DR.DOCBENCH: A Comprehensive Benchmark for Expert-Level and Difficult Document Parsing

Anonymous ACL submission

Abstract

Document parsing and recognition are fundamental capabilities for vision-language models (VLMs) and document processing systems. However, existing Optical Character Recognition (OCR) and document parsing benchmarks are increasingly limited in coverage and difficulty: many focus on common document genres or uniformly sampled pages where modern parsers already perform strongly, while offering limited annotation for expert-domain structures such as chemical formula, music notation, complex tables, and cross-page layouts. We introduce **DR.DOCBENCH**, a difficulty-aware benchmark for expert-level document parsing. Built from a large-scale multilingual book corpus, **DR.DOCBENCH** spans 52 BISAC subject domains and selects challenging documents through parser-failure-based sampling, targeting cases where multiple state-of-the-art systems struggle. It contains 4,514 annotated pages from long documents averaging around 100 pages, with 70k high-quality page- and block-level annotations for layout, reading order, hierarchical relations, and domain-specific visual contents. Evaluations of pipeline-based parsers and general-purpose VLMs show that strong performance on existing benchmarks does not transfer to our expert-level document parsing. Our analysis reveals substantial failures across subjects, content types, and structural attributes, highlighting **DR.DOCBENCH** as a comprehensive testbed for diagnosing and advancing document intelligence.

1 Introduction

Document parsing is a prerequisite for the daily usage of vision language models in the real world. Robust document understanding requires recovering both content and structure from visually complex pages, enabling downstream tasks such as visual document retrieval (Faysse et al., 2025; Choi et al., 2025; Zhang and Wu, 2026), question answering (Cho et al., 2024; Gong et al., 2025; Dong et al., 2025), and scientific reasoning (Luo et al., 2026; Chen et al., 2026).

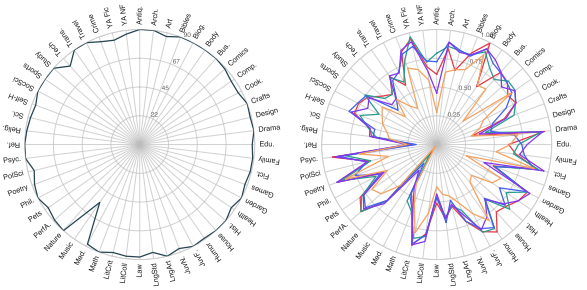


Figure 1: **Overview of DR.DOCBENCH across 52 BISAC subject domains.** *Left*: annotated pages per subject with (4,514 pages total). *Right*: per-subject overall score for five frontier general VLMs. **Claude**, **GPT-5.5**, **Gemini**, **Kimi**, **GPT-4o**, across the 51 subjects with overall-metric coverage.

While recent benchmarks (Pfitzmann et al., 2022; Ouyang et al., 2025; Fu et al., 2025) have driven significant progress in evaluating document parsing for VLMs, they primarily focus on common genres. In familiar domains, models have already achieved ceiling performance (Duan et al., 2026; Wei et al., 2026). Consequently, current evaluations obscure important failures in the long tail of real-world documents. Furthermore, existing benchmarks often assess document understanding at a coarse, page-level granularity, lacking the detailed annotations required to penalize mistakes on complex, localized components (e.g., dense formulas, chemical structures, rotated tables, or domain-specific layouts). This creates an evaluation gap: although VLMs appear nearly saturated on standard tasks, their true capacity to parse expert-level, complex documents remains unmeasured.

Therefore, we introduce **DR.DOCBENCH** (Figure 2), a subject-diverse, difficulty-aware, expert-level benchmark designed to close this gap. First, **DR.DOCBENCH** broadens data coverage to 52 BISAC subject domains with 4,514 annotated

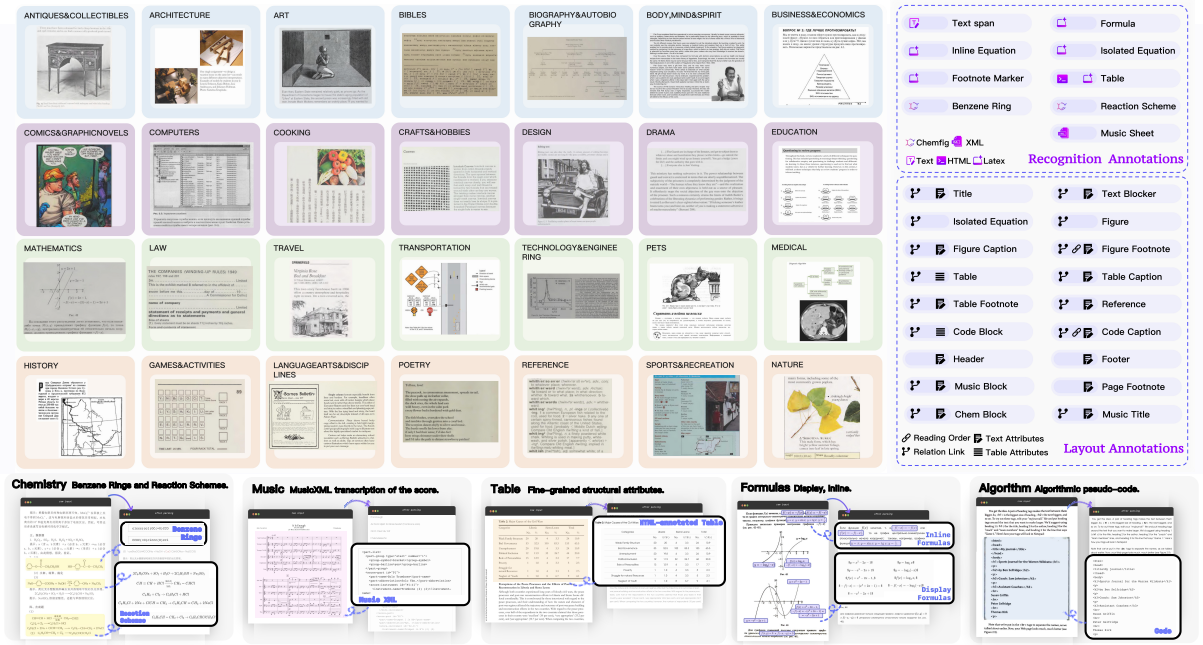


Figure 2: Overview of **DR.DOCBENCH**. The benchmark spans diverse BISAC subject domains and provides fine-grained annotations for layout, recognition, and expert-domain structures, including chemistry diagrams, music notation, complex tables, formulas, and pseudo-code.

pages, spanning both common knowledge areas and long-tail expert domains. Second, it is constructed through difficulty-based data sourcing: candidate documents are first filtered using four state-of-the-art parsers, and challenging pages with high cross-parser disagreement are filtered for real human annotation. Third, we introduce expert-level annotations for domain-specific visual content. For example, chemistry pages may contain benzene rings and reaction schemes; music documents may contain visual scores that must be transcribed into structured music representations; and technical documents may contain dense formulas and complex tables. These contents cannot be captured by generic OCR labels alone, so we design annotation standards that translate expert visual languages into unified machine-understandable forms. Finally, the benchmark consists of ultra-long documents with an average length of ∼100 pages, enabling long-document parsing evaluation, such as reading order and cross-page continuity, that conventional single-page OCR benchmarks rarely address.

We conduct a systematic evaluation of both general-purpose and document-specialized VLMs on **DR.DOCBENCH**. We provide fine-grained analyses across subjects, content types, and structural attributes. Our results reveal that model strengths are highly uneven: some models perform relatively well on literature-style documents but struggle on technical and symbolic domains, while subjects such as REFERENCE and MUSIC remain difficult for nearly all systems. We further identify component-level failure modes, including color-dependent tables, dense formulas, and other visually complex structures. These findings show that strong performance on existing document benchmarks does not necessarily imply robust expert-level document parsing under our difficulty-aware setting. Overall, **DR.DOCBENCH** provides a comprehensive and fine-grained testbed for diagnosing the capabilities and limitations of current models and guiding future progress in document intelligence.[1]

2 Related Work

Document parsing benchmarks. Several recent benchmarks have probed different facets of document recognition. OmniDocBench (Ouyang et al., 2025) assembles pages from nine genres (academic papers, textbooks, exam sheets, slides, magazines, research reports, newspapers, handwritten notes, and colorful textbooks) to evaluate parsers across heterogeneous layouts. OCRBench (Liu et al., 2024), together with its v2 successor (Fu et al., 2025), expands the evaluation surface beyond plain

[1]Data and code will be released upon acceptance.

text recognition to include table parsing, formula recognition, and reading-order prediction. Layout-centric efforts such as DocLayNet (Pfitzmann et al., 2022) and M^{6}Doc (Cheng et al., 2023) instead concentrate on region detection. MusiXQA (Chen et al., 2025) evaluates MLLMs on music sheets via VQA over MusiXTeX (Taupin et al., 1993) targets; we instead target end-to-end MusicXML (Good, 2001; Hewlett and Selfridge-Field, 2001) transcription. Our benchmark departs from prior work in two respects: rather than labeling pages with coarse genre buckets, we organize the corpus by *document subject* using the BISAC heading system (Book Industry Study Group, 2025), a 52-category taxonomy maintained for the North American book trade. After collapsing redundant codes, 52 BISAC top-level categories remain in our corpus, and we find that it is precisely the long tail of this distribution where current parsers degrade most sharply.

Document Parsing Systems. Document parsing systems have evolved from modular OCR pipelines to end-to-end VLM-based parsers. Traditional OCR systems (Smith, 2007; Du et al., 2020) decompose parsing into a multi-step pipeline of text detection, geometric correction, and text recognition. While efficient for standard OCR tasks, this approach is less flexible for interpreting complex document structures. Recent VLM-based specialized parsers (Cui et al., 2026; Niu et al., 2025; Wang et al., 2026) are proposed to handle richer document structures, and therefore are strong on targeted parsing tasks such as layout analysis, table extraction, and formula recognition, but are usually optimized around predefined parsing pipelines or output formats. Concurrently, general-purpose VLMs such as the GPT family models (OpenAI, 2024, 2026) and Qwen family models (Bai et al., 2023; Qwen Team, 2026) are trained on large-scale multimodal corpora and further aligned for instruction following, where a variety of OCR-style data are often included in their training data mixtures. This not only enables instruction-guided, flexible OCR but also further unlocks advanced OCR-based multimodal reasoning, although their behavior can remain less predictable than specialized parsers.

3 DR.DOCBENCH

We design **DR.DOCBENCH** to evaluate document parsing beyond common, increasingly saturated document genres and content types. **DR.DOCBENCH** targets documents whose information is encoded through expert visual conventions, dense structure, and long-range layouts. We combine broad subject coverage, difficulty-aware sampling, and expert-level annotation to carefully curate such a benchmark.

3.1 Difficulty-Aware Sampling

The source images of **DR.DOCBENCH** are from a large-scale book corpus, and further organised into using the BISAC taxonomy, spanning 52 subject domains across scientific, technical, artistic, educational, historical, literary, and reference materials.

Uniformly sampled pages are insufficient, since many common layouts are already handled well by current sota parsing systems. To focus annotation on genuinely difficult cases, **DR.DOCBENCH** uses a parser-failure-based sampling pipeline.

For each candidate book PDF, we run multiple strong document parsing systems independently, including PaddleOCR-VL-1.5 (Cui et al., 2026), MinerU-VLM (Wang et al., 2026), and Gemini-3.1-pro-preview (Gemini Team, 2026), to get inter-parser disagreement score using normalized edit distance between parser outputs. Documents with low disagreement are discarded as these systems have converged on so less challenging, while high-disagreement documents are retained within each subject domain for next-step expert annotation. In practice, high-disagreement samples are often noticed to contain dense formulas, unusual layouts, multilingual content, complex tables, music notation, chemical diagrams, or visually crowded reference pages.

This filtering strategy makes **DR.DOCBENCH** difficulty-aware: the benchmark concentrates on documents where existing systems are unreliable, while maintaining balanced subject coverage. The parsers used for sampling are used only for data selection and are distinct from the broader evaluation suite in Sec. 4, so the filter neither advantages nor disadvantages any benchmarked models. Final evaluation is performed against human-verified ground truth.

3.2 Annotation Schema

DR.DOCBENCH provides *page-* and *block-level annotations* that support both end-to-end evaluation and fine-grained error analysis. Each page is represented as a structured document object containing block regions, block types, transcribed content, reading order, and structural relations.

The schema covers common document elements such as paragraphs, titles, figures, tables, formulas, code blocks, algorithms, etc. It also captures relations that are critical for long-document parsing, including caption-to-object links, parent–child relations, and continuation links across columns or pages. These annotations allow evaluation to go beyond plain OCR accuracy and measure whether a system preserves document structure.

