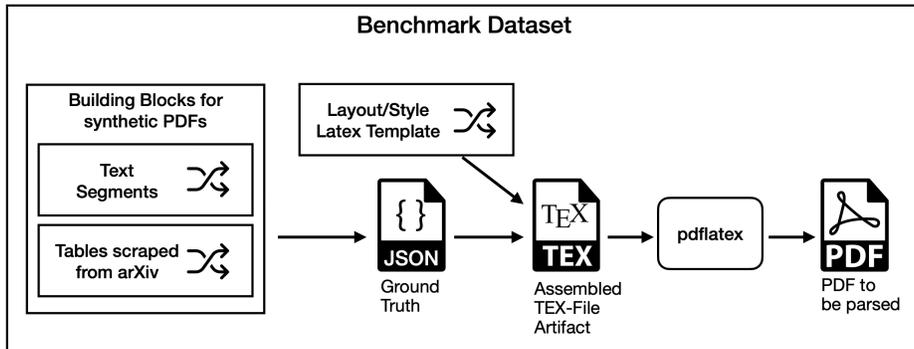


Fig. 1. Overview of the benchmark generation pipeline. Randomly sampled content blocks and layout templates yield a JSON ground truth, which is assembled into $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ and compiled to PDF.

3.2 Evaluation Pipeline: Table Matching

Before tables can be evaluated, each ground truth table must be matched to its counterpart in the parser output. This is non-trivial because parsers produce tables in diverse formats (HTML, Markdown, LaTeX, plain text), may split or merge tables, reorder content in multi-column layouts, or fail to recognize tables entirely.

We address this with an LLM-based matching pipeline using Gemini-3-Flash-Preview [8]. Given the list of ground truth tables (as LaTeX) and the full parser output, the model identifies and extracts the corresponding parsed representation for each ground truth table. Since the LLM may introduce minor artifacts such as whitespace changes, a rule-based post-validation step verifies and corrects each returned table against the original parser output, yielding a robust mapping between ground truth and parsed tables.

4 Assessment of Table Evaluation Approaches

Once ground truth and parsed tables have been aligned, the central question becomes how to quantify extraction quality. This section exposes the limitations of existing rule-based metrics on concrete parser outputs, introduces LLM-based semantic evaluation as an alternative, and validates both approaches against human judgment.

4.1 Limitations of Rule-based Metrics

Since TEDS, GriTS, and SCORE compare structure and content purely syntactically, they cannot distinguish between discrepancies that alter the semantic content of a table and those that are merely representational. Figure 2 illustrates this with a constructed example highlighting discrepancy patterns we frequently observed in parser outputs. Several difference types are semantically insignificant yet incur large edit distances:

- *Structural reorganization*: parsers flatten multi-level headers or resolve row-spans—often because the output format lacks support for spanning cells (e.g., Markdown, unlike HTML, has no `colspan/rowspan` mechanism), leading to workarounds such as repeating values or inserting empty padding cells.
- *Symbol encoding*: formulas appear as Unicode symbols instead of LaTeX commands (e.g., α instead of `\alpha`).
- *Value equivalence*: “85.0%” vs. “85%” or “—” vs. “N/A”.
- *Markup artifact*: visual attributes are encoded as raw commands (e.g., `\textbf{}`).

In contrast, the only semantically critical differences, *content errors* such as a lost decimal point ($1.12 \rightarrow 112$) and a flipped sign ($+2.8 \rightarrow -2.8$), change merely a single character each, contributing minimally to the edit distance. String-based metrics consequently assign low scores driven by harmless representational variation, while the few-character errors that fundamentally alter the table’s meaning are barely reflected.

