How Far Is Document Parsing from Solved? PureDocBench: A Source-Traceable Benchmark across Clean, Degraded, and Real-World Settings

Anonymous authors
Paper under double-blind review

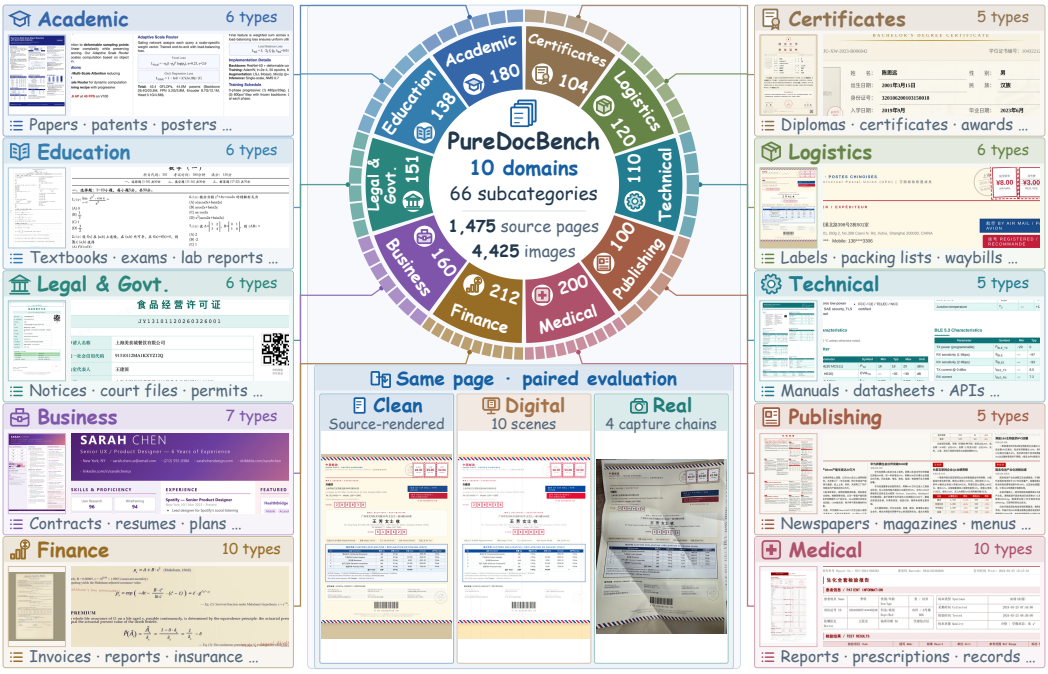


Figure 1: **PureDocBench overview.** The benchmark covers 10 domains and 66 subcategories, with 1,475 source pages and 4,425 matched Clean, Digital, and Real images. The three versions of each page share content and source-linked annotations for text, formulas, tables, and reading order.

Abstract

Document parsing models are improving rapidly, but existing benchmarks are still insufficient for a comprehensive evaluation of their capabilities. We introduce **PureDocBench**, a benchmark with **10 domains, 66 subcategories, and 1,475 pages**. Each page has matched Clean, Digital, and Real versions, covering clean renders, simulated degradation, and real capture or sharing conditions. The resulting 4,425 images have source-derived annotations for text, formulas, tables, and reading order, checked through automated validation and human review. Keeping page content and annotations fixed allows direct comparisons across image conditions. We evaluate 58 models, including pipeline-based parsers, end-to-end parsers, and general vision-language models. The best three-track average score is 79.69 out of 100, and the mean across models is 63.84. Scores drop by 3.34 points on Digital and 11.61 points on Real on average, relative to Clean. TeleOCR leads on Clean, GLM-5.3-Flash on Digital, and Gemini 3.6 Flash on Real. We *release* the benchmark, source documents, generation tools, and evaluation code.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Document parsing recovers text, table structure, formula LaTeX, and reading order from document images (Cui et al., 2021). Specialist models follow two broad designs. **Multi-stage systems** analyze layout before recognizing regions, using separate components (PaddleOCR-VL, GLM-OCR) (Cui et al., 2026; Duan et al., 2026) or a shared VLM (MinerU2.5, Dolphin-v2) (Niu et al., 2025; Wang et al., 2026a; Feng et al., 2026). **End-to-end systems** directly map page images to document markup, as in FireRed-OCR, HunyuanOCR, WeVisDoc, and DeepSeek-OCR/-2 (Wu et al., 2026; Hunyuan Vision Team et al., 2025; Yu et al., 2026; Wei et al., 2025; 2026). General VLMs, including Qwen3-VL, Qwen3.5, and Gemini-3.1-Pro (Bai et al., 2025; Qwen Team, 2026; Gemini Team, 2026), provide another widely used approach. Rapid releases across these families increase the need for reliable evaluation across document types and acquisition conditions.

Recent benchmarks, including OmniDocBench (Ouyang et al., 2025), DocPTBench (Du et al., 2025a), and other task-focused resources (Zhou et al., 2026b; Li et al., 2026; Fu et al., 2025; Poznanski et al., 2025b; Huang et al., 2026b), provide valuable foundations for measuring document understanding. Collecting documents across domains is labor-intensive, and annotating their text, formulas, tables, and reading order requires substantial manual effort. Extending a benchmark to new document types or domains often means repeating the collection and annotation process.

We introduce **PureDocBench** (Figure 1), built from generated HTML/CSS documents with source-linked annotations and matched image tracks. Its generation and annotation pipeline can be reused across document types and domains (Section 2).

Our 58-model evaluation (Figure 3) yields four findings. **(1) Overall accuracy:** best Avg_3 79.69, mean 63.84, and a 50.2-point best-to-worst spread. **(2) Competitive compact specialists:** MinerU2.5-Pro (1.2B) closely matches Kimi K2.6 (1T/32B active). **(3) Robustness differences:** digital/real Overall losses average 1.84/9.30 for general VLMs versus 5.17/14.23 for pipeline specialists. **(4) Component differences:** E2E specialists achieve the lowest mean text and reading-order errors, while general VLMs achieve the highest mean formula and table scores.

Our contributions are threefold:

- **Source-traceable benchmark.** We introduce 1,475 pages across 10 domains and 66 subcategories, with source-linked annotations and matched Clean, Digital, and Real images for each page.
- **Comprehensive evaluation.** We compare 58 specialist parsers and general VLMs across three tracks, examining overall accuracy, robustness, and performance on individual parsing components.
- **Open resources.** We release document images, sources, annotations, generation and evaluation tools, and available model predictions for reproducible evaluation and benchmark extension.

2 PUREDOCBENCH

2.1 DATA CONSTRUCTION

PureDocBench links rendered images and annotations to shared HTML/CSS sources (Figure 2). Compared with collecting and annotating public pages, generating document content from scratch and rendering images directly helps reduce the risk of overlap with model training data. Newly generated content and source-traceable annotations motivate *Pure*.

Taxonomy (Figure 1). Ten domains and 66 subcategories cover financial invoices, medical records, legal contracts, and logistics documents (Fortune Business Insights, 2026; Grand View Research, 2025).

Paired image tracks (2a). Each of 1,475 pages has clean, digitally degraded, and real-degraded versions: 4,425 images sharing GT. Ten digital scene templates include aging, binding, photocopying, compression, lighting, and geometric artifacts. Four real-track chains comprise phone capture of prints, phone capture of photocopies, screen photography, and screenshots with social-media compression. Pairing fixes content while varying acquisition; Appendix F lists operations and weights.

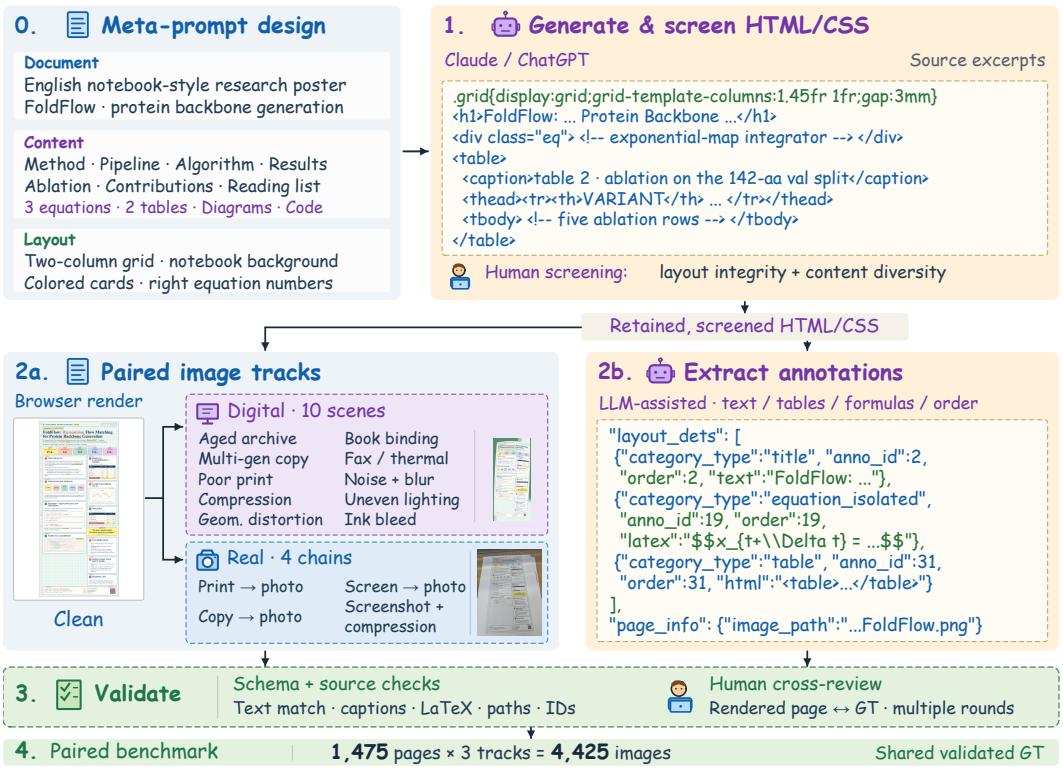


Figure 2: **Data construction and source-linked annotation.** Generated HTML/CSS documents provide a common source for page images and annotations of text, tables, formulas, and reading order. Matched Clean, Digital, and Real versions enable comparisons across image conditions, with automated checks and human review supporting annotation quality across all three image tracks.

Source and annotation pipeline (0–4). Subcategory prompts specify document type, layout, domain, language, and element composition. LLMs emit self-contained HTML/CSS with text, structured tables, and LaTeX formulas; human screening rejects collapsed layouts and repetitive content. A browser renders high-resolution PNGs. LLM-assisted extraction reads the same HTML to produce OmniDocBench-format text, table HTML, formulas, and reading order. Deterministic schema/source checks and multiple human cross-review rounds validate all 1,475 GT pages.

Retained sources support annotation inspection and regeneration and remain hidden from evaluated models. New prompts extend pages and document types through the same validation/review pipeline, enabling refreshed future splits.

2.2 Comparison with Related Benchmarks

Table 1 organizes benchmarks and related corpora by task; Appendix A records comparison definitions and version-specific evidence.

Component recognition and source supervision. The component datasets in Table 1 cover layout, table structure, and printed formulas, establishing PDF–XML, Word, and LaTeX supervision. FinTabNet (Zheng et al., 2021) adds financial tables; CROHME (Mouchère et al., 2011), ReadingBank (Wang et al., 2021), and HRDoc (Ma et al., 2023) address handwritten mathematics, reading order, and document hierarchy. FUNSD, SROIE, and CORD (Jaume et al., 2019; Huang et al., 2019; Park et al., 2019) target form/receipt fields. Source-derived supervision also underlies Nougat (Blecher et al., 2024) and READoc; CoSyn (Yang et al., 2025a) adds code-generated text-rich instruction-tuning data.

Document understanding. DocVQA (Mathew et al., 2021), InfographicVQA (Mathew et al., 2022), DUDE (Van Landeghem et al., 2023), and MMLongBench-Doc (Ma et al., 2024) evaluate

Table 1: **Comparison of document benchmarks and corpora.** Resources span component recognition, document understanding, and full-page parsing. PureDocBench combines four parsing targets, source-linked annotations, and matched digital and physical image variants across 10 domains.

Resource	Corpus size	Parsing targets				Label source	Paired image variants	
		Text	Tables	Formulas	Reading order		Digital	Physical
Component recognition								
im2latex-100k (2017)	104k formulas	—	—	✓	—	</>	—	—
PubLayNet (2019)	360k+ pages	—	—	—	—	</>	—	—
DocBank (2020b)	500k pages	—	—	—	—	</>	—	—
TableBank (2020a)	417k tables	—	✓	—	—	</>	—	—
PubTabNet (2020)	568k+ tables	◐	✓	—	—	</>	—	—
DocLayNet (2022)	80,863 pages	—	—	—	—	human	—	—
PubTables-1M (2022)	948k tables	—	✓	—	—	</>	—	—
Document understanding and synthetic supervision								
DocVQA (2021)	12,767 images	—	—	—	—	human	—	—
OCRBench (2024)	1k QA pairs	✓	—	◐	—	inherited + human	—	—
OCRBench v2 (2025)	10k QA pairs	✓	✓	✓	—	inherited + human	—	—
CC-OCR (2025b)	7,058 images	✓	✓	✓	—	inherited + model + human	—	—
CC-OCR v2 (2026)	7,093 samples	✓	✓	✓	—	inherited + model + human	—	—
DOGR-Bench (2025)	3.6k samples	—	—	—	—	</> + model	—	—
CoSyn-400K (2025a)	400k images	—	—	—	—	</> + model	—	—
MADQA (2026)	800 PDFs	—	—	—	—	human	—	—
SynthDocBench (2026)	200 docs	—	—	—	—	</> + model + human	—	—
Full-document parsing								
READoc (2025c)	3,576 docs	✓	✓	✓	✓	</>	—	—
OmniDocBench v1.5 (2025)	1,355 pages	✓	✓	✓	✓	model + human	—	—
olmOCR-bench (2025a)	1,403 pages	✓	◐	✓	✓	</> + model + human	—	—
DocPTBench (2025a)	981 base pages	✓	✓	✓	✓	inherited	✓	◐
Real5-OmniDocBench (2026b)	1,355 × 5	✓	✓	✓	✓	inherited	—	✓
MDPBench (2026)	850 × 4	✓	✓	✓	✓	model + human	—	✓
LogicsDocBench (2026)	900 pages	✓	✓	✓	✓	—	—	—
ParseBench (2026)	2,078 pages	✓	✓	◐	✓	model + human	—	—
MORE (2026)	∼1.3k images	✓	✓	✓	✓	model + human	—	—
MPDocBench-Parse (2026a)	3,246 pages	✓	✓	✓	✓	model + human	—	—
Dr. DocBench (2026)	4,514 pages	✓	✓	✓	✓	model + human	—	—
PureDocBench (Ours)	**1,475 × 3**	✓	✓	✓	✓	</> + model + human	✓	✓

✓ reported; ◐ restricted scope; — unreported. </> source markup; model; human labels/review; inherited labels.
Size: original corpus units; × = base pages × views (Real5: captures only; MDPBench/PureDocBench: original included).

answers grounded in documents. SynthDocBench (Verma et al., 2026) uses controlled synthetic long documents. OCRBench/v2 and CC-OCR/v2 (Liu et al., 2024; Fu et al., 2025; Yang et al., 2025b; Peng et al., 2026) probe broader OCR capabilities, including parsing subtasks; OCR-Reasoning (Huang et al., 2026b) tests text-rich reasoning, MMDocBench (Zhu et al., 2026) adds fine-grained perception, DOGR-Bench (Zhou et al., 2025) targets grounding and referring, and MADQA (Borchmann et al.,

2026) studies agentic search over document collections. These task-specific evaluations complement full-document structured recovery.

Full-document parsing. Table 1 spans structured extraction, targeted unit tests, and semantic utility. DocPTBench, Real5-OmniDocBench, and MDPBench provide matched acquisition stress, while MDPBench and MORE broaden language coverage. MPDocBench-Parse and Dr. DocBench extend evaluation to document hierarchy and expert content; FinixDocBench (Wang et al., 2026b) adds financial workflows and camera-captured inputs. Building on this lineage, PureDocBench combines freshly generated, retained HTML/CSS sources with human-reviewed annotations and matched clean/digital/real images across 10 domains. Authentic-document and multi-page benchmarks provide complementary coverage.

3 Experiments

3.1 Setup

Metrics. We adopt the OmniDocBench evaluation framework (Ouyang et al., 2025): TextEdit↓ (page-level normalized edit distance), FormulaCDM↑ (Wang et al., 2024) (render-based character detection matching for formulas), TableTEDS↑ (Zhong et al., 2020) (tree-edit-distance similarity on table HTML), and ReadOrderEdit↓ (edit distance on reading order). Per-track **Overall**↑ is the mean of TextEdit, FormulaCDM, and TableTEDS (ROEdit is reported separately): $\mathrm{Overall} = \big((1-\mathrm{TextEdit}) \times 100 + \mathrm{FormulaCDM} + \mathrm{TableTEDS}\big)/3$. $\mathbf{Avg}_3$ is the mean of the three track Overalls (Clean / Digital / Real).

Models (58). We prioritize recent releases and compare **13 pipeline / multi-stage specialists**, **19 end-to-end specialists**, and **26 general-purpose VLMs** (Table 2). Pipeline specialists include PaddleOCR-VL-1.6, MinerU2.5-Pro, MonkeyOCRv2-B-Parsing, and TeleOCR; end-to-end specialists include OvisOCR2, WeVisDoc-2B/4B, and HunyuanOCR-1.5. General VLMs span the Qwen and MiniCPM families alongside Claude Opus 5, Gemini 3.6 Flash, GLM-5.3-Flash, GPT-5.6 Sol, and Kimi K2.6. These models cover releases through September 2026 and multiple parameter scales.

3.2 Main Results

Document parsing still has room for improvement. Across 58 models, the best Avg_3 is 79.69 out of 100 (GLM-5.3-Flash), the model mean is 63.84, and the best-to-worst spread is 50.2 points. The median Avg_3 is 66.20. Only seven models exceed 75, while 41 of the 58 models score below 70.

Compact specialist parsers remain competitive. MinerU2.5-Pro (1.2B, Avg_3 70.07) closely matches Kimi K2.6 (1T total / 32B active, 70.10) and Gemini-3.1-Pro (70.43). WeVisDoc-4B achieves an Avg_3 of 75.57, exceeding Qwen3.5-122B-A10B (74.11) by 1.46 points. WeVisDoc-2B scores 73.69, outperforming Qwen3.5-27B (69.57) by 4.12 points with only 7.4% of the parameters. FD-RL and Logics-Parsing-v2, both 4B specialists, score 73.92 and 72.61, surpassing Kimi K2.6 by 3.82 and 2.51 points, respectively.

Degradation changes model rankings. Across all 58 models, mean Overall losses are 3.34/11.61 points under digital/real degradation. TeleOCR leads Clean (87.16), GLM-5.3-Flash leads Digital (81.03) and Avg_3 (79.69), and Gemini 3.6 Flash rises from Clean rank 6 to Real rank 1 (75.73). By architecture, Pipeline specialists lose 5.17/14.23 points, E2E specialists 4.14/12.99, and general VLMs 1.84/9.30 on average.

3.3 Component Analysis

Component strengths vary across models and architectures. In the three-track averages of Figure 4, Qwen3.8-27B leads Text (78.00), Gemini 3.6 Flash leads Formula (77.81), GLM-5.3-Flash leads Table (87.35), and TeleOCR leads Reading (75.14). Averaging equally over models within each architecture, E2E specialists achieve the highest mean Text/Reading scores (66.18/57.44), while general VLMs lead Formula/Table (62.48/70.07). The general-VLM advantage over E2E specialists is larger for Formula (7.39 points) than for Table (1.83 points). At the model level, the Avg_3 leader GLM-5.3-Flash scores 60.37 on Reading, 14.77 points below TeleOCR. Reading order is evaluated

Table 2: **Three-track parsing accuracy for 58 models across architectures.** Clean, Digital, and Real report per-track Overall on the same 1,475 source pages; the final column is their mean (Avg_3). Colored changes are relative to Clean (decrease/increase). **Bold**/underline: best/runner-up.

Model (release)	Clean↑	Digital↑		Real↑		Overall↑
Pipeline / multi-stage specialists						
PaddleOCR-VL-1.6 (2026-05)	64.46	59.29	(−5.17)	54.24	(−10.22)	59.33
YouTu-Parsing (2026-01)	75.02	69.66	(−5.36)	60.29	(−14.73)	68.32
DotsMOCR (2026-03)	76.27	73.16	(−3.11)	61.73	(−14.54)	70.39
GLM-OCR (2026-02)	68.65	63.06	(−5.59)	58.31	(−10.34)	63.34
MinerU2.5-Pro (2026-04)	75.87	71.77	(−4.10)	62.56	(−13.31)	70.07
MinerU2.5 (2025-09)	74.90	68.92	(−5.98)	59.15	(−15.75)	67.66
MonkeyOCRv2-B-Parsing (2026-07)	69.82	66.08	(−3.74)	58.83	(−10.99)	64.91
MonkeyOCR-pro-3B (2025-08)	62.23	57.40	(−4.83)	46.49	(−15.74)	55.37
MonkeyOCR-pro-1.2B (2025-07)	61.09	55.72	(−5.37)	43.82	(−17.27)	53.54
OpenDoc-0.1B (UniRec) (2025-12)	60.28	52.46	(−7.82)	44.27	(−16.01)	52.34
OpenOCR (2024-11)	32.70	30.03	(−2.67)	25.73	(−6.97)	29.49
Dolphin-v2 (2025-12)	65.90	60.24	(−5.66)	44.92	(−20.98)	57.02
TeleOCR (2026-08)	**87.16**	79.40	(−7.76)	69.08	(−18.08)	78.55
End-to-end specialists						
OvisOCR2 (2026-07)	81.53	77.29	(−4.24)	66.49	(−15.04)	75.10
ABot-OCR (2026-05)	76.92	72.43	(−4.49)	61.77	(−15.15)	70.37
Qianfan-OCR (2026-03)	57.22	50.85	(−6.37)	45.06	(−12.16)	51.04
Unlimited-OCR (2026-06)	72.05	65.30	(−6.75)	53.71	(−18.34)	63.69
WeVisDoc-4B (2026-09)	79.59	77.99	(−1.60)	69.14	(−10.45)	75.57
WeVisDoc-2B (2026-09)	79.08	76.33	(−2.75)	65.68	(−13.40)	73.69
HunyuanOCR-1.5 (2026-07)	73.09	69.36	(−3.73)	60.28	(−12.81)	67.58
HunyuanOCR (2025-11)	65.61	61.49	(−4.12)	54.58	(−11.03)	60.56
dots.ocr (2025-07)	72.01	65.95	(−6.06)	55.68	(−16.33)	64.55
FireRed-OCR (2026-02)	70.81	68.49	(−2.32)	57.42	(−13.39)	65.57
DeepSeek-OCR-2 (2026-01)	55.53	49.41	(−6.12)	43.60	(−11.93)	49.51
DeepSeek-OCR (2025-10)	53.50	46.95	(−6.55)	40.48	(−13.02)	46.98
olmOCR-2-7B (2025-10)	69.36	65.87	(−3.49)	56.10	(−13.26)	63.78
olmOCR-7B (2025-08)	62.56	57.84	(−4.72)	47.30	(−15.26)	55.90
OCRVerse (2026-01)	73.18	71.36	(−1.82)	63.66	(−9.52)	69.40
FD-RL (2025-11)	78.38	76.33	(−2.05)	67.04	(−11.34)	73.92
Nanonets-OCR2 (2025-10)	64.83	61.23	(−3.60)	49.03	(−15.80)	58.36
OCRFlux-3B (2025-06)	47.14	41.82	(−5.32)	37.21	(−9.93)	42.06
Logics-Parsing-v2 (2026-02)	76.35	73.85	(−2.50)	67.64	(−8.71)	72.61
General-purpose VLMs						
Qwen3.8-27B (2026-08)	80.45	78.86	(−1.59)	73.59	(−6.86)	77.63
Qwen3.8-Flash-Next (2026-08)	72.23	70.96	(−1.27)	62.42	(−9.81)	68.54
Qwen3.6-35B-A3B (2026-04)	72.14	69.16	(−2.98)	60.12	(−12.02)	67.14
Qwen3.6-27B (2026-04)	70.20	67.49	(−2.71)	59.18	(−11.02)	65.62
Qwen3.5-397B-A17B (2026-02)	69.12	68.34	(−0.78)	62.70	(−6.42)	66.72
Qwen3.5-122B-A10B (2026-02)	76.14	76.34	(+0.20)	69.85	(−6.29)	74.11
Qwen3.5-35B-A3B (2026-02)	68.40	68.04	(−0.36)	60.59	(−7.81)	65.68
Qwen3.5-27B (2026-02)	72.07	70.73	(−1.34)	65.92	(−6.15)	69.57
Qwen3.5-9B (2026-03)	73.87	73.34	(−0.53)	65.45	(−8.42)	70.89
Qwen3.5-4B (2026-03)	73.45	72.53	(−0.92)	63.47	(−9.98)	69.82
Qwen3.5-2B (2026-03)	66.24	65.22	(−1.02)	55.92	(−10.32)	62.46
Qwen3.5-0.8B (2026-03)	60.77	59.28	(−1.49)	47.93	(−12.84)	55.99
Qwen3-VL-8B (2025-10)	72.44	72.03	(−0.41)	62.73	(−9.71)	69.07
Qwen3-VL-4B (2025-10)	72.04	70.84	(−1.20)	59.61	(−12.43)	67.50
Qwen3-VL-2B (2025-10)	66.37	65.81	(−0.56)	54.09	(−12.28)	62.09
DeepSeek-V4-Flash-Vision-Exp (2026-08)	54.69	52.98	(−1.71)	42.68	(−12.01)	50.12
GLM-5.3-Flash (2026-08)	83.04	**81.03**	(−2.01)	74.99	(−8.05)	**79.69**
MiniCPM-V-4.6 (2026-05)	64.47	59.04	(−5.43)	48.01	(−16.46)	57.17
MiniCPM-V-4.5 (2025-08)	51.81	49.38	(−2.43)	37.59	(−14.22)	46.26
Gemini 3.6 Flash (2026-07)	79.97	79.11	(−0.86)	**75.73**	(−4.24)	78.27
Gemini-3.1-Pro (2026-02)	70.04	69.28	(−0.76)	71.98	(+1.94)	70.43
Claude Opus 5 (2026-07)	82.99	72.94	(−10.05)	75.27	(−7.72)	77.07
GPT-5.6 Sol (2026-07)	74.15	71.73	(−2.42)	61.59	(−12.56)	69.16
MiniMax-M3 (2026-06)	55.78	53.85	(−1.93)	42.54	(−13.24)	50.72
Kimi K2.6 (2026-04)	72.32	69.95	(−2.37)	68.02	(−4.30)	70.10
Step3-VL (2026-01)	53.65	52.74	(−0.91)	45.06	(−8.59)	50.48

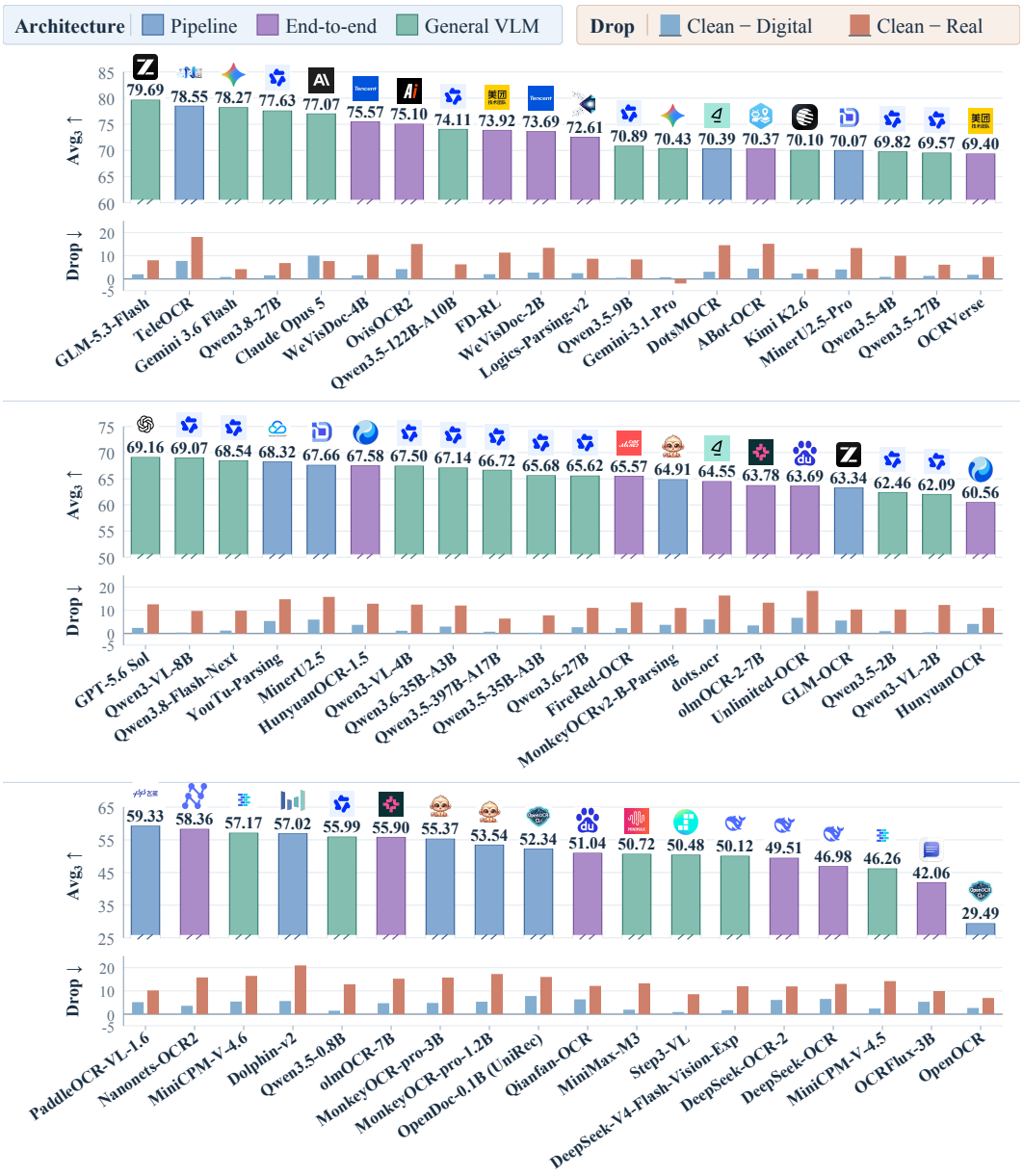


Figure 3: **Accuracy and robustness of 58 models.** Three-track mean accuracy (Avg_3) and losses relative to Clean reveal differences in parsing accuracy and robustness across the three architectural families. Losses are Clean−Digital and Clean−Real in points; negative values indicate gains.

separately from Overall and measures whether content is recovered in the correct sequence, including transitions between columns, text blocks, and other page elements.