A central contribution of **DR.DOCBENCH** is its annotation of expert-domain content that prior benchmarks often ignore or collapse into generic categories. *Chemical structures*, including benzene rings, fused-ring systems, and reaction schemes, are explicitly annotated rather than treated as ordinary figures. Particularly, *Music sheets* are preserved using MusicXML (Good, 2001; Hewlett and Selfridge-Field, 2001), while titles and credits are annotated as regular text. *Mathematical* formulas are transcribed into normalized LaTeX with a fixed delimiter convention, ensuring that edit distance and structure-aware metrics remain comparable. For content-specific structures, tables are represented with HTML-style structure with cell spans, borders, and fine-grained layout attributes. Algorithms and pseudo-code are distinguished from regular code or text, and cross-block relations are used to connect each component that continues across columns or pages.

3.3 Doctor Annotation and Quality Control

Because many pages require domain knowledge (e.g., translate visual chemical formulas into text representations), **DR.DOCBENCH** uses a mixed annotation workflow. General document elements are annotated by general trained annotators, while specialized content is routed to domain experts with MS/PhD level education. For example, chemistry-majored annotators review molecular structures and reaction schemes, music-literate-majored annotators transcribe music scores, and LaTeX-fluent annotators handle dense mathematical content.

The annotation process begins with parser-generated sparse pre-labels, which are corrected and completed by human annotators. Annotators adjust bounding boxes, assign block categories, transcribe content, and add reading-order and structural relations. Expert-content blocks receive additional review by domain specialists.

We apply multiple quality-control checks before release. Most pages is double-annotated to identify ambiguous cases, and disagreements are resolved through adjudication. This process ensures that **DR.DOCBENCH** provides reliable ground truth not only for text recognition, but also for structural and domain-specific document understanding.

3.4 Final Dataset

The final **DR.DOCBENCH** release contains **312 PDFs** and **4,514 annotated pages**, balanced at six PDFs per BISAC subject across **52 domains** and spanning **14 written languages** (predominantly English at ∼68%). In total it carries over 70k fine-grained annotations: ∼65k block-level labels across 21 block categories. We use publicly available books and scanned documents solely for benchmark construction and research evaluation under fair-use principles, with attention to applicable access and licensing constraints. Together with the expert-content tracks of Sec. 3.2, these properties make **DR.DOCBENCH** a challenging benchmark for diagnosing the limitations of current VLMs and document parsing systems on the structures that prior benchmarks under-represent.

4 Experiments

4.1 Models

We categorize document content extraction models into two main classes:

Specialized VLMs. These are large multimodal models specifically trained for document parsing tasks. We experiment with PaddleOCR-VL-1.5 (Cui et al., 2026) and MinerU (Niu et al., 2025; Wang et al., 2026)[2] through their official APIs, and they do not accept user prompts.

General VLMs. This category comprises general-purpose large multimodal models inherently capable of document parsing. Our evaluation covers both open-source and proprietary models. The open-source selection includes Kimi-K2.5 (Kimi Team et al., 2026), Nemotron-nano-12b-v2-VL (NVIDIA, 2025), and three from the Qwen 3.5 family (Qwen Team, 2026): Qwen-3.5-flash (a boosted 35B-A3B), Qwen-3.5-122B-A10B, and Qwen-3.5-plus (a boosted 397B-A17B). Additionally, we evaluate proprietary closed-source models: Gemini 3.1 pro (Gemini Team, 2026), GPT-4o (OpenAI, 2024), GPT-5.5 (OpenAI, 2026), Claude Opus 4.6 (Anthropic, 2026), and Doubao-Seed 1.6-

[2] https://mineru.net/apiManage/docs

Model	Size†	Access	Text	Formula		Table		Order	Overall↑
			Edit↓	Edit↓	CDM↑	TEDS↑	$TEDS_S$↑	Edit↓	
Specialized VLMs									
MinerU 2.5	1.2B	Open	0.33	0.51	24.15	**55.85**	**63.70**	0.30	54.37
PaddleOCR	0.9B	Open	0.73	0.42	30.73	51.79	59.94	0.71	34.78
General VLMs									
Qwen3.5-Flash	3B	Open	0.26	0.32	35.69	46.24	54.38	0.26	57.51
Qwen3.5-122B-A10B	10B	Open	0.23	0.32	32.54	38.31	44.99	0.23	56.32
Qwen3.5-Plus	17B	Open	0.25	**0.31**	30.40	49.77	58.23	0.26	57.32
Nemotron-Nano-12B	12B	Open	0.62	0.76	12.82	27.03	34.09	0.564	30.33
Kimi-K2.5	32B	Open	0.19	0.35	27.04	51.98	61.09	0.18	60.38
Claude-Opus 4.6	-	Closed	0.22	0.37	32.02	49.21	58.12	0.19	60.19
Doubao-Seed-1.6-Vision	-	Closed	0.34	0.31	**41.28**	33.93	42.16	0.28	53.14
Gemini-3.1 Pro	-	Closed	0.22	0.35	32.22	51.26	59.03	0.21	60.13
GPT-4o	-	Closed	0.37	0.47	36.02	30.14	38.62	0.31	49.73
GPT-5.5	-	Closed	**0.19**	0.36	34.55	48.90	58.96	**0.17**	**61.94**

Table 1: Overall Evaluation Results. †: activated Parameters. **bold**: **best**, underlined: worst.

vision (Bytedance Seed, 2025). Our inference prompt is in Appendix G (Figure 9,10).

4.2 Evaluation Metrics

Following Ouyang et al. (2025), we employ different evaluation metrics tailored to different document elements:

Pure Text. We calculate the Normalized Levenshtein Edit Distance (Levenshtein, 1966).

Tables. We convert all tables to HTML format and report both Tree-Edit-Distance-based Similarity (TEDS) and TEDS_S (Zhong et al., 2020).

Formulas. We evaluate the formula using Character Detection Matching (CDM; Wang et al. (2025)) and Normalized Levenshtein Edit Distance.

We report an *Overall Score* (the higher the better) by mapping each component metric to a 0–100 accuracy scale: text and reading-order scores are $(1 - \text{edit distance}) \times 100$, formula scores are $\text{CDM} \times 100$, and table scores are $\text{TEDS} \times 100$. The overall score is the mean of the available component scores for each sample, with missing elements excluded from the denominator.

4.3 Overall Results and Findings

Table 1 shows that no frontier VLM dominates expert-level document parsing. GPT-5.5, Kimi-K2.5, Claude Opus 4.6, and Gemini 3.1 Pro form a tight leading group, but their strengths vary by component: GPT-5.5 performs best on text extraction and reading order, while Kimi-K2.5 is more competitive on tables. Nemotron-Nano-12B, despite being a large pretrained VLM, is a clear low-performing outlier. Specialized parsers such as MinerU and PaddleOCR remain valuable for structured elements despite their smaller scale: MinerU achieves the best table scores, outperforming all general VLMs on both TEDS and $TEDS_S$. However, because these systems are typically user-prompt-incapable, they perform worse on prompt-dependent metrics such as reading order and instruction-following text extraction.

Model	Best 3		Worst 3	
	Score	Subject Names	Score	Subject Names
Paddleocr	56.94	YA Fic, Soc Sci, Hist	13.79	Drama, Ref, Fam
Nemotron	59.15	YA Fic, Edu, Biog	14.41	Antiq, Crafts, Ref
Claude	96.13	Biog, YA Fic, Drama	35.07	Ref, Des, Med
Doubao	91.01	YA Fic, Biog, Fic	25.15	Ref, Games, Antiq
Gemini	90.84	Biog, Arch, Lit Crit	33.84	Ref, Study, Law
GPT-4o	86.83	Biog, Fic, Poet	27.16	Ref, Games, Antiq
GPT-5.5	93.08	YA Fic, Drama, Fic	38.24	Ref, Des, Law
Kimi	94.26	Fic, Biog, Drama	36.00	Ref, Des, Med
MinerU	91.77	Drama, Fic, Biog	23.13	Comics, Ref, Law
Qwen3.5-Flash	87.37	Biog, YA Fic, Psych	31.84	Ref, Law, Games
Qwen3.5-122B	92.35	YA Fic, Humor, Fic	29.58	Study, Des, Games
Qwen3.5-Plus	87.94	Biog, YA Fic, Arch	29.96	Ref, Des, Law

Table 2: Model Best 3 and Worst 3 Subjects. Subject name abbreviations are listed in Appendix F.

4.4 Per-Subject Breakdown and Analysis

Table 2 shows the best 3 subjects and the worst 3 subjects of VLMs on **DR.DOCBENCH**. Nearly all models are strongest on narrative or text-dense subjects such as BIOGRAPHY & AUTOBIOGRAPHY, FICTION, and YOUNG ADULT FICTION, suggesting that current VLMs and parsers handle linear prose relatively well.

In contrast, REFERENCE is the most consistently difficult subject, appearing in the worst group for almost every model. DESIGN, GAMES, MEDICAL, and ANTIQUES & COLLECTIBLES recur as hard

Model	Type	All	Acad. Lit.	Book	Color. TB	Exam	Mag.	News.	Note	Res. Rep.	PPT
MinerU	Spec.	0.325	0.227	0.321	0.347	0.276	0.392	0.524	0.417	0.417	0.245
PaddleOCR		0.729	0.773	0.726	0.685	0.744	0.707	0.817	0.744	0.543	0.790
Qwen3.5-Flash	Gen.	0.258	0.209	0.261	0.184	0.335	0.216	0.471	0.278	0.802	0.163
Qwen3.5-122B-A10B		0.226	0.183	0.228	0.163	0.277	0.192	0.370	0.275	0.904	0.225
Qwen3.5-Plus		0.249	0.223	0.251	0.162	0.342	0.196	0.266	0.248	**0.383**†	0.317
Nemotron		0.622	0.577	0.625	0.546	0.601	0.621	0.656	0.597	1.000	0.394
Kimi		0.193	**0.174**	**0.190**	0.147	0.291	0.177	0.385	0.230	0.567	0.162
Claude		0.217	0.220	0.222	**0.129**†	0.255	**0.152**	**0.193**	0.223	0.411	**0.101**
Doubao		0.344	0.332	0.346	0.258	0.420	0.325	0.576	0.280	0.913	0.475
Gemini		0.219	0.214	0.216	0.160	0.332	0.183	0.352	0.315	0.567	0.223
GPT-4o		0.366	0.311	0.372	0.249	0.308	0.415	0.564	0.289	1.000	0.503
GPT-5.5		**0.189**	0.180	0.191	**0.129**†	**0.199**	0.165	0.268	**0.200**	**0.383**†	0.152

Table 3: Text extraction edit distance by document data source (↓ lower is better). **Bold**: best per column; † denotes a tie. *Color. TB*: colorful textbook; *Acad. Lit.*: academic literature; *Res. Rep.*: research report; *PPT*: PPT2PDF.

cases. These categories combine dense layouts, domain-specific notation, tables or lists that span multiple pages, or a visually irregular structure, indicating that expert-level document parsing remains limited less by ordinary text recognition than by structural and subject-specific understanding.

We provide the details of each models' worst five subjects in Appendix H.2.

4.5 Per-data-source Breakdown

We show the performance on different document source types in Table 3. Research reports (Res. Rep.) are the most challenging document type for text extraction across nearly all models. Doubao records the worst edit distance, while other frontier models, including GPT-4o (1.000), Gemini (0.567), and Kimi (0.567), also struggle on this document type. Only GPT-5.5 and Qwen3.5-plus maintain edit distances below 0.4. In the meantime, presentation slides (PPT2PDF) show the largest inter-model performance variation among frontier VLMs: Claude achieves 0.101 while Doubao scores 0.475, corresponding to a 4.7× gap.

4.6 Per-block Breakdown

The block-level breakdown in Table 4 shows that many errors come from layout and visual-structure ambiguity rather than plain text recognition alone. For reading order, single-column pages are consistently the easiest, while multi-column layouts substantially increase error for nearly all models. Similarly, rotated text is harder than normally oriented text, with 90-degree rotation often especially challenging, suggesting strong orientation biases in current models.

For tables, the clearest stressor is visual complexity and the presence of borders. Models also find tables with colored backgrounds harder to transcribe, as we observe a large and consistent drop in model performance, suggesting that background shading interferes with cell localization and content-structure alignment. Border style is more nuanced, implying that models rely on different cues for table reconstruction. In general, non-standard table layouts, especially list-style or vertical layouts, are more consistently challenging because they break the regular row-column assumptions used by most parsers.

5 Additional Analysis

5.1 Sensitivity to Multi-page Context Length

As each document spans ∼100 pages, we study how parsing and recognition performance change as the sliding window grows. We experiment with a sliding window from 1 to 15 pages. Figure 3 shows that larger windows do not reliably improve long-document recognition. Reading order consistently degrades for nearly all models (e.g., edit distance: Claude 0.24 → 0.69; GPT-5.5: 0.16→ 0.47 at $n = 15$), since the model must order many more blocks while separating page-local contents from cross-page continuations. Text extraction and table parsing also worsen. Formula metrics are comparatively stable, suggesting that formulas are more spatially local (details in Appendix H.5).