4.2 LLM-as-a-Judge for Table Evaluation

Building on the LLM-as-a-judge paradigm [44,14], we propose using LLMs to assess table extraction quality semantically. Given a ground truth table and its

(a) Ground Truth

Group	Method	Task 1		Task 2	
		Score	Diff	Score	Diff
Group 1	Baseline	85.0%	—	0.72 ± 0.03	—
	Method α	91.2%	+6.2 ($p \leq 0.1$)	1.12	+0.17
Group 2	Baseline	79.3%	—	0.65	—
	Method β	82.1%	+2.8	1.31	+0.66

(b) Parser Output

Group	Method	Task 1 Score	Task 1 Diff	Task 2 Score	Task 2 Diff
Group 1	Baseline	85%	N/A	0.72 ± 0.03	N/A
Group 1	Method α	91.2%	+6.2 ($p \leq 0.1$)	112	+0.17
Group 2	Baseline	79.3%	N/A	0.65	N/A
Group 2	Method β	82.1%	-2.8	1.31	+0.66

Fig. 2. Structural metrics penalize harmless variation while overlooking critical errors. The parser output (b) largely preserves the semantics of (a), yet incurs heavy edit distance from representational differences (structural reorganization , symbol encoding , value equivalence , markup artifact). The only meaning-altering errors—a lost decimal and a sign flip (content error)—barely affect the score.

parsed counterpart, an LLM evaluates the pair on a 0–10 scale for content accuracy and structural preservation, i.e., whether every cell value can be unambiguously mapped to its row and column headers. We evaluate four popular LLMs as judges: DeepSeek-v3.2 [21], GPT-5-mini [34], Gemini-3-Flash-Preview [8], and Claude Opus 4.6 [2], selected for their strong performance on public benchmarks across different price points.

4.3 Human Evaluation Protocol

To validate automated metrics against human judgment, we collected over 1,500 human ratings covering 518 pairs of ground truth and parsed tables. Each pair was rated on a 0–10 scale reflecting whether the semantic content of the table—all values, headers, and their associations—has been correctly, completely, and unambiguously preserved. The pairs were sampled across all parsers whose outputs span diverse formats (HTML, Markdown, LaTeX, plain text) and table complexities, ensuring broad coverage. Since tables can be large and discrepancies subtle, we prompted Claude Opus 4.6 to pre-identify potential differences in each pair. Evaluators were then presented with a web interface showing both

tables alongside these LLM-generated hints on potential discrepancies, ensuring that subtle issues are surfaced for human judgment while the final score remains entirely a human decision.

Inter-annotator agreement. To assess the reliability of the human reference scores, we report agreement among the three independent evaluators who each rated all 518 pairs. Krippendorff’s α (interval) is 0.77, indicating acceptable agreement [12]. Average pairwise Pearson correlation between annotators is $r = 0.85$, with individual pairs ranging from 0.81 to 0.91 and a mean absolute score difference of 1.2 on the 0–10 scale. As a human performance ceiling, the leave-one-out correlation of each annotator with the mean of the other two yields an average Pearson $r = 0.89$.

4.4 Correlation with Human Judgment

We compute Pearson, Spearman, and Kendall correlations between each automated metric and the human reference scores to quantify how well each approach captures human notions of table extraction quality. All metrics are scaled to a 0–10 range for comparability; Table 1 summarizes the results and Figure 3 visualizes the relationship for a subset of metrics.

Table 1. Correlation of automated metrics with averaged human scores ($n = 518$ table pairs, each rated by three evaluators).

Metric	Type	Pearson r	Spearman ρ	Kendall τ
TEDS	Rule-based	0.684	0.717	0.557
GriTS _{Top}	Rule-based	0.633	0.735	0.597
GriTS _{Con}	Rule-based	0.700	0.742	0.595
GriTS-Avg	Rule-based	0.698	0.763	0.604
SCORE Index	Rule-based	0.558	0.681	0.558
SCORE Content	Rule-based	0.641	0.654	0.522
SCORE-Avg	Rule-based	0.637	0.684	0.539
DeepSeek-v3.2	LLM	0.802	0.827	0.713
GPT-5-mini	LLM	0.888	0.827	0.739
Gemini-3-Flash-Preview	LLM	0.927	0.889	0.799
Claude Opus 4.6 [†]	LLM	0.939	0.890	0.804

[†] Also used to generate error hints shown to evaluators; see text.