3.4 Case Study

Figure 5 shows omissions, numerical errors, and failures in table structure, formulas, and symbol recognition across the three tracks. Appendix B provides 20 additional cases on different document pages.

Clean: omissions, values, and symbols. In Figure 5(b), Logics-Parsing-v2 omits the Social label and its 25% weight but retains the neighboring labels and all three ESG scores. In Figure 5(d),

Pipeline / multi-stage specialists

Model	Text	Formula	Table	Reading
PaddleOCR-VL-1.6	51.5	49.9	76.6	41.0
YouTu-Parsing	71.3	60.3	73.3	60.8
DotsMOCR	78.0	61.6	71.5	67.5
GLM-OCR	62.3	53.8	73.8	48.9
MinerU2.5-Pro	71.0	59.9	79.3	61.0
MinerU2.5	73.4	56.0	73.6	61.8
MonkeyOCRv2-B-Parsing	62.5	58.3	74.0	56.9
MonkeyOCR-pro-3B	58.2	44.1	63.9	46.1
MonkeyOCR-pro-1.2B	55.7	42.8	62.2	45.5
OpenDoc-0.1B (UniRec)	51.4	46.7	59.0	43.4
OpenOCR	58.3	30.1	0.0	45.4
Dolphin-v2	57.1	50.7	63.3	51.7
TeleOCR	77.8	74.4	83.5	75.1

End-to-end specialists

OvisOCR2	77.8	64.6	82.9	66.5
ABot-OCR	71.3	61.2	78.7	55.7
Qianfan-OCR	56.6	44.4	52.1	52.1
Unlimited-OCR	66.9	56.1	68.1	58.4
WeVisDoc-4B	74.3	67.8	84.6	63.3
WeVisDoc-2B	72.2	66.9	82.0	61.6
HunyuanOCR-1.5	70.2	62.9	69.6	57.3
HunyuanOCR	66.7	51.9	63.1	58.6
dots.ocr	68.1	55.2	70.3	57.9
FireRed-OCR	66.0	59.4	71.4	56.9
DeepSeek-OCR-2	58.3	41.4	48.9	50.3
DeepSeek-OCR	52.2	39.8	48.9	45.4
olmOCR-2-7B	66.0	53.4	71.9	60.8
olmOCR-7B	54.5	53.5	59.7	48.9
OCRVerse	68.7	61.6	77.9	58.0
FD-RL	76.5	64.7	80.5	64.2
Nanonets-OCR2	66.8	42.0	66.3	58.2
OCRFlux-3B	50.0	34.3	41.8	54.9
Logics-Parsing-v2	74.5	65.6	77.8	62.2

General-purpose VLMs

Model	Text	Formula	Table	Reading
Qwen3.8-27B	78.0	70.7	84.2	66.9
Qwen3.8-Flash-Next	63.5	69.3	72.8	56.0
Qwen3.6-35B-A3B	64.8	65.9	70.7	57.4
Qwen3.6-27B	60.7	64.1	72.1	53.1
Qwen3.5-397B-A17B	74.5	63.3	62.4	62.0
Qwen3.5-122B-A10B	75.8	66.0	80.6	61.9
Qwen3.5-35B-A3B	73.8	63.1	60.1	60.9
Qwen3.5-27B	75.1	64.1	69.5	62.7
Qwen3.5-9B	71.8	65.2	75.7	59.3
Qwen3.5-4B	68.8	66.7	74.0	56.7
Qwen3.5-2B	62.1	57.4	68.0	51.0
Qwen3.5-0.8B	58.0	51.1	58.9	47.8
Qwen3-VL-8B	71.0	61.8	74.3	57.7
Qwen3-VL-4B	69.6	61.3	71.6	55.9
Qwen3-VL-2B	65.3	56.8	64.2	53.4
DeepSeek-V4-Flash-Vision-Exp	41.9	53.6	54.8	41.0
GLM-5.3-Flash	74.3	77.4	87.3	60.4
MiniCPM-V-4.6	58.2	52.3	61.1	51.5
MiniCPM-V-4.5	50.6	40.3	48.0	49.3
Gemini 3.6 Flash	72.3	77.8	84.7	60.7
Gemini-3.1-Pro	69.1	66.7	75.5	59.6
Claude Opus 5	72.1	73.5	85.5	61.5
GPT-5.6 Sol	59.8	70.0	77.7	53.7
MiniMax-M3	44.5	50.6	57.1	43.9
Kimi K2.6	68.0	64.7	77.6	52.6
Step3-VL	47.0	50.8	53.7	46.3

Three-track mean score ↑

0 25 50 75 100

Text / Reading: 100 × (1 − Edit)

Formula: CDM · Table: TEDS

Figure 4: **Component profiles of 58 models.** Three-track means reveal distinct strengths in text, formulas, tables, and reading order. Text and Reading use $100(1-\text{Edit})$, Formula uses CDM, and Table uses TEDS. All scores range from 0 to 100, with higher values indicating better performance.

GLM-OCR changes a bill-of-lading total from USD 426,370.00 to USD 420,370.00 while preserving the correct amount in words.

In Figure 5(e), HunyuanOCR omits the $3,845.00 amount cell from the medical bill's *Itemized Charges* total row and expands the preceding label cell from five to six columns. The same amount remains in the following table, localizing the omission to the first table's total row. Nanonets-OCR2 converts the time subscript of ${}_t\bar{V}$ into division by t in an actuarial formula (Figure 5(f)). FD-RL reads the thickness notation "$\delta 3$–$\delta 50$" as "53-550" (Figure 5(g)).

Digital: percentage substitution. In a hotel reservation dashboard with uneven lighting, Gemini-3.1-Pro changes a clearly readable revenue-growth percentage from +11% to 112% (Figure 5(c)). Its paired Clean and Real predictions preserve +11%.

Real: omitted content. On an oblique photograph of a creased menu, MinerU2.5-Pro omits "Hazelnut" from "Jerusalem Artichoke / Hazelnut / Truffle" (Figure 5(a)). Its corresponding Clean and Digital predictions preserve the complete ingredient list.

4 Conclusion and Future Work

Conclusion. PureDocBench provides source-linked annotations and matched Clean, Digital, and Real images across 10 domains and 66 subcategories. Results from 58 models show room for improvement across parsing components and image conditions. Compact specialists can match or outperform much larger general models. We release the benchmark, document sources, generation tools, and evaluation code to support reproducible comparisons and further research.

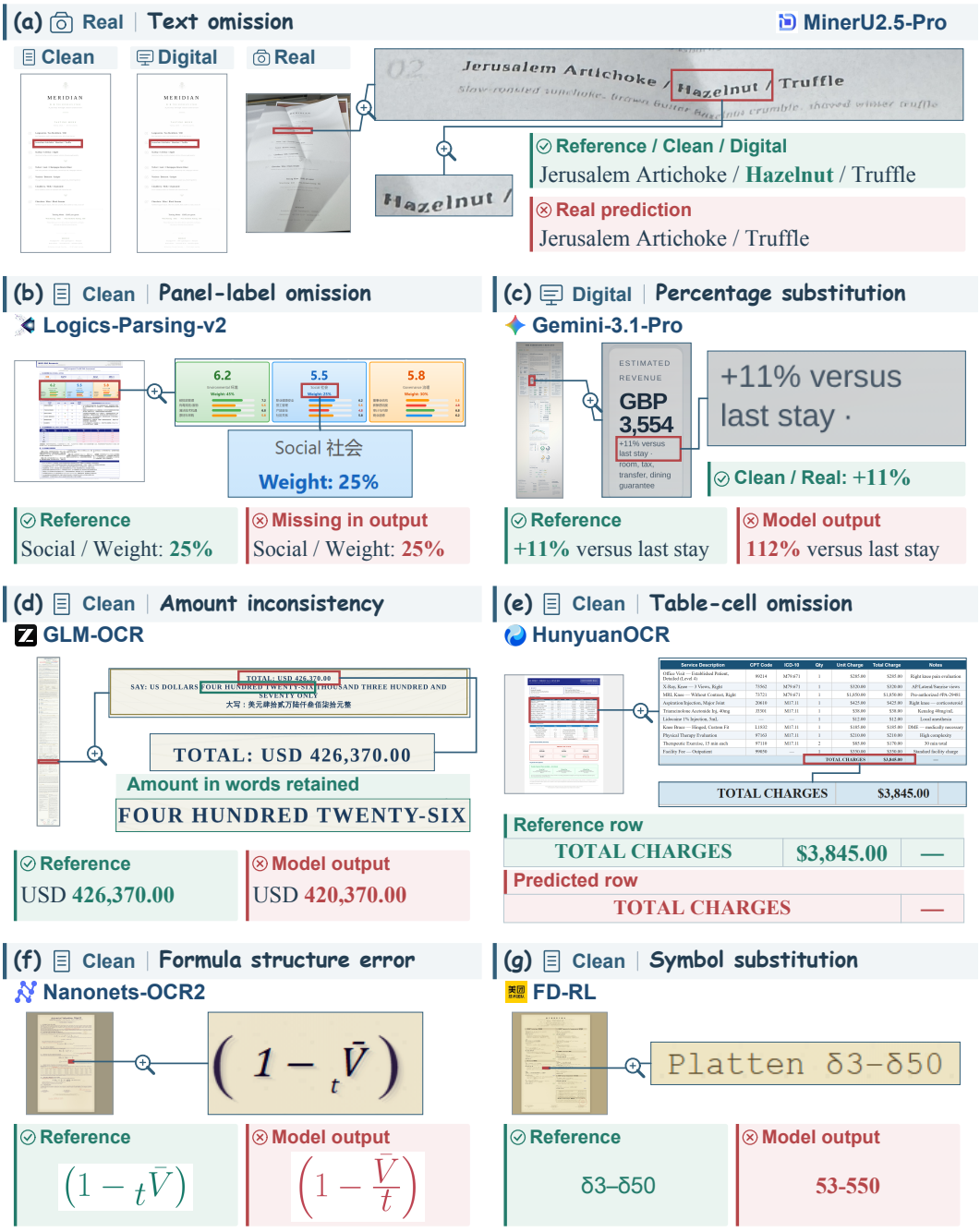


Figure 5: **Parsing failures across three image tracks.** Examples include (a) omitted ingredients, (b) missing panel labels, (c) altered percentages, (d) inconsistent numeric and written amounts, (e) a missing table cell, (f) a subscript rewritten as division, and (g) Greek symbols read as digits. These errors affect content completeness, numerical accuracy, and document structure.

Limitations and Future Work. (i) We adopt OmniDocBench's evaluation framework, whose metrics need refinement to better capture semantic omissions and structural consistency. (ii) HTML rendering has limited coverage of handwriting and historical books; these settings require additional real-document data. (iii) Human review identified 83 Real images that are very difficult to read, mainly recaptures of long pages. Appendix C analyzes these cases separately: excluding them leaves model rankings largely stable and preserves our main comparative conclusions. Future acquisition should retain readable detail and record quality flags to separate capture limits from model errors.

AI Use Statement. AI tools were used for manuscript formatting, data organization, and discussion of research ideas.

Ethics Statement. PureDocBench uses generated documents and recaptures of those pages for research evaluation. Applications involving sensitive information should include privacy safeguards and human verification.

Reproducibility Statement. The code link is provided in the abstract.

REFERENCES

Xin An, Jingyi Cai, Xiangyang Chen, Huayao Liu, Peiting Liu, Peng Wang, Bei Yang, Xiuwen Zhu, Yongfan Chen, Baoyu Hou, Shuzhao Li, Weidong Ren, Fan Yang, Jiangtao Zhang, Xiaoxiao Xu, and Lin Qu. Logics-Parsing-Omni technical report. *arXiv preprint arXiv:2603.09677*, 2026. URL https://arxiv.org/abs/2603.09677.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fUtxNAKpdV.

Łukasz Borchmann, Jordy Van Landeghem, Michał Turski, Shreyansh Padarha, Ryan Othniel Kearns, Adam Mahdi, Niels Rogge, Clémentine Fourrier, Siwei Han, Huaxiu Yao, Artemis Llabrés, Yiming Xu, Dimosthenis Karatzas, Hao Zhang, and Anupam Datta. Strategic navigation or stochastic search? how agents and humans reason over document collections, 2026. URL https://arxiv.org/abs/2603.12180.

ChatDOC Team. OCRFlux-3B. https://huggingface.co/ChatDOC/OCRFlux-3B, 2025. Hugging Face model card.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. PaddleOCR-VL-1.5: Towards a multi-task 0.9B VLM for robust in-the-wild document parsing. *arXiv preprint arXiv:2601.21957*, 2026.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document AI: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*, 2021.

Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush. Image-to-markup generation with coarse-to-fine attention. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 980–989. PMLR, 2017. URL https://proceedings.mlr.press/v70/deng17a.html.

Daxiang Dong, Mingming Zheng, Dong Xu, Chunhua Luo, Bairong Zhuang, Yuxuan Li, Ruoyun He, Haoran Wang, Wenyu Zhang, Wenbo Wang, et al. Qianfan-OCR: A unified end-to-end model for document intelligence. *arXiv preprint arXiv:2603.13398*, 2026.

Yongkun Du, Zhineng Chen, and Yu-Gang Jiang. OpenOCR: An open-source toolkit for general-OCR research and applications. https://github.com/Topdu/OpenOCR, 2024.

Yongkun Du, Pinxuan Chen, Xuye Ying, and Zhineng Chen. DocPTBench: Benchmarking end-to-end photographed document parsing and translation. *arXiv preprint arXiv:2511.18434*, 2025a.

Yongkun Du, Zhineng Chen, Yazhen Xie, Weikang Bai, Hao Feng, Wei Shi, Yuchen Su, Can Huang, and Yu-Gang Jiang. OpenDoc-0.1B: An ultra-lightweight document parsing system. https://github.com/Topdu/OpenOCR/blob/main/docs/opendoc.md, 2025b.

Yongkun Du, Zhineng Chen, Yazhen Xie, Weikang Bai, Hao Feng, Wei Shi, Yuchen Su, Can Huang, and Yu-Gang Jiang. UniRec-0.1B: Unified text and formula recognition with 0.1B parameters. *arXiv preprint arXiv:2512.21095*, 2025c.

Shuaiqi Duan, Yadong Xue, Weihan Wang, Zhe Su, Huan Liu, Sheng Yang, Guobing Gan, Guo Wang, Zihan Wang, Shengdong Yan, et al. GLM-OCR technical report. *arXiv preprint arXiv:2603.10910*, 2026.

Hao Feng, Wei Shi, Ke Zhang, Xiang Fei, Lei Liao, Dingkang Yang, Yongkun Du, Xuecheng Wu, Jingqun Tang, Yang Liu, Hong Chen, and Can Huang. Dolphin-v2: Universal document parsing via scalable anchor prompting. *arXiv preprint arXiv:2602.05384*, 2026.

Fortune Business Insights. Intelligent document processing market size, share & industry analysis. Technical Report FBI108590, Fortune Business Insights, 2026. Available at https://www.fortunebusinessinsights.com/intelligent-document-processing-market-108590. Accessed: 2026-05-07.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3253. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8c2e6bb15be1894b8fb4e0f9bcad1739-Abstract-Datasets_and_Benchmarks_Track.html.

Gemini Team. Gemini 3.1 Pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, February 2026. Google Blog post.

Grand View Research. Intelligent document processing market size report, 2030. Technical Report GVR-4-68040-023-3, Grand View Research, 2025. Available at https://www.grandviewresearch.com/industry-analysis/intelligent-document-processing-market-report. Accessed: 2026-05-07.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, et al. STEP3-VL-10B technical report. *arXiv preprint arXiv:2601.09668*, 2026a.

Mingxin Huang, Yongxin Shi, Dezhi Peng, Songxuan Lai, Zecheng Xie, and Lianwen Jin. OCR-Reasoning Benchmark: Unveiling the true capabilities of MLLMs in complex text-rich image reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=aH7eyx64pC.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. ICDAR2019 competition on scanned receipt OCR and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019. doi: 10.1109/ICDAR.2019.00244. URL https://arxiv.org/abs/2103.10213.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, Qi Yang, et al. HunyuanOCR technical report. *arXiv preprint arXiv:2511.19575*, 2025.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pp. 1–6, 2019. doi: 10.1109/ICDARW.2019.10029. URL https://arxiv.org/abs/1905.13538.

Kimi Team. Kimi K2.6. https://platform.kimi.com/docs/guide/kimi-k2-6-quickstart, 2026. Kimi API documentation.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1918–1925. European Language Resources Association, 2020a. URL https://aclanthology.org/2020.lrec-1.236/.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 949–960, 2020b. doi: 10.18653/v1/2020.coling-main.82. URL https://aclanthology.org/2020.coling-main.82/.

Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. dots.ocr: Multilingual document layout parsing in a single vision-language model. *arXiv preprint arXiv:2512.02498*, 2025a.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Biao Yang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. MonkeyOCR: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218*, 2025b.

Zhang Li, Zhibo Lin, Qiang Liu, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiajun Song, Jiarui Zhang, Xiang Bai, and Yuliang Liu. MDPBench: A benchmark for multilingual document parsing in real-world scenarios. *arXiv preprint arXiv:2603.28130*, 2026.

Zichao Li, Aizier Abulaiti, Yaojie Lu, Xuanang Chen, Jia Zheng, Hongyu Lin, Xianpei Han, Shanshan Jiang, Bin Dong, and Le Sun. READoc: A unified benchmark for realistic document structured extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21889–21905, 2025c. doi: 10.18653/v1/2025.findings-acl.1128. URL https://aclanthology.org/2025.findings-acl.1128/.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. doi: 10.1007/s11432-024-4235-6. URL https://doi.org/10.1007/s11432-024-4235-6.

Logics-MLLM Team. Logics-Parsing-v2. https://huggingface.co/Logics-MLLM/Logics-Parsing-v2, 2026. Hugging Face model card; accessed 2026-05-07.

Jiefeng Ma, Jun Du, Pengfei Hu, Zhenrong Zhang, Jianshu Zhang, Huihui Zhu, and Cong Liu. HRDoc: Dataset and baseline method toward hierarchical reconstruction of document structures. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1870–1877, 2023. doi: 10.1609/aaai.v37i2.25277. URL https://ojs.aaai.org/index.php/AAAI/article/view/25277.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95963–96010, 2024. doi: 10.52202/079017-3041. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ae0e43289bffea0c1fa34633fc608e92-Abstract-Datasets_and_Benchmarks_Track.html.

Souvik Mandal, Ashish Talewar, Siddhant Thakuria, Paras Ahuja, and Prathamesh Juvatkar. Nanonets-OCR2: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging. https://huggingface.co/nanonets/Nanonets-OCR2-3B, 2025. Hugging Face model card.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Mathew_InfographicVQA_WACV_2022_paper.html.

Harold Mouchère, Christian Viard-Gaudin, Dae Hwan Kim, Jin Hyung Kim, and Utpal Garain. CROHME2011: Competition on recognition of online handwritten mathematical expressions. In *2011 International Conference on Document Analysis and Recognition*, pp. 1497–1500, 2011. doi: 10.1109/ICDAR.2011.297. URL https://doi.org/10.1109/ICDAR.2011.297.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, et al. MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*, 2025.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, et al. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. CORD: A consolidated receipt dataset for post-OCR parsing. In *Document Intelligence Workshop at Neural Information Processing Systems*, 2019. URL https://openreview.net/forum?id=SJl3z659UH.

Chunyi Peng, Zhipeng Xu, Yuqi Xiong, Zulong Chen, Junhao Ji, Qing Liu, Zhenghao Liu, Zubao Qin, Bing Zhao, Jianqiang Wan, Jun Tang, Zhibo Yang, Shuai Bai, Dayiheng Liu, and Maosong Sun. CC-OCR V2: Fine-grained attribution of LMM failures in real-world visual document understanding. *arXiv preprint arXiv:2605.03903*, 2026. URL https://arxiv.org/abs/2605.03903v2.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter W. J. Staar. DocLayNet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. doi: 10.1145/3534678.3539043. URL https://arxiv.org/abs/2206.01062.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmOCR: Unlocking trillions of tokens in PDFs with vision language models. *arXiv preprint arXiv:2502.18443*, 2025a.

Jake Poznanski, Luca Soldaini, and Kyle Lo. olmOCR 2: Unit test rewards for document OCR. *arXiv preprint arXiv:2510.19817*, 2025b.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026. Model card citation; accessed 2026-05-07.

Brandon Smock, Rohith Pesala, and Robin Abraham. 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. URL https://arxiv.org/abs/2110.00061.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickael Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanislawek. Document understanding dataset and evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19528–19540, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Van_Landeghem_Document_Understanding_Dataset_and_Evaluation_DUDE_ICCV_2023_paper.html.

Abhigya Verma, Khyati Mahajan, Amit Kumar Saha, Shruthan Radhakrishna, Sagar Davasam, Vikas Yadav, and Sai Rajeswar. SynthDocBench: Controlled benchmark for long-context visual document understanding. *arXiv preprint arXiv:2607.10400*, 2026. URL https://arxiv.org/abs/2607.10400.

Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Bo Zhang, and Conghui He. Image over text: Transforming formula recognition evaluation with character detection matching. *arXiv preprint arXiv:2409.03643*, 2024.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, et al. MinerU2.5-Pro: Pushing the limits of data-centric document parsing at scale. *arXiv preprint arXiv:2604.04771*, 2026a.

Hang Wang, Jin Zhang, Guoliang Xu, Pengyue Lu, Yao Li, Zijiao Zhang, Tianyu Huang, Weiqi Xiong, Yulong Wang, Chuqiao Lu, Wenkang Huang, Kai Yang, Yadong Li, Hui Li, Xingzhong Xu, and Xiao Xu. FinixDoc: Rethinking financial document parsing beyond saturated benchmarks. *arXiv preprint arXiv:2608.22842*, 2026b. URL https://arxiv.org/abs/2608.22842.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4735–4744, 2021. doi: 10.18653/v1/2021.emnlp-main.389. URL https://aclanthology.org/2021.emnlp-main.389/.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*, 2026.

Hao Wu, Haoran Lou, Xinyue Li, Zuodong Zhong, Zhaojun Sun, Phellon Chen, Xuanhe Zhou, Kai Zuo, Yibo Chen, Xu Tang, et al. FireRed-OCR technical report. *arXiv preprint arXiv:2603.01840*, 2026.

Long Xu, Binghong Wu, Tinghao Yu, Hao Feng, Zhenyu Huang, Haoqing Jiang, Yunhao Wang, Shuo Huang, and Feng Zhang. MORE: A multilingual document parsing benchmark and evaluation. *arXiv preprint arXiv:2607.02956*, 2026. URL https://arxiv.org/abs/2607.02956.

Minglai Yang, Xinyan Velocity Yu, Pengyuan Li, Xinyu Guo, Zhenting Qi, Konwoo Kim, Longtian Ye, Xiaolong Luo, Jinhe Bi, Henry Zhang, Haris Riaz, Xuan Zhang, Yunze Xiao, Bangya Liu, Tom Tang, Yunfei Zhao, Qunshu Lin, Zihan Wang, Minghao Liu, Michael Lingzhi Li, Yilun Du, Jesse Thomason, Rogerio Feris, Alex Pentland, and Zexue He. Dr. DocBench: A comprehensive benchmark for expert-level and difficult document parsing. *arXiv preprint arXiv:2606.01393*, 2026. URL https://arxiv.org/abs/2606.01393.

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, and Christopher Clark. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17486–17505, 2025a. doi: 10.18653/v1/2025.acl-long.855. URL https://aclanthology.org/2025.acl-long.855/.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, Lianwen Jin, and Junyang Lin. CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21744–21754, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_CC-OCR_A_Comprehensive_and_Challenging_OCR_Benchmark_for_Evaluating_Large_ICCV_2025_paper.html.

Kun Yin, Yunfei Wu, Bing Liu, Zhongpeng Cai, Xiaotian Li, Huang Chen, Xin Li, Haoyu Cao, Yinsong Liu, Deqiang Jiang, Xing Sun, Yunsheng Wu, Qianyu Li, Antai Guo, Yanzhen Liao, Yanqiu Qu, Haodong Lin, Chengxu He, and Shuangyin Liu. Youtu-parsing: Perception, structuring and recognition via high-parallelism decoding. *arXiv preprint arXiv:2601.20430*, 2026.

Hao Yu, Kang Liu, Linnan Zhao, Jiabo Zhan, Chong Sun, Chen Li, and Jing Lyu. WeVisDoc: From coverage to capability for robust end-to-end document parsing. *arXiv preprint arXiv:2609.20423*, 2026. URL https://arxiv.org/abs/2609.20423.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Boyang Zhang, Sebastián G. Acosta, Preston Carlson, Sacha Bron, Pierre-Loïc Doulcet, Daniel B. Ospina, and Simon Suo. ParseBench: A document parsing benchmark for AI agents. *arXiv preprint arXiv:2604.08538*, 2026. URL https://arxiv.org/abs/2604.08538.

Handong Zheng, Yumeng Li, Kaile Zhang, Liang Xin, Guangwei Zhao, Hao Liu, Jiayu Chen, Jie Lou, Qi Fu, Rui Yang, et al. Multimodal OCR: Parse anything from documents. *arXiv preprint arXiv:2603.13032*, 2026.

Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 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. URL https://arxiv.org/abs/2005.00589.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. PubLayNet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1015–1022, 2019. doi: 10.1109/ICDAR.2019.00166. URL https://arxiv.org/abs/1908.07836.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: Data, model, and evaluation. In *Computer Vision – ECCV 2020*, pp. 564–580, 2020. doi: 10.1007/978-3-030-58589-1_34. URL https://arxiv.org/abs/1911.10683.

Yufeng Zhong, Lei Chen, Zhixiong Zeng, Xuanle Zhao, Deyang Jiang, Liming Zheng, Jing Huang, Haibo Qiu, Peng Shi, Siqi Yang, and Lin Ma. Reading or reasoning? format decoupled reinforcement learning for document OCR. *arXiv preprint arXiv:2601.08834*, 2025.

Yufeng Zhong, Lei Chen, Xuanle Zhao, Wenkang Han, Liming Zheng, Jing Huang, Deyang Jiang, Yilin Cao, Lin Ma, and Zhixiong Zeng. OCRVerse: Towards holistic OCR in end-to-end vision-language models. *arXiv preprint arXiv:2601.21639*, 2026.

Bangbang Zhou, Hangdi Xing, Yifan Chen, Jianjun Xu, Qi Zheng, Feiyu Gao, Zhibo Yang, Shuai Bai, Ming Yan, Jieping Ye, and Hongtao Xie. MPDocBench-Parse: Benchmarking practical multi-page document parsing. *arXiv preprint arXiv:2605.22100*, 2026a. URL https://arxiv.org/abs/2605.22100.

Changda Zhou, Ziyue Gao, Xueqing Wang, Tingquan Gao, Cheng Cui, Jing Tang, and Yi Liu. Real5-OmniDocBench: A full-scale physical reconstruction benchmark for robust document parsing in the wild. *arXiv preprint arXiv:2603.04205*, 2026b.

Yinan Zhou, Yuxin Chen, Haokun Lin, Yichen Wu, Shuyu Yang, Zhongang Qi, Chen Ma, and Li Zhu. DOGR: Towards versatile visual document grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3596–3606, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhou_DOGR_Towards_Versatile_Visual_Document_Grounding_and_Referring_ICCV_2025_paper.html.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. MMDocBench: Benchmarking large vision-language models for fine-grained visual document understanding and grounding. In *MultiMedia Modeling*, volume 16412 of *Lecture Notes in Computer Science*, pp. 74–88. Springer, 2026. doi: 10.1007/978-981-95-6950-2_6. URL https://doi.org/10.1007/978-981-95-6950-2_6.

Appendix Contents

A Benchmark Comparison: Scope, Versions, and Evidence 16
B Additional Case-Study Evidence 19
C Real-Track Readability Audit 29
D Full Component Results 31
E OmniDocBench Annotation Error Examples 34
F Degradation Parameter Details 35
G Full Per-Category Results 37

A Benchmark Comparison: Scope, Versions, and Evidence

Table 1 summarizes complementary datasets using the papers and author-maintained releases reviewed on September 25, 2026. Its grouping separates component recognition, document understanding and synthetic supervision, and full-document parsing. The layout is inspired by the task-oriented comparison in MADQA (Borchmann et al., 2026); the dimensions here are selected for parsing and annotation provenance.

Definitions. *Corpus size* retains the units of the cited publication or named release. Historical training-and-test corpora and evaluation-only sets are both included; their sizes are not directly comparable measures of evaluation strength. Rounded "k" counts indicate thousands. The four *parsing target* columns cover text transcription, table structure, formula recognition, and reading order. Layout detection and question answering provide complementary context outside the displayed task columns. A check denotes an explicitly evaluated task; a half-filled circle denotes a restricted subtask or subset; a dash denotes a dimension not reported in the reviewed protocol. These symbols summarize scope, not metric quality. QA answers and auxiliary OCR strings alone do not count as a text-transcription task. The *label source* icons denote markup-derived targets, model assistance, human annotation/review, and inherited labels; combinations summarize contributing processes. Markup derivation does not imply that every resource retains regenerable sources. The *paired image variants* columns distinguish digital transformations from physical acquisitions of the same underlying page. Naturally noisy images alone do not establish pairing. Multipliers denote base pages times views: Real5 counts five physical captures; MDPBench and PureDocBench include the original in their four and three views, respectively.

Restricted task scopes. PubTabNet's text task covers table cells. OCRBench's formula subset contains 100 handwritten expressions. olmOCR-bench explicitly tests content, formulas, and reading order; its table tests concern cell relations. Its markup icon refers to the arXiv-math subset. ParseBench evaluates text, tables, reading order, and layout grounding; its formula check concerns LaTeX notation rather than complete formula-content attribution. MPDocBench-Parse additionally evaluates figure extraction, hierarchy, and cross-page continuity. DocPTBench's physical captures cover a 100-page subset of its 981 base pages. A missing label-source entry for LogicsDocBench records undisclosed benchmark-specific construction details. Additional capabilities, including layout, question answering, hierarchy, translation, charts, code, and catalogs, are detailed below.