5.2 Token Efficiency

Figure 4 compares each model's average output-token cost against its overall parsing performance. Among the leading models, GPT-5.5 achieves the highest overall score with only a moderate output budget. Kimi K2.5 is more efficient: it recovers approximately 98% of GPT-5.5's overall performance

		Text						Table			
		RO Dist ↓		*Text Edit Dist* ↓				*TEDS (%)* ↑		*TEDS (%)* ↑	
Model	Type	Single Col.	≥2 Col.	Rot. Norm.	Rot. 90°	Rot. 270°	Rot. Horiz.	Full Line	Wireless	BG Yes	BG No
MinerU	Spec.	0.237	0.357	0.354	0.539	0.563	0.546	**54.0**	**78.8**	32.3	**67.9**
PaddleOCR		0.687	0.813	0.941	0.971	0.966	0.854	43.9	75.9	29.0	62.4
Qwen3.5-Flash	Gen.	0.197	0.357	0.408	0.518	0.576	0.478	39.8	55.1	26.3	49.1
Qwen3.5-122B-A10B		0.175	0.339	0.319	0.496	0.481	0.432	32.0	15.2	22.3	27.8
Qwen3.5-Plus		0.222	0.329	0.416	0.502	0.527	0.370	49.1	77.0	29.5	63.7
Nemotron		0.508	0.645	0.708	0.880	0.750	0.964	9.6	10.3	9.0	15.7
Kimi		**0.133**	0.263	0.259	0.672	0.432	0.357	49.1	21.4	33.1	40.1
Claude		0.158	0.256	0.257	0.269	0.293	0.413	43.0	46.9	26.4	49.3
Doubao		0.213	0.384	0.477	0.761	0.757	0.545	29.0	8.1	16.7	22.6
Gemini		0.177	0.269	0.319	0.446	**0.277**	0.341	41.6	77.1	29.2	60.8
GPT-4o		0.233	0.420	0.479	0.814	0.758	0.763	17.1	11.2	12.8	18.7
GPT-5.5		0.143	**0.226**	**0.215**	**0.240**	0.419	**0.318**	51.5	74.9	**35.4**	62.3

Table 4: Per block type breakdown. For **Text**, we report *Reading-Order (RO) edit distance* (↓) by layout (Col.: columns) and *text edit distance* (↓) by block rotation (Norm.: normal, Horiz.: horizontal). For **Table**, we report TEDS (%, ↑) by border type and by table background color. **Bold**: best per column; † denotes a tie.

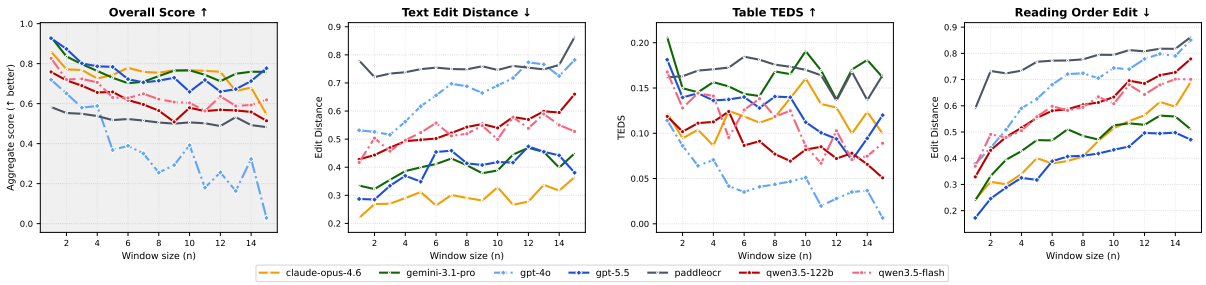


Figure 3: Impact of sliding window size (n pages) on overall and per-metric performance across seven models. The top-left panel shows a normalized aggregate score; the remaining panels show individual metrics. Reading order degrades most consistently with increasing window size, while formula and table metrics are comparatively stable.

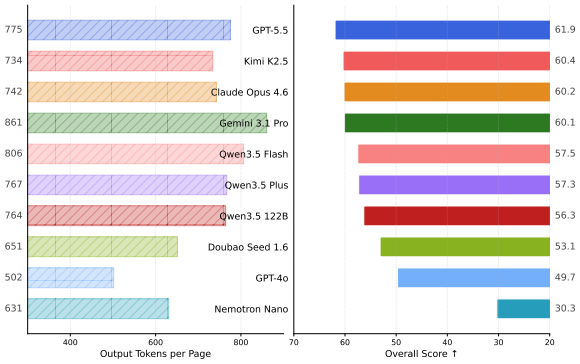


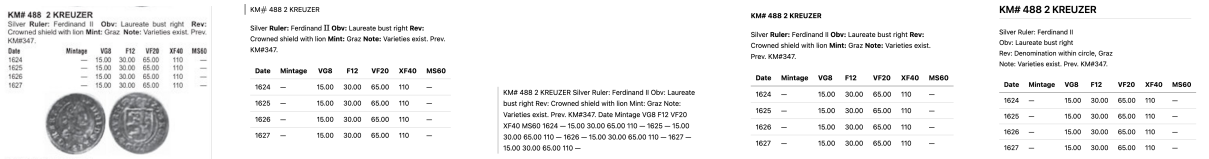
Figure 4: Token efficiency vs overall score.

while using the lowest token budget among the top-performing group. In contrast, Gemini 3.1 Pro appears more token-hungry, paying a roughly 17% token-cost more over Kimi K2.5 while yielding even lower accuracy. A closer inspection suggests that this additional generation often comes from redundant text, over-expanded structures, or less compact formatting rather than more faithful parsing. Nemotron Nano illustrates the opposite signal: despite its high output-token cost, it obtains the lowest overall score, indicating that longer outputs may reflect noisy extraction or poorly organized structure instead of better document understanding. Overall, these results suggests that the most effective parsers are those that produce compact, well-structured representations.

5.3 Case study: Table

In Table 4, we notice an interesting performance drop for wireless (borderless) in table recognition. From full line to wireless, Kimi drops from 49.1 → 21.4 TEDS; Doubao collapses to 8.1, which is near-random, revealing that most models rely heavily on visible grid lines for structure reconstruction. In this case study, we show an example of the borderless table as well as the rendered model output.

Figure 5 shows the source table (5a), the ground-truth HTML rendering (5b), and three model outputs. Doubao (5e) collapses under page-positioning: in the previous window (where the table falls on the second page), it attends almost entirely to prior-page content, recovering 5 spurious tables while ignoring the 21 on the second page. In



(a) Source image (b) Ground-truth (c) Kimi (curr. window) (d) Kimi (prev. window) (e) Doubao (curr. window)

Figure 5: Table case study. (5a) source image, (5b) ground-truth HTML rendering. (5c) Kimi in the current window collapses to plain text (5d). Kimi in the previous window produces correct content but ignores the HTML format instruction (5e). Doubao attends almost entirely to prior-page content and misses the target table.

the current window (where the table is on the first page), Doubao correctly gets all 4 rows, confirming that the failure is positional, not an intrinsic weakness with borderless tables. Kimi (5c, 5d) ignores the HTML format instruction: in the previous window, it produces a correct markdown rendering, but in the current window, it collapses to plain text.

5.4 Case study: Optical Music Recognition

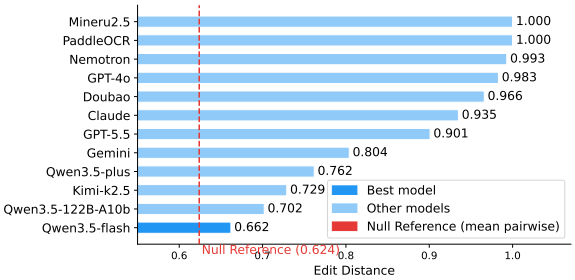


Figure 6: Music score → MusicXML transcription results (1-page window). Lower edit distance is better. The baseline is the mean pairwise edit distance across all 6 ground-truth documents.

The subject of Optical Music Recognition (OMR) is a brand-new subject not present in prior work (Ouyang et al., 2025; Liu et al., 2024; Fu et al., 2025). We are the first to evaluate this subject on a variety of frontier VLMs and specialized VLMs in the task of transcribing it to MusicXML. It exposes a failure mode that general document benchmarks do not: simultaneous visual comprehension of musical notation (clefs, noteheads, key/time signatures, articulations) and structured, schema-faithful XML generation. To anchor difficulty, we compute a cross-document null reference, the mean pairwise edit distance across all GT MusicXML files, yielding 0.624: the score expected from emitting an arbitrary unrelated score. As shown in Figure 6, no model beats this null, so as a schema-faithful transcription task, music remains unsolved for all systems evaluated. Pipeline OCR systems (MinerU-2.5, PaddleOCR) score 1.0 by design, since they cannot follow user instructions. For additional music results, see Appendix H.1.1.

6 Conclusion

Dr.DocBench is a difficulty-aware benchmark for expert-level document parsing, built from long-form multilingual books across 52 BISAC subject domains. Unlike prior OCR and document parsing benchmarks that mainly emphasize common genres or page-level recognition, **Dr.DocBench** targets challenging documents selected through parser-failure-based sampling and provides fine-grained annotations for layout, reading order, structural relations, and expert-domain content such as chemical diagrams, music notation, formulas, algorithms, and complex tables. Experiments on both specialized document parsers and frontier general-purpose VLMs show that strong performance on existing benchmarks does not reliably transfer to expert-level document parsing, with persistent failures on structurally complex subjects such as reference, design, medical, and music documents.

Ethical considerations

Dr.DocBench is intended for research evaluation of document parsing systems. During dataset construction and quality control, we removed pages containing clearly sensitive personal information, private identifying information, or inappropriate content when identified, and annotators were instructed to flag such cases for review. The released annotations are intended for benchmarking and analysis, while the underlying source documents may remain subject to their original access, copyright, and licensing conditions. Users should follow the applicable terms of the original sources when using the benchmark.

Limitations

While **DR.DOCBENCH** provides broad subject coverage and fine-grained annotations for difficult document parsing, it is designed primarily as a diagnostic benchmark for parsing and structural recognition rather than the full range of document-intelligence tasks. We use a unified inference prompt to ensure consistent comparison across models, but do not perform model-specific prompt engineering or extensive prompt optimization; the reported results therefore reflect standardized benchmark performance rather than the best achievable performance for each model. Certain specialized parsers, such as MinerU and PaddleOCR, do not favor user prompts or strictly follow format-specific instructions, so their scores on prompt-dependent outputs should be interpreted with caution and are reported mainly as reference baselines. The benchmark focuses on challenging pages selected from long-form book documents, which makes it especially useful for stress-testing current parsers on expert-level content; broader coverage of document sources, application scenarios, and downstream tasks remains an important direction.

Although BISAC subjects provide a useful organizing taxonomy, document difficulty also depends on fine-grained factors such as layout density, scan quality, visual artifacts, multilingual content, cross-page continuation, and domain-specific structures. We therefore view subject-level analysis as complementary to block-level and attribute-level analysis. In addition, expert-domain contents such as chemistry diagrams, music notation, formulas, algorithms, and complex tables are represented through structured formats such as normalized LaTeX, HTML, and MusicXML. These formats enable scalable automatic evaluation, but they do not exhaustively cover all domain-specific knowledge representations or specialized package/library-dependent formats. Semantic, execution-based, or domain-specific metrics could further improve the evaluation of equivalent representations.

References

Anthropic. 2026. Claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Book Industry Study Group. 2025. BISAC subject headings list, 2024 edition. https://www.bisg.org/complete-bisac-subject-headings-list.

Bytedance Seed. 2025. Doubao seed 1.6 vision. https://seed.bytedance.com/en/seed1_6.

Jian Chen, Wenye Ma, Penghang Liu, Wei Wang, Tengwei Song, Ming Li, Chenguang Wang, Jiayu Qin, Ruiyi Zhang, and Changyou Chen. 2025. Musixqa: Advancing visual music understanding in multimodal large language models. *Preprint*, arXiv:2506.23009.

Ziyu Chen, Yilun Zhao, Chengye Wang, Rilyn Han, Manasi Patwardhan, and Arman Cohan. 2026. Scimdr: Advancing scientific multimodal document reasoning. *Preprint*, arXiv:2603.12249.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. 2023. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15138–15147.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *Preprint*, arXiv:2411.04952.

Yejin Choi, Jaewoo Park, Janghan Yoon, Saejin Kim, Jaehyun Jeon, and Youngjae Yu. 2025. Zero-shot multimodal document retrieval via cross-modal question generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

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. 2026. Paddleocr-vl-1.5: Towards a multi-task 0.9b vlm for robust in-the-wild document parsing. *Preprint*, arXiv:2601.21957.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. Benchmarking retrieval-augmented multimodal generation for document question answering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2020. Pp-ocr: A practical ultra lightweight ocr system. *Preprint*, arXiv:2009.09941.

Shuaiqi Duan, Yadong Xue, Weihan Wang, Zhe Su, Huan Liu, Sheng Yang, Guobing Gan, Guo Wang, Zihan Wang, Shengdong Yan, Dexin Jin, Yuxuan Zhang, Guohong Wen, Yanfeng Wang, Yutao Zhang, Xiaohan Zhang, Wenyi Hong, Yukuo Cen, Da Yin, and 4 others. 2026. Glm-ocr technical report. *Preprint*, arXiv:2603.10910.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

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, and 5 others. 2025. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Gemini Team. 2026. Gemini 3.1 pro. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Ziyu Gong, Chengcheng Mai, and Yihua Huang. 2025. Mhier-rag: Multi-modal rag for visual-rich document question-answering via hierarchical and multi-granularity reasoning. *Preprint*, arXiv:2508.00579.