Rule-based metrics achieve only moderate correlation with human judgment. Since GriTS and SCORE each decompose into separate structure and content sub-metrics, unlike TEDS and LLM-based judges which assess both aspects jointly, we additionally compute their arithmetic means (GriTS-Avg, SCORE-Avg) to enable direct comparison. All rule-based metrics, whether structure-focused, content-focused, or averaged, fall within a narrow band of $r = 0.56$ – 0.70 (Table 1), confirming the limitations analyzed in Section 4.1.

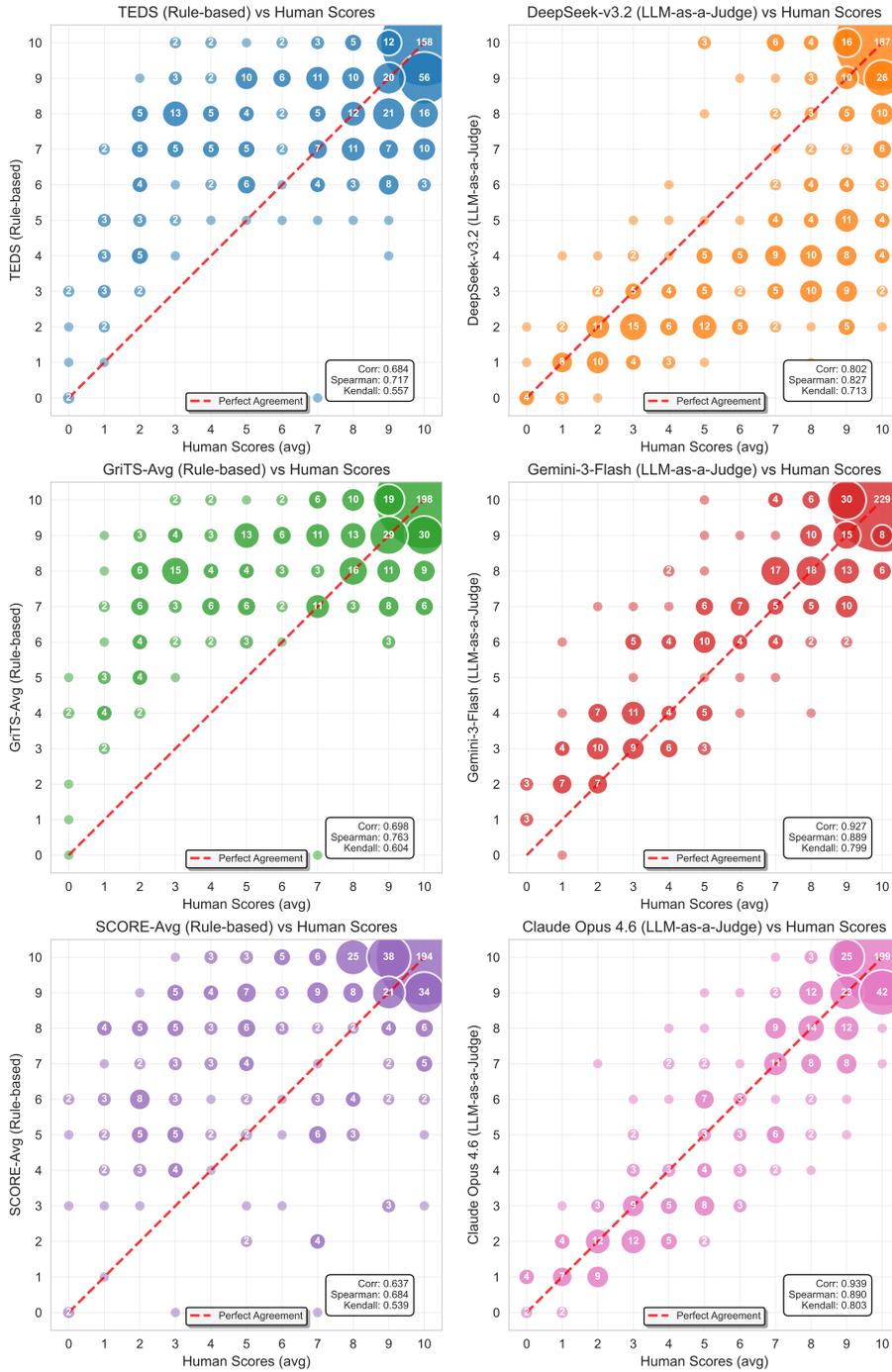


Fig. 3. Scatter plots comparing automated metrics with human scores. Left column: rule-based metrics (TEDS, GrITS-Avg, SCORE-Avg); right column: LLM judges (DeepSeek-v3.2, Gemini-3-Flash-Preview, Claude Opus 4.6). Bubble size indicates point count.