Component datasets. PubLayNet (Zhong et al., 2019) reports over 360,000 document images with PDF–XML alignment, while DocBank (Li et al., 2020b) uses LaTeX to label tokens on 500,000 pages. DocLayNet (Pfitzmann et al., 2022) provides 80,863 human-annotated pages for layout segmentation. TableBank (Li et al., 2020a) contains 417,234 table instances on 278,582 page images for detection; its separate structure-recognition task contains 145,463 table images and predicts structural tags. PubTabNet (Zhong et al., 2020) pairs approximately 568,000 table crops with HTML and introduces TEDS. PubTables-1M (Smock et al., 2022) contains 947,642 tables on 575,305 pages; canonicalization and functional labels enrich its structural evaluation. Its original detection-based recognition model uses separately supplied text for text-bearing outputs. The im2latex corpus (Deng et al., 2017) contains 103,556 rendered expressions, rounded to 104k in the table. These resources establish clear precedents for source-derived supervision. ReadingBank (Wang et al., 2021) extends this lineage to Word-derived reading order, and HRDoc (Ma et al., 2023) targets document hierarchies.

Form and receipt datasets (Jaume et al., 2019; Huang et al., 2019; Park et al., 2019) and financial tables (Zheng et al., 2021) provide further task/domain coverage.

Understanding tasks. DocVQA (Mathew et al., 2021) uses 12,767 document images for question answering. OCRBench v2 (Fu et al., 2025) has 10,000 public instruction–answer samples and includes full-page OCR, element parsing, and reasoning; its table entry refers to this public set. Construction reuses existing dataset GT, adds human labels where needed, and checks instructions and answers. MADQA (Borchmann et al., 2026) uses 800 PDFs and 2,250 human-authored questions for retrieval and agentic answering. Other benchmarks discussed in Section 2.2 test infographic reasoning (Mathew et al., 2022), heterogeneous multi-page QA (Van Landeghem et al., 2023), long-document QA (Ma et al., 2024), or visual text reasoning and grounded perception (Liu et al., 2024; Huang et al., 2026b; Zhu et al., 2026). Answer correctness and evidence grounding complement the content/structure recovery measured here; our parsing scores do not establish downstream QA or agent performance.

OCR and grounded understanding. The original OCRBench (Liu et al., 2024) contains 1,000 manually checked QA pairs, including 300 text-recognition and 100 handwritten-formula samples drawn from existing datasets. CC-OCR (Yang et al., 2025b) contains 7,058 images across 39 subsets, covering scene and multilingual OCR, document/table/formula parsing, and information extraction. Its labels combine inherited annotations, model preannotations, and manual correction; reading order has no separate evaluation metric. DOGR-Bench (Zhou et al., 2025) contains 3,600 samples across seven grounding and referring tasks, built with rendered-source/PDF annotations and model-generated instructions. Its grounded-answer metrics assess text spans and their boxes; its four parsing columns therefore remain unreported under our task definitions. The larger DOGR training corpus is a separate resource. Physical captures in CC-OCR do not establish a paired acquisition protocol.

CC-OCR v2. The August 2026 revision of CC-OCR v2 (Peng et al., 2026) reports 7,093 evaluation samples across five tracks and 16 subtasks, with ten document categories, 32 recognition languages, and ten diagnostic factors. It reconstructs CC-OCR through sample refinement, new collection, inherited labels, model-assisted quality control, and human annotation/review. Text recognition uses multi-set F1; document/formula parsing uses edit similarity and tables use TEDS. Reading order is discussed within parsing but has no separately reported score. Acquisition-factor labels characterize observed conditions; the protocol does not report paired digital/physical views of identical content.

Synthetic supervision and controlled evaluation. CoSyn-400K (Yang et al., 2025a) is a **training corpus**: 400,000 code-rendered images with 2.7 million instruction-tuning examples. Its source-code-conditioned generation supplies related prior art for synthetic supervision; task scores are evaluated on separate benchmarks. SynthDocBench (Verma et al., 2026) evaluates 1,788 questions over 200 synthetic reports, averaging 51.1 pages, using controlled HTML layouts and source-backed chart data. Labels combine model generation, metadata-based numerical checks, consistency filtering, and a 100-question human review. It evaluates long-document chart reading, cross-modal reasoning, and multi-hop QA. Its optional scanned-PDF renderer and resolution ablations do not define a matched digital/physical degradation suite, so paired-variant checks are left unreported.

READoc and OmniDocBench. We use the ACL 2025 version of READoc (Li et al., 2025c), with 3,576 documents from arXiv, GitHub, and Zenodo. Its pipeline derives Markdown targets from LaTeX, Markdown, DOCX, and HTML, including source-to-PDF rendering. Together with Nougat (Blecher et al., 2024), it is relevant prior art for markup-aligned parsing. For OmniDocBench (Ouyang et al., 2025), the table uses **v1.5 / 1,355 pages**. Illustrative annotation errors from this version appear in Appendix E.

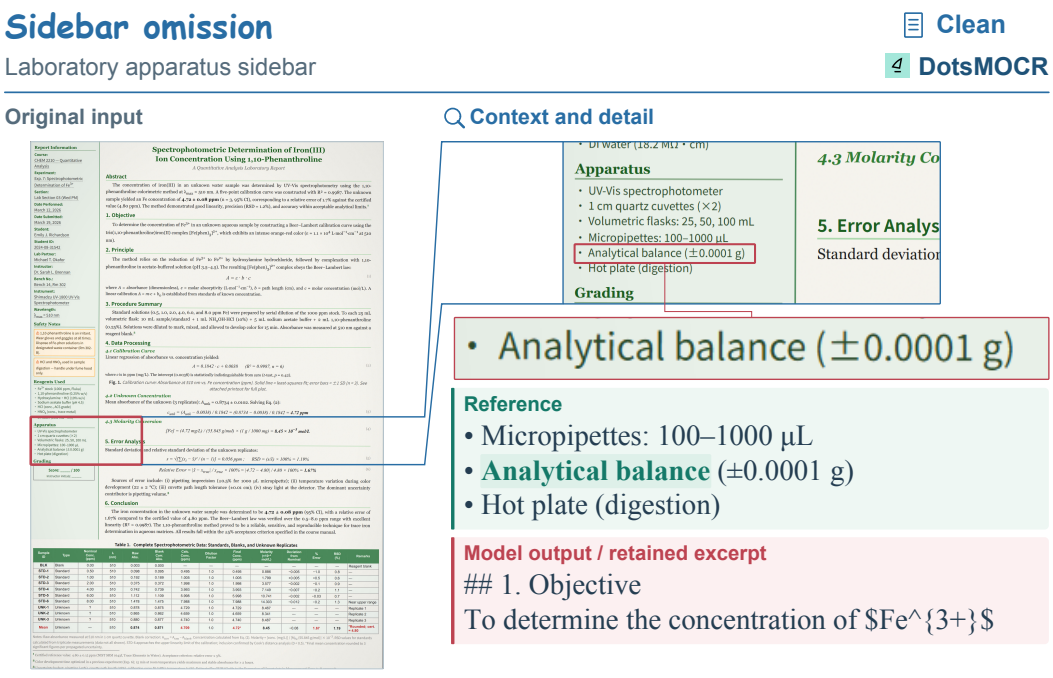
Matched robustness datasets. DocPTBench (Du et al., 2025a) starts with 981 OmniDocBench pages, adds 981 digitally simulated photographs, and physically captures a 100-page subset under four conditions (400 images). Thus 1,381 counts altered images, while 2,362 counts originals plus altered images; the additionally released rectified derivatives are a separate set. Its parsing targets inherit the base annotations, and translations undergo checking. Real5-OmniDocBench (Zhou et al., 2026b) applies five physical conditions to all 1,355 v1.5 pages, yielding 6,775 variants with inherited GT. MDPBench (Li et al., 2026) contains 850 original pages in 17 languages and three photographs

per page: 3,400 images including originals. These are concrete precedents for paired acquisition evaluation. PureDocBench pairs all 1,475 freshly generated pages across clean, digital, and real tracks and retains their common HTML/CSS sources for inspection and regeneration.

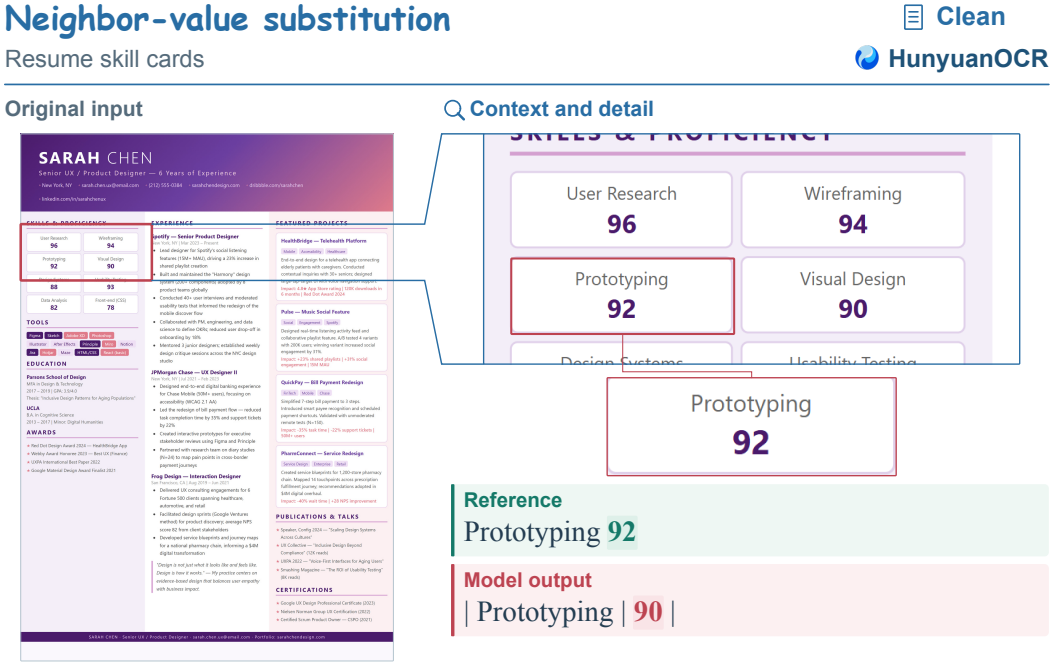
Recent parsing suites. olmOCR-bench is introduced in the June 2025 revision of olmOCR (Poznanski et al., 2025a); the 1,403-page / 7,010-test count follows Appendix F, Table 10. Its unit tests cover content, mathematical expressions, table relations, and order; targets combine model generation, human design, and review. The arXiv-math subset aligns to original TeX, with manual verification, as described in the benchmark documentation. The 900-page document module in the Logics-Parsing-Omni report (An et al., 2026) is named *LogicsDocBench*; OmniParsingBench denotes the broader suite. Its document evaluation reports text edit distance, formula CDM, table TEDS/TEDS-S, and reading order; its training-data annotation description does not by itself establish the benchmark GT procedure. ParseBench (Zhang et al., 2026) v3 reports 2,078 pages and 169,011 rules across tables, charts, content, semantic formatting, and grounding, with VLM-assisted targets and human verification.

Languages, multiple pages, and specialized content. MORE (Xu et al., 2026) reports 149 languages and six tasks. Its official repository flags a count discrepancy between 1,237 pages in the collection description and 1,288 images in the statistics table; we use the transparent approximation $\sim$1.3k. MPDocBench-Parse (Zhou et al., 2026a) contains 433 documents / 3,246 pages, with model preannotations, human correction, and structural checks, and evaluates hierarchy and cross-page continuity alongside content recovery. Dr. DocBench (Yang et al., 2026) contains 4,514 annotated pages across 52 subjects, selected using parser disagreement and corrected by humans with expert review for specialized content. FinixDocBench (Wang et al., 2026b), discussed in the text, covers approximately 5,000 pages across digital, photographed, large-page, and internal financial-workflow subsets; its report releases a subset, and paired original/photo identities are not specified. These settings complement our predominantly Chinese–English, page-level, source-rendered evaluation.

B Additional Case-Study Evidence



The apparatus list in the left sidebar is absent from the output. The main-body Objective section is retained.



The output assigns 90 to Prototyping, whose card reads 92. The value 90 belongs to the adjacent Visual Design card.

Figure 6: **Clean-track omissions and association errors.** A laboratory sidebar is omitted, and a skill score is copied from its neighboring label, affecting content completeness and field accuracy.

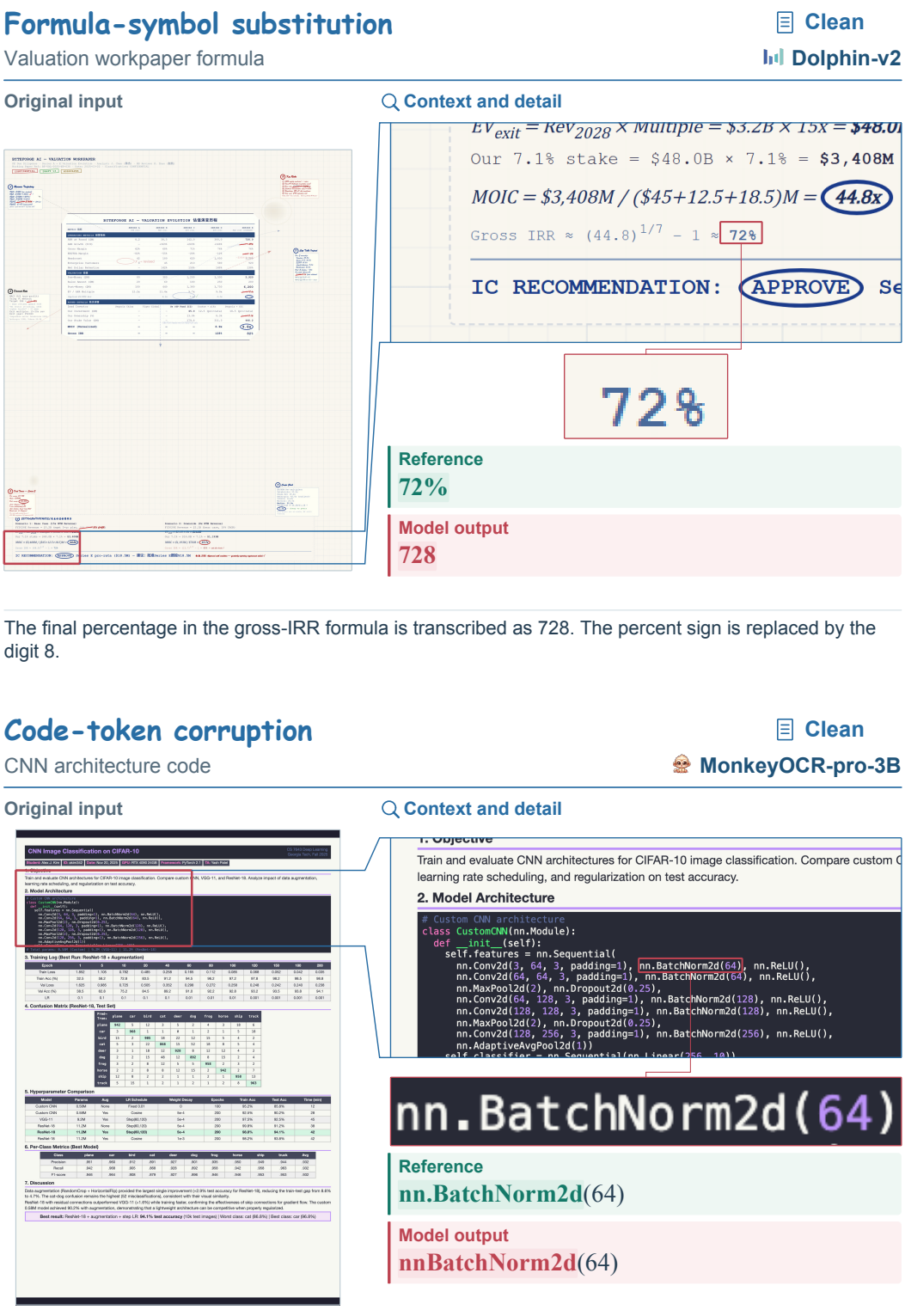


Figure 7: **Clean-track symbol and code errors.** A percent sign becomes a digit in a financial formula, and a code identifier loses its separator, altering a numerical value and a callable identifier.

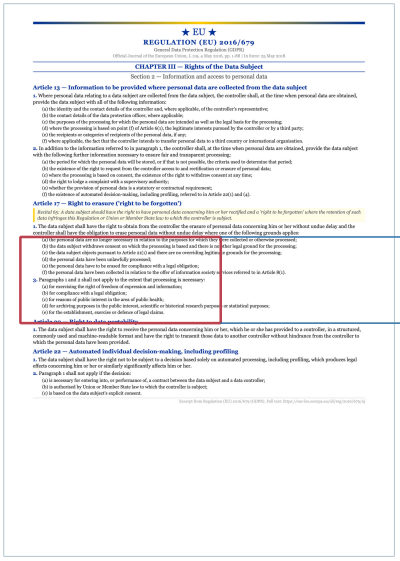
Paragraph-number substitution

Legal paragraph numbering

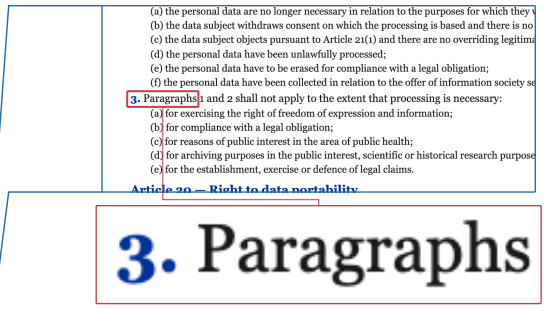
Clean

Qwen3.5-9B

Original input



Context and detail



Reference

3. Paragraphs 1 and 2 shall not apply to the extent that processing is necessary:

Model output

2. Paragraphs 1 and 2 shall not apply to the extent that processing is necessary:

Paragraph 3 under Article 17 is relabeled as paragraph 2. The wording of its introductory sentence is retained.

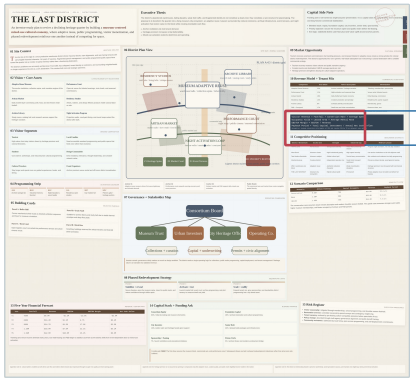
Percentage substitution

Museum revenue assumptions

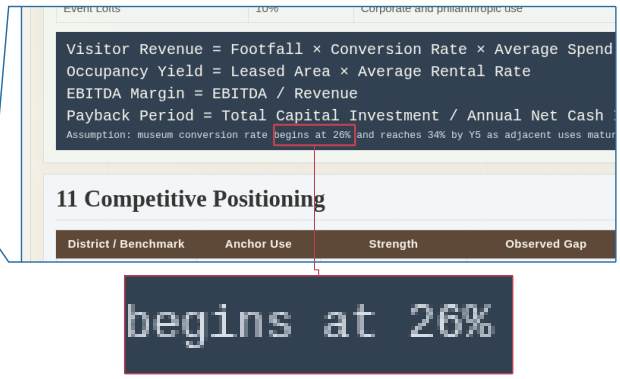
Clean

Gemini-3.1-Pro

Original input



Context and detail



Reference

conversion rate begins at **26%** and reaches 34%

Model output

conversion rate begins at **20%** and reaches 34%

The starting museum conversion rate changes from 26% to 20%. The later target of 34% is preserved.

Figure 8: **Clean-track numbering and value errors.** A legal paragraph number and a business-plan percentage are altered, corrupting document references and a stated revenue assumption.

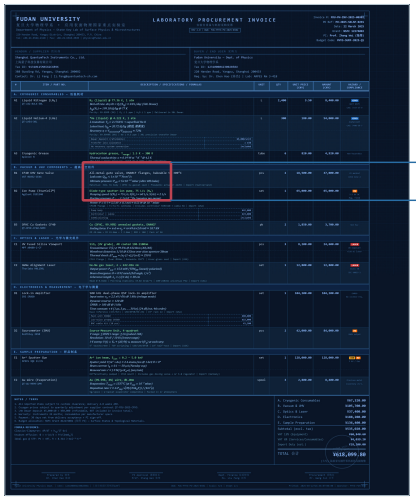
Exponent substitution

Vacuum-valve specifications

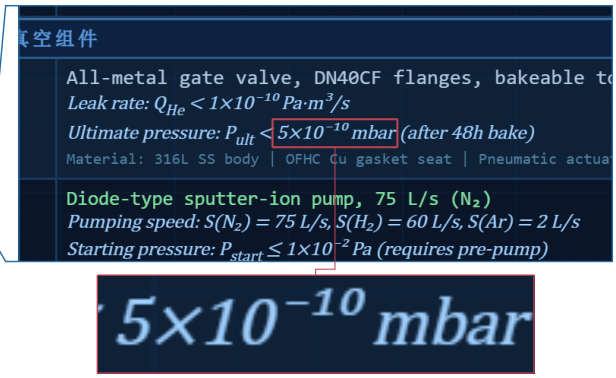
Clean

HunyuanOCR

Original input



Context and detail



Reference

5×10^{-10} mbar

Model output

5×10^{-4} mbar

The ultimate-pressure exponent changes from −10 to −4. This changes the specified pressure by six orders of magnitude.

Extra-digit insertion

Campus incident log

Clean

OCRVerse

Original input



Context and detail



Reference

Authentication timeout after **30000ms**

Model output

Authentication timeout after **300000ms**

An extra zero changes the authentication timeout from 30,000 ms to 300,000 ms. The event identity and surrounding log text are retained.

Figure 9: **Clean-track numerical errors.** A pressure value shifts by six orders of magnitude, and an extra digit increases a timeout tenfold, altering scientific specifications and software settings.

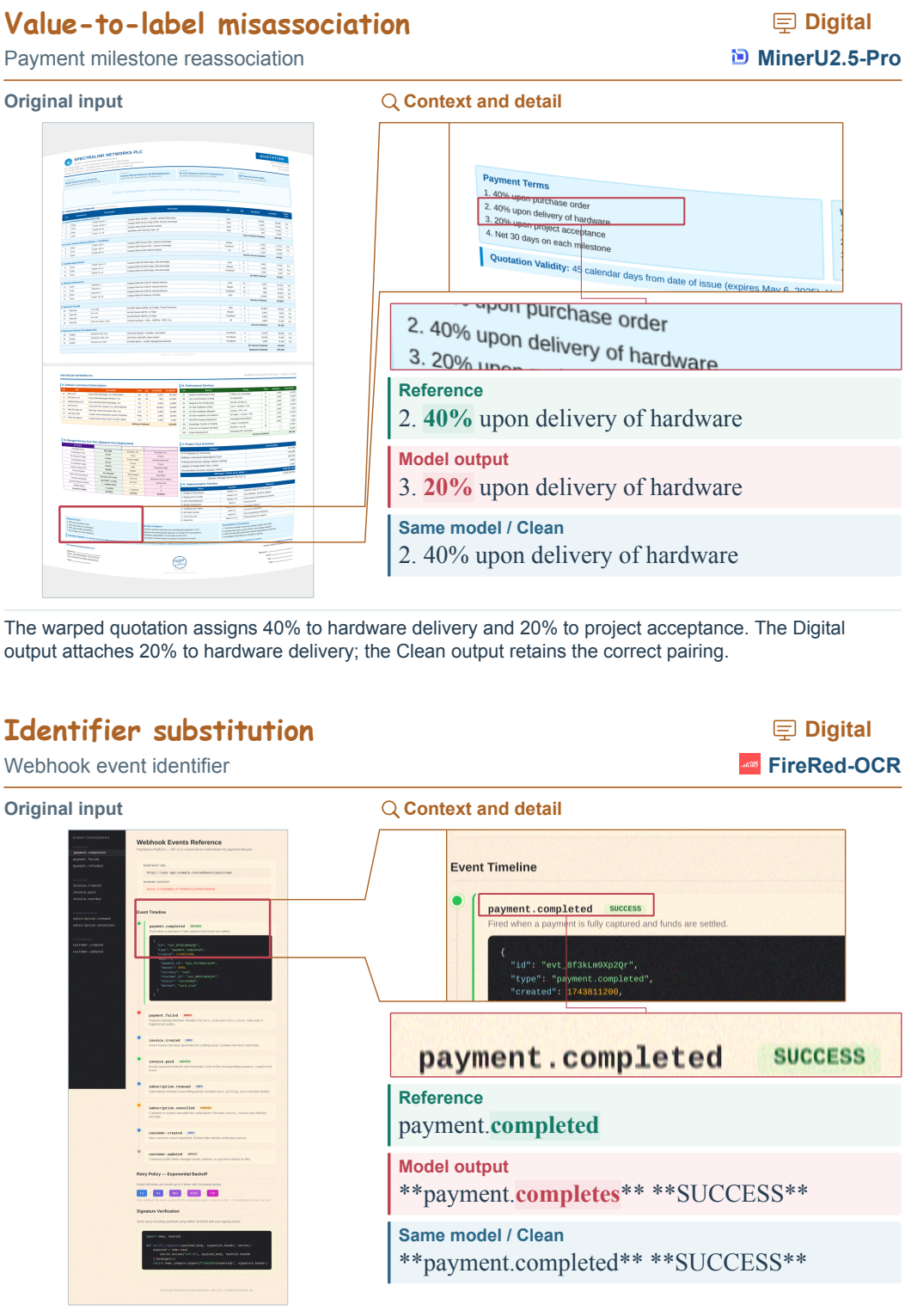


Figure 10: **Digital-track association and identifier errors.** Payment percentages are misassigned and an API event name changes; both fields are recovered on the corresponding Clean pages.

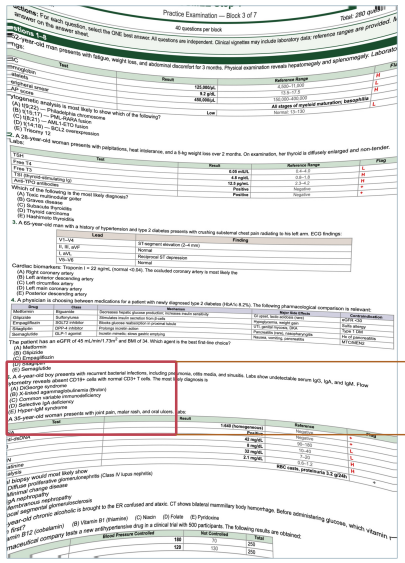
Answer-label shift

Multiple-choice answer label

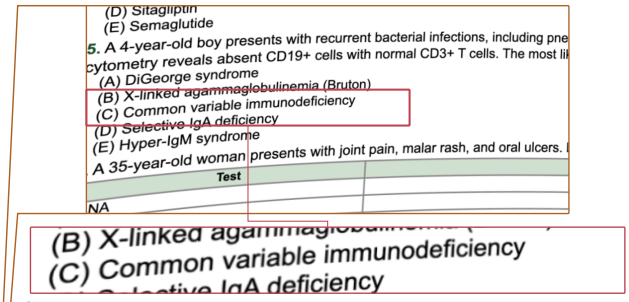
Digital

MinerU2.5-Pro

Original input



Context and detail



Reference

(C) Common variable immunodeficiency

Model output

B) Common variable immunodeficiency

Same model / Clean

(C) Common variable immunodeficiency

The source assigns Common variable immunodeficiency to choice C. MinerU2.5-Pro labels the same option B in the Digital output; its paired Clean output retains C.

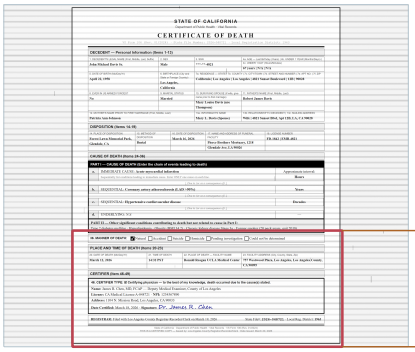
Localized text omission

Registrar filing statement

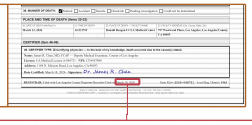
Digital

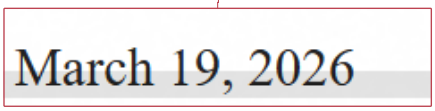
FireRed-OCR

Original input



Context and detail





Reference

REGISTRAR: Filed with Los Angeles County Registrar-Recorder/Clerk on **March 19, 2026**

Model output / retained excerpt

REGISTRAR:
State File#: 2026-048721 · Local Reg. District: 1963

Same model / Clean

REGISTRAR: Filed with Los Angeles County Registrar-Recorder/Clerk on March 19, 2026

FireRed-OCR retains the REGISTRAR label and adjacent State File number but omits the filing statement and its March 19 date. The paired Clean prediction retains the full statement.

Figure 11: **Digital-track label and omission errors.** An answer choice is relabeled and a registrar filing statement is omitted; paired Clean outputs preserve both pieces of information.

Year substitution

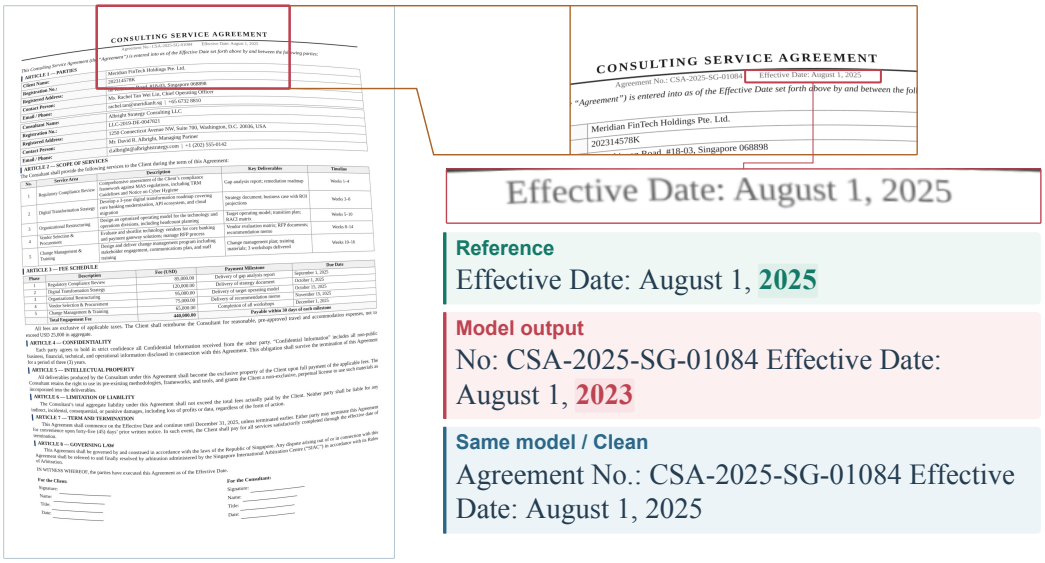
Agreement effective date

Digital

MinerU2.5-Pro

Original input

Context and detail



Reference

Effective Date: August 1, **2025**

Model output

No: CSA-2025-SG-01084 Effective Date: August 1, **2023**

Same model / Clean

Agreement No.: CSA-2025-SG-01084 Effective Date: August 1, 2025

The curved agreement header clearly states August 1, 2025. MinerU2.5-Pro changes the effective year to 2023 in the Digital output; its Clean output retains 2025.