Michael Good. 2001. Musicxml for notation and analysis. *The virtual score: representation, retrieval, restoration*, 12(113-124):160.

Walter B. Hewlett and Eleanor Selfridge-Field. 2001. The virtual score : representation, retrieval, restoration.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 307 others. 2026. Kimi k2.5: Visual agentic intelligence. *arXiv Preprint*.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Yun Luo, Futing Wang, Qianjia Cheng, Fangchen Yu, Haodi Lei, Jianhao Yan, Chenxi Li, Jiacheng Chen, Yufeng Zhao, Haiyuan Wan, and 1 others. 2026. P1-vl: bridging visual perception and scientific reasoning in physics olympiads. *arXiv preprint arXiv:2602.09443*.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

NVIDIA. 2025. Nvidia nemotron nano v2 vl. *arXiv Preprint*.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

OpenAI. 2026. Introducing gpt-5.5. https://openai.com/index/introducing-gpt-5-5/.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. *Preprint*, arXiv:2412.07626.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter W J Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout analysis.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Daniel Taupin, Ross Mitchell, and Andreas Egler. 1993. Musixtex. using tex to write polyphonic or instrumental music. *EuroTEX*, 92:257–272.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, and 1 others. 2026. Mineru2. 5-pro: Pushing the limits of data-centric document parsing at scale. *arXiv preprint arXiv:2604.04771*.

Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Botian Shi, Bo Zhang, and Conghui He. 2025. Image over text: Transforming formula recognition evaluation with character detection matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19681–19690.

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

Yongyue Zhang and Yaxiong Wu. 2026. Mldocrag: Multimodal long-context document retrieval augmented generation. *Preprint*, arXiv:2602.10271.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer.

Content of Appendix

A Data Sourcing and Pipeline
B Multi-Parser Disagreement Score
C Annotation Schema Reference
D Full Per-Domain Statistics
E Language Coverage
F Subject Name Abbreviations
G Inference Prompt
H Additional Results

A Data Sourcing and Pipeline

The seed corpus is sourced from a publicly accessible digital library aggregating trade and academic books with peer-reviewed research journals across a wide range of professional and consumer domains. Every document is assigned to the BISAC 2024 subject taxonomy and *human-verified* by a trained reviewer; where multiple BISAC codes apply, the primary code is kept.

After classification we apply difficulty-aware sampling (Appendix B) and take the top-k books per BISAC subject, allocating exactly 6 PDFs per subject so that no single subject can dominate aggregate scores.

License. The **DR.DOCBENCH** annotations are released under the Creative Commons CC0 1.0 Universal license (public domain dedication).

Annotators and Compensation. The released **DR.DOCBENCH** labels were produced by a team of annotators recruited under the mixed workflow described in Sec. 3.3. General document elements (text blocks, titles, simple tables, reading order, and structural relations) were handled by trained annotators holding at least an undergraduate degree, while expert-content blocks were routed to domain specialists with MS- or PhD-level training in the relevant field: chemistry-majored annotators for molecular structures and reaction schemes, music-literate annotators for MusicXML transcription of scores, and LaTeX-fluent mathematicians for dense-formula pages. Annotators were compensated per annotated block rather than at a fixed hourly or per-page rate, with the per-unit price tiered by block complexity so that domain-expert blocks (chemistry, music, dense formulas, complex tables) paid materially more per unit than plain text or simple structural elements. All compensation was at or above the prevailing local market rate for skilled annotation work.

B Multi-Parser Disagreement Score

To bias annotation toward pages that current parsers cannot already solve, we score each candidate book by the disagreement of three frontier parsers—PaddleOCR-VL-1.5, MinerU-VLM, and Gemini-3.1-pro-preview—on the Markdown transcription of all pages. Let $T_{m,b}$ denote the concatenation of model-m Markdown outputs over book b. The text-disagreement term is the mean normalized Levenshtein distance over the six parser pairs,

$$D_b^{\text{text}} = \frac{1}{6}\sum_{(i,j)} \frac{\text{Lev}(T_{m_i,b}, T_{m_j,b})}{\max(|T_{m_i,b}|, |T_{m_j,b}|)}. \quad (1)$$

A composite score $D_b = \alpha D_b^{\text{text}} + \gamma D_b^{\text{struct}} + \delta D_b^{\text{conf}}$ optionally adds a structural-tag-mismatch term D_b^{struct} (headings, list items, table cells) and a low-confidence penalty D_b^{conf} for empty/malformed outputs. In our release we use $\alpha = 0.6$, $\gamma = 0.3$, $\delta = 0.1$ and admit books with $D_b > \tau = 0.5$ into the annotation queue.

C Annotation Schema Reference

C.1 OmniDocJSON Page Record

```
{
  "page_info": {
    "page_no": 7,
    "height": 1500, "width": 2000,
    "image_path": "<file>.jpg",
    "page_attribute": {
      "subject": "<BISAC code>",
      "challenge_type": "perception" |
        "structural_reconstruction" |
        "domain_reasoning",
      "data_source": "academic_literature",
      "language": "english",
      "layout": "double_column",
      "special_issue": ["table_full_line"]
    }
  },
  "layout_dets": [ ... ],
  "extra": { "relation": [ ... ] }
}
```

C.2 Page-Level Attributes

subject. One of 52 BISAC top-level domains.

challenge_type. Perception, structural reconstruction, or domain reasoning.

data_source. One of nine genres: academic literature, book, colorful textbook, exam paper, note, magazine, research report, newspaper, and PPT-to-PDF.

language. English, simplified Chinese, en–ch mixed, or other.

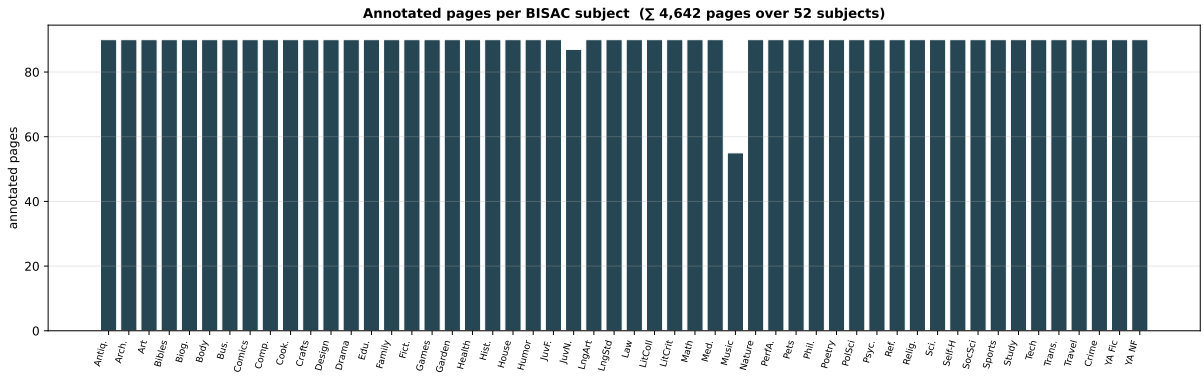


Figure 7: Number of annotated pages per BISAC subject in the full **DR.DOCBENCH** release. Each subject contains 6 source PDFs, for a total of 4,514 annotated pages. The distribution reflects the natural variation in document length after applying our difficulty-aware sampling process.

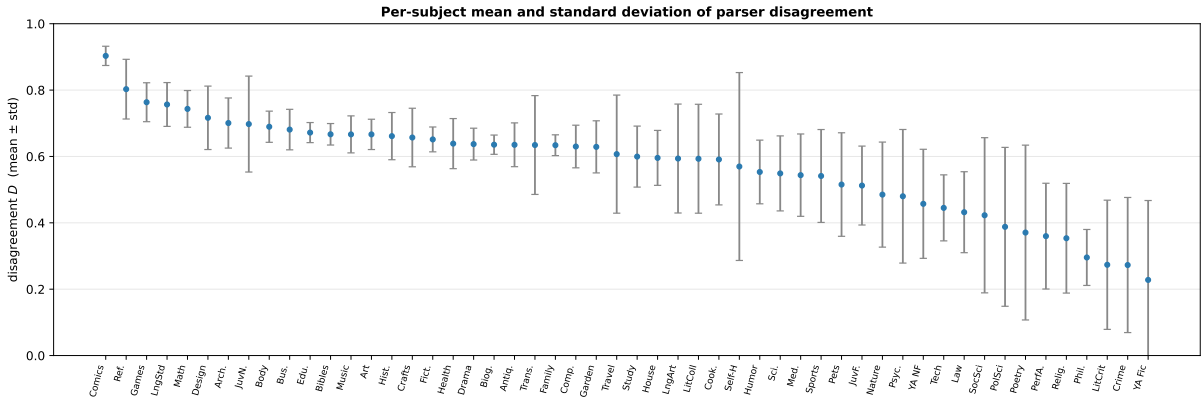


Figure 8: Mean and standard deviation of four-parser disagreement for each BISAC subject. The mean captures the overall parsing difficulty of a subject, while the standard deviation reflects within-subject heterogeneity across sampled documents. Subjects with both high mean and high variance, such as DESIGN and REFERENCE, serve as particularly discriminative stress tests because they contain diverse layouts and failure modes.

layout. Single-, double-, three-, or 1-and-more-column, plus an "other" bucket.

***special_issue* (multi-select).** e.g. colorful background, horizontal table, table span, table with formula, fuzzy scan.

C.3 Block-Level Categories

Quadrilateral block annotations fall into five groups:

Textual blocks. Title, text block, reference, and the caption / footnote / explanation variants attached to figures, tables, equations, and code (including *code_txt* and *code_algorithm* families).

Page-element blocks (no reading order). Header, footer, page number, and page footnote.

Tables. LaTeX-annotated, described by *table_layout*, *line*, *language*, *with_span*, *include_background*, *include_equation*, *include_photo*, and *with_structured_text*.

Formulas. Isolated equations with *formula_type* $\in$ {print, handwritten}; chemical structures are flagged via *need_mask*.

Reading-order-only blocks. Figure, table mask, text mask.

No-property blocks. Ignore-formula/abandon.

C.4 Inter-Block Relations

Each record in *extra.relation* has the form (*source_anno_id*, *target_anno_id*, *relation_type*). Two relation types are defined:

parent–son. Links a figure, table, equation, or code parent to its caption or footnote child.

truncated. Marks two blocks that form one logical unit; for cross-page truncation, *target_anno_id* is set to -1.

A *merge_list* attribute on a parent text block stores the original child blocks (e.g. TOC entries), while the parent's *text* field carries the flattened content.

D Full Per-Domain Statistics

To provide a complete view of the benchmark composition and sampling difficulty, we report per-domain statistics for all 52 BISAC subjects. These statistics complement the aggregate analysis in the main paper by showing how annotated pages and parser-disagreement scores vary across subjects.

Together, these statistics show that **DR.DOCBENCH** is not only broad in subject coverage, but also heterogeneous in parsing difficulty. Some domains are consistently challenging across parsers, while others exhibit large within-domain variation, suggesting that expert-document parsing requires robustness to both domain-specific visual conventions and document-level layout diversity. The complete 52-row per-subject table, including PDF counts, annotated page counts, and mean, minimum, maximum, and standard deviation of parser disagreement.

E Language Coverage

Language	Pages	Share
English	3,083	68.30%
Other	518	11.48%
Russian	334	7.40%
Simplified Chinese	152	3.37%
German	102	2.26%
Japanese	74	1.64%
French	72	1.60%
Turkish	46	1.02%
Spanish	45	1.00%
Traditional Chinese	41	0.91%
English–Chinese mixed	20	0.44%
Portuguese	15	0.33%
Egyptian Arabic	9	0.20%
Arabic	3	0.07%
Total	**4,514**	**100%**

Table 5: Per-page primary-language distribution across **DR.DOCBENCH**. *Other* aggregates pages whose primary language was tagged as "other" by the annotator.

Every page in **DR.DOCBENCH**carries a per-page `language` attribute. Tab. 5 reports the page counts across all fourteen distinct languages present in the release. English accounts for roughly two-thirds of pages, but the remaining $\sim$32% spans a long tail of major world languages (Russian, Simplified and Traditional Chinese, German, Japanese, French, Turkish, Spanish, Portuguese, and Arabic) which makes the benchmark multilingual.

F Subject Name Abbreviations

To save horizontal space, BISAC subject domains are abbreviated in the main-body tables and figures. The mapping below lists every abbreviation used together with its full BISAC name: **Antiq** (Antiques & Collectibles); **Arch** (Architecture); **Biog** (Biography & Autobiography); **Comics** (Comics & Graphic Novels); **Crafts** (Crafts & Hobbies); **Des** (Design); **Edu** (Education); **Fam** (Family & Relationships); **Fic** (Fiction); **Games** (Games & Activities); **Hist** (History); **Lit Crit** (Literary Criticism); **Med** (Medical); **Poet** (Poetry); **Psych** (Psychology); **Ref** (Reference); **Soc Sci** (Social Science); **Study** (Study Aids); **YA Fic** (Young Adult Fiction).