LLM-based evaluation substantially outperforms all rule-based metrics. Even the weakest LLM judge (DeepSeek-v3.2, $r = 0.80$) exceeds the best rule-based metric (GriTS-Avg, $r = 0.70$), confirming that semantic assessment captures dimensions of table quality that string-level comparison systematically misses. Claude Opus 4.6 achieves the highest correlation ($r = 0.94$), though the human reference scores may be skewed toward its assessments since it also generated the error hints shown to evaluators ([†] in Table 1). Gemini-3-Flash-Preview ($r = 0.93$) and GPT-5-mini ($r = 0.89$), which had no role in the annotation process, still far surpass all rule-based metrics, confirming that the LLM advantage is independent of this confound.

For the parser benchmark in Section 5, we adopt Gemini-3-Flash-Preview as it offers near-ceiling correlation at substantially lower inference cost than Claude Opus 4.6.

5 Experiments and Results

Using the validated Gemini-3-Flash-Preview judge, we evaluate 21 parsers on 100 synthetic PDF pages containing 451 tables with diverse structural characteristics.

We selected 21 parsers spanning the full spectrum of contemporary document parsing approaches. Among specialized OCR models, we evaluate Chandra [30], DeepSeek-OCR [42], dots.ocr [17], GOT-OCR2.0 [41], LightOnOCR-2-1B [39], Mathpix [24], MinerU2.5 [40], Mistral OCR 3 [25], MonkeyOCR-3B [18], Nanonets-OCR-s [23], and olmOCR-2-7B [31]. These range from compact end-to-end vision-language models with under 1B parameters (LightOnOCR, DeepSeek-OCR) to full-page decoders built on larger VLM backbones (Chandra on Qwen3-VL, MonkeyOCR) and commercial API services (Mathpix, Mistral OCR 3).

We also evaluate general-purpose multimodal models including Gemini 3 Pro and Flash [8], Gemini 2.5 Flash [5], GLM-4.5V [9], Qwen3-VL-235B [3], GPT-5 mini and nano [34], and Claude Sonnet 4.6 [2]; since these models lack a dedicated document parsing mode, they were prompted to convert each page to Markdown with tables rendered as HTML.

Additionally, we include PyMuPDF4LLM [32], a rule-based tool that extracts text directly from the PDF text layer, and the scientific document parser GROBID [22].

All 100 pages are processed through each parser and the extracted tables are evaluated against ground truth using the LLM-based pipeline described in Sections 3.2 and 4.2. Table 2 reports the resulting scores alongside the approximate cost or time for parsing all 100 pages: API pricing in USD at the time of writing or wall-clock time on a single NVIDIA RTX 4090. As most models offer multiple deployment options and we did not use a uniform inference framework (e.g., vLLM or Hugging Face Transformers), reported runtimes are rough estimates. Our code repository provides ready-to-use implementations for all 21 parsers together with the exact prompts, configurations, and software versions used to produce the leaderboard results, enabling full reproducibility.