Premature output termination

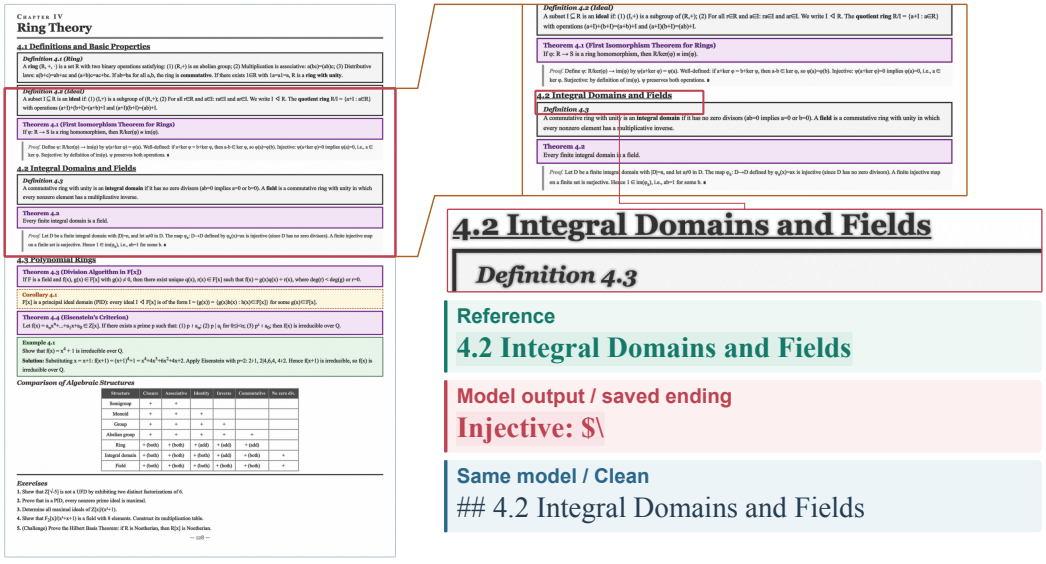
Output stops before the next section

Digital

Gemini-3.1-Pro

Original input

Context and detail



Reference

4.2 Integral Domains and Fields

Model output / saved ending

Injective: $

Same model / Clean

4.2 Integral Domains and Fields

The saved Digital prediction ends inside the preceding proof, immediately after "Injective: $\". The visible next section is absent, although the paired Clean prediction includes it.

Figure 12: **Digital-track date and completeness errors.** A contract's effective year changes and a prediction ends before the next section; the paired Clean outputs preserve both details.

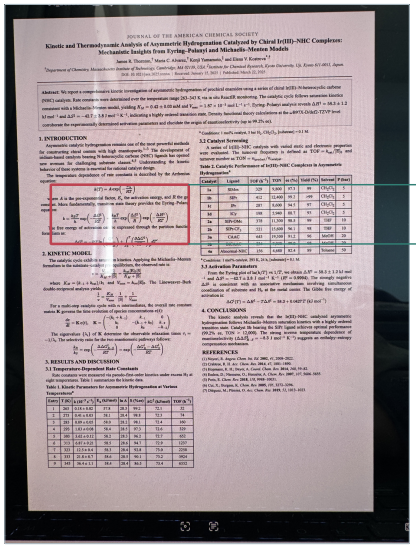
Formula sign substitution

Sign error in an exponential factor

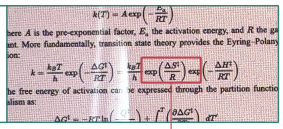
Real

Nanonets-OCR2

Original input



Context and detail





Reference

$$\exp\left(\frac{\Delta S^{\ddagger}}{R}\right)$$

Model output

$$\exp\left(-\frac{\Delta S^{*}}{R}\right)$$

Same model / Clean

$$\exp(\frac{\Delta S^{\ddagger}}{R})$$

The Real output adds a minus sign before the entropy term in the exponential. Its Clean output preserves the positive term.

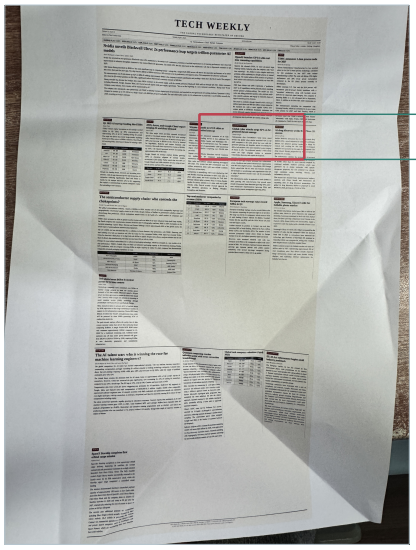
Numeric substitution

Percentage changed in a headline

Real

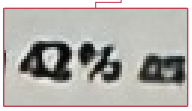
Kimi K2.6

Original input



Context and detail





Reference

Global cyber attacks surge **42%** as AI-powered threats emerge

Model output

Global cyber attacks surge **43%** as AI-powered threats emerge

Same model / Clean

Global cyber attacks surge 42% as AI-powered threats emerge

The printed headline reports a 42% increase, while the Real output changes the value to 43%. The paired Clean output preserves 42%.

Figure 13: **Real-track mathematical and numerical errors.** A formula gains a minus sign and a headline percentage changes, altering an exponential factor and a quantitative news claim.

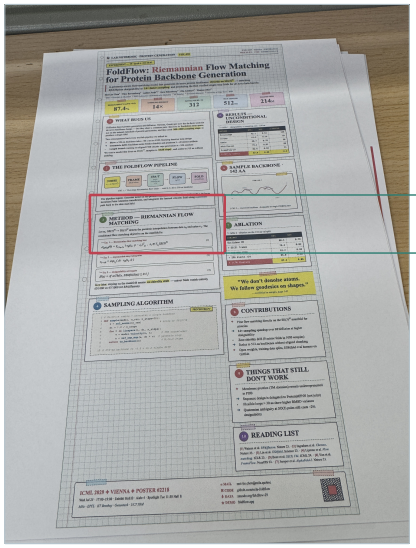
Heading truncation

A second heading line disappears

Real

GLM-OCR

Original input



Context and detail

Reference

3 METHOD — RIEMANNIAN FLOW **MATCHING**

Model output / retained excerpt

METHOD — RIEMANNIAN FLOW

Same model / Clean

3 METHOD — RIEMANNIAN FLOW MATCHING

The source heading continues with MATCHING on its second line. The Real output drops that word from the heading, while the paired Clean output retains it.

Numeric range substitution

Subscription-price range

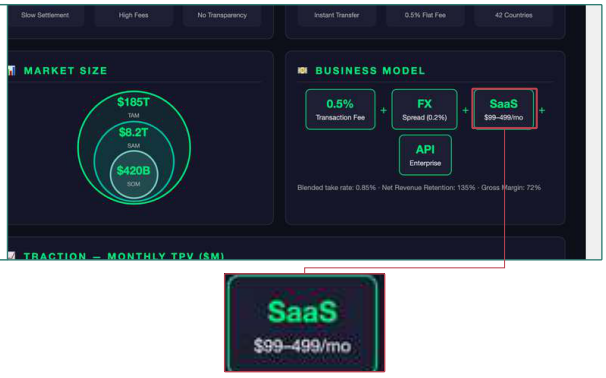
Real

Qwen3.5-397B-A17B

Original input



Context and detail



Reference

SaaS **$99–499/mo**

Model output

SaaS **$39-$99/mo**

The compressed screenshot gives a SaaS subscription range of $99–499 per month. The Real prediction changes both endpoints to $39–$99; the paired Clean prediction retains the source range.

Figure 14: **Real-track heading and range errors.** A poster heading loses part of a method name, while altered endpoints of a subscription-price range misstate the service's pricing.

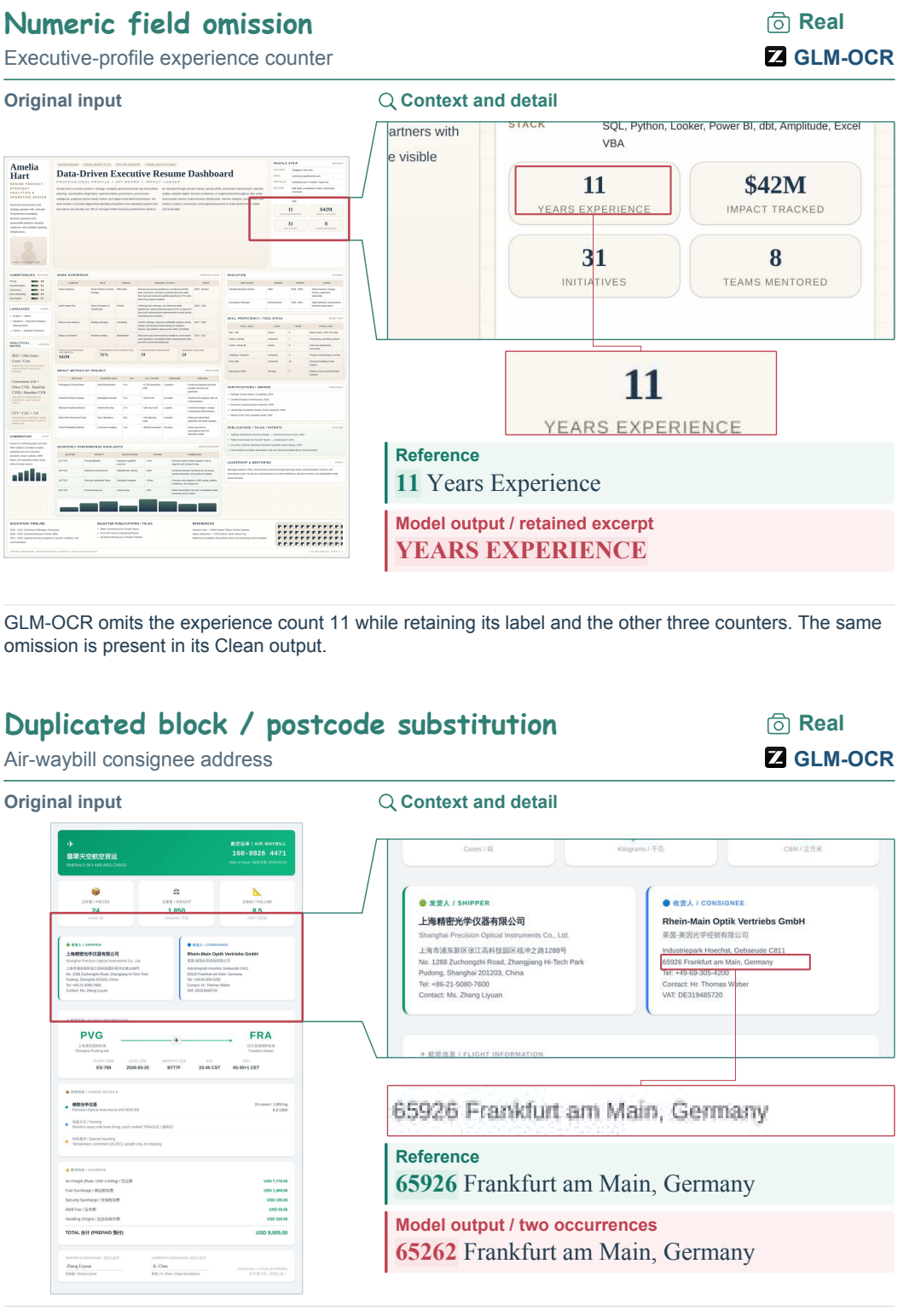


Figure 15: **Real-track omission and duplication errors.** An experience count is omitted, and an address repeats with an incorrect postcode, corrupting document completeness and structure.

C Real-Track Readability Audit

Screening and human review. We analyzed the geometry of all 1,475 base pages. A hypothetical resize to a longest side of 1,024 pixels flagged 492 candidates whose effective width would fall below 512 pixels; 104 fell below 256 pixels. These thresholds prioritize long pages for inspection. Two annotators independently reviewed each candidate's Clean, Digital, and Real images at original resolution with zoom and reconciled their decisions. The 1,024-pixel setting is a geometric screening diagnostic; the final judgments use original-resolution content. The audit identified 83 Real images (5.63%) whose text could not be reliably verified against its transcription. Their Clean and Digital counterparts remained readable. Figure 16 illustrates how fitting a long page into a recaptured image can remove fine textual detail. The released page-level decision list records image identifiers, geometry, track-specific readability, and the final decision.

Sensitivity protocol. Table 3 compares full and filtered evaluation for 51 models. The filtered Real set excludes the 83 flagged images, retaining 1,392 pages. Clean and Digital retain all 1,475 pages and their original scores; each model's Avg_3 is recomputed with its filtered Real score. This analysis measures the effect of the Real exclusion on aggregate scores and model rankings. Both rankings are computed within the same 51-model cohort. The full-track results in Table 2 retain their original evaluation scope.

Effect on scores and rankings. Exclusion increases every model's Real score, with a mean gain of 2.32 points and an Avg_3 gain of 0.77 points. Spearman correlations are ρ=0.9983 for Real and ρ=0.9991 for Avg_3; the maximum absolute rank change is two positions in both cases. Both top-ten sets remain unchanged, as do the Real leader (Gemini 3.6 Flash) and Avg_3 leader (GLM-5.3-Flash). Thus, acquisition-related legibility loss depresses absolute scores, while relative rankings remain stable within this cohort. This motivates improved capture and readability screening alongside evaluation on difficult but legible inputs.

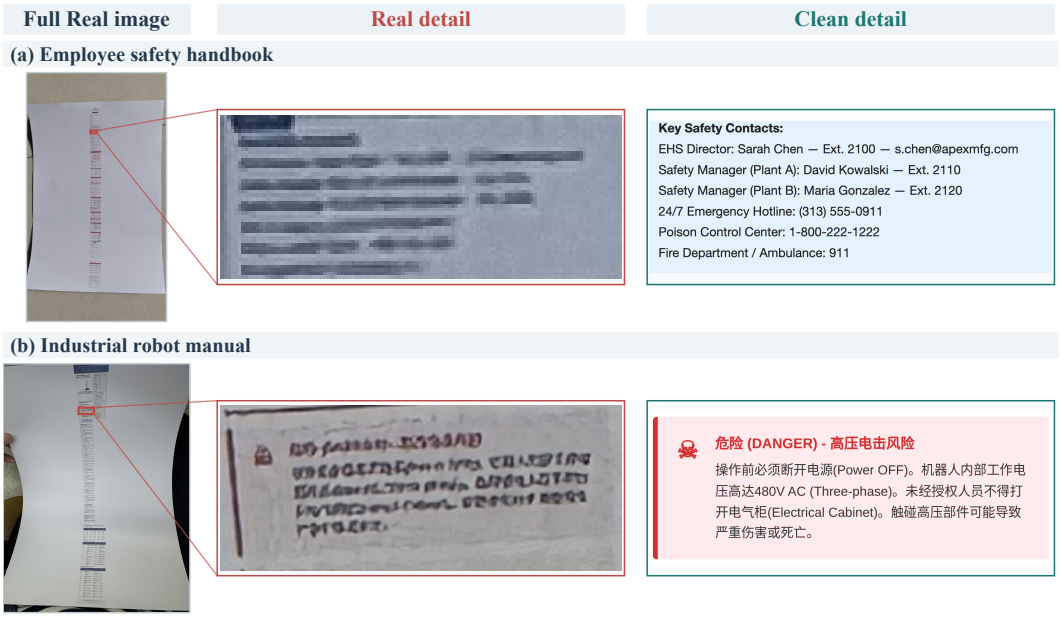


Figure 16: **Readability limits in Real recaptures of long pages.** Fine text remains readable in Clean but loses essential detail during recapture, obscuring emergency contacts and a high-voltage warning. Both Real images belong to the 83 pages identified as unreadable in human review.

Table 3: **Readability sensitivity for 51 models.** Removing 83 unreadable Real images preserves top-ten membership. Full and Filtered use 1,475 and 1,392 Real pages; Clean and Digital stay fixed. Avg_3 averages the three tracks, and Δ is Filtered minus Full. **Bold**/underline: best/runner-up.

Model	Real↑			Avg_3↑		
	Full	Filtered	Δ	Full	Filtered	Δ
Pipeline / multi-stage specialists						
PaddleOCR-VL-1.6	54.24	56.08	+1.83	59.33	59.94	+0.61
YouTu-Parsing	60.29	63.26	+2.97	68.32	69.31	+0.99
DotsMOCR	61.73	64.57	+2.84	70.39	71.33	+0.95
GLM-OCR	58.31	60.82	+2.52	63.34	64.18	+0.84
MinerU2.5-Pro	62.56	65.36	+2.80	70.07	71.00	+0.93
MinerU2.5	59.15	61.94	+2.80	67.66	68.59	+0.93
MonkeyOCRv2-B-Parsing	58.83	60.80	+1.97	64.91	65.57	+0.66
MonkeyOCR-pro-3B	46.49	48.76	+2.27	55.37	56.13	+0.76
MonkeyOCR-pro-1.2B	43.82	46.11	+2.29	53.54	54.31	+0.76
OpenOCR	25.73	27.02	+1.29	29.49	29.92	+0.43
Dolphin-v2	44.92	47.23	+2.31	57.02	57.79	+0.77
TeleOCR	69.08	71.60	+2.52	78.55	79.39	+0.84
End-to-end specialists						
OvisOCR2	66.49	69.22	+2.73	75.10	76.01	+0.91
ABot-OCR	61.77	64.05	+2.29	70.37	71.13	+0.76
Qianfan-OCR	45.06	47.48	+2.42	51.04	51.85	+0.81
Unlimited-OCR	53.71	56.06	+2.34	63.69	64.47	+0.78
WeVisDoc-4B	69.14	71.41	+2.27	75.57	76.33	+0.76
WeVisDoc-2B	65.68	68.11	+2.43	73.69	74.50	+0.81
HunyuanOCR-1.5	60.28	62.72	+2.44	67.58	68.39	+0.81
dots.ocr	55.68	58.69	+3.01	64.55	65.55	+1.00
FireRed-OCR	57.42	60.16	+2.75	65.57	66.49	+0.92
DeepSeek-OCR	40.48	42.76	+2.28	46.98	47.74	+0.76
olmOCR-2-7B	56.10	58.90	+2.80	63.78	64.71	+0.93
olmOCR-7B	47.30	49.64	+2.34	55.90	56.68	+0.78
OCRVerse	63.66	66.25	+2.59	69.40	70.26	+0.86
FD-RL	67.04	69.71	+2.67	73.92	74.81	+0.89
Nanonets-OCR2	49.03	51.70	+2.67	58.36	59.25	+0.89
OCRFlux-3B	37.21	39.05	+1.83	42.06	42.67	+0.61
Logics-Parsing-v2	67.64	70.21	+2.56	72.61	73.47	+0.85
General-purpose VLMs						
Qwen3.8-Flash-Next	62.42	64.57	+2.15	68.54	69.25	+0.72
Qwen3.6-35B-A3B	60.12	62.16	+2.04	67.14	67.82	+0.68
Qwen3.6-27B	59.18	61.18	+2.00	65.62	66.29	+0.67
Qwen3.5-397B-A17B	62.70	64.60	+1.90	66.72	67.35	+0.63
Qwen3.5-122B-A10B	69.85	72.21	+2.36	74.11	74.90	+0.79
Qwen3.5-35B-A3B	60.59	62.59	+2.01	65.68	66.34	+0.67
Qwen3.5-27B	65.92	68.13	+2.21	69.57	70.31	+0.74
Qwen3.5-9B	65.45	67.76	+2.32	70.89	71.66	+0.77
Qwen3.5-4B	63.47	66.02	+2.56	69.82	70.67	+0.85
Qwen3.5-2B	55.92	58.39	+2.47	62.46	63.28	+0.82
Qwen3.5-0.8B	47.93	50.56	+2.62	55.99	56.87	+0.87
Qwen3-VL-8B	62.73	65.31	+2.58	69.07	69.93	+0.86
Qwen3-VL-4B	59.61	62.30	+2.69	67.50	68.39	+0.90
Qwen3-VL-2B	54.09	56.60	+2.50	62.09	62.93	+0.83
DeepSeek-V4-Flash-Vision-Exp	42.68	44.09	+1.42	50.12	50.59	+0.47
GLM-5.3-Flash	74.99	77.02	+2.03	**79.69**	**80.36**	+0.68
MiniCPM-V-4.6	48.01	49.85	+1.84	57.17	57.79	+0.61
MiniCPM-V-4.5	37.59	39.54	+1.95	46.26	46.91	+0.65
Gemini 3.6 Flash	**75.73**	**77.30**	+1.57	78.27	78.79	+0.52
Gemini-3.1-Pro	71.98	74.03	+2.05	70.43	71.12	+0.68
Kimi K2.6	68.02	70.56	+2.54	70.10	70.94	+0.85
Step3-VL	45.06	46.96	+1.90	50.48	51.12	+0.63

D FULL COMPONENT RESULTS

Table 4: **Clean component results for 58 models.** Text and Reading Edit are normalized edit distances (↓); Formula CDM, Table TEDS, and Overall use a shared 0–100 scoring scale (↑). **Bold**/underline: best/runner-up.

Model	Overall↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓
Pipeline / multi-stage specialists					
PaddleOCR-VL-1.6	64.46	0.432	53.91	82.70	0.556
YouTu-Parsing	75.02	0.230	67.34	80.74	0.358
DotsMOCR	76.27	0.151	66.23	77.65	0.273
GLM-OCR	68.65	0.314	57.89	79.44	0.470
MinerU2.5-Pro	75.87	0.222	65.14	84.68	0.346
MinerU2.5	74.90	0.184	62.08	81.04	0.327
MonkeyOCRv2-B-Parsing	69.82	0.328	62.27	80.05	0.409
MonkeyOCR-pro-3B	62.23	0.346	48.46	72.83	0.492
MonkeyOCR-pro-1.2B	61.09	0.358	47.43	71.60	0.498
OpenDoc-0.1B (UniRec)	60.28	0.411	53.09	68.86	0.519
OpenOCR	32.70	0.354	33.50	0.00	0.507
Dolphin-v2	65.90	0.342	59.80	72.12	0.429
TeleOCR	**87.16**	**0.120**	**82.33**	91.12	**0.157**
End-to-end specialists					
OvisOCR2	81.53	0.152	69.66	90.10	0.282
ABot-OCR	76.92	0.224	66.81	86.39	0.397
Qianfan-OCR	57.22	0.370	49.79	58.83	0.443
Unlimited-OCR	72.05	0.243	63.68	76.81	0.367
WeVisDoc-4B	79.59	0.204	69.95	89.23	0.332
WeVisDoc-2B	79.08	0.217	71.25	87.64	0.347
HunyuanOCR-1.5	73.09	0.242	67.54	75.94	0.391
HunyuanOCR	65.61	0.269	55.74	68.02	0.382
dots.ocr	72.01	0.248	61.37	79.51	0.379
FireRed-OCR	70.81	0.287	63.86	77.23	0.396
DeepSeek-OCR-2	55.53	0.354	46.00	56.01	0.466
DeepSeek-OCR	53.50	0.419	45.39	57.06	0.514
olmOCR-2-7B	69.36	0.284	56.89	79.59	0.358
olmOCR-7B	62.56	0.388	58.69	67.77	0.466
OCRVerse	73.18	0.273	63.78	83.09	0.393
FD-RL	78.38	0.193	68.21	86.22	0.334
Nanonets-OCR2	64.83	0.254	44.98	74.94	0.377
OCRFlux-3B	47.14	0.454	38.35	48.46	0.424
Logics-Parsing-v2	76.35	0.213	67.67	82.67	0.342
General-purpose VLMs					
Qwen3.8-27B	80.45	0.180	71.68	87.69	0.305
Qwen3.8-Flash-Next	72.23	0.325	72.28	76.93	0.420
Qwen3.6-35B-A3B	72.14	0.302	70.39	76.25	0.397
Qwen3.6-27B	70.20	0.352	69.36	76.47	0.446
Qwen3.5-397B-A17B	69.12	0.233	65.26	65.40	0.366
Qwen3.5-122B-A10B	76.14	0.226	67.96	83.03	0.375
Qwen3.5-35B-A3B	68.40	0.232	64.94	63.45	0.374
Qwen3.5-27B	72.07	0.227	66.36	72.51	0.362
Qwen3.5-9B	73.87	0.254	67.60	79.39	0.388
Qwen3.5-4B	73.45	0.276	69.96	78.02	0.410
Qwen3.5-2B	66.24	0.348	62.84	70.70	0.473
Qwen3.5-0.8B	60.77	0.376	54.39	65.54	0.500
Qwen3-VL-8B	72.44	0.261	65.10	78.35	0.411
Qwen3-VL-4B	72.04	0.262	65.10	77.17	0.418
Qwen3-VL-2B	66.37	0.300	59.04	70.03	0.439
DeepSeek-V4-Flash-Vision-Exp	54.69	0.540	57.47	60.55	0.567
GLM-5.3-Flash	83.04	0.218	80.16	90.76	0.375
MiniCPM-V-4.6	64.47	0.343	58.05	69.72	0.437
MiniCPM-V-4.5	51.81	0.439	45.97	53.36	0.481
Gemini 3.6 Flash	79.97	0.251	78.29	86.76	0.381
Gemini-3.1-Pro	70.04	0.306	65.63	75.08	0.409
Claude Opus 5	82.99	0.191	75.92	**92.17**	0.336
GPT-5.6 Sol	74.15	0.354	74.96	82.87	0.440
MiniMax-M3	55.78	0.506	54.41	63.53	0.529
Kimi K2.6	72.32	0.303	66.93	80.30	0.466
Step3-VL	53.65	0.496	53.41	57.16	0.509

Table 5: **Digital component results for 58 models.** Text and Reading Edit are normalized edit distances (↓); Formula CDM, Table TEDS, and Overall use a shared 0–100 scoring scale (↑). **Bold**/underline: best/runner-up.

Model	Overall↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓
Pipeline / multi-stage specialists					
PaddleOCR-VL-1.6	59.29	0.494	51.78	75.46	0.596
YouTu-Parsing	69.66	0.270	61.44	74.49	0.388
DotsMOCR	73.16	0.198	64.32	74.95	0.309
GLM-OCR	63.06	0.383	53.23	74.21	0.520
MinerU2.5-Pro	71.77	0.272	61.79	80.73	0.378
MinerU2.5	68.92	0.245	56.99	74.24	0.374
MonkeyOCRv2-B-Parsing	66.08	0.356	59.06	74.74	0.426
MonkeyOCR-pro-3B	57.40	0.397	45.57	66.32	0.526
MonkeyOCR-pro-1.2B	55.72	0.416	43.91	64.83	0.529
OpenDoc-0.1B (UniRec)	52.46	0.501	48.41	59.04	0.577
OpenOCR	30.03	0.410	31.09	0.00	0.541
Dolphin-v2	60.24	0.393	52.20	67.86	0.461
TeleOCR	79.40	0.209	76.12	82.94	**0.237**
End-to-end specialists					
OvisOCR2	77.29	**0.196**	66.49	84.94	0.316
ABot-OCR	72.43	0.269	63.16	80.99	0.435
Qianfan-OCR	50.85	0.438	44.41	51.96	0.485
Unlimited-OCR	65.30	0.310	57.63	69.32	0.412
WeVisDoc-4B	77.99	0.234	70.75	86.65	0.356
WeVisDoc-2B	76.33	0.255	70.34	84.09	0.374
HunyuanOCR-1.5	69.36	0.275	64.63	70.99	0.414
HunyuanOCR	61.49	0.308	51.62	63.68	0.400
dots.ocr	65.95	0.307	56.67	71.86	0.417
FireRed-OCR	68.49	0.319	62.64	74.77	0.422
DeepSeek-OCR-2	49.41	0.412	40.78	48.67	0.493
DeepSeek-OCR	46.95	0.478	39.99	48.64	0.548
olmOCR-2-7B	65.87	0.318	54.57	74.81	0.378
olmOCR-7B	57.84	0.436	55.44	61.66	0.499
OCRVerse	71.36	0.302	63.95	80.36	0.415
FD-RL	76.33	0.214	67.16	83.22	0.350
Nanonets-OCR2	61.23	0.307	45.40	68.97	0.408
OCRFlux-3B	41.82	0.486	31.90	42.17	0.437
Logics-Parsing-v2	73.85	0.248	67.33	79.02	0.375
General-purpose VLMs					
Qwen3.8-27B	78.86	0.206	71.98	85.15	0.325
Qwen3.8-Flash-Next	70.96	0.346	72.87	74.59	0.430
Qwen3.6-35B-A3B	69.16	0.329	67.67	72.66	0.412
Qwen3.6-27B	67.49	0.375	66.54	73.42	0.457
Qwen3.5-397B-A17B	68.34	0.244	63.91	65.53	0.376
Qwen3.5-122B-A10B	76.34	0.220	67.82	83.21	0.366
Qwen3.5-35B-A3B	68.04	0.245	64.78	63.86	0.379
Qwen3.5-27B	70.73	0.236	64.61	71.17	0.367
Qwen3.5-9B	73.34	0.260	67.00	79.01	0.396
Qwen3.5-4B	72.53	0.281	68.88	76.78	0.412
Qwen3.5-2B	65.22	0.350	58.30	72.36	0.477
Qwen3.5-0.8B	59.28	0.386	54.22	62.22	0.510
Qwen3-VL-8B	72.03	0.266	64.88	77.82	0.409
Qwen3-VL-4B	70.84	0.272	63.54	76.13	0.425
Qwen3-VL-2B	65.81	0.314	60.25	68.52	0.448
DeepSeek-V4-Flash-Vision-Exp	52.98	0.561	57.24	57.77	0.576
GLM-5.3-Flash	**81.03**	0.242	78.51	**88.84**	0.391
MiniCPM-V-4.6	59.04	0.395	54.59	62.00	0.473
MiniCPM-V-4.5	49.38	0.461	42.79	51.50	0.489
Gemini 3.6 Flash	79.11	0.268	**78.88**	85.31	0.391
Gemini-3.1-Pro	69.28	0.322	65.81	74.24	0.417
Claude Opus 5	72.94	0.351	73.43	80.45	0.420
GPT-5.6 Sol	71.73	0.380	72.83	80.33	0.447
MiniMax-M3	53.85	0.532	54.15	60.55	0.548
Kimi K2.6	69.95	0.322	64.69	77.31	0.475
Step3-VL	52.74	0.516	53.62	56.15	0.529

Table 6: **Real component results for 58 models.** Text and Reading Edit are normalized edit distances (↓); Formula CDM, Table TEDS, and Overall use a shared 0–100 scoring scale (↑). **Bold**/underline: best/runner-up.