G Inference Prompt

Unified Inference Prompt (1/2)

```
You are an AI assistant specialized in converting document images to Markdown format. You are
↪  given {num_images} consecutive page image(s). Convert all pages in order into a single
↪  continuous Markdown document. Please follow these instructions for the conversion:

1. Text Processing:
   - Accurately recognize all text content in the image without guessing or inferring.
   - Transcribe text in its original language and script — do not translate. This includes but
   ↪  is not limited to Chinese, English, Japanese, Russian, Arabic, and other languages.
   - For degraded, blurry, or noisy scans: transcribe what is legible; do not invent or guess
   ↪  characters that are not clearly readable.
   - Maintain the original document structure, including headings, paragraphs, lists, etc.
   - For multi-column layouts, output columns in natural reading order (left-to-right,
   ↪  top-to-bottom).

2. Mathematical Formula Processing:
   - Convert all mathematical formulas to LaTeX format.
   - Enclose inline formulas with \( \). For example: This is an inline formula \( E = mc^2 \)
   - Enclose block formulas with \[ \]. For example: \[ \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]

3. Chemical Notation:
   - For chemical reaction equations and ionic equations, use the \ce{} syntax from the mhchem
   ↪  package.
     Example inline: \( \ce{H2O} \)
     Example block: \[ \ce{2H2 + O2 -> 2H2O} \]
   - For 2D molecular structures given as SMILES strings, preserve the SMILES notation exactly
   ↪  inside a smiles block:
     ```smiles
     CCO
     ```
   - For chemical symbols, element names, subscripts, and superscripts that appear in running
   ↪ text (e.g. HO, CO), render them using \ce{} or standard LaTeX subscript/superscript
   ↪ notation.
   - Preserve bond notation (single, double, triple), stereochemical indicators, reaction
   ↪ arrows, catalyst labels, and condition markers exactly as they appear.

4. Domain-Specific Notation:
   - Legal documents: preserve clause and sub-clause numbering exactly (e.g. Article
   ↪ 3.2(a)(i)).
   - Biomedical text: preserve gene names, protein names, dosage units, and clinical notation
   ↪ without alteration.
   - Financial documents: preserve ticker symbols, currency symbols, and numerical formats
   ↪ (e.g. $1,234.56, €, ¥).
   - Code and pseudocode: wrap in fenced code blocks with the appropriate language tag.
     Example:
     ```python
     def foo():
         pass
     ```
   - Diagram labels, callout text, and annotation text that is legible within a figure:
   ↪ transcribe as a plain text paragraph immediately after the figure's position; do not
   ↪ describe the diagram itself.

5. Table Processing:
   - Convert tables to HTML format.
   - Wrap the entire table with <table> and </table>.
   - For long tables that span multiple pages, output all rows in a single continuous HTML
   ↪ table.
   - For tables with merged cells, use rowspan and colspan attributes.

6. Comics and Dialog Balloons:
   - Extract all readable text in visual reading order: speech bubbles, thought bubbles,
   ↪ caption boxes, narration boxes, and any text visible within panels.
   - Output each piece of extracted text as a separate paragraph.
   - Do not add speaker attributions, panel labels, or separators. Do not output placeholder
   ↪ text for panels with no text.

7. Forms:
   - Render form fields and their labels as key-value pairs using a definition list or plain
   ↪ text, preserving the original field order.
   - For filled-in values, transcribe the handwritten or typed content exactly.
   - For checkboxes or radio buttons, indicate the selection state: [x] for checked, [ ] for
   ↪ unchecked.
```

Figure 9: Unified inference prompt used in our evaluation (Part 1).

Unified Inference Prompt (2/2)

```
8. Figure Handling:
   - Ignore purely visual figures (photographs, illustrations, charts) where no text is
   ↪  embedded. Do not describe or caption them.
   - Exception: extract any embedded text labels, axis labels, legend text, or annotations that
   ↪  are legible (see rule 4 above).

9. Music:
   - If the image contains a music score (sheet music with staves, notes, clefs, time
   ↪  signatures, etc.), render it in MusicXML format inside a fenced code block:
     ```musicxml
     <part-list>
       <score-part id="P1">
         <part-name>...</part-name>
         <part-abbreviation>...</part-abbreviation>
       </score-part>
     </part-list>
     <part id="P1">
       <measure number="1">
         ...
       </measure>
     </part>
     ```
   - Do not include an XML declaration (`<?xml ... ?>`) or `<score-partwise>` wrapper.
   - Include `<part-list>` with `<part-name>` and `<part-abbreviation>` for any instrument
   ↪ names or abbreviations visible in the score margin.
   - Transcribe all visible elements: clef, key signature, time signature, notes, rests, ties,
   ↪ slurs, dynamics, tempo markings, articulations, lyrics, stave count, rehearsal marks,
   ↪ repeat signs, barline types, ottava markings (8va/8vb), fermatas, trills, and other
   ↪ ornament symbols.
   - Key signatures must use the `<fifths>` integer encoding: positive = sharps (e.g.
   ↪ `<fifths>1</fifths>` = G major), negative = flats, zero = C major.
   - For keyboard instruments or any score with multiple staves per part, use `<staves>` and
   ↪ add a `<staff>` number attribute to each note to indicate which staff it belongs to.
   - Omit all non-visual metadata that cannot be read from the image: `<print>`,
   ↪ `<system-layout>`, `<system-margins>`, `<staff-layout>`, pixel/point measurements,
   ↪ `<divisions>`, `<midi-device>`, `<midi-instrument>`, and any rendering coordinates.
   - For text-only music content (song lyrics without notation, chord charts, tablature in
   ↪ plain text), transcribe as regular Markdown text, preserving chord symbols above lyrics
   ↪ lines.

10. Output Format:
   - Ensure the output Markdown document has a clear structure with appropriate line breaks
   ↪ between elements.
   - For complex layouts, maintain the original document's reading order as closely as possible.
   - Do not wrap the entire output in a code block. Do not add explanations, comments, or
   ↪ meta-commentary.
```

Figure 10: Unified inference prompt used in our evaluation (Part 2).

The full inference prompt is in Figure 9 and Figure 10.

H Additional Results

The top four general VLMs (GPT-5.5, Kimi, Claude, Gemini) cluster within 1.8 overall points, with no single model dominant across all metrics; GPT-5.5 leads on text and reading order, while Kimi leads on tables.

GPT-5.5 vs. GPT-4o. GPT-5.5 (61.94) outperforms GPT-4o (49.73) by 12.2 points overall, with the largest gains on text_Edit (0.189 vs. 0.366) and formula_Edit (0.359 vs. 0.468).

Formula and table metrics. For Kimi, CDM and edit distance diverge: it records the lowest formula_CDM (27.04) among frontier models while remaining competitive on formula_Edit (0.350), indicating that the two metrics capture different aspects of formula quality. On tables, MinerU 2.5 (TEDS 55.85, TEDS_S 63.70) outperforms all general VLMs despite not supporting user prompts.

Prompt-incapable systems and outliers. Since MinerU 2.5 and PaddleOCR do not support user prompts, their reading order and text scores are not interpretable as capability measures. PaddleOCR's formula_Edit (0.415) is further inflated by a delimiter mismatch (`$$...$$` vs. ground truth `\[...\]`). Nemotron-Nano-12B (30.33) falls 19 points below the next weakest general VLM (GPT-4o), making it a clear outlier.

H.1 Per-subject breakdown

Table 6 presents the per-subject text edit distance across 52 subject categories. We provide a heatmap of each subject's dominant components and data types in Figure 11. We separately evaluate Music in Section H.1.1.

Subject-level wins. Kimi achieves the lowest edit distance on 17 of the 51 non-Music subjects, with particular strength on text-heavy genres such as FICTION (0.023), HEALTH & FITNESS (0.039), and PSYCHOLOGY (0.163). Claude and GPT-5.5 rank second and third with 13 and 11 subject wins, respectively. GPT-5.5 marks a substantial improvement over GPT-4o: GPT-4o does not lead on any subject and records edit distances 1.5–3× higher on several categories (e.g., COMICS & GRAPHIC NOVELS: 0.811 vs. 0.295; TRANSPORTATION: 0.686 vs. 0.383).

Model specializations. GPT-5.5 leads on subjects with substantial equation content, including MATHEMATICS (0.302), the highest-equation subject (23.9% of annotated elements), as well as MEDICAL (0.162) and TECHNOLOGY & ENGINEERING (0.151); Qwen-3.5-122B leads on Science (0.203), the second-highest in equation density (17.5%). By contrast, all 13 of Claude's subject wins occur in almost equation-free content (mean equation share: 0.08%, maximum: 1.0% in Study Aids), including PERFORMING ARTS (0.049), DRAMA (0.103), TRANSPORTATION (0.122), LITERARY CRITICISM (0.182), and BIOGRAPHY & AUTOBIOGRAPHY (0.001). This correlation suggests Claude is well-calibrated for prose reconstruction but less competitive when mathematical notation is present.

Subject difficulty. The lowest-error subjects are text-dense narrative categories: BIOGRAPHY & AUTOBIOGRAPHY (Claude: 0.001), BUSINESS & ECONOMICS (Kimi: 0.022), and FICTION (Kimi: 0.023).

H.1.1 Music Analysis

Table 7 presents the edit distance results for the music score to MusicXML transcription task. This task is not a component of the previous OmniDocBench benchmark, and is uniquely demanding: it requires simultaneous visual reading of musical notation (clefs, key/time signatures, noteheads, articulations, etc.) and structured XML generation that matches the ground-truth schema (see Section 2 for positioning against prior MLLM music-understanding work). We note up front that our metric is a normalized string-level edit distance over the serialized MusicXML, so it primarily reflects schema-faithful structured generation rather than musical content in isolation; we scope our conclusions accordingly and discuss a content-level metric as a next step below. We treat this subject as a first-of-its-kind exploratory probe over 6 documents rather than a population-level capability estimate.

User-prompt-incapable baselines (minerU2.5, PaddleOCR). Both score 1 (complete failure) by design. These are pipeline OCR API systems that do not allow a user prompt. Without instruction, they output plain text, Markdown, or JSON, which is maximally distant from MusicXML ground truth. These results are not meaningful as music transcription scores and should be treated as excluded base-