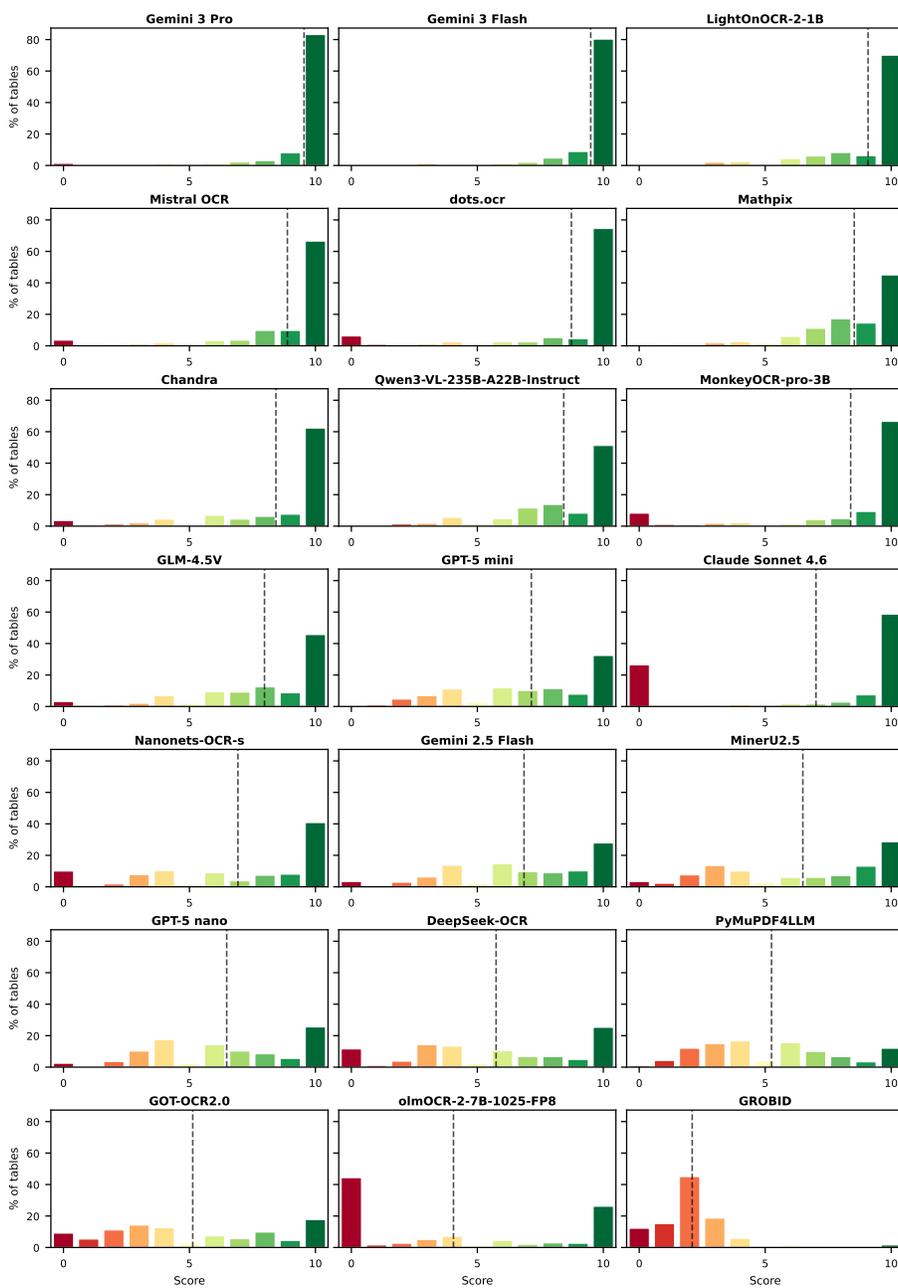


Fig. 4. Per-parser score distributions across 451 tables. Each subplot shows the percentage of tables receiving each integer score (0–10); the dashed line marks the mean. Parsers are ordered by mean score (top-left to bottom-right).

Table 2. Table extraction performance across 451 tables from 100 synthetic pages, scored 0–10 by Gemini-3-Flash-Preview and broken down by structural complexity, with TEDS scores (0–1 scale) for comparison. Parsers are ranked by overall score.