Model	Overall↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓
Pipeline / multi-stage specialists					
PaddleOCR-VL-1.6	54.24	0.529	43.90	71.76	0.619
YouTu-Parsing	60.29	0.360	52.20	64.69	0.430
DotsMOCR	61.73	0.312	54.39	61.97	0.393
GLM-OCR	58.31	0.433	50.34	67.83	0.543
MinerU2.5-Pro	62.56	0.375	52.70	72.47	0.446
MinerU2.5	59.15	0.370	49.01	65.41	0.446
MonkeyOCRv2-B-Parsing	58.83	0.442	53.48	67.22	0.458
MonkeyOCR-pro-3B	46.49	0.511	38.18	52.43	0.600
MonkeyOCR-pro-1.2B	43.82	0.556	36.94	50.07	0.609
OpenDoc-0.1B (UniRec)	44.27	0.547	38.46	49.06	0.603
OpenOCR	25.73	0.486	25.81	0.00	0.591
Dolphin-v2	44.92	0.553	39.98	50.04	0.558
TeleOCR	69.08	0.338	64.69	76.32	**0.352**
End-to-end specialists					
OvisOCR2	66.49	0.320	57.72	73.70	0.407
ABot-OCR	61.77	0.369	53.51	68.73	0.498
Qianfan-OCR	45.06	0.494	39.08	45.53	0.509
Unlimited-OCR	53.71	0.440	46.97	58.13	0.469
WeVisDoc-4B	69.14	0.333	62.74	77.92	0.411
WeVisDoc-2B	65.68	0.364	59.21	74.21	0.432
HunyuanOCR-1.5	60.28	0.376	56.62	61.81	0.476
HunyuanOCR	54.58	0.421	48.30	57.54	0.459
dots.ocr	55.68	0.403	47.70	59.63	0.467
FireRed-OCR	57.42	0.415	51.60	62.16	0.474
DeepSeek-OCR-2	43.60	0.486	37.30	42.06	0.533
DeepSeek-OCR	40.48	0.537	34.04	41.12	0.575
olmOCR-2-7B	56.10	0.417	48.79	61.25	0.439
olmOCR-7B	47.30	0.542	46.26	49.80	0.568
OCRVerse	63.66	0.363	57.03	70.30	0.452
FD-RL	67.04	0.298	58.82	72.08	0.391
Nanonets-OCR2	49.03	0.435	35.50	55.09	0.468
OCRFlux-3B	37.21	0.559	32.65	34.87	0.491
Logics-Parsing-v2	67.64	0.304	61.65	71.64	0.416
General-purpose VLMs					
Qwen3.8-27B	73.59	**0.274**	68.56	79.63	0.364
Qwen3.8-Flash-Next	62.42	0.424	62.84	66.85	0.472
Qwen3.6-35B-A3B	60.12	0.426	59.66	63.27	0.469
Qwen3.6-27B	59.18	0.453	56.51	66.34	0.504
Qwen3.5-397B-A17B	62.70	0.287	60.70	56.12	0.399
Qwen3.5-122B-A10B	69.85	0.281	62.19	75.44	0.401
Qwen3.5-35B-A3B	60.59	0.310	59.68	53.07	0.419
Qwen3.5-27B	65.92	0.283	61.23	64.82	0.390
Qwen3.5-9B	65.45	0.332	60.91	68.59	0.437
Qwen3.5-4B	63.47	0.380	61.27	67.17	0.477
Qwen3.5-2B	55.92	0.440	50.99	60.79	0.521
Qwen3.5-0.8B	47.93	0.498	44.60	48.98	0.557
Qwen3-VL-8B	62.73	0.342	55.55	66.81	0.448
Qwen3-VL-4B	59.61	0.378	55.15	61.47	0.480
Qwen3-VL-2B	54.09	0.428	51.05	53.99	0.511
DeepSeek-V4-Flash-Vision-Exp	42.68	0.642	46.21	46.00	0.626
GLM-5.3-Flash	74.99	0.311	73.59	82.45	0.423
MiniCPM-V-4.6	48.01	0.517	44.23	51.53	0.544
MiniCPM-V-4.5	37.59	0.583	32.01	39.06	0.552
Gemini 3.6 Flash	**75.73**	0.310	**76.27**	81.96	0.406
Gemini-3.1-Pro	71.98	0.300	68.62	77.26	0.386
Claude Opus 5	75.27	0.294	71.24	**84.01**	0.400
GPT-5.6 Sol	61.59	0.472	62.26	69.75	0.502
MiniMax-M3	42.54	0.628	43.10	47.29	0.605
Kimi K2.6	68.02	0.335	62.44	75.14	0.481
Step3-VL	45.06	0.579	45.42	47.66	0.573

E OmniDocBench Annotation Error Examples

Figure 17 illustrates annotation discrepancies in OmniDocBench v1.5 (Ouyang et al., 2025). Source-image regions are shown alongside annotation excerpts, covering extra words, hallucinated text, missing words, table structure errors, category mismatches, and OCR residual errors.

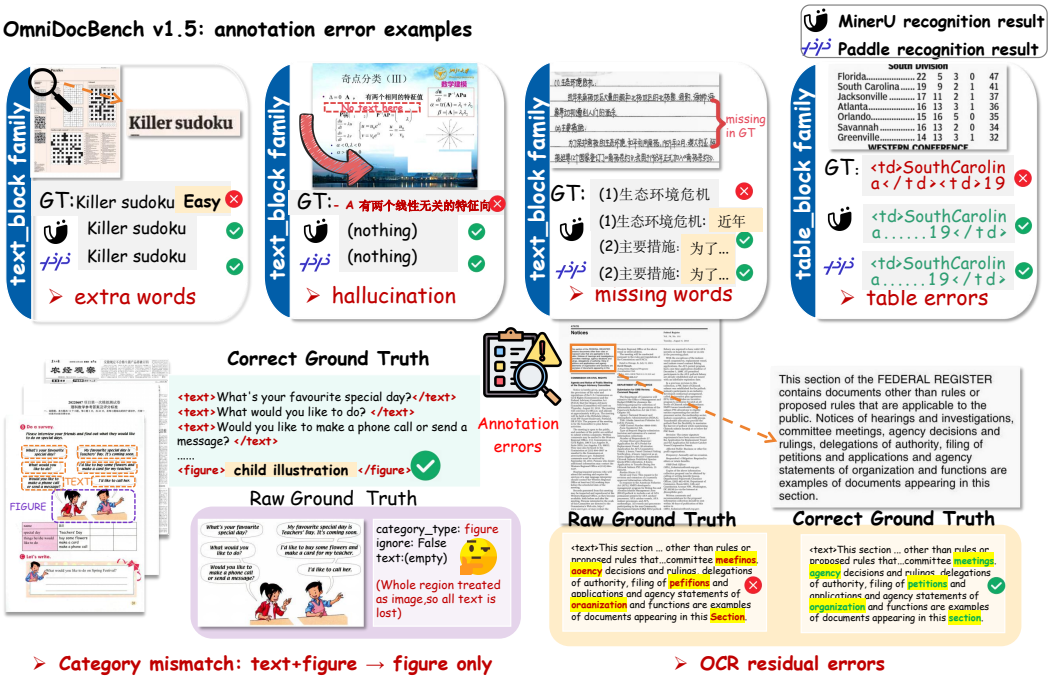


Figure 17: **Annotation errors in OmniDocBench v1.5.** The examples cover extra words, hallucinated text, missing words, table structure errors, category mismatches, and residual OCR errors. Such discrepancies can affect the evaluation of otherwise correct predictions and change benchmark scores.

F DEGRADATION PARAMETER DETAILS

Released degradation tracks. The released dataset keeps the three image tracks as parallel directory trees: clean pages under `images/`, digital degradation under `degraded_v2_merged/`, and physical recapture under `real_degraded/`. All three tracks are GT-preserving: the file identity is mapped back to the same clean-page annotation, so differences in score isolate image-condition robustness rather than content changes. The digital generation manifest records `source=images`, `output=degraded_v2`, `severity=moderate`, `severity_val=0.6`, and zero generation failures. The ten digital scene families are first materialised separately under `degraded_v2/<scene>/` with a per-image JSON parameter log under `degraded_v2/params/<scene>/`; the evaluation track `degraded_v2_merged/` then selects exactly one scene version per clean page and records the provenance in `degraded_v2_merged/scene_mapping.json`. This mapping is intentionally explicit: each evaluated digital image can be traced to its source clean page, scene name, generated PNG, and parameter JSON.

Digital scenes and operations. The digital track covers ten GT-preserving scene profiles spanning archival aging, bound-book scans, multi-generation copy chains, fax/thermal pages, low-quality printing, ink bleed, heavy compression, uneven lighting, geometric distortion, and noise–blur mixtures. These correspond to the release scene identifiers `aged_archive`, `book_binding`, `multi_gen_copy`, `fax_thermal`, `low_quality_print`, `ink_bleed`, `heavy_compression`, `uneven_lighting`, `geometric_distort`, and `noise_blur_combo`. Each scene composes low-level OpenCV / NumPy operations such as paper grain, yellowing, foxing spots, faded ink, page curl, binding shadow, bleed-through, photocopier contrast, dirty-drum streaks, thermal/fax binarization, printer banding, ink dilation, JPEG compression, down-up sampling, lighting gradients, vignetting, perspective warp, lens distortion, rotation, Gaussian noise, motion blur, and warm/cool color shift. To avoid confounding content with image quality, no operation changes the underlying HTML source or GT ordering. The scene release is balanced by construction: round-robin assignment yields 147–148 pages per digital scene in the released `degraded_v2` tree, and the merged track stores the selected scene for every clean page.

Physical recapture and ablations. The real-degraded track is built from physical or screen-mediated recaptures of the rendered pages rather than synthetic filters. Filenames encode the acquisition condition after the `__real_` marker, e.g. phone/screen/screenshot source, paper state such as flat, creased, bent, cover, or photocopy, lighting such as normal, low-light, flash, shadow, outdoor, or screen, and view geometry such as top-down, oblique, handheld, or screenshot. The mapping files in the real-degraded evaluation workspace map each recaptured file back to the original clean relative path, again allowing the same GT to score all tracks. In addition to the released digital and real tracks, we keep an ablation generator that decomposes the degradation process into **15 fundamental operations** over a four-stage causal pipeline (*Print* $\rightarrow$ *Paper* $\rightarrow$ *Capture* $\rightarrow$ *Digital*). Figure 18 lists these operations and their concrete parameter values at $s = 0.7$ (single-factor ablation) and $s = 1.0$ (peak stress). Figure 19 summarizes the higher-level capture scenarios and their dominant operation weights; these figures document the controllable design space used for sensitivity analysis, while `degraded_v2_merged/` is the canonical digital track used in the main benchmark.

15 fundamental degradation operations across 4 sequential pipeline stages

Operation	Mechanism	Default (s = 0.7)	Peak (s = 1.0)
Stage 1: Print Simulation			
ink_bleed	Dilate ink mask, blend toward dark gray	k = 2 px, blend up to 42%	k = 3 px, blend up to 60%
toner_variation	Low-frequency Gaussian density modulation	ink x (1 +/- 0.19)	ink x (1 +/- 0.24)
Stage 2: Paper Condition			
paper_aging	Warm yellow tint blended over the page	alpha = 0.24	alpha = 0.30
stains	Soft Gaussian-blurred ellipse stains	4 stains, alpha up to 0.27	6 stains, alpha up to 0.36
wrinkles	Faint blurred line shading	4 lines, amp +/- 18	6 lines, amp +/- 23
fold_lines	Strong folds with local shading dip	2 folds, dip 31	2 folds, dip 40
Stage 3: Capture Simulation			
skew	Affine rotation around the page center	+/- 3.1 deg	+/- 4.0 deg
perspective	Random 4-corner projective warp	shift up to 3.1% of min(h,w)	shift up to 4.0% of min(h,w)
uneven_lighting	Linear lighting gradient across page	+/- 24% (clipped to [.70,1.30])	+/- 35% (clipped to [.70,1.30])
defocus_blur	Gaussian blur (k odd, k <= 15)	k = 7 px	k = 11 px
motion_blur	Directional 1D averaging kernel	k = 4 px	k = 7 px
Stage 4: Digitization Artifacts			
resolution_loss	Down-then-upsample (effective DPI loss)	scale = 0.72x	scale = 0.60x
jpeg_compression	JPEG encode-decode round trip	quality = 42	quality = 30 (floor 15)
sensor_noise	Per-pixel Gaussian noise on RGB	sigma = 14	sigma = 20
color_shift	Random warm/cool channel bias	R / B +/- 7% / +/- 5%	R / B +/- 10% / +/- 8%

Figure 18: **Fifteen degradation operations across printing, paper, capture, and digital processing.** The operations model changes in document appearance, with parameter values at ablation strength $s = 0.7$ and peak stress $s = 1.0$ supporting controlled tests of individual factors.

10 capture scenarios (operations are mixed with per-scene weights)

Scenario	Description	Strength	Top-3 operations (weight in mixture)
Scanner-based (3)			
flatbed_scan	Clean flatbed scan	0.22 - 0.56	scanner_banding(.28), uneven_lighting(.18) resolution_loss(.18)
archival_scan	Aged / yellowed scan	0.30 - 0.74	paper_aging(.55), bleed_through(.40) stains(.36)
bound_manual_scan	Book scan, gutter shadow	0.28 - 0.76	binding_shadow(.58), bleed_through(.26) border_crop(.22)
Photocopy chain (2)			
office_copy_scan	Print-copy-scan chain	0.30 - 0.78	resolution_loss(.48), photocopy_mottle(.42) sharpen_halo(.42)
harsh_photocopy	Multi-gen copy / fax wear	0.34 - 0.84	photocopy_mottle(.58), resolution_loss(.34) toner_variation(.28)
Phone-camera capture (5)			
mobile_capture	Handheld phone capture	0.34 - 0.84	perspective(.46), shadow_glare(.42) page_warp(.34)
glossy_phone_capture	Glossy glare / color cast	0.32 - 0.82	shadow_glare(.62), color_shift(.34) perspective(.34)
low_light_phone	Low light, blur, noise	0.44 - 1.00	sensor_noise(.58), shadow_glare(.54) defocus_blur(.42)
folded_receipt_capture	Wrinkled / folded receipt	0.34 - 0.90	fold_lines(.58), page_warp(.46) wrinkles(.42)
thermal_receipt_capture	Faded thermal receipt	0.30 - 0.88	thermal_fade(.62), fold_lines(.26) perspective(.20)

Figure 19: **Ten capture scenarios combining multiple degradation factors.** Scenario-specific strength ranges and weighted operations represent different document acquisition conditions. The three dominant operations identify the main sources of degradation in each scenario.

G Full Per-Category Results

PureDocBench pages are organized into ten top-level domains: *academic*, *education*, *legal_gov*, *business*, *finance*, *medical*, *publishing*, *technical*, *logistics*, and *certificate*, with 180/138/151/160/212/200/100/110/120/104 pages respectively (sum: 1,475). For each (model, category, track) triple we recompute Overall as $((1-\text{TextEdit}) \times 100 + \text{FormulaCDM} + \text{TableTEDS})/3$ on the subset of samples in that category, and treat the entry as defined only when all three components exist (i.e. the category contains at least one text page, one formula sample, and one table sample). Per-category Avg_3 is then the mean of the three track Overalls.

Tables 7–16 report results for 53 models across all ten domains. Each table includes Overall↑, TextEdit↓, FormulaCDM↑, TableTEDS↑, and ROEdit↓ for Clean, Digital, and Real, followed by the three-track mean Avg_3. Models follow the architecture groups and within-series order of Table 2, with linked names, icons, and release months. Full-benchmark component scores appear in Appendix D. The category-level scores are provided in the supplementary file `category_benchmark_all_tracks.csv`.

Cross-category observations. Domain difficulty varies substantially across the 53 models. Mean Avg_3 ranges from 42.0 on *certificate* to 82.9 on *academic*; individual scores span 9.4–68.9 and 36.4–92.4, respectively. Model dispersion is largest on *certificate* (σ=14.4), followed by *legal_gov* (13.5), *business* (12.6), and *medical* (12.4). These differences motivate evaluating document parsers across domains with varied layouts and content.

Domain leaders and architecture mix. GLM-5.3-Flash leads five domains by Avg_3: *academic*, *education*, *legal_gov*, *business*, and *medical*. Gemini 3.6 Flash leads *publishing*; TeleOCR leads *finance* and *technical*; DotsMOCR leads *logistics*; and FD-RL leads *certificate*. Across the 50 top-five positions (10 domains $\times$ 5 ranks), general VLMs occupy 58%, E2E specialists 30%, and Pipeline specialists 12%. The domain winners comprise six VLM, three Pipeline, and one E2E result, showing that leadership depends on the document domain.

2030 2029 2028 2027 2026 2025 2024 2023 2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 2005 2004 2003 2002 2001 2000 1999 1998

Table 7: **Academic: three-track component results** (180 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	86.11	0.170	81.59	93.69	0.256	83.90	0.210	80.41	92.25	0.292	76.49	0.252	66.55	88.13	0.338	82.16
YouTu-Parsing (2026-01)	90.58	0.137	91.44	93.99	0.206	89.48	0.150	90.98	92.46	0.218	81.44	0.210	81.71	83.60	0.248	87.17
DotsMOCR (2026-03)	94.57	0.079	95.19	96.39	0.153	92.11	0.084	91.19	93.58	0.153	83.92	0.155	82.05	85.22	0.196	90.20
GLM-OCR (2026-02)	91.87	0.114	91.79	95.25	0.202	88.86	0.157	87.64	94.62	0.240	81.93	0.197	76.90	88.62	0.261	87.55
MinerU2.5-Pro (2026-04)	93.36	0.111	93.51	**97.68**	0.155	89.66	0.145	89.95	93.51	0.185	83.81	0.191	82.41	88.09	0.217	88.94
MinerU2.5 (2025-09)	93.61	**0.075**	91.20	97.09	0.139	89.67	0.114	86.76	93.68	0.174	82.22	0.177	80.62	83.78	0.216	88.50
MonkeyOCRv2-B-Parsing (2026-07)	90.08	0.120	86.69	95.54	0.182	87.49	0.127	82.99	92.16	0.191	79.96	0.206	75.29	85.15	0.231	85.84
MonkeyOCR-pro-3B (2025-08)	85.26	0.127	77.85	90.67	0.212	82.38	0.146	74.43	87.28	0.227	69.81	0.252	59.41	75.20	0.319	79.15
MonkeyOCR-pro-1.2B (2025-07)	85.79	0.127	79.65	90.43	0.212	80.06	0.174	72.87	84.73	0.246	67.02	0.306	61.64	70.01	0.357	77.62
OpenOCR (2024-11)	39.53	0.233	41.85	0.00	0.379	37.00	0.285	39.51	0.00	0.401	32.74	0.337	31.88	0.00	0.444	36.42
Dolphin-v2 (2025-12)	91.41	0.115	88.88	96.87	0.155	86.06	0.150	80.55	92.65	0.182	66.88	0.292	57.28	72.52	0.296	81.45
TeleOCR (2026-08)	**94.59**	0.077	94.59	96.83	**0.101**	91.24	0.104	90.57	93.58	**0.119**	84.96	0.182	82.29	90.77	0.194	90.26
End-to-end specialists																
OvisOCR2 (2026-07)	93.89	0.078	93.46	95.98	0.137	92.82	**0.083**	91.88	94.91	0.143	85.58	0.152	84.76	87.18	0.206	90.76
ABot-OCR (2026-05)	92.52	0.109	92.42	96.00	0.212	91.38	0.116	90.43	95.31	0.229	82.41	0.187	80.93	84.98	0.288	88.77
Qianfan-OCR (2026-03)	84.86	0.176	85.76	86.44	0.216	77.31	0.237	77.19	78.40	0.270	75.64	0.261	76.72	76.31	0.281	79.27
Unlimited-OCR (2026-06)	91.30	0.086	91.17	91.29	0.148	87.04	0.116	84.87	87.88	0.182	74.90	0.218	70.11	76.37	0.246	84.41
WeVisDoc-4B (2026-09)	94.11	0.081	93.98	96.42	0.150	93.60	0.096	94.15	96.24	0.156	89.06	0.148	91.41	90.52	0.195	92.26
WeVisDoc-2B (2026-09)	94.23	0.093	95.04	96.94	0.155	93.25	0.103	93.67	**96.37**	0.165	85.24	0.169	85.55	87.08	0.222	90.91
HunyuanOCR-1.5 (2026-07)	89.64	0.124	90.33	90.94	0.199	89.48	0.120	90.56	89.87	0.192	82.54	0.175	81.99	83.12	0.238	87.22
HunyuanOCR (2025-11)	89.65	0.117	90.64	90.03	0.194	87.35	0.123	85.89	88.46	0.189	78.31	0.191	77.26	76.73	0.251	85.10
dots.ocr (2025-07)	91.96	0.095	93.22	92.19	0.158	87.03	0.118	84.88	88.05	0.181	79.47	0.187	77.50	79.60	0.220	86.15
FireRed-OCR (2026-02)	89.64	0.133	92.71	89.52	0.224	88.66	0.149	89.57	91.27	0.234	83.95	0.186	85.91	84.57	0.251	87.42
DeepSeek-OCR-2 (2026-01)	84.85	0.121	78.98	87.65	0.194	79.81	0.150	73.97	80.47	0.222	65.88	0.400	66.75	70.93	0.312	76.85
DeepSeek-OCR (2025-10)	60.39	0.292	59.78	50.55	0.384	56.78	0.292	55.90	43.63	0.380	40.22	0.414	38.75	23.26	0.501	52.46
olmOCR-2-7B (2025-10)	89.53	0.177	92.35	93.94	0.220	88.35	0.193	92.61	91.74	0.226	82.78	0.230	86.20	85.13	0.263	86.89
olmOCR-7B (2025-08)	83.44	0.272	94.18	83.28	0.281	81.79	0.274	92.03	80.71	0.279	74.19	0.331	83.79	71.84	0.325	79.81
OCRVerse (2026-01)	92.57	0.120	93.11	96.61	0.162	91.79	0.122	92.06	95.53	0.166	85.23	0.171	84.56	88.25	0.206	89.86
FD-RL (2025-11)	93.07	0.091	92.80	95.48	0.147	92.24	0.101	92.13	94.66	0.151	86.61	0.154	88.41	86.81	**0.187**	90.64
Nanonets-OCR2 (2025-10)	84.38	0.134	78.71	87.86	0.216	84.94	0.126	78.68	88.73	0.213	73.62	0.223	66.92	76.23	0.279	80.98
OCRFlux-3B (2025-06)	55.85	0.385	50.45	55.55	0.428	52.65	0.402	48.68	49.51	0.427	46.57	0.466	47.11	39.19	0.445	51.69
Logics-Parsing-v2 (2026-02)	94.18	0.095	95.25	96.79	0.159	92.44	0.111	92.34	96.07	0.171	87.58	0.147	88.90	88.50	0.222	91.40

Table 7: Academic component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	88.58	0.151	89.44	91.36	0.257	87.37	0.176	90.69	88.97	0.277	82.85	0.214	86.73	83.21	0.266	86.26
Qwen3.6-35B-A3B (2026-04)	88.30	0.158	90.08	90.63	0.243	87.01	0.172	89.52	88.73	0.223	81.15	0.213	84.87	79.93	0.270	85.48
Qwen3.6-27B (2026-04)	87.67	0.159	90.03	88.92	0.230	87.44	0.152	89.42	88.15	0.220	80.47	0.210	80.30	82.15	0.246	85.19
Qwen3.5-397B-A17B (2026-02)	85.19	0.123	94.28	73.61	0.187	81.56	0.141	95.28	63.54	0.206	76.10	0.149	92.06	51.16	0.227	80.95
Qwen3.5-122B-A10B (2026-02)	90.99	0.132	92.82	93.39	0.242	**94.08**	**0.083**	**95.50**	94.98	0.189	90.30	**0.121**	92.17	90.80	0.224	91.79
Qwen3.5-35B-A3B (2026-02)	80.32	0.121	93.67	59.36	0.237	78.64	0.141	92.59	57.42	0.245	76.18	0.167	91.07	54.20	0.260	78.38
Qwen3.5-27B (2026-02)	85.41	0.114	94.34	73.28	0.192	83.46	0.114	94.69	67.03	0.200	82.54	0.137	90.26	71.04	0.208	83.80
Qwen3.5-9B (2026-03)	91.55	0.128	93.80	93.69	0.198	91.61	0.124	95.47	91.74	0.211	86.35	0.167	88.54	87.23	0.246	89.84
Qwen3.5-4B (2026-03)	91.34	0.109	94.60	90.30	0.194	90.43	0.107	93.84	88.18	0.205	85.67	0.186	90.68	84.91	0.280	89.15
Qwen3.5-2B (2026-03)	82.83	0.232	89.00	82.75	0.361	84.20	0.218	90.53	83.85	0.347	78.35	0.255	83.02	77.53	0.347	81.79
Qwen3.5-0.8B (2026-03)	81.23	0.243	90.67	77.32	0.385	78.50	0.256	85.79	75.33	0.386	70.35	0.325	76.50	67.02	0.423	76.69
Qwen3-VL-8B (2025-10)	91.31	0.118	92.96	92.79	0.224	89.66	0.124	90.27	91.15	0.230	85.58	0.181	90.49	84.39	0.264	88.85
Qwen3-VL-4B (2025-10)	87.60	0.181	93.68	87.18	0.294	88.52	0.172	93.32	89.41	0.288	79.17	0.249	84.09	78.32	0.336	85.10
Qwen3-VL-2B (2025-10)	85.81	0.202	94.31	83.29	0.318	84.80	0.209	91.51	83.80	0.327	75.11	0.296	83.97	71.01	0.382	81.91
DeepSeek-V4-Flash-Vision-Exp (2026-08)	81.02	0.223	81.50	83.86	0.317	78.93	0.247	79.28	82.21	0.306	66.78	0.366	70.63	66.30	0.393	75.57
GLM-5.3-Flash (2026-08)	93.97	0.092	**95.66**	95.44	0.190	93.40	0.096	95.06	94.70	0.201	89.94	0.144	92.61	91.60	0.221	**92.43**
MiniCPM-V-4.6 (2026-05)	84.98	0.155	85.16	85.25	0.225	81.34	0.190	80.55	82.43	0.247	69.47	0.302	68.98	69.63	0.327	78.60
MiniCPM-V-4.5 (2025-08)	83.43	0.161	80.26	86.17	0.226	80.75	0.180	76.46	83.80	0.232	66.46	0.330	65.30	67.10	0.346	76.88
Gemini 3.6 Flash (2026-07)	92.64	0.101	94.23	93.78	0.217	92.40	0.109	94.19	93.95	0.227	**91.53**	**0.121**	**94.41**	92.27	0.204	92.19
Gemini-3.1-Pro (2026-02)	80.77	0.215	85.81	77.94	0.268	80.88	0.213	83.34	80.59	0.275	89.82	0.158	92.61	**92.65**	0.233	83.82
Kimi K2.6 (2026-04)	93.58	0.103	95.29	95.73	0.245	92.36	0.108	94.37	93.48	0.251	90.03	0.134	92.43	91.03	0.271	91.99
Step3-VL (2026-01)	75.69	0.299	82.66	74.29	0.373	71.99	0.331	77.89	71.20	0.392	66.78	0.379	70.45	67.83	0.417	71.49

Table 8: **Education: three-track component results** (138 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	67.93	0.436	65.51	81.90	0.568	60.38	0.518	59.97	72.99	0.621	53.90	0.543	50.06	65.89	0.642	60.73
YouTu-Parsing (2026-01)	81.19	0.240	81.59	85.98	0.386	76.55	0.273	76.08	80.92	0.405	59.23	0.414	56.95	62.11	0.459	72.32
DotsMOCR (2026-03)	84.97	0.172	84.44	87.68	0.276	80.71	0.214	81.15	82.38	0.299	64.75	0.343	65.79	62.72	0.411	76.81
GLM-OCR (2026-02)	76.00	0.293	73.14	84.18	0.455	69.63	0.384	66.80	80.47	0.517	65.40	0.411	65.65	71.62	0.534	70.34
MinerU2.5-Pro (2026-04)	85.13	0.214	84.26	**92.55**	0.320	80.19	0.274	81.25	86.68	0.358	65.23	0.392	61.61	73.29	0.453	76.85
MinerU2.5 (2025-09)	84.16	0.173	83.41	86.40	0.301	79.04	0.243	81.39	80.06	0.338	62.13	0.399	61.72	64.63	0.458	75.11
MonkeyOCRv2-B-Parsing (2026-07)	75.06	0.315	75.24	81.41	0.383	71.24	0.357	73.20	76.21	0.413	60.83	0.457	62.10	66.11	0.460	69.04
MonkeyOCR-pro-3B (2025-08)	73.06	0.291	68.58	79.64	0.441	69.04	0.338	69.12	71.75	0.465	53.95	0.490	56.52	54.34	0.557	65.35
MonkeyOCR-pro-1.2B (2025-07)	68.27	0.309	57.72	77.99	0.462	64.61	0.372	60.92	70.14	0.488	47.65	0.557	49.24	49.46	0.582	60.18
OpenOCR (2024-11)	34.53	0.391	42.75	-0.02	0.545	32.53	0.455	43.09	0.00	0.591	26.57	0.531	32.80	0.00	0.637	31.21
Dolphin-v2 (2025-12)	79.98	0.275	81.31	86.13	0.354	70.66	0.379	72.11	77.77	0.418	53.27	0.492	53.26	55.79	0.498	67.97
TeleOCR (2026-08)	84.75	0.174	83.61	88.08	**0.195**	78.67	0.250	78.52	82.45	0.283	66.20	0.411	66.40	73.27	0.398	76.54
End-to-end specialists																
OvisOCR2 (2026-07)	87.35	**0.125**	83.99	90.52	0.225	81.09	**0.183**	78.10	83.42	**0.271**	66.96	0.352	68.11	67.93	0.401	78.47
ABot-OCR (2026-05)	84.30	0.198	85.31	87.39	0.361	78.83	0.237	79.03	81.11	0.394	60.98	0.409	61.64	62.17	0.519	74.70
Qianfan-OCR (2026-03)	65.60	0.397	74.89	61.63	0.482	57.76	0.482	68.55	52.90	0.521	50.83	0.510	54.28	49.24	0.533	58.06
Unlimited-OCR (2026-06)	76.50	0.249	78.48	75.88	0.360	66.44	0.334	66.73	66.02	0.412	53.63	0.478	56.47	52.26	0.505	65.52
WeVisDoc-4B (2026-09)	85.75	0.169	84.05	90.11	0.279	84.92	0.190	84.44	89.29	0.301	70.10	0.337	71.47	72.56	0.364	80.26
WeVisDoc-2B (2026-09)	85.50	0.178	85.18	89.12	0.289	82.25	0.230	83.75	85.98	0.331	67.11	0.367	67.62	70.45	0.422	78.28
HunyuanOCR-1.5 (2026-07)	79.63	0.227	83.15	78.45	0.385	74.26	0.277	75.49	74.94	0.407	62.78	0.401	66.53	61.93	0.496	72.22
HunyuanOCR (2025-11)	75.17	0.279	78.33	75.08	0.377	70.88	0.318	75.02	69.47	0.404	55.12	0.441	54.92	54.57	0.492	67.06
dots.ocr (2025-07)	80.59	0.223	79.84	84.22	0.358	71.90	0.306	72.74	73.55	0.421	59.79	0.405	60.88	58.96	0.471	70.76
FireRed-OCR (2026-02)	79.05	0.309	86.28	81.73	0.394	78.63	0.329	87.78	81.04	0.407	60.70	0.456	65.90	61.80	0.504	72.79
DeepSeek-OCR-2 (2026-01)	64.28	0.351	66.91	61.04	0.468	53.74	0.436	55.67	49.18	0.513	42.59	0.627	46.30	44.21	0.591	53.54
DeepSeek-OCR (2025-10)	43.40	0.478	49.65	28.38	0.551	35.40	0.568	41.99	20.99	0.613	33.16	0.636	41.91	21.19	0.665	37.32
olmOCR-2-7B (2025-10)	81.44	0.287	82.22	90.84	0.335	76.98	0.325	78.09	85.38	0.362	60.85	0.450	61.24	66.29	0.452	73.09
olmOCR-7B (2025-08)	73.97	0.387	85.87	74.70	0.466	67.32	0.447	80.64	66.01	0.513	51.18	0.563	60.82	49.03	0.600	64.16
OCRVerse (2026-01)	84.43	0.227	84.77	91.17	0.338	82.50	0.256	83.81	89.28	0.361	67.04	0.384	66.20	73.36	0.436	77.99
FD-RL (2025-11)	86.79	0.173	86.47	91.24	0.294	84.87	0.195	85.54	88.57	0.318	69.75	0.324	66.95	74.67	0.379	80.47
Nanonets-OCR2 (2025-10)	70.91	0.289	58.86	82.80	0.374	70.31	0.314	64.41	77.93	0.393	52.17	0.475	45.55	58.44	0.492	64.46
OCRFlux-3B (2025-06)	46.06	0.469	44.50	40.57	0.501	43.68	0.498	45.17	35.68	0.514	34.24	0.600	37.75	24.94	0.559	41.33
Logics-Parsing-v2 (2026-02)	**87.84**	0.166	89.04	91.05	0.290	84.77	0.210	87.83	87.51	0.333	75.47	0.267	76.97	76.17	0.377	82.69