Subject	Claude	Doubao	Gemini	GPT-4o	GPT-5.5	Kimi	MinerU	Nemotron	PaddleOCR	Qwen-122B	Qwen-Flash	Qwen-Plus	Avg
Antiques & Collectibles	0.299	0.593	0.302	0.666	**0.173**	0.414	0.658	0.870	0.885	0.354	0.390	0.345	0.496
Architecture	0.208	0.264	0.133	0.403	0.153	**0.095**	0.206	0.591	0.715	0.123	0.212	0.140	0.270
Art	0.311	0.279	0.192	0.336	0.196	0.193	0.355	0.485	0.790	0.201	0.219	**0.188**	0.312
Bibles	0.203	0.374	0.192	0.408	0.205	**0.182**	0.329	0.763	0.612	0.215	0.214	0.237	0.328
Biography & Autobiography	**0.001**	0.111	0.103	0.110	0.092	0.102	0.118	0.506	0.735	0.102	0.133	0.118	0.186
Body, Mind & Spirit	0.510	0.203	0.159	0.445	0.218	**0.094**	0.150	0.713	0.494	0.122	0.170	0.181	0.288
Business & Economics	0.077	0.271	0.133	0.337	0.089	**0.028**	0.098	0.661	0.853	0.181	0.132	0.165	0.252
Comics & Graphic Novels	0.176	0.320	**0.152**	0.866	0.198	0.278	0.962	0.761	0.724	0.277	0.258	0.157	0.427
Computers	0.357	0.401	0.239	0.332	0.201	**0.189**	0.399	0.692	0.828	0.294	0.284	0.277	0.374
Cooking	0.198	0.391	0.434	0.300	**0.075**	0.107	0.171	0.655	0.834	0.345	0.413	0.515	0.370
Crafts & Hobbies	**0.174**	0.663	0.358	0.583	0.245	0.212	0.511	0.814	0.863	0.397	0.427	0.434	0.473
Design	0.634	0.647	**0.423**	0.694	0.532	0.620	0.593	0.788	0.710	0.619	0.603	0.549	0.618
Drama	**0.076**	0.414	0.339	0.345	0.098	0.110	0.092	0.548	0.871	0.263	0.336	0.451	0.329
Education	0.226	0.316	0.141	0.272	0.114	**0.083**	0.136	0.463	0.868	0.153	0.180	0.175	0.261
Family & Relationships	0.161	0.270	0.163	0.306	0.136	0.143	**0.130**	0.385	0.764	0.214	0.225	0.160	0.255
Fiction	0.112	0.185	0.130	0.151	0.092	**0.039**	0.092	0.601	0.703	0.088	0.228	0.208	0.219
Games & Activities	0.341	0.639	**0.340**	0.654	0.384	0.375	0.580	0.795	0.804	0.465	0.504	0.435	0.526
Gardening	0.087	0.597	0.107	0.530	0.110	**0.078**	0.140	0.671	0.798	0.122	0.270	0.258	0.314
Health & Fitness	0.053	0.421	0.152	0.385	0.070	**0.035**	0.128	0.739	0.782	0.119	0.162	0.271	0.276
History	**0.177**	0.538	0.293	0.497	0.207	0.187	0.364	0.743	0.644	0.186	0.300	0.242	0.365
House & Home	**0.097**	0.531	0.122	0.432	0.100	0.108	0.181	0.710	0.716	0.151	0.179	0.229	0.296
Humor	0.105	0.174	0.152	0.239	0.138	0.247	0.207	0.621	0.626	**0.096**	0.191	0.184	0.248
Juvenile Fiction	**0.105**	0.257	0.152	0.342	0.135	0.151	0.544	0.548	0.572	0.146	0.162	0.126	0.270
Juvenile Nonfiction	0.230	0.396	**0.189**	0.398	0.310	0.295	0.469	0.605	0.777	0.338	0.442	0.368	0.401
Language Arts & Disciplines	0.645	0.693	0.704	0.645	**0.633**	0.682	0.774	0.816	0.916	0.651	0.673	0.663	0.708
Language Study	0.402	0.495	**0.315**	0.624	0.363	0.343	0.630	0.920	0.644	0.365	0.440	0.355	0.491
Law	**0.277**	0.470	0.374	0.500	0.383	0.420	0.649	0.789	0.630	0.345	0.483	0.366	0.474
Literary Collections	0.256	0.322	0.182	0.496	0.212	**0.156**	0.386	0.808	0.719	0.236	0.258	0.289	0.360
Literary Criticism	**0.131**	0.258	0.144	0.229	0.190	0.140	0.271	0.662	0.651	0.178	0.186	0.274	0.276
Mathematics	0.234	0.439	0.251	0.333	**0.216**	0.288	0.270	0.688	0.700	0.263	0.264	0.236	0.348
Medical	0.261	0.544	0.271	0.542	**0.187**	0.515	0.552	0.772	0.827	0.542	0.524	0.506	0.503
Music[†]	0.935	0.966	0.804	0.983	0.901	0.729	1.000	0.993	1.000	0.702	**0.662**	0.762	0.870
Nature	0.306	0.473	0.326	0.374	0.278	**0.244**	0.576	0.685	0.708	0.373	0.352	0.296	0.416
Performing Arts	**0.049**	0.325	0.166	0.353	0.081	0.143	0.141	0.687	0.798	0.169	0.206	0.206	0.277
Pets	**0.083**	0.339	0.152	0.214	0.113	0.129	0.199	0.754	0.840	0.153	0.159	0.211	0.279
Philosophy	0.212	0.269	0.212	0.263	0.160	**0.133**	0.203	0.633	0.753	0.197	0.160	0.209	0.284
Poetry	0.076	0.171	0.234	0.174	0.129	**0.069**	0.221	0.582	0.618	0.125	0.218	0.256	0.239
Political Science	0.209	0.279	0.202	0.474	0.198	0.157	0.249	0.741	0.610	0.214	**0.144**	0.215	0.308
Psychology	0.111	0.400	0.173	0.332	0.185	**0.110**	0.211	0.753	0.809	0.138	0.158	0.171	0.296
Reference	0.833	0.909	0.860	0.906	0.788	0.819	0.769	0.926	0.938	**0.768**	0.840	0.883	0.853
Religion	0.174	0.340	0.168	0.427	0.216	0.144	0.229	0.719	0.684	0.194	0.150	**0.133**	0.298
Science	0.283	0.345	0.339	0.470	**0.240**	0.264	0.273	0.633	0.684	0.258	0.258	0.348	0.366
Self-Help	0.068	0.297	**0.042**	0.435	0.112	0.076	0.214	0.824	0.810	0.169	0.198	0.121	0.281
Social Science	0.118	0.575	**0.057**	0.384	0.370	0.252	0.065	0.661	0.540	0.128	0.300	0.222	0.306
Sports & Recreation	0.116	0.206	**0.090**	0.322	0.133	0.120	0.566	0.681	0.608	0.156	0.238	0.153	0.282
Study Aids	**0.270**	0.661	0.452	0.678	0.280	0.413	0.341	0.761	0.666	0.593	0.583	0.368	0.506
Technology & Engineering	0.328	0.433	0.360	0.425	**0.294**	0.330	0.506	0.723	0.752	0.409	0.397	0.320	0.440
Transportation	**0.105**	0.512	0.188	0.467	0.128	0.151	0.256	0.829	0.717	0.226	0.259	0.235	0.339
Travel	0.467	0.493	**0.277**	0.741	0.412	0.322	0.438	0.890	0.724	0.370	0.393	0.310	0.486
True Crime	0.164	0.309	0.187	0.368	0.224	**0.148**	0.288	0.658	0.630	0.242	0.212	0.184	0.301
Young Adult Fiction	**0.050**	0.068	0.114	0.475	0.053	0.085	0.633	0.453	0.362	0.073	0.160	0.156	0.224
Young Adult Nonfiction	0.189	0.355	0.223	0.252	**0.168**	0.187	0.281	0.655	0.747	0.203	0.195	0.235	0.307

Table 6: Text block edit distance by subject (lower is better). **Bold** = best per row. Full model IDs: Claude = claude-opus-4-6; Doubao = doubao-seed-1-6-vision-250815; Gemini = gemini-3.1-pro-preview; GPT-4o = gpt-4o; GPT-5.5 = gpt-5.5; Kimi = kimi-k2.5; MinerU = mineru2.5; Nemotron = nemotron-nano-12b-v2-vl; PaddleOCR = paddleocr-api; Qwen-122B = qwen3.5-122b-a10b; Qwen-Flash = qwen3.5-flash; Qwen-Plus = qwen3.5-plus. [†]Music uses a 1-page sliding window; all other subjects use 2 pages.

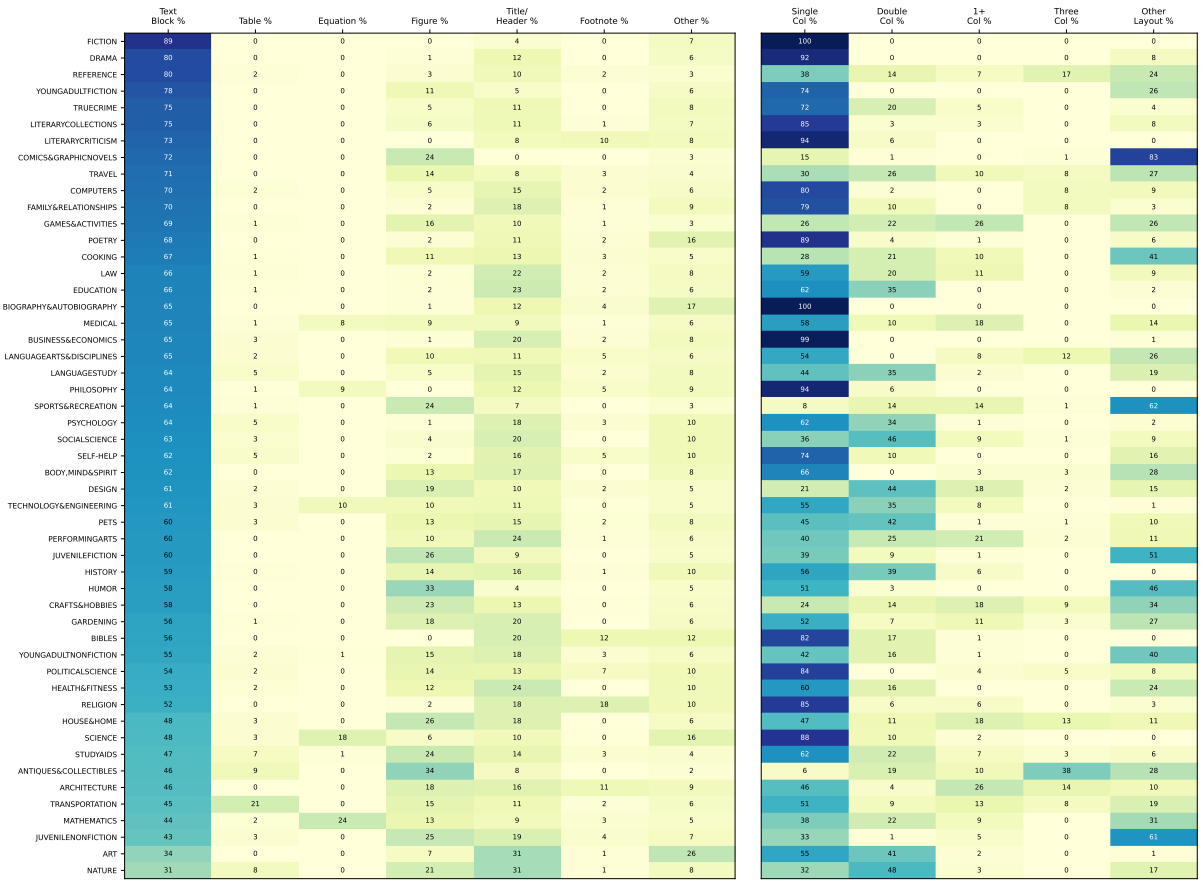


Figure 11: Left: the composition of the data elements per subject. Right: the composition of the layout types per subject. Each individual row sums up to 100%.

Model	Dist	Dist (w/o outlier)
Baseline	0.624	0.513
Qwen3.5-flash	**0.662**	**0.598**
Qwen3.5-122B-A10b	0.702	0.661
Kimi-k2.5	0.729	0.702
Qwen3.5-plus	0.762	0.741
Gemini	0.804	0.792
GPT-5.5	0.901	0.895
Claude	0.935	0.925
Doubao	0.966	0.964
GPT-4o	0.983	0.983
Nemotron	0.993	0.994
Mineru2.5	1.000	1.000
PaddleOCR	1.000	1.000

Table 7: Music score → MusicXML transcription results (1-page window). Lower edit distance is better. The baseline is the mean pairwise edit distance across all 6 ground-truth documents. The outlier is the single document with anomalous length (6.1M chars vs. 419K–1.3M for the rest).

lines. We still report the numbers for completeness.

GT cross-document baseline. To anchor the difficulty of this task, we compute the pairwise edit distance between all 6 groundtruth files (15 pairs), yielding a mean of 0.624 (std 0.194, min 0.378, max 0.932). This is a *null reference* rather than a quality ceiling: it is the score one would obtain by submitting an arbitrary *different* real score, so any model scoring above it produces output no closer to the target than an unrelated GT document would. The high standard deviation (0.194) is driven by extreme length variation across the 6 GT files, which range from 419K to 6.1M characters (a 15× spread), with one outlier nearly 5× larger than the next longest, inflating all pairwise distances that involve it. Excluding this outlier yields a mean of 0.513 (std 0.127, 10 pairs), yet the best model score (0.662) still exceeds even this lower null, so the conclusion is robust to the outlier's presence.

Frontier VLMs. Among the prompt-following models, the Qwen3.5 family dominates the top positions: Qwen3.5-Flash (0.662) achieves the best score. However, this barely clears the null reference (0.624), so even the top model is only marginally closer to the target than a random different GT file. Notably, Qwen3.5-Flash is the smallest but best-performing Qwen variant. Together with a likely "overthinking penalty" in larger models, which emit additional commentary, reformatting, or structural variation that inflates string-level edit distance, this points to MusicXML transcription being gated primarily by compact, schema-faithful generation rather than by raw model scale. Gemini-3.1 Pro (0.804) trails the Qwen models but outperforms GPT-5.5 (0.901) and Claude-Opus 4.6 (0.935), indicating that the ability to emit schema-faithful MusicXML is unevenly distributed across frontier models. Doubao (0.966) and GPT-4o (0.983) are near-collapsed, approaching the failure threshold of the prompt-incapable baselines.

Effect of the outlier document. Model rankings are fully preserved with and without the outlier, so all qualitative conclusions stand. However, removing the outlier reveals that the full-corpus scores were understating model failure: the gap between the best model and the null widens from 0.038 (0.662 vs. 0.624) to 0.085 (0.598 vs. 0.513), because the outlier inflated both model scores and the null mean proportionally — but masked more of the model failure in the process. The Qwen models benefit most from the outlier's removal (Qwen3.5-Flash: -0.064, Qwen3.5-122B: -0.041), whereas GPT and Claude frontier models drop by ≤ 0.012 and near-failed models (GPT-4o, MinerU2.5, PaddleOCR) are unaffected (≤ 0.001), indicating they fail uniformly regardless of document length.