Parser	LLM Score (0–10)				TEDS	Inference	Cost / Time
	Overall	Simple	Moderate	Complex			
Gemini 3 Pro	9.55	9.58	9.57	9.49	0.85	API	\$10.00
Gemini 3 Flash	9.50	9.53	9.38	9.61	0.85	API	\$0.57
LightOnOCR-2-1B	9.08	9.41	8.90	8.91	0.83	GPU	30 min
Mistral OCR 3	8.89	8.92	8.69	9.07	0.88	API	\$0.20
dots.ocr	8.73	9.01	8.43	8.76	0.81	GPU	20 min
Mathpix	8.53	9.32	8.40	7.77	0.74	API	\$0.35–0.50
Chandra	8.43	8.96	8.14	8.15	0.77	GPU	4 h
Qwen3-VL-235B	8.43	9.23	8.27	7.67	0.78	API/GPU	\$0.20
MonkeyOCR-3B	8.39	8.60	8.10	8.47	0.80	GPU	20 min
GLM-4.5V	7.98	9.19	7.59	7.00	0.78	API	\$0.60
GPT-5 mini	7.14	8.03	6.82	6.48	0.68	API	\$1.00
Claude Sonnet 4.6	7.02	6.94	7.10	7.01	0.63	API	\$3.00
Nanonets-OCR-s	6.92	8.27	6.51	5.82	0.69	GPU	50 min
Gemini 2.5 Flash	6.85	7.93	6.52	5.94	0.72	API	\$0.40
MinerU2.5	6.49	7.07	6.03	6.35	0.78	API/GPU	— [‡]
GPT-5 nano	6.48	7.63	6.18	5.47	0.32	API	\$0.35
DeepSeek-OCR	5.75	7.45	5.34	4.20	0.66	GPU	4 min
PyMuPDF4LLM	5.25	6.78	4.86	3.91	— [§]	CPU	30 s
GOT-OCR2.0	5.13	5.89	4.95	4.45	0.58	GPU	20 min
olmOCR-2-7B	4.05	4.64	3.78	3.68	0.35	GPU	25 min
GROBID	2.10	2.27	1.94	2.09	— [§]	CPU	2 min

Cost: API pricing (USD) for 100 pages. Time: wall-clock on a single NVIDIA RTX 4090.

[‡] Tested via free-tier API; also available for local GPU deployment.

[§] TEDS not applicable; output lacks tabular structure entirely.

6 Discussion

LLM scores vs. TEDS. The TEDS scores in Table 2 reinforce the metric limitations discussed in Section 4.1. When parsers that frequently fail to detect tables entirely are excluded, TEDS clusters within 22% of its scale (0.66–0.88), painting a misleading picture of comparable quality. LLM-based scores, by contrast, span 38% (5.75–9.55), far better reflecting the substantial quality differences visible upon manual inspection.

Parser performance patterns. Overall scores range from 2.10 to 9.55, revealing that parser choice can largely determine whether extracted tables are usable or nearly unusable. The top-performing systems are the Gemini 3 models, which are general-purpose multimodal models rather than dedicated OCR tools, suggesting that broad visual-linguistic capabilities transfer well to table extraction. However, targeted design can rival much larger models: the specialized LightOnOCR-2-1B achieves 9.08 with only 1B parameters, and dots.ocr (8.73) and MonkeyOCR-3B (8.39) also run on a single consumer GPU, narrowing the gap between proprietary API services and self-hosted pipelines for applications with data privacy constraints or limited API budgets. At the other end of the

spectrum, rule-based tools (PyMuPDF4LLM, GROBID) require no GPU but lag substantially behind all learning-based approaches. Beyond overall ranking, the complexity breakdown in Table 2 reveals that table complexity affects parsers unevenly: while most show declining scores from simple to complex tables, the magnitude varies widely, from negligible drops (Gemini 3 Flash actually scores higher on complex tables) to severe degradation (GLM-4.5V: -2.19 , Qwen3-VL: -1.56 , Mathpix: -1.55), indicating that handling multi-dimensional cell merging remains a key differentiator. Even the top-scoring Gemini 3 models exhibit errors upon manual inspection, including misaligned spanning cells, subtly altered values, and incorrect header-cell associations, confirming that accurate table extraction from PDFs remains an unsolved problem.

Score distributions. The per-parser histograms in Figure 4 expose failure patterns that mean scores obscure. Top parsers (Gemini 3, LightOnOCR) concentrate $>70\%$ of tables at score 10, while Claude Sonnet 4.6 and olmOCR show strongly bimodal distributions: they frequently omit tables entirely (score 0) but extract them near-perfectly when they do. Mid-tier parsers such as GPT-5 mini and Gemini 2.5 Flash produce broad distributions centered around scores 5–8, indicating pervasive partial errors rather than clean successes or outright failures. Depending on the application, a missed table may be preferable to a corrupted one, making bimodal parsers with high-quality successes more useful than those with uniformly mediocre output.