Table 8: Education component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	80.72	0.274	89.30	80.24	0.386	78.67	0.295	88.97	76.54	0.396	64.29	0.435	73.10	63.24	0.490	74.56
Qwen3.6-35B-A3B (2026-04)	77.88	0.305	88.77	75.38	0.400	75.77	0.323	83.48	76.13	0.418	61.26	0.458	70.59	58.99	0.505	71.64
Qwen3.6-27B (2026-04)	78.64	0.291	83.61	81.39	0.398	74.72	0.338	80.21	77.77	0.438	60.48	0.456	62.28	64.74	0.512	71.28
Qwen3.5-397B-A17B (2026-02)	78.27	0.230	87.13	70.67	0.350	79.16	0.228	88.25	72.03	0.359	70.98	0.275	82.05	58.37	0.387	76.14
Qwen3.5-122B-A10B (2026-02)	85.37	0.220	89.60	88.48	0.373	84.85	0.233	90.37	87.46	0.383	76.53	0.287	80.88	77.41	0.404	82.25
Qwen3.5-35B-A3B (2026-02)	77.80	0.233	88.16	68.56	0.383	76.88	0.235	88.17	66.00	0.382	70.71	0.314	81.12	62.42	0.418	75.13
Qwen3.5-27B (2026-02)	84.78	0.191	89.89	83.58	0.332	83.41	0.193	88.19	81.36	0.340	76.33	**0.260**	79.55	75.41	**0.356**	81.51
Qwen3.5-9B (2026-03)	83.44	0.234	89.31	84.46	0.370	80.70	0.280	86.89	83.20	0.406	73.15	0.315	81.53	69.44	0.408	79.10
Qwen3.5-4B (2026-03)	82.35	0.256	89.39	83.25	0.405	81.67	0.261	88.13	82.95	0.413	69.97	0.371	78.40	68.63	0.468	78.00
Qwen3.5-2B (2026-03)	76.00	0.312	82.19	77.04	0.442	73.54	0.349	79.32	76.13	0.461	59.06	0.470	68.34	55.88	0.527	69.53
Qwen3.5-0.8B (2026-03)	71.30	0.368	80.70	69.95	0.504	71.60	0.371	79.21	72.71	0.517	51.73	0.506	57.01	48.81	0.574	64.88
Qwen3-VL-8B (2025-10)	83.17	0.228	89.85	82.45	0.407	81.84	0.242	87.91	81.84	0.405	67.83	0.361	73.62	66.02	0.474	77.61
Qwen3-VL-4B (2025-10)	82.97	0.229	90.26	81.56	0.394	81.41	0.251	90.19	79.13	0.414	65.38	0.397	74.26	61.57	0.472	76.59
Qwen3-VL-2B (2025-10)	77.42	0.283	84.29	76.32	0.459	77.30	0.300	86.06	75.82	0.467	59.70	0.425	67.77	53.78	0.524	71.47
DeepSeek-V4-Flash-Vision-Exp (2026-08)	56.89	0.558	69.80	56.70	0.591	58.44	0.560	74.20	57.14	0.595	44.30	0.667	57.94	41.71	0.668	53.21
GLM-5.3-Flash (2026-08)	87.01	0.203	**90.47**	90.81	0.337	**86.55**	0.208	**90.79**	**89.67**	0.361	77.25	0.319	83.44	80.20	0.411	**83.60**
MiniCPM-V-4.6 (2026-05)	71.15	0.335	76.79	70.12	0.424	65.07	0.395	70.03	64.63	0.470	49.19	0.546	54.36	47.80	0.568	61.80
MiniCPM-V-4.5 (2025-08)	62.75	0.424	73.00	57.60	0.482	59.86	0.459	70.30	55.20	0.488	43.25	0.587	48.38	40.04	0.585	55.29
Gemini 3.6 Flash (2026-07)	84.44	0.238	89.86	87.25	0.383	84.19	0.246	89.37	87.80	0.371	78.81	0.294	84.56	81.27	0.408	82.48
Gemini-3.1-Pro (2026-02)	76.67	0.253	77.10	78.23	0.358	76.85	0.273	81.18	76.72	0.378	77.80	0.310	81.87	**82.54**	0.409	77.11
Kimi K2.6 (2026-04)	82.72	0.263	86.32	88.14	0.430	83.39	0.265	88.94	87.72	0.428	**78.91**	0.280	**87.02**	77.74	0.408	81.67
Step3-VL (2026-01)	62.77	0.472	74.89	60.65	0.519	64.89	0.453	81.51	58.48	0.531	52.08	0.569	65.27	47.92	0.594	59.91

Table 9: **Legal and Government: three-track component results** (151 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	55.18	0.284	0.00	93.96	0.422	50.29	0.339	0.00	84.76	0.459	52.58	0.361	6.05	87.75	0.453	52.68
YouTu-Parsing (2026-01)	90.68	0.175	98.95	90.58	0.247	86.68	0.178	94.85	83.01	0.260	55.87	0.182	6.05	79.78	0.228	77.74
DotsMOCR (2026-03)	**94.48**	0.093	**100.00**	92.71	0.194	92.09	0.108	**100.00**	87.01	0.200	57.80	0.147	6.05	82.03	0.242	81.46
GLM-OCR (2026-02)	60.86	0.204	13.75	89.29	0.327	70.27	0.255	51.80	84.46	0.380	61.20	0.275	23.25	87.84	0.373	64.11
MinerU2.5-Pro (2026-04)	91.82	0.157	96.15	94.96	0.176	87.65	0.188	92.50	89.26	0.192	58.48	0.251	11.60	88.92	0.250	79.32
MinerU2.5 (2025-09)	92.39	0.093	92.50	93.94	0.157	85.24	0.137	87.30	82.10	0.199	59.32	0.214	16.65	82.71	0.246	78.98
MonkeyOCRv2-B-Parsing (2026-07)	86.99	0.199	90.90	89.93	0.269	62.73	0.189	24.25	82.83	0.264	62.04	0.257	30.10	81.72	0.281	70.59
MonkeyOCR-pro-3B (2025-08)	58.72	0.202	10.10	86.29	0.313	81.34	0.265	92.55	78.03	0.359	51.56	0.327	12.80	74.55	0.413	63.87
MonkeyOCR-pro-1.2B (2025-07)	56.85	0.225	7.60	85.41	0.334	70.58	0.283	62.60	77.42	0.371	47.65	0.354	3.05	75.28	0.418	58.36
OpenOCR (2024-11)	33.93	0.307	32.45	0.00	0.355	32.63	0.333	31.15	0.00	0.369	23.03	0.370	6.05	0.00	0.389	29.86
Dolphin-v2 (2025-12)	88.93	0.186	95.65	89.77	0.226	84.69	0.222	93.50	82.76	0.246	53.31	0.402	34.80	65.30	0.396	75.64
TeleOCR (2026-08)	64.68	**0.059**	3.45	96.45	**0.087**	69.29	0.121	32.30	87.64	**0.129**	61.51	0.190	15.70	87.82	**0.183**	65.16
End-to-end specialists																
OvisOCR2 (2026-07)	94.13	0.089	94.80	**96.47**	0.145	90.48	**0.095**	88.00	92.92	0.146	63.73	0.151	17.80	88.48	0.187	82.78
ABot-OCR (2026-05)	89.35	0.129	88.00	92.98	0.294	86.06	0.165	88.00	86.73	0.325	61.85	0.182	18.10	85.67	0.326	79.09
Qianfan-OCR (2026-03)	80.15	0.177	77.90	80.27	0.232	73.85	0.256	76.80	70.37	0.291	50.33	0.293	11.20	69.08	0.316	68.11
Unlimited-OCR (2026-06)	90.41	0.116	90.95	91.86	0.176	83.29	0.159	85.55	80.24	0.213	55.90	0.234	11.80	79.29	0.248	76.53
WeVisDoc-4B (2026-09)	91.15	0.136	91.40	95.64	0.257	90.21	0.146	92.50	92.75	0.259	60.20	0.164	6.05	90.96	0.269	80.52
WeVisDoc-2B (2026-09)	89.28	0.171	92.50	92.44	0.286	88.85	0.167	92.50	90.79	0.275	57.74	0.209	6.05	88.04	0.302	78.63
HunyuanOCR-1.5 (2026-07)	86.53	0.141	87.90	85.75	0.256	83.08	0.150	85.65	78.54	0.270	52.13	0.212	0.00	77.64	0.308	73.91
HunyuanOCR (2025-11)	86.90	0.126	89.05	84.25	0.213	83.91	0.149	90.15	76.51	0.239	52.32	0.246	10.15	71.45	0.286	74.38
dots.ocr (2025-07)	89.64	0.134	90.30	92.07	0.213	86.63	0.184	90.30	88.03	0.273	55.48	0.222	9.10	79.54	0.299	77.25
FireRed-OCR (2026-02)	56.18	0.197	0.00	88.21	0.245	53.85	0.217	0.00	83.31	0.249	51.14	0.287	4.70	77.42	0.297	53.72
DeepSeek-OCR-2 (2026-01)	71.10	0.166	45.90	83.95	0.270	64.93	0.218	48.90	67.70	0.320	47.14	0.424	14.85	68.95	0.364	61.06
DeepSeek-OCR (2025-10)	39.75	0.322	3.85	47.57	0.408	36.34	0.316	3.55	37.08	0.371	31.52	0.383	5.00	27.81	0.402	35.87
olmOCR-2-7B (2025-10)	89.15	0.175	94.80	90.11	0.218	86.56	0.177	94.80	82.53	0.232	56.14	0.219	9.40	80.92	0.251	77.28
olmOCR-7B (2025-08)	82.65	0.240	94.80	77.20	0.344	79.46	0.269	94.80	70.43	0.358	45.27	0.326	0.00	68.44	0.397	69.13
OCRVerse (2026-01)	91.09	0.207	**100.00**	93.97	0.260	89.11	0.221	**100.00**	89.39	0.270	58.43	0.207	8.00	87.98	0.270	79.54
FD-RL (2025-11)	93.54	0.130	**100.00**	93.68	0.201	**92.58**	0.139	**100.00**	91.61	0.212	58.47	0.155	3.15	87.78	0.211	81.53
Nanonets-OCR2 (2025-10)	88.43	0.120	90.30	87.00	0.181	84.45	0.163	90.30	79.34	0.219	51.68	0.221	0.00	77.11	0.234	74.85
OCRFlux-3B (2025-06)	44.38	0.317	0.00	64.84	0.328	58.85	0.354	61.30	50.65	0.333	35.42	0.390	0.00	45.22	0.357	46.22
Logics-Parsing-v2 (2026-02)	91.78	0.120	93.70	93.63	0.241	89.32	0.146	93.70	88.81	0.267	77.98	0.152	58.95	90.18	0.254	86.36

Table 9: Legal_Gov component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg$_3$)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	83.80	0.133	74.25	90.42	0.225	82.99	0.149	77.25	86.57	0.240	52.50	0.235	0.00	80.94	0.283	73.09
Qwen3.6-35B-A3B (2026-04)	87.38	0.131	85.15	90.06	0.206	84.10	0.139	83.45	82.70	0.214	54.30	0.220	6.40	78.50	0.265	75.26
Qwen3.6-27B (2026-04)	82.80	0.178	79.25	87.00	0.236	82.21	0.173	78.15	85.74	0.228	54.54	0.245	5.00	83.09	0.273	73.18
Qwen3.5-397B-A17B (2026-02)	84.45	0.127	92.50	73.58	0.203	84.19	0.116	92.50	71.66	0.200	78.41	0.131	**82.70**	65.61	0.209	82.35
Qwen3.5-122B-A10B (2026-02)	90.42	0.185	**100.00**	89.73	0.266	91.29	0.127	94.80	91.75	0.210	62.53	0.147	9.65	**92.62**	0.216	81.41
Qwen3.5-35B-A3B (2026-02)	88.83	0.113	**100.00**	77.74	0.202	88.12	0.133	**100.00**	77.64	0.222	52.57	0.153	4.10	68.93	0.228	76.51
Qwen3.5-27B (2026-02)	91.51	0.096	92.50	91.60	0.170	91.50	0.113	97.85	87.98	0.198	81.07	**0.121**	79.65	75.69	0.195	88.03
Qwen3.5-9B (2026-03)	87.75	0.136	90.30	86.55	0.239	88.06	0.124	90.30	86.26	0.212	81.35	0.187	77.70	85.01	0.262	85.72
Qwen3.5-4B (2026-03)	88.38	0.152	90.30	90.06	0.223	88.28	0.150	90.30	89.55	0.230	58.84	0.191	9.65	86.00	0.264	78.50
Qwen3.5-2B (2026-03)	86.85	0.204	97.85	83.11	0.313	85.13	0.216	95.65	81.34	0.326	53.18	0.242	4.70	79.04	0.338	75.05
Qwen3.5-0.8B (2026-03)	81.52	0.202	90.30	74.51	0.322	78.75	0.228	90.30	68.72	0.335	45.26	0.296	0.00	65.39	0.372	68.51
Qwen3-VL-8B (2025-10)	86.54	0.150	90.30	84.37	0.258	87.67	0.149	95.65	82.29	0.245	72.13	0.173	53.25	80.43	0.258	82.11
Qwen3-VL-4B (2025-10)	58.87	0.135	4.05	86.04	0.273	88.12	0.142	94.80	83.81	0.280	57.11	0.210	10.65	81.72	0.296	68.03
Qwen3-VL-2B (2025-10)	54.64	0.169	0.00	80.79	0.288	52.84	0.176	0.00	76.08	0.287	49.93	0.250	1.55	73.23	0.357	52.47
DeepSeek-V4-Flash-Vision-Exp (2026-08)	49.34	0.332	0.00	81.18	0.345	48.12	0.345	3.75	75.08	0.380	41.08	0.450	4.60	63.65	0.441	46.18
GLM-5.3-Flash (2026-08)	93.21	0.111	94.80	95.90	0.202	91.72	0.136	94.80	**93.96**	0.233	**81.89**	0.152	70.60	90.29	0.227	**88.94**
MiniCPM-V-4.6 (2026-05)	82.97	0.213	90.80	79.43	0.307	77.99	0.226	88.70	67.92	0.331	46.35	0.330	4.55	67.51	0.395	69.10
MiniCPM-V-4.5 (2025-08)	55.30	0.222	9.40	78.70	0.272	54.28	0.215	6.05	78.32	0.246	40.75	0.400	6.05	56.20	0.344	50.11
Gemini 3.6 Flash (2026-07)	91.58	0.128	92.55	94.94	0.205	90.44	0.139	93.10	92.09	0.208	80.91	0.157	66.40	91.98	0.208	87.64
Gemini-3.1-Pro (2026-02)	86.09	0.184	90.30	86.33	0.211	88.13	0.184	94.80	87.97	0.229	81.72	0.157	70.65	90.19	0.194	85.31
Kimi K2.6 (2026-04)	89.84	0.189	94.80	93.64	0.292	89.03	0.182	94.80	90.45	0.291	60.42	0.174	7.90	90.77	0.288	79.76
Step3-VL (2026-01)	41.61	0.387	0.00	63.57	0.400	69.19	0.436	90.20	60.94	0.422	34.82	0.460	0.00	50.47	0.420	48.54

Table 10: **Business: three-track component results** (160 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	58.90	0.368	24.24	89.29	0.521	50.42	0.473	19.90	78.72	0.587	48.78	0.475	17.96	75.88	0.588	52.70
YouTu-Parsing (2026-01)	75.37	0.209	61.40	85.64	0.320	65.37	0.296	49.34	76.37	0.377	57.46	0.336	39.48	66.48	0.408	66.07
DotsMOCR (2026-03)	75.03	0.146	55.31	84.41	0.236	70.04	0.192	50.32	79.04	0.274	56.43	0.297	34.62	64.32	0.379	67.17
GLM-OCR (2026-02)	64.85	0.262	33.84	86.89	0.444	56.53	0.367	27.97	78.33	0.528	56.13	0.359	28.35	75.95	0.505	59.17
MinerU2.5-Pro (2026-04)	69.49	0.225	37.90	93.11	0.345	66.20	0.284	37.50	89.47	0.380	55.56	0.374	24.66	79.43	0.447	63.75
MinerU2.5 (2025-09)	69.30	0.151	33.34	89.63	0.292	61.58	0.245	28.65	80.58	0.360	54.78	0.317	23.41	72.63	0.401	61.89
MonkeyOCRv2-B-Parsing (2026-07)	63.06	0.287	33.18	84.66	0.368	56.32	0.348	27.37	76.37	0.400	51.61	0.400	27.30	67.55	0.418	56.99
MonkeyOCR-pro-3B (2025-08)	58.38	0.285	20.60	83.01	0.455	53.40	0.339	18.42	75.71	0.489	45.53	0.450	19.08	62.54	0.563	52.44
MonkeyOCR-pro-1.2B (2025-07)	59.70	0.284	24.96	82.52	0.452	53.61	0.348	23.33	72.27	0.488	46.83	0.474	20.92	66.94	0.556	53.38
OpenOCR (2024-11)	30.58	0.350	26.72	0.00	0.544	26.52	0.422	21.79	0.00	0.570	23.51	0.481	18.68	0.00	0.604	26.87
Dolphin-v2 (2025-12)	62.98	0.296	39.68	78.87	0.391	56.33	0.366	31.16	74.44	0.433	49.17	0.490	34.66	61.88	0.522	56.16
TeleOCR (2026-08)	**88.72**	**0.104**	**83.56**	93.00	**0.128**	77.79	0.201	69.05	84.47	**0.214**	67.49	0.297	54.66	77.49	**0.286**	78.00
End-to-end specialists																
OvisOCR2 (2026-07)	74.16	0.128	41.20	94.13	0.237	68.08	**0.181**	35.26	87.06	0.275	54.57	0.321	27.17	68.61	0.396	65.60
ABot-OCR (2026-05)	68.76	0.175	33.78	89.96	0.361	63.60	0.221	28.64	84.25	0.413	52.80	0.338	21.74	70.41	0.476	61.72
Qianfan-OCR (2026-03)	49.17	0.368	16.79	67.54	0.439	41.08	0.450	13.62	54.60	0.502	38.53	0.496	13.98	51.21	0.503	42.93
Unlimited-OCR (2026-06)	61.36	0.224	28.11	78.37	0.380	53.67	0.324	26.60	66.82	0.431	44.44	0.424	17.90	57.79	0.476	53.15
WeVisDoc-4B (2026-09)	72.58	0.164	42.51	91.63	0.280	72.06	0.194	45.91	89.70	0.314	61.11	0.275	31.63	79.23	0.362	68.58
WeVisDoc-2B (2026-09)	73.20	0.178	44.69	92.68	0.306	70.56	0.220	44.91	88.77	0.341	58.11	0.319	28.90	77.35	0.390	67.29
HunyuanOCR-1.5 (2026-07)	66.73	0.214	41.87	79.76	0.355	61.30	0.267	37.33	73.29	0.383	51.48	0.353	28.10	61.67	0.456	59.84
HunyuanOCR (2025-11)	58.71	0.236	25.43	74.34	0.341	54.70	0.282	23.08	69.21	0.364	41.43	0.413	13.92	51.67	0.439	51.61
dots.ocr (2025-07)	67.40	0.222	37.94	86.41	0.361	63.75	0.270	39.24	78.95	0.379	52.55	0.362	29.10	64.78	0.443	61.23
FireRed-OCR (2026-02)	67.00	0.254	39.56	86.80	0.366	64.11	0.294	38.17	83.53	0.397	51.82	0.398	22.40	72.90	0.441	60.98
DeepSeek-OCR-2 (2026-01)	45.55	0.332	13.95	55.93	0.470	40.87	0.397	12.66	49.67	0.479	30.55	0.606	8.90	43.34	0.566	38.99
DeepSeek-OCR (2025-10)	24.52	0.548	5.87	22.45	0.626	20.88	0.584	4.58	16.50	0.631	18.33	0.649	5.62	14.27	0.640	21.24
olmOCR-2-7B (2025-10)	60.57	0.272	21.71	87.23	0.341	57.44	0.306	20.51	82.38	0.362	48.64	0.407	17.29	69.37	0.433	55.55
olmOCR-7B (2025-08)	54.81	0.382	29.74	72.93	0.451	49.51	0.439	27.45	64.93	0.499	39.32	0.555	23.39	50.02	0.566	47.88
OCRVerse (2026-01)	68.96	0.245	41.90	89.51	0.345	66.77	0.287	45.11	83.94	0.384	61.22	0.319	38.12	77.47	0.418	65.65
FD-RL (2025-11)	76.61	0.190	56.72	92.14	0.318	74.25	0.215	54.73	89.50	0.343	61.45	0.304	39.12	75.60	0.409	70.77
Nanonets-OCR2 (2025-10)	56.53	0.263	14.80	81.05	0.392	51.66	0.321	12.31	74.77	0.427	40.49	0.448	7.05	59.20	0.488	49.56
OCRFlux-3B (2025-06)	41.38	0.449	21.19	47.80	0.442	36.96	0.492	17.64	42.49	0.452	31.73	0.534	13.16	35.41	0.483	36.69
Logics-Parsing-v2 (2026-02)	75.18	0.170	51.74	90.80	0.286	74.07	0.208	54.91	88.13	0.324	69.33	**0.261**	54.94	79.12	0.371	72.86

Table 10: Business component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	68.51	0.319	55.59	81.85	0.398	67.45	0.331	56.93	78.56	0.405	53.81	0.443	40.44	65.29	0.469	63.25
Qwen3.6-35B-A3B (2026-04)	66.55	0.309	49.62	80.94	0.404	61.11	0.347	41.36	76.70	0.445	49.82	0.456	33.74	61.30	0.482	59.16
Qwen3.6-27B (2026-04)	65.42	0.326	46.51	82.39	0.426	61.62	0.377	47.17	75.38	0.457	50.48	0.469	35.07	63.30	0.507	59.17
Qwen3.5-397B-A17B (2026-02)	72.39	0.208	52.78	85.22	0.365	67.81	0.234	47.35	79.48	0.383	68.82	0.270	59.94	73.52	0.393	69.67
Qwen3.5-122B-A10B (2026-02)	75.93	0.204	57.72	90.52	0.347	73.20	0.233	54.25	88.61	0.370	71.77	0.274	57.96	84.77	0.385	73.63
Qwen3.5-35B-A3B (2026-02)	69.67	0.214	47.54	82.86	0.353	69.74	0.236	52.16	80.65	0.365	63.79	0.282	47.60	71.99	0.381	67.73
Qwen3.5-27B (2026-02)	68.48	0.246	55.31	74.71	0.379	66.56	0.255	47.95	77.29	0.367	66.83	0.262	52.16	74.56	0.367	67.29
Qwen3.5-9B (2026-03)	75.01	0.243	59.28	90.04	0.380	72.91	0.256	56.16	88.14	0.380	70.25	0.307	59.73	81.70	0.417	72.72
Qwen3.5-4B (2026-03)	75.49	0.244	63.91	86.98	0.397	71.85	0.278	57.54	85.78	0.417	65.59	0.353	55.28	76.77	0.456	70.98
Qwen3.5-2B (2026-03)	64.88	0.328	46.00	81.45	0.462	61.00	0.357	38.81	79.87	0.487	54.15	0.433	38.89	66.87	0.517	60.01
Qwen3.5-0.8B (2026-03)	53.51	0.365	21.90	75.13	0.495	50.06	0.404	21.53	69.02	0.518	43.22	0.479	21.39	56.18	0.556	48.93
Qwen3-VL-8B (2025-10)	68.95	0.262	45.95	87.09	0.415	67.55	0.290	48.85	82.77	0.427	53.94	0.370	27.36	71.46	0.472	63.48
Qwen3-VL-4B (2025-10)	67.77	0.274	48.38	82.35	0.428	63.16	0.303	40.04	79.77	0.441	56.64	0.387	38.61	70.00	0.483	62.52
Qwen3-VL-2B (2025-10)	59.43	0.311	31.99	77.39	0.444	57.73	0.342	33.55	73.84	0.458	48.69	0.423	25.54	62.84	0.490	55.28
DeepSeek-V4-Flash-Vision-Exp (2026-08)	43.18	0.596	34.01	55.10	0.623	42.53	0.616	37.35	51.81	0.623	32.84	0.678	24.94	41.41	0.655	39.52
GLM-5.3-Flash (2026-08)	83.92	0.186	74.97	**95.45**	0.347	**79.27**	0.216	67.44	**92.00**	0.355	71.54	0.285	59.85	83.27	0.394	**78.24**
MiniCPM-V-4.6 (2026-05)	54.49	0.337	22.24	74.96	0.451	50.38	0.395	24.80	65.79	0.486	37.81	0.525	15.01	50.91	0.559	47.56
MiniCPM-V-4.5 (2025-08)	40.37	0.471	13.82	54.42	0.476	38.70	0.492	13.38	51.89	0.502	28.66	0.611	8.01	39.08	0.571	35.91
Gemini 3.6 Flash (2026-07)	81.12	0.214	70.46	94.31	0.328	78.63	0.251	72.14	88.81	0.354	**73.96**	0.301	**65.50**	**86.49**	0.378	77.90
Gemini-3.1-Pro (2026-02)	76.44	0.305	76.15	83.65	0.393	77.85	0.305	**80.84**	83.21	0.389	71.11	0.287	62.06	79.96	0.327	75.13
Kimi K2.6 (2026-04)	69.19	0.307	51.03	87.23	0.477	68.71	0.323	52.22	86.25	0.487	67.13	0.347	52.31	83.82	0.487	68.34
Step3-VL (2026-01)	51.10	0.452	34.39	64.14	0.496	50.52	0.472	30.44	68.34	0.508	38.82	0.575	23.20	50.76	0.572	46.81