Overall takeaway. On this exploratory probe, no model produces MusicXML closer to the reference than an unrelated score: every system sits at or above the cross-document null under both conditions, so as a schema-faithful transcription task, music remains unsolved for current frontier and specialized systems. Removing the single outlier strengthens this reading, since the best model (Qwen3.5-Flash, 0.598) remains 0.085 above the outlier-excluded null (0.513), a larger margin than the 0.038 observed with all documents. The task cleanly filters out non-prompt-capable systems and separates frontier VLMs along schema-faithful structured generation, a dimension that general document benchmarks do not expose, with the smaller Qwen3.5-Flash leading, suggesting that optimization for structured output generation matters more than scale here. We deliberately scope this probe to schema-faithful generation; isolating musical content from XML verbosity requires a structure-aware score over the parsed MusicXML tree (e.g., TEDS, as we use for tables) or a note-level F1 over (pitch, onset, duration) tuples. We leave this content-level evaluation to future work.

H.2 The Worst Five Subjects

The detailed breakdown of the worst five subjects and the individual performance on these five subjects can be found in Table 8.

Every single model (all 12) has REFERENCE in its worst 5. This is the one consistent failure across pipeline tools and frontier LLMs, likely due to dense bibliographic entries, unusual list formatting, and high text volume.

DESIGN is also near-universal (10/12 models). Only Nemotron and PaddleOCR escape it, and that's probably because their failure modes are so severe elsewhere that Design doesn't even rank in their worst 5. Heavy visual layout with irregular text placement appears to be a systematic challenge.

Mineru's worst subject is COMICS, with an overall score of 8.39. Since Mineru can't receive the user prompt, it never sees our specific prompt for comics. GPT-4o also has Comics in its worst 5, suggesting even with the instruction, it's hard, but the gap between prompted and unprompted models here is enormous.

PaddleOCR's failure mode is categorically different. Its worst subjects are DRAMA, FAMILY & RELATIONSHIPS, COOKING, PHILOSOPHY, which are more of consumer prose/layout categories rather than the structurally complex subjects like REFERENCE or LAW that trip LLMs. Without user prompts. It has no semantic guidance and falls apart on visually rich or narrative-heavy content. Its worst 5 don't overlap much with any other model.

7 out of the 10 LLM-based models have GAMES in their worst 5. Rulebooks and activity guides likely mix visual layouts, tables, and lists in non-standard ways that confuse even strong models.

H.3 Additional Analyses

Page degradation (Table 9). General VLMs consistently score *better* on `colorful_background` pages than on `fuzzy_scan` pages, suggesting color complexity is less disruptive than scan quality degradation for capable models. MinerU is the only exception: it has a higher edit distance when the background is colorful, making it the only model that degrades more on colored than blurry pages.

Reading order by layout (Table 10). GPT-5.5 leads on every multi-column layout variant (double 0.194, three 0.180, $\geq$ 0.226, other 0.222), demonstrating strong robustness across complex layouts. Conversely, Doubao degrades sharply as column count increases (single 0.213 $\rightarrow$ double 0.409 $\rightarrow$ three 0.454), suggesting poor generalization to multi-column arrangements. MinerU exhibits the steepest absolute performance decline among non-PaddleOCR models, dropping from single_column (0.237) to other_layout (0.463), indicating distinct challenges with non-standard page configurations. However, PaddleOCR's other_layout score (0.602) surpasses its double_column (0.795) and three_column (0.853) performance.

Text element attributes by edit distance (Table 11). Rotation is a sharp model differentiator. GPT-4o is the most rotation-sensitive frontier VLM, with rotate90 (0.814) and rotate270 (0.758) far exceeding its normal score (0.479). Doubao similarly collapses on rotate90 (0.761) and rotate27z0 (0.757). Kimi shows an asymmetric pattern: rotate90 (0.672) is catastrophic while rotate270 (0.432) is tolerated, suggesting partial handling. GPT-5.5 is the most rotation-robust frontier model (0.240 / 0.419). Chinese-script text elements (`text_simplified_chinese`) are substantially harder than English elements across all models: kimi's ZH score (0.572) is 2.2$\times$ its EN score (0.264), and claude's ZH (0.588) is 2.2$\times$ its EN (0.269). This is in apparent tension with Table 14, where Claude and Kimi score far better on Chinese-language *pages*; the discrepancy arises because the two breakdowns measure different slices — element-level language vs. page-level document language — and is discussed further in the context of Table 14. MinerU shows the largest degradation under multi-colored text backgrounds (0.637 vs. 0.363 for white), steeper than any frontier VLM.

Formula performance by document type and layout (Table 12). GPT-4o's formula recognition degrades most severely on `colorful_textbook` (0.906), followed closely by Claude (0.891) and `exam_paper` (0.888) — a 3.5$\times$ decline from Claude's `academic_literature` score (0.253). Claude's other_layout formula edit distance (0.940) is the worst among frontier VLMs (GPT-5.5: 0.860, kimi: 0.842, gemini: 0.775), indicating a particular weakness with formulas in non-standard page arrangements. PaddleOCR leads on `colorful_textbook` (0.315) and `other_layout` (0.398) despite its `$$` delimiter penalty (Section 4), further confirming that its formula recognition module is genuinely strong on these conditions.

Challenge type (Table 13). Most models score better on `domain_reasoning` than on `structural_reconstruction` despite the former being semantically harder by design — including PaddleOCR (domain 0.668 vs. structural 0.738), which has no language understanding. This suggests the category labels partly reflect the visual and structural complexity of individual elements: domain_reasoning elements may be shorter or less layout-intensive, making them easier to transcribe accurately regardless of semantic content. Doubao, Qwen3.5+, and Qwen3.5-122B are exceptions that show domain > structural (worse performance on the semantically harder type), contrasting with claude and GPT-5.5, where domain_reasoning is the *best*-scoring challenge type.

Page-level language (Table 14). Models are split into three groups based on Chinese-page performance. Claude, Kimi, and Gemini score dramatically lower in edit distance (better) on Chinese pages than English (Claude: 0.218 → 0.053; Kimi: 0.173 → 0.059; Gemini: 0.200 → 0.055 — approximately 4× improvement). GPT-5.5 and most other models show the opposite or no advantage: GPT-5.5 is slightly worse on Chinese (0.156 → 0.179), as are doubao and mineru. Separately, the element-level breakdown (Table 11) shows that Chinese-script text elements are substantially *harder* for all models — Kimi's ZH element score (0.572) is 2.2× its EN score (0.264). These two observations are not directly comparable: Table 14 aggregates all content on a page by that page's primary language, while Table 11 groups individual blocks by their own script label. The underlying causes of the page-level advantage for the first model group and the element-level hardness of Chinese-script text are open questions that require inspecting the document-type distribution of Chinese-language pages to resolve.

H.4 Scaling of model size

We analyze the impact of model size within the Qwen3.5 family, comparing three variants ordered by scale: flash (smallest), 122B-A10B, and plus (largest). Figure 12 reports performance across five metrics. The overall scores vary within a narrow range (57.5, 56.3, and 57.3 for the three variants, respectively), but individual metrics exhibit larger variation.

Formula CDM decreases monotonically with scale. Formula recognition accuracy decreases as model size increases: 35.7 (flash) → 32.5 (122B-A10B) → 30.4 (plus), a cumulative drop of 5.3 points. This trend suggests a trade-off between formula-level precision and other capabilities.

Per-metric trade-offs across model variants. Table TEDS drops from 46.2 (flash) to 38.3 (122B-A10B), a decrease of 7.9 points, before recovering to 49.8 (plus). Conversely, 122B-A10B achieves the highest text accuracy (77.4 vs. 74.2 for flash) and reading-order accuracy (77.1 vs. 73.9 for flash). These inverse rankings indicate that the mid-tier model performs better on text transcription and ordering than on structured layout recognition.

These results indicate that scaling model size does not uniformly improve document understanding; per-metric trends can move in opposite directions even as the overall score remains stable. We note a potential confound: Qwen3.5-Flash and Qwen3.5-Plus are production API variants with possible serving-side enhancements and extended context window, [3] whereas the 122B-A10B variant (32k context length) is served without such optimizations, which may partly account for the observed differences.

H.5 Scaling of sliding window size

We evaluate how varying the sliding window size n (from 1 to 15 pages) affects nine models: Claude, Gemini, GPT-4o, GPT-5.5, Qwen-3.5 flash (smallest), 122B-A10B, and plus (largest), as well as two specialized VLMs (MinerU2.5 and PaddleOCR). Based on the subject hardness, we arbitrarily choose to focus on two subjects: law and medicine.Figure 3 shows the aggregate score and per-metric trends, Figure 13 shows the formula metrics trends.

Reading order degrades most consistently. The most robust correlation across all models and subjects is between window size and reading order edit distance. Aggregated over all subjects, Claude's reading order error rises from 0.24 at n=1 to 0.69 at n=15; GPT-5.5 climbs from 0.17 to 0.47; Gemini from 0.24 to 0.51. GPT-4o shows the sharpest degradation among frontier LLMs, rising from 0.38 to 0.85. The effect is especially pronounced in the law subject, where Claude's reading order edit distance increases from 0.077 at n=1 to 0.694 at n=15. Larger windows require the model to track

[3] Qwen3.5-Plus: https://qwen.ai/blog?id=qwen3.5, Qwen3.5-Flash: https://huggingface.co/Qwen/Qwen3.5-35B-A3B. Both with a 1M context length.

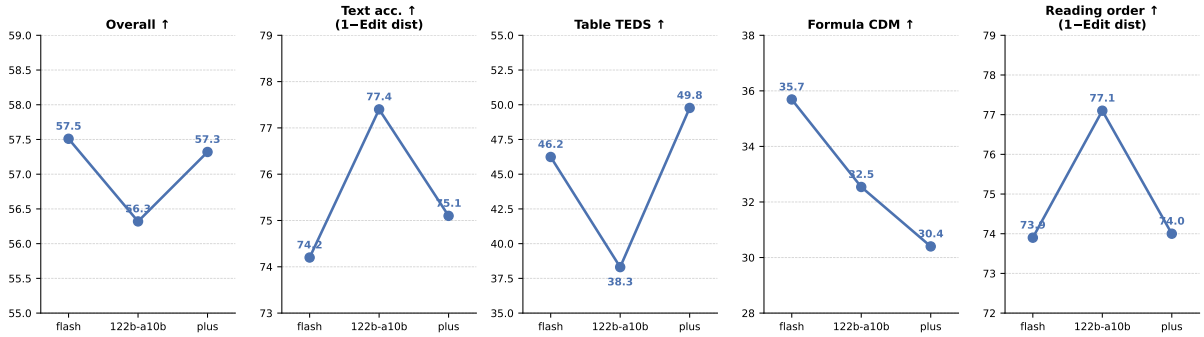


Figure 12: Scaling per model size

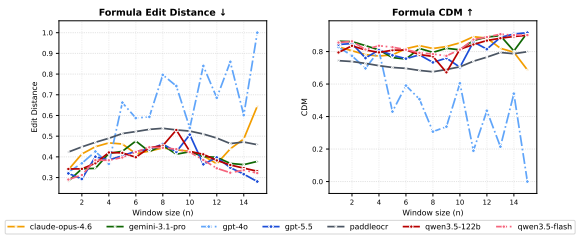


Figure 13: Impact of sliding window size n pages) on the formula metrics across seven models.

ordering cues across more pages, and accuracy tends to fall as this span grows.

Text edit distance shows a mixed early-window benefit. For text edit distance, GPT-5.5 and Gemini both improve marginally from n=1 to n=2 (GPT-5.5: 0.287 $\rightarrow$ 0.285; Gemini: 0.335 $\rightarrow$ 0.322), suggesting a slight cross-page context benefit for resolving boundary-spanning content. Claude generally worsens with context size (0.220 at n=1 to 0.365 at n=15), while GPT-4o shows the most severe degradation among frontier models, rising from 0.531 to 0.782 at n=15. Qwen-Plus shows relatively flat behavior throughout (0.345 at n=1 to 0.369 at n=15). Specialized VLMs (MinerU2.5 and PaddleOCR) are nearly invariant across all window sizes, consistent with pipeline architectures that do not leverage cross-page context. Beyond n=2, text accuracy generally worsens across all models, consistent with longer contexts introducing more opportunities for content confusion or truncation.

Formula metrics show no systematic degradation. Formula CDM stays within a moderate band for most models across small-to-moderate window sizes (e.g., Qwen-Plus: 0.817 at n=1 to a range of 0.763–0.867 through n=12). Claude shows a substantial drop at very large windows (0.687 at n=15). GPT-4o begins failing erratically from n=5 onward—oscillating between near-zero values and partial recoveries before collapsing completely to 0.000 at n=15. In stark contrast, GPT-5.5 remains broadly stable throughout (0.704–0.916), suggesting a significant difference in how the two models handle structured element prediction under long contexts.

The n=2 default. The two-page window used throughout our main evaluation reflects a practical trade-off. It captures enough cross-page context to benefit models that improve from n=1 to n=2, while staying well below the window sizes where reading order and text accuracy degrade substantially. Larger windows yield diminishing returns at best and measurably worse performance at realistic document lengths.