Limitations. Synthetic PDFs do not capture the full diversity of real-world tables (such as scanned documents or non-standard layouts), and the table dataset is sourced exclusively from arXiv, which may bias toward scientific table formats, leaving domains such as financial reports or medical records unrepresented. While LLM-as-a-judge substantially outperforms rule-based metrics, it is not infallible and requires proprietary models, though evaluation costs remain modest: scoring all 451 tables costs approximately \$0.20, and a full benchmark run for one parser totals roughly \$1 in API costs.

Future Work. Future work includes incorporating more diverse document formats and layouts, evaluating parsers’ ability to extract information from figures, and extending the benchmark toward holistic document parsing covering tables, formulas, and text jointly.

Code and Data Availability. The synthetic PDF generation pipeline, ready-to-use configurations for all 21 parsers, the evaluation pipeline, and the benchmark dataset (100 pages with ground truth) are publicly available.³ The meta-evaluation of table extraction metrics, including all metric implementations and the human evaluation study, is provided in a separate repository.⁴

³ <https://github.com/phorn1/pdf-parse-bench>

⁴ <https://github.com/phorn1/table-metric-study>

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments. We thank Sarah Cebulla and Martin Spitznagel for their patience and thoroughness in rating table extraction quality across hundreds of table pairs. This work has been supported by the German Federal Ministry of Research, Technology, and Space (BMFTR) in the program “Forschung an Fachhochschulen in Kooperation mit Unternehmen (FH-Kooperativ)” within the joint project *LLMpraxis* under grant 13FH622KX2.

References

1. Adhikari, N.S., Agarwal, S.: A comparative study of pdf parsing tools across diverse document categories. arXiv preprint arXiv:2410.09871 (2024)
2. Anthropic: Introducing Claude Opus 4.6. <https://www.anthropic.com/news/claude-opus-4-6> (2025), accessed: 2026-01-10
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Chi, Z., Huang, H., Xu, H.D., Yu, H., Yin, W., Mao, X.L.: Complicated table structure recognition. arXiv preprint arXiv:1908.04729 (2019)
5. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
6. Gao, L., Huang, Y., Déjean, H., Meunier, J.L., Yan, Q., Fang, Y., Kleber, F., Lang, E.: Icdar 2019 competition on table detection and recognition (ctdar). In: Proceedings of the 15th International Conference on Document Analysis and Recognition (ICDAR). pp. 1510–1515 (2019)
7. Göbel, M., Hassan, T., Oro, E., Orsi, G.: Icdar 2013 table competition. In: Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR). pp. 1449–1453 (2013)
8. Google DeepMind: A new era of intelligence with Gemini 3. <https://blog.google/products/gemini/gemini-3/> (2025), accessed: 2026-01-10
9. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
10. Horn, P., Keuper, J.: Benchmarking document parsers on mathematical formula extraction from pdfs. arXiv preprint arXiv:2512.09874 (2025)
11. Kayal, P., Anand, M., Desai, H., Singh, M.: Icdar 2021 competition on scientific table image recognition to latex. In: Proceedings of the 16th International Conference on Document Analysis and Recognition (ICDAR). pp. 754–766 (2021)
12. Krippendorff, K.: Computing krippendorff’s alpha-reliability (2011)
13. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**(8), 707–710 (1966)
14. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: Llm-as-judges: a comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579 (2024)

15. Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.: Tablebank: Table benchmark for image-based table detection and recognition. In: Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC). pp. 1918–1925 (2020)
16. Li, R., Yepes, A.J., You, Y., Pluciński, K., Operlejn, M., Wolfe, C.: Score: A semantic evaluation framework for generative document parsing. arXiv preprint arXiv:2509.19345 (2025)
17. Li, Y., Yang, G., Liu, H., Wang, B., Zhang, C.: dots. ocr: Multilingual document layout parsing in a single vision-language model. arXiv preprint arXiv:2512.02498 (2025)
18. Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., Bai, X.: Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. arXiv preprint arXiv:2506.05218 (2025)
19. Li, Z., Abulaiti, A., Lu, Y., Chen, X., Zheng, J., Lin, H., Han, X., Sun, L.: READoc: A unified benchmark for realistic document structured extraction. In: Findings of the Association for Computational Linguistics (ACL). pp. 21889–21905 (2025)
20. Ling, J., Qi, Y., Huang, T., Zhou, S., Huang, Y., Yang, J., Song, Z., Zhou, Y., Yang, Y., Shen, H.T., Wang, P.: Table2latex-rl: High-fidelity latex code generation from table images via reinforced multimodal language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
21. Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al.: Deepseek-v3. 2: Pushing the frontier of open large language models. arXiv preprint arXiv:2512.02556 (2025)
22. Lopez, P.: Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL). pp. 473–474 (2009)
23. Mandal, S., Talewar, A., Ahuja, P., Juvatkar, P.: Nanonets-ocr-s: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging (2025)
24. Mathpix, Inc.: Mathpix: Document conversion for stem. <https://mathpix.com> (2025), accessed: 2026-02-10
25. Mistral AI: Introducing mistral ocr 3. <https://mistral.ai/news/mistral-ocr-3> (2025), accessed: 2026-02-01
26. Nassar, A., Livathinos, N., Lysak, M., Staar, P.: Tableformer: Table structure understanding with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4614–4623 (2022)
27. OmniAI Technology, Inc.: Omni OCR Benchmark. Hugging Face Dataset, <https://huggingface.co/datasets/getomni-ai/ocr-benchmark> (2025), accessed: 2026-01-10
28. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., et al.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24838–24848 (2025)
29. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311–318 (2002)
30. Paruchuri, V., Datalab Team: Chandra: OCR model for complex documents with full layout preservation. GitHub repository, <https://github.com/datalab-to/chandra> (2025), accessed: 2026-02-10

31. Poznanski, J., Soldaini, L., Lo, K.: olmocr 2: Unit test rewards for document ocr. arXiv preprint arXiv:2510.19817 (2025)
32. PyMuPDF Contributors: PyMuPDF4LLM: Pdf extraction for large language models. GitHub repository, <https://github.com/pymupdf/PyMuPDF4LLM> (2025), accessed: 2026-02-01
33. Salaheldin Kasem, M., Abdallah, A., Berendeyev, A., Elkady, E., Mahmoud, M., Abdalla, M., Hamada, M., Vascon, S., Nurseitov, D., Taj-Eddin, I.: Deep learning for table detection and structure recognition: A survey. *ACM Computing Surveys* **56**(12), 1–41 (2024)
34. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
35. Smock, B., Faucon-Morin, V., Sokolov, M., Liang, L., Khanam, T., Courtland, M.: Pubtables-v2: A new large-scale dataset for full-page and multi-page table extraction. arXiv preprint arXiv:2512.10888 (2025)
36. Smock, B., Pesala, R., Abraham, R.: Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4634–4642 (2022)
37. Smock, B., Pesala, R., Abraham, R.: Grits: Grid table similarity metric for table structure recognition. In: *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*. pp. 535–549 (2023)
38. Soric, M., Gracianne, C., Manolescu, I., Senellart, P.: Benchmarking table extraction from heterogeneous scientific extraction documents. arXiv preprint arXiv:2511.16134 (2025)
39. Taghadouini, S., Cavallès, A., Aubertin, B.: Lightocr: A 1b end-to-end multilingual vision-language model for state-of-the-art ocr. arXiv preprint arXiv:2601.14251 (2026)
40. Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et al.: Mineru: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839 (2024)
41. Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al.: General ocr theory: Towards ocr-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704 (2024)
42. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
43. Zhang, Q., Wang, B., Huang, V.S.J., Zhang, J., Wang, Z., Liang, H., He, C., Zhang, W.: Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. arXiv preprint arXiv:2410.21169 (2024)
44. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* **36**, 46595–46623 (2023)
45. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 697–706 (2021)
46. Zhong, X., ShafieiBavani, E., Jimeno Yepes, A.: Image-based table recognition: data, model, and evaluation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 564–580 (2020)