Table 11: **Finance: three-track component results** (212 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	52.74	0.589	42.35	74.78	0.696	47.87	0.666	41.98	68.24	0.733	47.49	0.667	40.98	68.24	0.743	49.37
YouTu-Parsing (2026-01)	62.22	0.251	33.48	78.31	0.425	56.49	0.314	29.51	71.38	0.464	50.69	0.423	29.35	65.00	0.527	56.47
DotsMOCR (2026-03)	60.85	0.164	31.36	67.58	0.302	59.24	0.253	34.63	68.36	0.371	51.18	0.370	33.45	57.08	0.477	57.09
GLM-OCR (2026-02)	51.80	0.458	28.23	73.02	0.617	46.91	0.544	24.81	70.28	0.656	51.81	0.572	43.24	69.40	0.678	50.17
MinerU2.5-Pro (2026-04)	63.10	0.290	39.84	78.43	0.446	56.62	0.367	31.07	75.51	0.498	53.42	0.451	34.19	71.12	0.558	57.71
MinerU2.5 (2025-09)	61.86	0.265	36.98	75.07	0.434	54.31	0.353	29.14	69.13	0.498	48.22	0.460	26.25	64.40	0.546	54.80
MonkeyOCRv2-B-Parsing (2026-07)	59.79	0.406	46.15	73.85	0.522	56.91	0.449	46.57	69.04	0.533	49.77	0.528	40.30	61.75	0.566	55.49
MonkeyOCR-pro-3B (2025-08)	46.53	0.493	28.56	60.37	0.660	37.20	0.595	19.83	51.31	0.700	38.85	0.638	31.38	49.00	0.731	40.86
MonkeyOCR-pro-1.2B (2025-07)	45.68	0.505	28.61	58.92	0.659	37.19	0.618	21.94	51.41	0.701	35.01	0.703	30.16	45.16	0.755	39.29
OpenOCR (2024-11)	28.64	0.336	19.50	0.00	0.537	24.81	0.414	15.86	0.00	0.582	22.24	0.490	15.69	0.00	0.634	25.23
Dolphin-v2 (2025-12)	48.27	0.466	29.29	62.14	0.605	41.75	0.575	28.28	54.44	0.641	36.72	0.641	30.07	44.20	0.650	42.25
TeleOCR (2026-08)	**82.36**	**0.072**	**61.56**	**92.72**	**0.138**	**70.55**	**0.225**	52.98	81.19	**0.264**	62.78	0.380	51.11	75.23	**0.414**	**71.90**
End-to-end specialists																
OvisOCR2 (2026-07)	70.36	0.250	51.78	84.24	0.442	65.86	0.315	49.21	79.86	0.474	59.49	0.387	44.79	72.36	0.523	65.24
ABot-OCR (2026-05)	65.60	0.373	53.51	80.61	0.537	57.49	0.457	44.38	73.78	0.595	50.87	0.501	40.57	62.11	0.630	57.99
Qianfan-OCR (2026-03)	44.18	0.474	24.08	55.83	0.560	37.38	0.564	19.43	49.11	0.611	41.32	0.598	37.30	46.42	0.608	40.96
Unlimited-OCR (2026-06)	61.69	0.316	46.75	69.96	0.488	50.38	0.430	35.07	59.08	0.544	47.98	0.512	43.30	51.82	0.571	53.35
WeVisDoc-4B (2026-09)	67.07	0.296	47.48	83.34	0.425	63.64	0.363	45.80	81.43	0.478	58.01	0.443	43.35	74.98	0.534	62.91
WeVisDoc-2B (2026-09)	65.88	0.325	48.10	82.02	0.473	61.11	0.398	47.03	76.13	0.513	55.48	0.480	44.82	69.63	0.573	60.83
HunyuanOCR-1.5 (2026-07)	61.27	0.352	52.22	66.80	0.502	53.99	0.420	43.02	60.95	0.552	50.46	0.502	46.55	55.06	0.600	55.24
HunyuanOCR (2025-11)	49.47	0.376	27.11	58.91	0.501	44.15	0.433	22.01	53.70	0.513	38.96	0.543	25.86	45.28	0.571	44.19
dots.ocr (2025-07)	57.80	0.316	33.82	71.19	0.473	49.74	0.403	26.75	62.82	0.525	45.79	0.498	28.20	58.95	0.583	51.11
FireRed-OCR (2026-02)	53.06	0.401	33.06	66.26	0.530	49.39	0.464	30.14	64.39	0.576	44.20	0.532	29.16	56.65	0.613	48.88
DeepSeek-OCR-2 (2026-01)	40.68	0.500	23.37	48.64	0.607	35.07	0.581	19.06	44.23	0.637	28.27	0.699	19.07	35.67	0.683	34.67
DeepSeek-OCR (2025-10)	17.02	0.659	7.38	9.58	0.684	16.18	0.636	5.88	6.32	0.683	15.15	0.685	6.67	7.23	0.673	16.12
olmOCR-2-7B (2025-10)	55.74	0.350	32.53	69.71	0.422	50.16	0.409	26.74	64.63	0.457	45.53	0.508	30.74	56.65	0.508	50.48
olmOCR-7B (2025-08)	47.17	0.469	28.68	59.71	0.560	42.30	0.538	25.12	55.59	0.587	36.36	0.633	22.36	50.05	0.656	41.94
OCRVerse (2026-01)	53.71	0.431	32.53	71.71	0.590	50.91	0.481	33.05	67.76	0.616	49.32	0.481	33.31	62.79	0.594	51.31
FD-RL (2025-11)	61.65	0.285	35.78	77.66	0.474	59.30	0.314	34.62	74.72	0.486	56.18	0.355	34.86	69.24	0.499	59.04
Nanonets-OCR2 (2025-10)	54.96	0.301	25.85	69.11	0.467	48.48	0.400	23.56	61.85	0.500	40.92	0.509	19.52	54.11	0.569	48.12
OCRFlux-3B (2025-06)	44.45	0.545	43.33	44.51	0.540	32.88	0.592	16.82	41.01	0.563	39.20	0.631	38.88	41.79	0.583	38.84
Logics-Parsing-v2 (2026-02)	58.92	0.331	36.39	73.52	0.470	55.14	0.393	36.66	68.03	0.522	61.58	0.403	54.58	70.42	0.534	58.55

Table 11: Finance component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	55.64	0.468	47.43	66.29	0.556	53.12	0.482	40.78	66.76	0.550	53.51	0.517	48.69	63.52	0.589	54.09
Qwen3.6-35B-A3B (2026-04)	59.51	0.365	44.47	70.61	0.470	54.79	0.407	39.48	65.60	0.504	53.34	0.491	49.70	59.39	0.558	55.88
Qwen3.6-27B (2026-04)	58.16	0.451	49.68	69.95	0.558	53.65	0.503	44.59	66.64	0.587	49.51	0.569	44.30	61.12	0.633	53.78
Qwen3.5-397B-A17B (2026-02)	52.68	0.266	32.43	52.19	0.449	52.67	0.296	28.33	59.29	0.468	56.07	0.322	44.92	55.53	0.465	53.81
Qwen3.5-122B-A10B (2026-02)	62.48	0.242	33.98	77.65	0.449	59.77	0.289	32.30	75.90	0.474	**67.14**	**0.305**	54.00	77.94	0.476	63.13
Qwen3.5-35B-A3B (2026-02)	53.33	0.257	31.32	54.33	0.448	52.24	0.310	30.43	57.27	0.477	57.44	0.359	53.30	54.87	0.510	54.34
Qwen3.5-27B (2026-02)	56.48	0.285	34.68	63.22	0.465	56.63	0.310	33.70	67.15	0.479	60.40	0.339	**56.07**	59.00	0.486	57.84
Qwen3.5-9B (2026-03)	57.53	0.305	34.61	68.50	0.490	54.87	0.366	33.85	67.34	0.525	55.52	0.406	40.61	66.58	0.546	55.97
Qwen3.5-4B (2026-03)	57.45	0.346	37.51	69.45	0.528	54.26	0.391	35.14	66.70	0.550	56.64	0.465	50.15	66.25	0.588	56.12
Qwen3.5-2B (2026-03)	53.14	0.363	34.35	61.32	0.526	50.88	0.399	30.03	62.51	0.546	52.25	0.492	45.46	60.51	0.599	52.09
Qwen3.5-0.8B (2026-03)	49.34	0.442	35.29	56.90	0.595	42.17	0.525	29.91	49.14	0.630	42.90	0.605	47.20	42.00	0.669	44.80
Qwen3-VL-8B (2025-10)	58.90	0.301	36.83	69.96	0.496	55.24	0.355	33.19	68.01	0.524	53.39	0.379	32.63	65.48	0.537	55.84
Qwen3-VL-4B (2025-10)	59.50	0.270	34.39	71.17	0.489	56.35	0.329	33.65	68.27	0.512	58.80	0.405	52.97	63.97	0.548	58.22
Qwen3-VL-2B (2025-10)	55.61	0.332	37.21	62.85	0.514	50.76	0.386	29.94	60.95	0.537	51.11	0.479	46.68	54.52	0.593	52.49
DeepSeek-V4-Flash-Vision-Exp (2026-08)	39.42	0.688	35.34	51.75	0.676	35.30	0.718	31.03	46.63	0.703	29.59	0.774	29.38	36.78	0.727	34.77
GLM-5.3-Flash (2026-08)	70.08	0.264	49.07	87.53	0.486	68.56	0.314	52.08	**85.00**	0.499	65.24	0.353	50.00	**81.00**	0.514	67.96
MiniCPM-V-4.6 (2026-05)	51.96	0.462	45.45	56.63	0.544	43.59	0.547	35.52	50.00	0.589	40.39	0.619	40.14	42.98	0.629	45.32
MiniCPM-V-4.5 (2025-08)	37.59	0.559	23.92	44.73	0.594	33.88	0.610	18.67	43.92	0.623	26.85	0.669	14.86	32.58	0.645	32.77
Gemini 3.6 Flash (2026-07)	65.88	0.350	53.44	79.21	0.509	63.61	0.380	51.25	77.54	0.520	63.90	0.393	53.82	77.21	0.516	64.46
Gemini-3.1-Pro (2026-02)	59.33	0.420	48.20	71.80	0.554	59.21	0.467	**54.30**	70.06	0.570	58.37	0.372	38.70	73.62	0.510	58.97
Kimi K2.6 (2026-04)	57.26	0.348	32.96	73.66	0.572	53.79	0.385	28.55	71.31	0.573	62.16	0.381	48.31	76.21	0.570	57.74
Step3-VL (2026-01)	39.77	0.598	26.25	52.87	0.594	39.39	0.622	26.88	53.54	0.616	37.40	0.654	26.81	50.82	0.644	38.85

Table 12: **Medical: three-track component results** (200 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	45.69	0.594	29.05	67.40	0.687	46.95	0.644	46.88	58.34	0.709	39.99	0.659	25.38	60.45	0.711	44.21
YouTu-Parsing (2026-01)	69.34	0.227	59.50	71.19	0.376	59.41	0.300	45.57	62.65	0.410	55.75	0.372	47.58	56.90	0.454	61.50
DotsMOCR (2026-03)	62.86	0.152	42.80	60.95	0.299	63.02	**0.235**	50.07	62.51	0.353	54.12	0.328	41.90	53.26	0.422	60.00
GLM-OCR (2026-02)	53.99	0.469	46.61	62.27	0.596	44.80	0.539	32.76	55.54	0.631	52.46	0.565	51.12	62.73	0.647	50.42
MinerU2.5-Pro (2026-04)	68.35	0.242	55.91	73.31	0.408	64.22	0.313	55.07	68.90	0.433	56.83	0.424	50.85	62.05	0.507	63.13
MinerU2.5 (2025-09)	66.86	0.217	52.61	69.65	0.393	61.74	0.295	52.40	62.37	0.444	49.96	0.429	37.69	55.05	0.531	59.52
MonkeyOCRv2-B-Parsing (2026-07)	63.61	0.404	63.80	67.43	0.491	57.61	0.437	52.62	63.94	0.520	53.89	0.513	55.51	57.48	0.534	58.37
MonkeyOCR-pro-3B (2025-08)	56.48	0.412	50.04	60.64	0.559	47.38	0.501	37.93	54.29	0.619	47.46	0.582	51.19	49.36	0.657	50.44
MonkeyOCR-pro-1.2B (2025-07)	56.58	0.419	52.03	59.66	0.564	42.87	0.518	26.95	53.45	0.606	41.24	0.650	43.67	45.01	0.677	46.90
OpenOCR (2024-11)	34.11	0.330	35.36	0.00	0.506	31.64	0.408	35.72	0.00	0.560	28.21	0.481	32.74	0.00	0.607	31.32
Dolphin-v2 (2025-12)	45.74	0.453	28.09	54.39	0.538	42.34	0.531	27.93	52.15	0.591	32.84	0.704	26.88	42.01	0.680	40.31
TeleOCR (2026-08)	**79.96**	**0.128**	66.20	**86.52**	**0.147**	73.20	0.246	69.34	74.85	**0.246**	65.75	0.375	64.65	70.12	**0.389**	72.97
End-to-end specialists																
OvisOCR2 (2026-07)	76.53	0.200	68.31	81.34	0.377	71.88	0.273	68.77	74.21	0.430	59.50	0.352	46.77	66.98	0.473	69.31
ABot-OCR (2026-05)	69.51	0.291	60.42	77.18	0.473	68.91	0.351	**74.01**	67.82	0.513	60.85	0.419	60.61	63.86	0.559	66.42
Qianfan-OCR (2026-03)	39.86	0.531	33.66	39.06	0.581	37.26	0.595	35.60	35.66	0.609	37.63	0.594	31.18	41.15	0.601	38.25
Unlimited-OCR (2026-06)	63.93	0.310	58.35	64.41	0.457	61.99	0.381	63.90	60.12	0.500	47.86	0.490	43.34	49.27	0.550	57.93
WeVisDoc-4B (2026-09)	74.52	0.229	63.63	82.83	0.364	71.69	0.267	63.49	78.27	0.412	64.02	0.390	61.10	69.94	0.457	70.08
WeVisDoc-2B (2026-09)	74.40	0.275	**70.84**	79.88	0.401	69.55	0.332	66.76	75.08	0.428	62.74	0.440	65.28	66.97	0.505	68.90
HunyuanOCR-1.5 (2026-07)	60.97	0.312	50.71	63.41	0.473	60.01	0.373	58.50	58.82	0.511	51.46	0.442	47.98	50.65	0.549	57.48
HunyuanOCR (2025-11)	54.70	0.357	43.33	56.49	0.485	50.92	0.423	43.27	51.77	0.510	46.00	0.470	44.44	40.59	0.536	50.54
dots.ocr (2025-07)	63.82	0.247	46.44	69.74	0.413	56.78	0.347	45.84	59.21	0.466	38.04	0.465	12.34	48.24	0.541	52.88
FireRed-OCR (2026-02)	59.65	0.388	52.44	65.33	0.482	57.52	0.423	54.68	60.15	0.514	37.61	0.503	15.64	47.46	0.564	51.59
DeepSeek-OCR-2 (2026-01)	48.59	0.401	41.96	43.89	0.522	35.41	0.503	24.02	32.46	0.564	32.33	0.684	35.78	29.64	0.635	38.78
DeepSeek-OCR (2025-10)	13.64	0.722	6.29	6.86	0.728	13.72	0.710	8.33	3.87	0.722	12.78	0.713	4.08	5.59	0.676	13.38
olmOCR-2-7B (2025-10)	58.35	0.355	41.49	69.08	0.410	53.32	0.412	37.82	63.37	0.446	45.28	0.474	33.96	49.29	0.493	52.32
olmOCR-7B (2025-08)	53.42	0.442	47.56	56.86	0.507	48.55	0.496	43.64	51.63	0.539	40.45	0.596	40.49	40.41	0.604	47.47
OCRVerse (2026-01)	63.85	0.344	53.53	72.37	0.492	61.78	0.366	52.66	69.27	0.498	59.06	0.407	56.77	61.10	0.511	61.56
FD-RL (2025-11)	70.98	0.220	55.03	79.94	0.382	67.57	0.262	55.43	73.45	0.404	61.82	0.320	53.41	64.07	0.427	66.79
Nanonets-OCR2 (2025-10)	50.84	0.314	21.89	62.05	0.436	46.14	0.396	24.55	53.52	0.503	35.47	0.514	17.95	39.81	0.536	44.15
OCRFlux-3B (2025-06)	44.19	0.503	35.12	47.78	0.483	39.32	0.545	30.95	41.49	0.500	33.38	0.596	25.34	34.43	0.536	38.96
Logics-Parsing-v2 (2026-02)	67.09	0.266	55.07	72.84	0.395	63.05	0.316	53.52	67.21	0.432	66.55	0.358	68.24	67.22	0.460	65.56

Table 12: Medical component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	60.90	0.421	60.20	64.58	0.480	59.13	0.455	61.77	61.15	0.495	51.91	0.515	50.77	56.48	0.552	57.31
Qwen3.6-35B-A3B (2026-04)	62.05	0.386	61.58	63.20	0.484	59.39	0.448	66.09	56.89	0.497	49.83	0.503	45.94	53.86	0.531	57.09
Qwen3.6-27B (2026-04)	56.90	0.470	53.93	63.73	0.553	45.48	0.521	30.73	57.81	0.571	48.58	0.545	46.98	53.29	0.591	50.32
Qwen3.5-397B-A17B (2026-02)	57.98	0.297	56.39	47.28	0.445	57.07	0.339	53.81	51.24	0.456	59.71	0.369	69.06	46.99	0.471	58.25
Qwen3.5-122B-A10B (2026-02)	66.29	0.245	49.18	74.14	0.420	66.89	0.263	53.85	73.15	0.425	68.38	0.320	63.72	73.43	0.462	67.19
Qwen3.5-35B-A3B (2026-02)	55.15	0.267	43.52	48.62	0.419	56.51	0.301	49.50	50.12	0.421	58.74	0.378	69.02	45.06	0.485	56.80
Qwen3.5-27B (2026-02)	64.54	0.256	58.01	61.18	0.408	61.62	0.294	55.23	59.01	0.416	64.77	0.356	68.48	61.39	0.464	63.64
Qwen3.5-9B (2026-03)	64.87	0.303	54.22	70.73	0.455	62.23	0.319	48.44	70.17	0.473	65.39	0.395	**70.93**	64.70	0.519	64.16
Qwen3.5-4B (2026-03)	62.55	0.343	54.62	67.34	0.501	55.89	0.379	43.59	62.02	0.494	60.83	0.454	65.20	62.70	0.554	59.76
Qwen3.5-2B (2026-03)	57.53	0.398	50.89	61.50	0.521	48.62	0.450	28.40	62.45	0.550	50.60	0.525	46.90	57.37	0.591	52.25
Qwen3.5-0.8B (2026-03)	44.94	0.431	21.01	56.86	0.552	47.17	0.482	38.98	50.76	0.586	39.25	0.554	28.81	44.38	0.602	43.79
Qwen3-VL-8B (2025-10)	63.51	0.275	46.23	71.79	0.447	63.03	0.293	48.83	69.55	0.450	51.68	0.390	37.09	56.96	0.493	59.41
Qwen3-VL-4B (2025-10)	64.16	0.284	47.61	73.27	0.468	60.35	0.325	47.57	65.96	0.488	54.18	0.411	46.87	56.80	0.540	59.56
Qwen3-VL-2B (2025-10)	57.60	0.326	41.38	63.99	0.476	53.92	0.381	42.66	57.19	0.505	52.39	0.469	54.00	50.03	0.542	54.64
DeepSeek-V4-Flash-Vision-Exp (2026-08)	35.03	0.703	35.84	39.56	0.672	29.57	0.732	24.96	36.92	0.680	23.49	0.753	15.10	30.63	0.714	29.36
GLM-5.3-Flash (2026-08)	75.43	0.257	70.04	81.92	0.438	**73.69**	0.294	70.29	**80.13**	0.457	**70.85**	**0.315**	68.08	**75.98**	0.456	**73.32**
MiniCPM-V-4.6 (2026-05)	53.24	0.430	43.61	59.16	0.500	49.30	0.485	47.35	49.04	0.554	34.11	0.601	22.61	39.80	0.605	45.55
MiniCPM-V-4.5 (2025-08)	30.99	0.574	17.44	32.90	0.574	28.24	0.616	16.52	29.75	0.585	21.69	0.682	7.41	25.86	0.630	26.97
Gemini 3.6 Flash (2026-07)	64.29	0.372	56.94	73.19	0.490	66.64	0.394	66.44	72.86	0.512	62.88	0.423	62.17	68.75	0.521	64.60
Gemini-3.1-Pro (2026-02)	62.57	0.369	50.28	74.31	0.481	64.32	0.408	63.40	70.33	0.504	59.37	0.389	51.88	65.10	0.465	62.09
Kimi K2.6 (2026-04)	62.26	0.338	53.07	67.52	0.540	61.45	0.360	53.28	67.09	0.547	67.65	0.359	67.42	71.43	0.551	63.79
Step3-VL (2026-01)	42.91	0.571	38.56	47.28	0.521	42.08	0.598	41.70	44.37	0.548	36.72	0.656	38.05	37.74	0.626	40.57

Table 13: **Publishing: three-track component results** (100 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	74.83	0.339	68.93	89.49	0.482	71.86	0.384	68.93	85.04	0.539	63.43	0.469	67.99	69.21	0.583	70.04
YouTu-Parsing (2026-01)	81.21	0.173	77.72	83.20	0.287	77.68	0.204	79.36	74.05	0.317	70.28	0.342	87.59	57.47	0.414	76.39
DotsMOCR (2026-03)	80.72	0.130	79.06	76.07	0.222	78.56	0.171	79.06	73.70	0.273	66.66	0.280	71.32	56.69	0.357	75.31
GLM-OCR (2026-02)	83.43	0.233	87.04	86.50	0.413	79.67	0.286	87.04	80.58	0.464	67.35	0.367	66.19	72.52	0.501	76.82
MinerU2.5-Pro (2026-04)	84.80	0.148	86.81	82.42	0.280	85.15	0.180	86.10	87.37	0.305	73.69	0.334	83.20	71.29	0.396	81.21
MinerU2.5 (2025-09)	81.39	0.161	81.98	78.30	0.274	79.53	0.203	83.80	75.04	0.331	67.98	0.393	83.38	59.87	0.438	76.30
MonkeyOCRv2-B-Parsing (2026-07)	81.90	0.290	90.56	84.18	0.359	79.34	0.321	90.54	79.56	0.383	72.25	0.396	87.67	68.65	0.392	77.83
MonkeyOCR-pro-3B (2025-08)	79.21	0.288	81.40	84.98	0.464	71.06	0.413	82.69	71.79	0.558	61.40	0.526	75.18	61.64	0.654	70.56
MonkeyOCR-pro-1.2B (2025-07)	54.19	0.300	9.46	83.12	0.484	70.34	0.410	80.48	71.52	0.556	43.22	0.586	34.83	53.48	0.682	55.92
OpenOCR (2024-11)	45.35	0.372	73.27	0.00	0.540	43.74	0.424	73.58	0.00	0.565	34.65	0.500	53.94	0.00	0.611	41.25
Dolphin-v2 (2025-12)	78.68	0.202	80.66	75.57	0.310	75.11	0.247	74.35	75.64	0.350	48.36	0.501	38.71	56.46	0.504	67.38
TeleOCR (2026-08)	87.84	**0.088**	78.46	93.82	**0.107**	84.27	0.161	79.56	89.38	**0.188**	70.23	0.333	73.81	70.17	0.334	80.78
End-to-end specialists																
OvisOCR2 (2026-07)	87.65	0.127	82.29	93.40	0.214	84.26	0.173	82.29	87.82	0.276	77.22	0.306	**91.03**	71.21	0.376	83.04
ABot-OCR (2026-05)	68.87	0.212	40.94	86.92	0.356	66.62	0.244	39.27	84.98	0.378	69.92	0.347	82.06	62.45	0.440	68.47
Qianfan-OCR (2026-03)	59.61	0.314	72.06	38.19	0.362	54.99	0.369	70.27	31.56	0.418	53.84	0.457	64.61	42.58	0.467	56.15
Unlimited-OCR (2026-06)	84.33	0.190	90.95	81.01	0.347	80.48	0.217	81.20	81.88	0.347	57.98	0.458	68.60	51.14	0.444	74.26
WeVisDoc-4B (2026-09)	86.92	0.217	90.41	92.07	0.332	84.61	0.246	90.41	87.99	0.335	78.51	0.330	90.41	78.17	0.426	83.35
WeVisDoc-2B (2026-09)	85.34	0.211	89.16	87.97	0.311	83.49	0.250	89.16	86.33	0.356	77.83	0.334	90.41	76.50	0.393	82.22
HunyuanOCR-1.5 (2026-07)	81.69	0.244	90.76	78.75	0.389	77.13	0.267	90.79	67.33	0.407	69.05	0.360	88.97	54.20	0.482	75.96
HunyuanOCR (2025-11)	69.46	0.280	78.25	58.09	0.364	69.08	0.289	78.25	57.87	0.388	57.84	0.419	84.88	30.58	0.437	65.46
dots.ocr (2025-07)	79.34	0.237	78.06	83.62	0.344	76.80	0.284	78.06	80.74	0.375	64.51	0.387	83.39	48.82	0.406	73.55
FireRed-OCR (2026-02)	82.45	0.214	84.60	84.19	0.354	80.60	0.253	85.11	81.99	0.381	70.94	0.353	83.58	64.55	0.425	78.00
DeepSeek-OCR-2 (2026-01)	56.14	0.365	72.11	32.77	0.474	56.75	0.396	77.39	32.50	0.493	41.42	0.634	67.53	20.13	0.597	51.44
DeepSeek-OCR (2025-10)	55.81	0.362	87.60	16.03	0.497	55.54	0.409	89.47	18.07	0.503	19.52	0.604	12.36	6.59	0.580	43.62
olmOCR-2-7B (2025-10)	72.05	0.339	85.51	64.54	0.414	71.43	0.352	85.51	63.95	0.420	59.44	0.450	84.38	38.94	0.473	67.64
olmOCR-7B (2025-08)	66.44	0.409	84.05	56.12	0.489	64.73	0.431	84.05	53.27	0.531	54.50	0.537	84.05	33.13	0.580	61.89
OCRVerse (2026-01)	79.31	0.211	70.25	88.75	0.318	77.96	0.235	70.25	87.16	0.358	71.51	0.346	76.88	72.29	0.428	76.26
FD-RL (2025-11)	83.84	0.144	82.16	83.73	0.277	83.83	0.169	82.16	86.23	0.311	74.82	0.247	79.57	69.54	0.339	80.83
Nanonets-OCR2 (2025-10)	76.92	0.196	79.44	70.92	0.348	71.67	0.270	77.26	64.76	0.396	59.89	0.422	80.21	41.60	0.462	69.49
OCRFlux-3B (2025-06)	33.34	0.460	3.41	42.59	0.398	29.91	0.477	2.59	34.83	0.396	22.42	0.567	1.39	22.58	0.456	28.56
Logics-Parsing-v2 (2026-02)	82.50	0.204	84.82	83.08	0.332	79.74	0.231	84.82	77.55	0.366	80.11	0.301	90.41	79.99	0.407	80.78

Table 13: Publishing component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	73.96	0.316	88.26	65.20	0.398	74.00	0.319	90.95	62.95	0.417	69.23	0.378	87.39	58.13	0.438	72.40
Qwen3.6-35B-A3B (2026-04)	74.21	0.294	89.10	62.92	0.372	73.77	0.298	**91.03**	60.12	0.386	66.35	0.404	88.75	50.65	0.446	71.45
Qwen3.6-27B (2026-04)	70.88	0.383	**91.03**	59.97	0.468	71.24	0.367	88.48	61.89	0.458	66.67	0.458	85.92	59.95	0.503	69.60
Qwen3.5-397B-A17B (2026-02)	80.42	0.164	85.51	72.11	0.275	73.41	0.212	85.51	55.91	0.320	74.68	0.247	86.84	61.91	0.360	76.17
Qwen3.5-122B-A10B (2026-02)	62.46	0.147	12.14	89.98	0.277	86.15	**0.152**	85.51	88.12	0.285	**83.54**	0.223	90.48	82.41	0.350	77.38
Qwen3.5-35B-A3B (2026-02)	80.29	0.145	85.51	69.90	0.269	80.10	0.179	85.51	72.67	0.307	77.08	0.228	86.84	67.24	0.352	79.16
Qwen3.5-27B (2026-02)	87.43	0.153	85.96	91.61	0.262	80.61	0.205	85.51	76.83	0.312	82.07	**0.201**	90.48	75.82	**0.320**	83.37
Qwen3.5-9B (2026-03)	80.85	0.178	82.84	77.50	0.300	82.71	0.191	85.76	81.45	0.350	76.92	0.270	89.68	68.12	0.378	80.16
Qwen3.5-4B (2026-03)	82.89	0.179	85.49	81.06	0.304	81.20	0.199	85.49	77.96	0.317	77.48	0.308	90.74	72.50	0.400	80.52
Qwen3.5-2B (2026-03)	74.81	0.257	84.97	65.12	0.411	74.23	0.269	85.96	63.64	0.411	70.46	0.367	89.19	58.84	0.462	73.17
Qwen3.5-0.8B (2026-03)	70.38	0.273	74.75	63.66	0.393	67.32	0.283	68.28	62.03	0.424	66.25	0.416	88.85	51.48	0.473	67.98
Qwen3-VL-8B (2025-10)	81.34	0.193	85.11	78.26	0.343	79.75	0.227	85.31	76.65	0.351	72.69	0.263	85.31	59.06	0.375	77.93
Qwen3-VL-4B (2025-10)	81.64	0.169	85.11	76.74	0.312	83.21	0.206	85.11	85.14	0.354	74.62	0.322	87.89	68.22	0.410	79.82
Qwen3-VL-2B (2025-10)	75.38	0.213	83.58	63.83	0.327	74.89	0.228	84.69	62.81	0.349	65.81	0.381	82.29	53.27	0.441	72.03
DeepSeek-V4-Flash-Vision-Exp (2026-08)	63.38	0.455	78.09	57.59	0.526	63.01	0.451	77.27	56.84	0.498	42.07	0.586	42.44	42.38	0.553	56.16
GLM-5.3-Flash (2026-08)	**89.20**	0.155	88.04	**95.05**	0.284	86.31	0.206	88.04	**91.47**	0.334	79.75	0.294	88.48	80.19	0.391	85.09
MiniCPM-V-4.6 (2026-05)	74.25	0.293	90.49	61.57	0.434	60.61	0.355	64.03	53.32	0.474	58.45	0.531	85.01	43.45	0.567	64.43
MiniCPM-V-4.5 (2025-08)	54.04	0.445	70.51	36.07	0.517	56.84	0.454	79.47	36.43	0.506	44.73	0.584	65.91	26.71	0.588	51.87
Gemini 3.6 Flash (2026-07)	86.68	0.206	88.21	92.46	0.328	**86.86**	0.196	90.59	89.59	0.314	82.54	0.267	**91.03**	83.31	0.357	**85.36**
Gemini-3.1-Pro (2026-02)	80.03	0.260	89.94	76.11	0.367	79.92	0.269	89.14	77.56	0.366	80.39	0.248	79.67	**86.25**	0.344	80.11
Kimi K2.6 (2026-04)	77.27	0.271	85.31	73.61	0.398	79.06	0.259	85.31	77.77	0.392	80.99	0.287	89.68	82.03	0.428	79.11
Step3-VL (2026-01)	55.19	0.505	77.20	38.85	0.489	57.72	0.506	83.75	40.03	0.516	49.68	0.572	78.06	28.18	0.529	54.20