Model / Subject	Overall↑	Text↓	Form.↓	CDM↑	TEDS↑	$TEDS_S$↑	Order↓
Claude-Opus-4.6	40.33	0.469	0.414	37.86	17.66	28.54	0.473
REFERENCE	20.17	0.833	–	–	23.12	32.67	0.793
DESIGN	32.06	0.634	–	–	15.75	18.49	0.561
MEDICAL	47.09	0.261	0.414	37.86	18.88	33.48	0.423
LAW	50.87	0.277	–	–	0.00	0.00	0.197
GAMES&ACTIVITIES	52.47	0.341	–	–	30.56	58.04	0.391
Doubao	29.31	0.652	–	–	5.89	9.67	0.528
REFERENCE	15.69	0.909	–	–	4.94	11.46	0.670
GAMES&ACTIVITIES	26.45	0.639	–	–	0.84	3.24	0.576
ANTIQUES&COLLECTIBLES	33.30	0.593	–	–	3.47	5.16	0.442
DESIGN	34.02	0.647	–	–	20.21	28.48	0.534
LAW	37.09	0.470	–	–	0.00	0.00	0.417
Gemini-3.1-Pro	36.88	0.563	0.431	18.31	28.20	31.94	0.427
REFERENCE	17.83	0.860	–	–	25.58	33.40	0.861
STUDYAIDS	44.89	0.452	0.431	18.31	46.33	50.92	0.398
LAW	45.47	0.374	–	–	0.00	0.00	0.262
DESIGN	48.33	0.423	–	–	31.25	33.33	0.440
LANGUAGEARTS&DISCIPLINES	49.98	0.704	–	–	37.82	42.07	0.175
GPT-4o	27.33	0.757	–	–	11.51	16.78	0.538
REFERENCE	25.83	0.906	–	–	12.81	19.23	0.447
GAMES&ACTIVITIES	27.73	0.654	–	–	3.88	8.32	0.553
ANTIQUES&COLLECTIBLES	27.91	0.666	–	–	9.78	15.15	0.594
COMICS&GRAPHICNOVELS	30.53	0.866	–	–	–	–	0.523
DESIGN	30.97	0.694	–	–	19.59	24.40	0.573
GPT-5.5	44.24	0.461	–	–	19.41	26.34	0.406
REFERENCE	24.98	0.788	–	–	27.82	37.61	0.741
DESIGN	43.72	0.532	–	–	33.28	33.33	0.489
LAW	46.01	0.383	–	–	0.00	0.00	0.237
RELIGION	52.97	0.216	–	–	0.00	0.00	0.195
GAMES&ACTIVITIES	53.52	0.384	–	–	35.97	60.77	0.370
Kimi-K2.5	38.19	0.550	0.367	35.76	23.73	33.36	0.517
REFERENCE	25.70	0.819	–	–	37.90	48.91	0.789
DESIGN	38.29	0.620	–	–	35.02	39.03	0.581
MEDICAL	42.13	0.515	0.367	35.76	30.38	45.70	0.461
LAW	42.89	0.420	–	–	0.00	0.00	0.294
GAMES&ACTIVITIES	43.90	0.375	–	–	15.36	33.14	0.462
MinerU2.5	30.15	0.705	0.323	31.42	25.07	32.52	0.654
COMICS&GRAPHICNOVELS	8.39	0.962	–	–	–	–	0.871
REFERENCE	27.58	0.769	–	–	33.63	45.73	0.740
LAW	30.62	0.649	–	–	0.00	0.00	0.432
DESIGN	37.25	0.593	–	–	33.14	36.79	0.621
MEDICAL	37.29	0.552	0.323	31.42	33.52	47.58	0.606
Nemotron	15.46	0.846	–	–	6.45	12.73	0.755
ANTIQUES&COLLECTIBLES	12.05	0.870	–	–	8.48	12.67	0.853
CRAFTS&HOBBIES	15.37	0.814	–	–	2.28	15.00	0.748
REFERENCE	15.81	0.926	–	–	11.37	18.78	0.713
LAW	16.37	0.789	–	–	0.00	0.00	0.720
TRANSPORTATION	17.69	0.829	–	–	10.13	17.20	0.742
PaddleOCR	14.43	0.832	0.690	8.76	16.47	29.35	0.843
DRAMA	12.13	0.871	–	–	–	–	0.886
REFERENCE	14.27	0.938	–	–	27.47	38.96	0.909
FAMILY&RELATIONSHIPS	14.44	0.764	–	–	0.00	0.00	0.802
COOKING	15.75	0.834	1.000	0.00	26.87	44.93	0.805
PHILOSOPHY	18.10	0.753	0.379	17.51	11.53	33.51	0.813
Qwen3.5-Flash	37.19	0.591	0.312	42.07	18.05	24.95	0.523
REFERENCE	24.87	0.840	–	–	21.00	27.90	0.624
LAW	33.45	0.483	–	–	0.00	0.00	0.513
GAMES&ACTIVITIES	37.20	0.504	–	–	10.78	21.23	0.488
DESIGN	40.17	0.603	–	–	32.85	35.60	0.521
MEDICAL	42.09	0.524	0.312	42.07	25.60	40.04	0.469
Qwen3.5-122B	34.62	0.597	0.381	22.99	22.05	28.66	0.468
STUDYAIDS	31.14	0.593	0.420	6.40	36.86	40.77	0.594
DESIGN	33.17	0.619	–	–	13.85	14.88	0.524
GAMES&ACTIVITIES	39.36	0.465	–	–	8.33	17.82	0.437
MEDICAL	39.38	0.542	0.342	39.58	20.36	30.79	0.482
REFERENCE	41.24	0.768	–	–	30.84	39.05	0.303
Qwen3.5-Plus	36.45	0.548	0.370	38.27	15.55	24.98	0.533
REFERENCE	16.77	0.883	–	–	26.81	37.52	0.882
DESIGN	32.50	0.549	–	–	1.13	1.28	0.488
LAW	40.61	0.366	–	–	0.00	0.00	0.416
MEDICAL	42.94	0.506	0.370	38.27	30.10	47.90	0.460
GAMES&ACTIVITIES	44.81	0.435	–	–	19.69	38.22	0.418

Table 8: Detailed breakdown per subject results for the worst 5 subjects per model.

Model	Fuzzy Scan↓	Color BG↓
Claude	0.256	**0.163**
Doubao	0.454	0.285
Gemini	0.269	0.172
GPT-4o	0.473	0.362
GPT-5.5	0.240	0.164
Kimi	**0.238**	0.189
MinerU	0.364	0.402
Nemotron	0.665	0.573
PaddleOCR	0.774	0.655
Qwen3.5-Flash	0.344	0.222
Qwen3.5-122B-A10B	0.288	0.200
Qwen3.5-Plus	0.310	0.194

Table 9: Text extraction edit distance by page degradation type (↓ lower is better). **Bold**: best per column.

Model	All	Single Col.	Double Col.	Three Col.	≥2 Col.	Other Layout
Claude	0.188	0.158	0.233	0.202	0.256	0.241
Doubao	0.283	0.213	0.409	0.454	0.384	0.326
Gemini	0.210	0.177	0.255	0.293	0.269	0.258
GPT-4o	0.307	0.233	0.388	0.508	0.420	0.391
GPT-5.5	**0.167**	**0.143**	**0.194**	**0.180**	**0.226**	**0.222**
Kimi	0.182	**0.133**	0.231	0.274	0.263	0.271
MinerU	0.300	0.237	0.340	0.336	0.357	0.463
Nemotron	0.564	0.508	0.646	0.769	0.645	0.600
PaddleOCR	0.705	0.687	0.795	0.853	0.813	0.602
Qwen3.5-Flash	0.261	0.197	0.385	0.397	0.357	0.295
Qwen3.5-122B	0.229	0.175	0.298	0.341	0.339	0.296
Qwen3.5-Plus	0.260	0.222	0.336	0.365	0.329	0.280

Table 10: Reading-order edit distance by page layout type (↓ lower is better). **Bold**: best per column.

Model	Lang EN	Lang ZH	Lang EN+ZH	Lang Other	BG White	BG Single	BG Multi	Rot. Norm.	Rot. 90°	Rot. 270°	Rot. Horiz.
Claude	0.269	0.588	0.453	0.202	0.282	0.238	0.312	0.257	0.269	0.293	0.413
Doubao	0.461	0.675	0.534	0.522	0.498	0.469	0.491	0.477	0.761	0.757	0.545
Gemini	0.314	0.590	0.381	0.297	0.318	0.314	0.369	0.319	0.446	**0.277**	0.341
GPT-4o	0.475	0.741	0.436	0.505	0.504	0.468	0.631	0.479	0.814	0.758	0.763
GPT-5.5	**0.214**	0.605	**0.168**	**0.197**	**0.253**	**0.191**	**0.269**	**0.215**	**0.240**	0.419	**0.318**
Kimi	0.264	**0.572**	0.270	0.236	0.274	0.248	0.406	0.259	0.672	0.432	0.357
MinerU	0.372	0.705	0.424	0.305	0.363	0.345	0.637	0.354	0.539	0.563	0.546
Nemotron	0.671	0.877	0.537	0.813	0.727	0.698	0.764	0.708	0.880	0.750	0.964
PaddleOCR	0.938	0.899	0.947	0.945	0.935	0.942	0.930	0.941	0.971	0.966	0.854
Qwen3.5-Flash	0.402	0.714	0.499	0.404	0.411	0.412	0.387	0.408	0.518	0.576	0.478
Qwen3.5-122B	0.321	0.654	0.328	0.299	0.344	0.302	0.398	0.319	0.496	0.481	0.432
Qwen3.5-Plus	0.393	0.675	0.430	0.448	0.437	0.398	0.389	0.416	0.502	0.527	0.370

Table 11: Text extraction edit distance by text-block element attributes (↓ lower is better). *Lang*: language label of the individual text block. *BG*: background color (White / Single-colored / Multi-colored). *Rot.*: text rotation (Normal / 90° / 270° / Horizontally flipped). **Bold**: best per column.

Model	All	Acad. Lit.	Color. TB	Exam Paper	Book	Single Col.	Double Col.	Other Layout
Claude	0.368	0.253	0.891	0.888	**0.259**	0.273	0.318	0.940
Doubao	0.313	0.208	0.719	0.527	0.277	0.279	**0.201**	0.545
Gemini	0.353	0.249	0.724	0.757	0.280	0.290	0.277	0.775
GPT-4o	0.468	0.381	0.906	0.641	0.440	0.413	0.645	0.708
GPT-5.5	0.359	0.250	0.449	0.803	0.289	0.294	0.263	0.860
Kimi	0.350	0.177	0.741	0.769	0.277	0.274	0.247	0.842
MinerU	0.506	0.470	0.518	0.885	0.439	0.454	0.288	0.952
Nemotron	0.759	0.604	0.844	0.955	0.751	0.724	0.703	0.969
PaddleOCR	0.415	0.395	**0.315**	**0.442**	0.417	0.412	0.461	**0.398**
Qwen3.5-Flash	0.317	0.179	0.773	0.678	0.262	**0.260**	0.232	0.697
Qwen3.5-122B	0.322	**0.173**	0.348	0.574	0.282	0.286	0.217	0.576
Qwen3.5-Plus	**0.307**	0.190	0.660	0.579	0.265	0.261	0.264	0.600

Table 12: Display formula edit distance broken down by data source and page layout (↓ lower is better). *Color. TB*: colorful textbook; *Acad. Lit.*: academic literature. **Bold**: best per column.

Model	All	Percep-tion	Structural Recon.	Domain Reasoning
Claude	0.217	0.158	0.256	**0.123**
Doubao	0.344	0.270	0.392	0.402
Gemini	0.219	0.167	0.255	0.192
GPT-4o	0.366	0.293	0.414	0.284
GPT-5.5	**0.189**	0.148	**0.223**	0.146
Kimi	0.193	**0.137**	0.234	0.165
MinerU	0.325	0.249	0.387	0.236
Nemotron	0.622	0.570	0.656	0.644
PaddleOCR	0.729	0.700	0.738	0.668
Qwen3.5-Flash	0.258	0.190	0.303	0.266
Qwen3.5-122B	0.226	0.163	0.271	0.292
Qwen3.5-Plus	0.249	0.203	0.280	0.285

Table 13: Text extraction edit distance by challenge type (↓ lower is better). **Bold**: best per column.

Model	All	English	Simplified Chinese	Other
Claude	0.217	0.218	**0.053**	0.211
Doubao	0.344	0.293	0.353	0.521
Gemini	0.219	0.200	0.055	0.264
GPT-4o	0.366	0.321	0.541	0.493
GPT-5.5	**0.189**	**0.156**	0.179	**0.204**
Kimi	0.193	0.173	0.059	0.205
MinerU	0.325	0.293	0.393	0.340
Nemotron	0.622	0.531	0.752	0.832
PaddleOCR	0.729	0.732	0.628	0.732
Qwen3.5-Flash	0.258	0.241	0.288	0.307
Qwen3.5-122B	0.226	0.210	0.210	0.270
Qwen3.5-Plus	0.249	0.222	0.248	0.300

Table 14: Text extraction edit distance by page-level language (↓ lower is better). **Bold**: best per column.