Table 14: **Technical: three-track component results** (110 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	54.13	0.389	13.27	87.97	0.516	52.18	0.437	18.75	81.46	0.551	46.14	0.527	21.43	69.67	0.608	50.82
YouTu-Parsing (2026-01)	67.62	0.150	31.35	86.50	0.267	68.81	0.180	41.98	82.43	0.311	54.79	0.260	22.67	67.68	0.325	63.74
DotsMOCR (2026-03)	73.56	0.161	47.96	88.80	0.268	69.12	0.174	39.40	85.39	0.286	53.28	0.303	28.51	61.64	0.355	65.32
GLM-OCR (2026-02)	64.01	0.244	27.63	88.85	0.394	59.69	0.293	25.98	82.36	0.441	54.60	0.391	30.47	72.45	0.499	59.43
MinerU2.5-Pro (2026-04)	69.15	0.188	33.96	92.28	0.307	66.77	0.219	35.40	86.77	0.332	57.66	0.331	31.76	74.31	0.385	64.53
MinerU2.5 (2025-09)	70.13	0.128	32.71	90.49	0.260	64.00	0.178	29.38	80.41	0.281	53.66	0.309	26.97	64.91	0.356	62.60
MonkeyOCRv2-B-Parsing (2026-07)	61.09	0.316	25.81	89.10	0.366	58.11	0.323	25.69	80.91	0.382	50.09	0.443	25.81	68.70	0.432	56.43
MonkeyOCR-pro-3B (2025-08)	63.94	0.274	34.47	84.73	0.433	57.43	0.351	31.55	75.81	0.500	39.20	0.547	21.45	50.89	0.622	53.52
MonkeyOCR-pro-1.2B (2025-07)	62.43	0.274	31.28	83.38	0.434	55.62	0.351	29.88	72.06	0.494	43.17	0.557	34.94	50.31	0.589	53.74
OpenOCR (2024-11)	30.84	0.248	17.33	0.00	0.372	27.42	0.321	14.31	0.00	0.439	20.49	0.469	8.36	0.00	0.533	26.25
Dolphin-v2 (2025-12)	64.49	0.228	33.34	82.91	0.359	59.53	0.285	26.42	80.69	0.410	41.42	0.514	16.12	59.52	0.556	55.15
TeleOCR (2026-08)	**80.80**	**0.054**	**53.74**	94.03	**0.118**	**75.77**	0.137	**54.42**	86.54	**0.204**	59.94	0.280	32.52	75.30	0.314	**72.17**
End-to-end specialists																
OvisOCR2 (2026-07)	73.28	0.088	33.38	95.23	0.207	74.42	**0.100**	40.48	92.75	0.217	60.65	0.237	28.15	77.48	0.308	69.45
ABot-OCR (2026-05)	63.20	0.213	21.32	89.55	0.391	61.18	0.232	17.87	88.91	0.410	51.38	0.346	21.81	66.89	0.475	58.58
Qianfan-OCR (2026-03)	65.29	0.233	37.44	81.69	0.335	57.39	0.290	25.53	75.65	0.369	46.50	0.380	13.77	63.74	0.425	56.39
Unlimited-OCR (2026-06)	68.80	0.151	33.96	87.53	0.270	63.57	0.215	31.74	80.50	0.333	46.55	0.400	19.53	60.15	0.421	59.64
WeVisDoc-4B (2026-09)	68.55	0.202	31.82	94.00	0.315	67.26	0.206	31.70	90.74	0.311	61.88	0.306	33.81	82.40	0.367	65.90
WeVisDoc-2B (2026-09)	68.75	0.199	33.32	92.85	0.330	67.56	0.217	35.67	88.75	0.349	59.98	0.331	36.89	76.17	0.385	65.43
HunyuanOCR-1.5 (2026-07)	67.83	0.227	40.51	85.66	0.370	65.66	0.227	38.22	81.44	0.372	52.99	0.345	29.68	63.84	0.424	62.16
HunyuanOCR (2025-11)	62.04	0.273	31.43	82.04	0.352	60.54	0.267	31.65	76.69	0.355	47.58	0.410	29.76	53.98	0.425	56.72
dots.ocr (2025-07)	70.05	0.216	43.63	88.15	0.370	68.44	0.257	46.88	84.12	0.388	53.59	0.356	29.72	66.66	0.430	64.03
FireRed-OCR (2026-02)	70.46	0.163	38.23	89.50	0.275	68.53	0.185	35.71	88.33	0.304	57.50	0.290	28.79	72.76	0.353	65.50
DeepSeek-OCR-2 (2026-01)	62.53	0.217	33.64	75.60	0.369	55.66	0.280	28.39	66.59	0.407	39.35	0.544	15.31	57.16	0.469	52.51
DeepSeek-OCR (2025-10)	48.63	0.335	33.40	46.00	0.468	48.39	0.379	37.58	45.47	0.478	29.29	0.501	7.87	30.14	0.518	42.10
olmOCR-2-7B (2025-10)	58.72	0.257	11.18	90.65	0.276	59.19	0.268	15.77	88.58	0.285	45.71	0.400	11.23	65.87	0.374	54.54
olmOCR-7B (2025-08)	63.41	0.332	40.19	83.24	0.379	60.64	0.372	41.26	77.85	0.416	40.77	0.539	21.36	54.88	0.534	54.94
OCRVerse (2026-01)	72.09	0.168	41.71	91.31	0.288	71.59	0.191	42.34	91.54	0.323	60.52	0.320	39.61	73.96	0.403	68.07
FD-RL (2025-11)	73.31	0.146	41.53	92.98	0.283	73.11	0.164	44.02	91.76	0.301	64.14	0.253	42.97	74.77	0.336	70.19
Nanonets-OCR2 (2025-10)	66.64	0.159	27.40	88.45	0.253	64.42	0.189	26.07	86.08	0.281	47.47	0.373	17.02	62.74	0.373	59.51
OCRFlux-3B (2025-06)	45.53	0.419	25.33	53.19	0.317	38.15	0.450	16.82	42.62	0.302	30.31	0.537	10.86	33.74	0.377	38.00
Logics-Parsing-v2 (2026-02)	69.89	0.187	36.82	91.57	0.309	72.66	0.203	47.89	90.43	0.329	64.69	0.265	37.98	82.57	0.351	69.08

Table 14: Technical component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	67.65	0.223	35.71	89.56	0.329	68.90	0.219	39.62	88.93	0.321	58.92	0.298	27.48	79.03	0.354	65.16
Qwen3.6-35B-A3B (2026-04)	70.54	0.144	35.49	90.53	0.273	71.70	0.159	41.35	89.65	0.272	56.83	0.291	22.84	76.77	0.350	66.36
Qwen3.6-27B (2026-04)	67.26	0.214	37.31	85.81	0.336	65.18	0.206	28.93	87.21	0.320	56.44	0.310	22.52	77.76	0.391	62.96
Qwen3.5-397B-A17B (2026-02)	72.81	0.164	43.34	91.52	0.280	72.72	0.151	42.94	90.34	0.283	66.59	0.188	34.46	84.09	0.313	70.71
Qwen3.5-122B-A10B (2026-02)	71.78	0.134	36.89	91.82	0.259	74.46	0.138	44.62	92.61	0.267	**69.88**	0.205	41.81	**88.37**	0.308	72.04
Qwen3.5-35B-A3B (2026-02)	68.90	0.142	38.51	82.39	0.256	67.12	0.183	41.51	78.18	0.296	66.05	0.228	43.70	77.23	0.315	67.36
Qwen3.5-27B (2026-02)	69.93	0.147	33.65	90.84	0.274	70.08	0.153	35.70	89.88	0.279	68.16	0.217	39.92	86.22	0.321	69.39
Qwen3.5-9B (2026-03)	71.31	0.188	42.38	90.32	0.294	71.82	0.218	48.39	88.83	0.331	64.76	0.256	33.90	85.99	0.360	69.30
Qwen3.5-4B (2026-03)	75.67	0.148	51.51	90.31	0.266	73.84	0.169	46.89	91.56	0.295	63.45	0.283	35.09	83.56	0.356	70.99
Qwen3.5-2B (2026-03)	69.30	0.302	51.19	86.93	0.386	67.04	0.303	43.52	87.88	0.397	57.88	0.376	34.56	76.72	0.442	64.74
Qwen3.5-0.8B (2026-03)	67.84	0.301	46.82	86.78	0.412	64.83	0.305	41.18	83.82	0.439	47.67	0.473	31.96	58.33	0.512	60.11
Qwen3-VL-8B (2025-10)	67.98	0.281	44.26	87.82	0.381	67.86	0.266	42.83	87.38	0.375	60.90	0.302	37.71	75.16	0.386	65.58
Qwen3-VL-4B (2025-10)	70.56	0.234	45.72	89.37	0.374	71.19	0.246	50.46	87.69	0.384	53.50	0.350	21.94	73.59	0.435	65.08
Qwen3-VL-2B (2025-10)	66.30	0.251	39.94	84.12	0.383	66.94	0.267	46.11	81.43	0.402	56.20	0.387	39.11	68.19	0.460	63.15
DeepSeek-V4-Flash-Vision-Exp (2026-08)	63.82	0.291	36.05	84.48	0.431	59.32	0.355	30.31	83.20	0.456	45.13	0.450	13.76	66.60	0.495	56.09
GLM-5.3-Flash (2026-08)	70.38	0.162	30.24	**97.10**	0.332	70.29	0.161	30.81	**96.13**	0.316	64.24	0.271	32.53	87.32	0.379	68.30
MiniCPM-V-4.6 (2026-05)	62.08	0.277	29.99	83.92	0.346	57.52	0.330	27.58	78.01	0.384	48.62	0.447	27.59	62.97	0.443	56.07
MiniCPM-V-4.5 (2025-08)	60.54	0.234	25.83	79.24	0.309	56.31	0.238	13.36	79.31	0.329	40.73	0.433	8.61	56.93	0.408	52.53
Gemini 3.6 Flash (2026-07)	72.97	0.151	40.24	93.79	0.266	71.59	0.153	37.03	92.99	0.257	66.47	0.222	34.62	86.97	0.306	70.34
Gemini-3.1-Pro (2026-02)	64.96	0.204	33.02	82.25	0.290	65.43	0.194	31.85	83.82	0.279	65.82	**0.176**	30.07	85.02	**0.241**	65.40
Kimi K2.6 (2026-04)	69.99	0.224	37.70	94.68	0.398	69.87	0.251	44.22	90.44	0.421	67.29	0.306	**45.34**	87.08	0.466	69.05
Step3-VL (2026-01)	57.99	0.376	35.76	75.78	0.419	60.78	0.342	40.57	75.96	0.422	49.08	0.448	28.19	63.90	0.501	55.95

Table 15: **Logistics: three-track component results** (120 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	53.35	0.566	38.79	77.84	0.722	46.23	0.643	31.07	71.90	0.770	40.29	0.687	28.23	61.35	0.786	46.62
YouTu-Parsing (2026-01)	68.08	0.374	71.28	70.34	0.553	67.15	0.411	76.38	66.12	0.599	52.68	0.533	60.47	50.87	0.648	62.64
DotsMOCR (2026-03)	**78.20**	0.243	**80.96**	77.96	0.435	70.51	0.292	**77.24**	63.46	0.488	54.62	0.507	**70.94**	43.59	0.602	**67.78**
GLM-OCR (2026-02)	66.35	0.391	54.39	83.77	0.653	63.38	0.489	58.94	80.06	0.707	52.83	0.581	47.52	69.08	0.733	60.85
MinerU2.5-Pro (2026-04)	68.54	0.329	66.36	72.11	0.548	66.36	0.393	67.79	70.61	0.596	52.85	0.514	52.22	57.73	0.646	62.58
MinerU2.5 (2025-09)	69.39	0.307	70.12	68.74	0.567	59.39	0.348	54.38	58.58	0.596	50.11	0.530	53.59	49.77	0.676	59.63
MonkeyOCRv2-B-Parsing (2026-07)	48.55	0.479	39.10	54.48	0.594	49.95	0.533	51.96	51.15	0.611	40.90	0.635	41.01	45.22	0.655	46.47
MonkeyOCR-pro-3B (2025-08)	46.69	0.521	36.40	55.76	0.697	47.24	0.519	39.65	53.99	0.726	33.91	0.687	18.85	51.53	0.793	42.61
MonkeyOCR-pro-1.2B (2025-07)	48.54	0.553	48.82	52.13	0.711	46.09	0.557	40.98	52.98	0.727	30.75	0.723	15.20	49.39	0.791	41.79
OpenOCR (2024-11)	34.43	0.519	55.21	0.00	0.697	31.70	0.534	48.43	0.09	0.720	27.15	0.634	44.82	0.00	0.776	31.09
Dolphin-v2 (2025-12)	49.73	0.608	59.17	50.83	0.705	58.05	0.551	71.54	57.70	0.691	28.55	0.761	25.78	35.94	0.769	45.44
TeleOCR (2026-08)	74.79	0.232	69.18	78.39	**0.318**	69.65	0.326	72.80	68.77	0.391	50.25	0.495	39.58	60.65	0.556	64.90
End-to-end specialists																
OvisOCR2 (2026-07)	73.78	**0.184**	53.37	86.40	0.397	68.56	**0.242**	53.05	76.79	0.454	49.56	0.496	36.79	61.49	0.624	63.97
ABot-OCR (2026-05)	69.94	0.247	52.29	82.19	0.501	64.52	0.320	51.59	74.00	0.571	47.73	0.505	32.13	61.57	0.660	60.73
Qianfan-OCR (2026-03)	42.23	0.479	46.90	27.65	0.606	36.13	0.538	32.87	29.31	0.621	22.81	0.666	12.78	22.29	0.696	33.72
Unlimited-OCR (2026-06)	55.61	0.404	51.57	55.70	0.544	49.19	0.469	48.56	45.86	0.610	34.10	0.636	28.20	37.73	0.636	46.30
WeVisDoc-4B (2026-09)	67.27	0.292	54.67	76.34	0.507	64.57	0.338	52.76	74.72	0.536	51.01	0.495	43.85	58.67	0.615	60.95
WeVisDoc-2B (2026-09)	69.19	0.253	55.80	77.09	0.477	64.06	0.316	53.64	70.14	0.532	46.93	0.517	37.08	55.45	0.572	60.06
HunyuanOCR-1.5 (2026-07)	51.85	0.282	31.49	52.25	0.505	52.74	0.302	39.43	48.95	0.544	39.92	0.493	31.92	37.12	0.633	48.17
HunyuanOCR (2025-11)	53.61	0.314	57.45	34.76	0.501	48.78	0.386	45.92	38.98	0.542	32.01	0.555	26.71	24.81	0.588	44.80
dots.ocr (2025-07)	62.64	0.463	68.95	65.33	0.628	58.46	0.487	68.85	55.21	0.644	39.53	0.629	45.60	35.96	0.684	53.54
FireRed-OCR (2026-02)	69.01	0.366	72.17	71.51	0.567	66.79	0.390	73.73	65.60	0.589	52.28	0.565	68.42	44.93	0.649	62.69
DeepSeek-OCR-2 (2026-01)	44.82	0.503	53.13	31.67	0.662	40.89	0.582	51.13	29.69	0.718	21.88	0.764	25.07	17.01	0.772	35.86
DeepSeek-OCR (2025-10)	18.00	0.638	11.00	6.74	0.704	18.44	0.682	20.05	3.45	0.710	13.09	0.761	12.19	3.18	0.731	16.51
olmOCR-2-7B (2025-10)	56.49	0.322	26.05	75.63	0.521	49.18	0.374	27.21	57.76	0.530	37.76	0.544	20.16	47.51	0.614	47.81
olmOCR-7B (2025-08)	56.42	0.481	55.24	62.15	0.634	43.29	0.587	48.72	39.80	0.702	29.64	0.734	28.76	33.56	0.779	43.12
OCRVerse (2026-01)	64.37	0.388	55.89	76.04	0.571	63.04	0.426	55.91	75.82	0.609	51.39	0.526	53.42	53.34	0.663	59.60
FD-RL (2025-11)	72.34	0.285	63.33	82.16	0.514	70.21	0.287	60.78	78.54	0.513	57.55	0.465	61.11	58.07	0.603	66.70
Nanonets-OCR2 (2025-10)	46.72	0.422	24.55	57.76	0.606	42.44	0.493	27.02	49.54	0.630	31.94	0.627	15.30	43.20	0.657	40.37
OCRFlux-3B (2025-06)	43.00	0.485	34.86	42.64	0.332	40.91	0.531	37.83	37.98	**0.386**	27.37	0.653	28.72	18.71	**0.527**	37.09
Logics-Parsing-v2 (2026-02)	71.04	0.290	70.27	71.83	0.489	68.28	0.326	68.40	69.07	0.538	62.76	**0.439**	64.41	67.76	0.607	67.36

Table 15: Logistics component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	57.41	0.469	55.61	63.54	0.612	55.85	0.507	55.33	62.90	0.620	41.55	0.604	36.27	48.77	0.657	51.60
Qwen3.6-35B-A3B (2026-04)	52.18	0.473	46.18	57.68	0.590	57.78	0.472	61.13	59.40	0.599	35.54	0.658	28.53	43.88	0.684	48.50
Qwen3.6-27B (2026-04)	57.86	0.529	55.54	70.96	0.656	50.97	0.538	43.44	63.25	0.667	41.42	0.650	34.45	54.83	0.736	50.09
Qwen3.5-397B-A17B (2026-02)	56.66	0.368	41.34	65.43	0.574	66.68	0.375	59.35	78.22	0.576	54.98	0.472	46.43	65.68	0.616	59.44
Qwen3.5-122B-A10B (2026-02)	66.27	0.365	65.13	70.20	0.563	70.07	0.359	74.18	71.97	0.549	59.75	0.496	55.81	73.06	0.620	65.36
Qwen3.5-35B-A3B (2026-02)	61.78	0.422	54.04	73.52	0.608	56.22	0.359	32.68	71.88	0.557	47.24	0.500	30.16	61.54	0.626	55.08
Qwen3.5-27B (2026-02)	59.53	0.389	52.83	64.64	0.593	64.97	0.380	63.18	69.68	0.577	55.28	0.459	40.86	70.90	0.600	59.93
Qwen3.5-9B (2026-03)	65.37	0.411	64.33	72.86	0.594	69.25	0.350	75.19	67.59	0.550	57.27	0.509	57.18	65.56	0.620	63.96
Qwen3.5-4B (2026-03)	56.83	0.520	61.81	60.66	0.674	69.12	0.434	76.55	74.26	0.621	51.74	0.623	61.05	56.43	0.719	59.23
Qwen3.5-2B (2026-03)	47.78	0.598	63.23	39.96	0.703	57.58	0.491	46.86	74.96	0.654	41.76	0.663	40.63	50.95	0.723	49.04
Qwen3.5-0.8B (2026-03)	32.10	0.618	24.73	33.36	0.698	50.31	0.500	51.34	49.60	0.660	33.75	0.694	33.06	37.60	0.731	38.72
Qwen3-VL-8B (2025-10)	53.32	0.457	54.55	51.15	0.618	**71.26**	0.378	73.31	78.27	0.586	48.01	0.536	46.14	51.45	0.656	57.53
Qwen3-VL-4B (2025-10)	50.65	0.455	55.30	42.17	0.603	57.36	0.392	50.47	60.80	0.579	44.33	0.564	40.96	48.41	0.659	50.78
Qwen3-VL-2B (2025-10)	37.14	0.489	24.53	35.84	0.623	52.74	0.434	41.78	59.85	0.598	42.21	0.607	33.23	54.05	0.688	44.03
DeepSeek-V4-Flash-Vision-Exp (2026-08)	34.49	0.758	36.90	42.35	0.771	30.44	0.758	29.02	38.13	0.776	21.23	0.842	19.01	28.89	0.817	28.72
GLM-5.3-Flash (2026-08)	65.93	0.425	53.62	**86.64**	0.629	64.90	0.438	56.12	**82.40**	0.630	56.13	0.547	45.01	78.05	0.678	62.32
MiniCPM-V-4.6 (2026-05)	46.47	0.480	27.64	59.72	0.596	38.77	0.533	22.84	46.80	0.623	28.82	0.665	13.72	39.26	0.714	38.02
MiniCPM-V-4.5 (2025-08)	36.91	0.634	40.73	33.38	0.704	31.79	0.660	32.06	29.31	0.685	15.52	0.783	11.36	13.50	0.740	28.07
Gemini 3.6 Flash (2026-07)	64.36	0.379	45.23	85.76	0.571	65.77	0.403	55.89	81.78	0.604	64.73	0.470	58.23	**82.94**	0.610	64.96
Gemini-3.1-Pro (2026-02)	64.28	0.409	60.60	73.18	0.606	66.53	0.424	63.86	78.13	0.602	**65.10**	0.447	68.78	71.23	0.576	65.30
Kimi K2.6 (2026-04)	61.30	0.540	64.99	72.91	0.688	47.84	0.602	67.57	36.13	0.729	58.15	0.578	57.75	74.48	0.711	55.76
Step3-VL (2026-01)	45.60	0.680	58.46	46.30	0.667	41.72	0.696	59.85	34.88	0.699	29.65	0.747	39.46	24.19	0.734	38.99

Table 16: **Certificate: three-track component results** (104 pages; 53 models). The final Overall averages Clean, Digital, and Real. Text and Reading Edit measure errors (↓); Formula CDM, Table TEDS, and Overall measure accuracy (↑). Reading order is evaluated separately. Colored track scores indicate losses/gains relative to Clean. **Bold**/underline: best/runner-up.

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
Pipeline / multi-stage specialists																
PaddleOCR-VL-1.6 (2026-05)	48.48	0.547	30.02	70.17	0.696	44.71	0.562	29.99	60.40	0.711	33.32	0.668	30.02	36.76	0.781	42.17
YouTu-Parsing (2026-01)	53.76	0.429	68.12	36.02	0.565	44.11	0.408	29.87	43.28	0.570	29.44	0.617	33.65	16.34	0.670	42.44
DotsMOCR (2026-03)	59.79	0.212	38.17	62.35	0.405	51.21	0.271	38.17	42.53	0.439	36.53	0.466	38.17	18.05	0.554	49.18
GLM-OCR (2026-02)	71.68	0.390	79.96	74.05	0.586	64.21	0.417	**80.08**	54.31	0.622	47.36	0.553	60.16	37.19	0.701	61.08
MinerU2.5-Pro (2026-04)	70.64	0.320	**79.98**	64.00	0.500	66.90	0.325	79.98	53.20	0.519	51.78	0.526	**79.98**	28.00	0.653	63.11
MinerU2.5 (2025-09)	53.95	0.286	52.41	38.07	0.490	51.37	0.301	39.65	44.55	0.535	40.98	0.528	52.07	23.69	0.654	48.77
MonkeyOCRv2-B-Parsing (2026-07)	37.53	0.532	28.15	37.65	0.592	32.30	0.529	17.76	32.01	0.603	27.56	0.692	28.15	23.76	0.681	32.47
MonkeyOCR-pro-3B (2025-08)	23.93	0.602	5.60	26.44	0.738	40.39	0.477	17.61	51.27	0.647	26.11	0.654	11.09	32.60	0.754	30.14
MonkeyOCR-pro-1.2B (2025-07)	22.39	0.623	8.57	20.88	0.725	35.31	0.492	6.30	48.80	0.644	30.57	0.655	20.68	36.48	0.713	29.42
OpenOCR (2024-11)	15.97	0.573	5.24	0.00	0.673	16.33	0.593	8.32	0.00	0.683	12.80	0.700	8.37	0.00	0.771	15.03
Dolphin-v2 (2025-12)	47.96	0.605	75.70	28.63	0.654	23.74	0.561	2.09	25.26	0.624	33.75	0.729	53.61	20.58	0.722	35.15
TeleOCR (2026-08)	49.32	0.286	18.02	58.55	**0.313**	52.23	0.354	38.99	53.16	0.407	33.70	0.512	18.02	34.33	0.532	45.08
End-to-end specialists																
OvisOCR2 (2026-07)	63.91	**0.203**	27.42	**84.64**	0.378	59.49	**0.254**	30.02	73.86	0.425	40.60	0.508	27.42	45.20	0.614	54.67
ABot-OCR (2026-05)	63.52	0.240	41.15	73.43	0.465	63.67	0.260	49.15	67.83	0.489	39.29	0.480	41.93	23.90	0.626	55.49
Qianfan-OCR (2026-03)	24.57	0.497	6.04	17.34	0.594	25.87	0.526	16.22	13.94	0.603	17.62	0.680	6.69	14.19	0.659	22.69
Unlimited-OCR (2026-06)	37.85	0.411	20.56	34.11	0.513	33.27	0.442	5.20	38.79	0.548	26.70	0.646	20.56	24.10	0.631	32.60
WeVisDoc-4B (2026-09)	69.30	0.280	54.16	81.73	0.477	64.09	0.297	54.54	67.41	0.486	51.24	0.487	54.16	48.23	0.589	61.54
WeVisDoc-2B (2026-09)	56.77	0.256	30.40	65.50	0.453	55.79	0.261	30.40	63.06	0.464	37.06	0.491	30.40	29.93	0.562	49.87
HunyuanOCR-1.5 (2026-07)	49.82	0.267	21.21	54.91	0.491	47.96	0.285	20.27	52.12	0.499	34.86	0.488	21.21	32.20	0.609	44.21
HunyuanOCR (2025-11)	47.35	0.321	35.78	38.40	0.487	40.22	0.386	25.90	33.35	0.495	31.48	0.573	35.78	15.93	0.589	39.68
dots.ocr (2025-07)	51.21	0.414	37.01	57.98	0.551	45.81	0.460	37.01	46.38	0.578	31.36	0.589	37.01	15.91	0.638	42.79
FireRed-OCR (2026-02)	62.99	0.380	65.02	61.95	0.493	61.44	0.421	62.03	64.42	0.536	40.50	0.577	67.95	11.26	0.648	54.98
DeepSeek-OCR-2 (2026-01)	20.54	0.563	1.85	16.09	0.636	18.25	0.582	0.73	12.25	0.660	12.74	0.776	7.64	8.21	0.759	17.18
DeepSeek-OCR (2025-10)	10.68	0.696	0.00	1.64	0.655	10.52	0.719	1.72	1.78	0.696	6.90	0.801	0.00	0.85	0.710	9.37
olmOCR-2-7B (2025-10)	70.21	0.299	76.05	64.48	0.467	70.66	0.331	76.98	68.09	0.484	44.92	0.547	76.05	13.45	0.591	61.93
olmOCR-7B (2025-08)	34.43	0.476	0.13	50.77	0.586	28.57	0.512	1.06	35.88	0.628	13.53	0.664	0.13	6.83	0.727	25.51
OCRVerse (2026-01)	71.81	0.321	75.80	71.75	0.510	69.95	0.369	72.55	74.18	0.539	49.64	0.512	77.51	22.56	0.637	63.80
FD-RL (2025-11)	**76.83**	0.238	77.51	76.81	0.450	**75.29**	0.256	77.51	73.97	0.464	54.52	0.452	77.51	31.22	0.569	**68.88**
Nanonets-OCR2 (2025-10)	67.97	0.350	75.73	63.13	0.518	62.14	0.379	74.35	49.91	0.528	43.76	0.595	73.83	16.99	0.625	57.96
OCRFlux-3B (2025-06)	29.86	0.474	3.79	33.23	0.324	31.91	0.463	6.65	35.36	**0.338**	20.63	0.636	4.91	20.54	**0.518**	27.47
Logics-Parsing-v2 (2026-02)	53.26	0.282	40.68	47.31	0.464	54.11	0.292	37.21	54.30	0.471	45.03	**0.443**	35.77	43.60	0.600	50.80

Table 16: Certificate component results (continued)

Model (release)	Clean					Digital					Real					Overall↑
	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	Overall ↑	Text Edit↓	Formula CDM↑	Table TEDS↑	Reading Edit↓	(Avg_3)
General-purpose VLMs																
Qwen3.8-Flash-Next (2026-08)	35.66	0.459	4.06	48.87	0.563	41.62	0.503	30.02	45.12	0.586	23.22	0.630	4.73	27.98	0.627	33.50
Qwen3.6-35B-A3B (2026-04)	34.55	0.473	7.06	43.92	0.549	37.12	0.506	21.05	40.90	0.574	27.25	0.615	16.99	26.29	0.642	32.97
Qwen3.6-27B (2026-04)	39.42	0.546	27.27	45.56	0.636	31.84	0.555	3.69	47.30	0.642	27.24	0.655	9.51	37.69	0.699	32.83
Qwen3.5-397B-A17B (2026-02)	49.56	0.409	36.62	52.91	0.554	52.15	0.335	18.10	71.83	0.515	34.58	0.451	10.13	38.69	0.574	45.43
Qwen3.5-122B-A10B (2026-02)	61.21	0.430	78.41	48.22	0.575	60.46	0.322	37.62	76.00	0.500	43.94	0.472	26.50	52.49	0.586	55.20
Qwen3.5-35B-A3B (2026-02)	57.97	0.463	71.84	48.32	0.602	67.51	0.373	69.11	70.76	0.532	44.39	0.491	41.43	40.88	0.618	56.62
Qwen3.5-27B (2026-02)	35.09	0.423	4.52	43.02	0.571	37.97	0.338	0.00	47.72	0.521	46.13	0.480	41.62	44.73	0.587	39.73
Qwen3.5-9B (2026-03)	40.45	0.444	16.01	49.70	0.575	52.92	0.347	17.12	76.37	0.518	34.81	0.501	13.09	41.41	0.598	42.73
Qwen3.5-4B (2026-03)	46.43	0.500	36.16	53.11	0.619	71.17	0.405	75.31	78.70	0.562	45.26	0.575	46.68	46.65	0.673	54.29
Qwen3.5-2B (2026-03)	43.36	0.555	39.33	46.23	0.649	52.46	0.439	36.20	65.03	0.590	34.05	0.592	21.72	39.66	0.658	43.29
Qwen3.5-0.8B (2026-03)	30.59	0.548	12.09	34.49	0.625	48.37	0.417	35.40	51.39	0.543	32.63	0.614	21.72	37.56	0.608	37.20
Qwen3-VL-8B (2025-10)	48.10	0.422	33.34	53.12	0.565	59.68	0.350	37.93	76.09	0.509	37.53	0.518	37.62	26.75	0.599	48.44
Qwen3-VL-4B (2025-10)	41.17	0.442	26.77	40.98	0.559	58.34	0.342	37.25	71.95	0.503	36.43	0.506	22.90	36.99	0.613	45.31
Qwen3-VL-2B (2025-10)	29.51	0.461	7.59	27.03	0.563	62.93	0.382	70.37	56.58	0.528	38.54	0.576	40.97	32.24	0.631	43.66
DeepSeek-V4-Flash-Vision-Exp (2026-08)	22.55	0.755	9.95	33.22	0.727	22.49	0.780	11.11	34.34	0.749	13.70	0.851	4.15	22.04	0.810	19.58
GLM-5.3-Flash (2026-08)	57.64	0.376	38.59	71.96	0.535	59.90	0.387	43.03	75.40	0.549	43.26	0.548	38.46	46.13	0.653	53.60
MiniCPM-V-4.6 (2026-05)	36.20	0.434	0.51	51.46	0.560	32.12	0.457	0.56	41.45	0.579	22.04	0.632	0.51	28.78	0.681	30.12
MiniCPM-V-4.5 (2025-08)	17.35	0.683	0.00	20.40	0.714	17.12	0.674	0.00	18.78	0.730	10.31	0.796	0.00	10.57	0.693	14.93
Gemini 3.6 Flash (2026-07)	47.98	0.347	5.41	73.20	0.489	51.37	0.360	6.96	**83.11**	0.506	38.66	0.466	6.69	**55.88**	0.554	46.00
Gemini-3.1-Pro (2026-02)	63.41	0.353	55.15	70.36	0.501	53.73	0.368	28.06	69.93	0.498	**54.54**	0.446	77.22	31.00	0.559	57.23
Kimi K2.6 (2026-04)	54.80	0.535	75.68	42.25	0.652	20.29	0.574	0.00	18.23	0.683	46.46	0.587	49.48	48.58	0.656	40.52
Step3-VL (2026-01)	17.89	0.651	0.00	18.79	0.636	17.91	0.705	0.00	24.26	0.668	11.03	0.763	0.00	9.36	0.717	15.61