GDP.pdf: Benchmarking Grounded Multimodal Reasoning over Professional PDF Documents

Suhaas Garre*, Emily Ritchie, Sushant Mehta, Edwin Chen
Surge AI

Abstract

A large share of day-to-day work in professional domains happens inside PDF files: benefits packets, leases, datasheets, clinical guidelines, construction plans. Benchmarks for document AI have generally measured the required capabilities in isolation: OCR, layout analysis, chart reasoning, table QA, document VQA. A high score on any one of them does not necessarily reveal whether a model can answer a realistic question that someone in the field would actually ask about a specific PDF. GDP.pdf is a benchmark built to measure this directly. It consists of question–document pairs authored by working professionals in ten fields, and a candidate question was kept only when two strong frontier multimodal models both failed it in a way that mattered: a wrong answer, missed decisive evidence, or a fabricated claim, rather than a superficial difference such as style. Each item comes with a rubric of atomic criteria, so we can report a graded rubric score as well as a strict task-level pass rate, and each item is tagged against a taxonomy of eleven capabilities in three tiers, spanning text extraction and grounding, table and chart comprehension, cross-referencing, spatial reasoning, and abstention on unsupported queries. We evaluated seven frontier models on the 100-item benchmark. The best model passed only 15% of the items and the worst passed 1%. Most errors trace back to a small set of recurring loss patterns: misaligned tables, misread charts, skipped footnotes and exclusions, miscounted floor-plan symbols, scan noise, and amendments that supersede earlier text. A 50-example public split is available at https://huggingface.co/datasets/surgeai/GDP.pdf.

1. Introduction

Multimodal models are usually evaluated on visual QA of a fairly academic kind. The document tasks people would actually like to hand to a model look different. Comparing plan tiers in a benefits packet, locating the indemnification clause in a lease, or counting fixtures on a floor plan all require working through a long and variably formatted file, and the fact that settles the answer may sit in a footnote three pages away from the flowchart it qualifies. Current models score well on the standard visual-reasoning suites, but we find those scores to be a poor guide to performance on the document workflows that power everyday economic activity.

Several separate problems are involved here, and prior benchmarks have tended to study each in isolation. The first is document structure (multi-page tables, sidebars, legends, footnotes, amendments appended at the end). The second is background knowledge: a benefits table assumes, for example, that the reader knows what a "tier" is. The third problem, which was also a major motivation of this work, is that the failures are not visible as failures. The model cites a clause that exists and a number that is on the page; the clause is simply not the one that governs the user's question.

GDP.pdf was constructed to evaluate all three problems jointly. The questions are phrased as real practitioners phrase them, the input is the original PDF rather than a cleaned-up extract, and the grading checks whether the response rests on the correct evidence. The current version contains 100 items covering ten domains (*Finance*, *Healthcare*, *Legal*, *STEM/Research*, *Engineering*, *Construction*, *Manufacturing/Supply Chains*, *Insurance*, *Real Estate*, and *Human Resources*). Every item has an expert rubric and capability tags, and an item was admitted only after two strong frontier models had failed it. A 50-example public split is released with source PDFs, prompts, rubrics, and domain labels.[1]

Contributions.

1. The GDP.pdf benchmark: expert-authored tasks over professional PDFs in ten workflow domains, screened so that every item defeats two frontier models.
2. A taxonomy of eleven capability axes in three tiers (foundational extraction and grounding; structural and multimodal comprehension; advanced reasoning), which is used to tag tasks and to organize the error analysis.

*Correspondence to: suhaas@surgehq.ai

[1] https://huggingface.co/datasets/surgeai/GDP.pdf

3. An evaluation protocol built on atomic rubric criteria, reported as a graded rubric score and as a strict pass rate.
4. A pilot study of seven frontier models, with an analysis of the failure patterns we expect will matter most in professional use.

2. Related Work

Document parsing, OCR, and layout analysis. The early document benchmarks were concerned with turning page images into structure. FUNSD [12] covered form parsing, CORD [24] covered receipts, PubLayNet [35] and DocLayNet [26] covered page layout, and DocILE [29] covered localization and extraction on business documents. Later work has continued in this direction with an emphasis on OCR and layout robustness; see OCRBench [14] and OCRBench v2 [9], OmniDocBench [23], Real5-OmniDocBench [36], RoDLA [4], and MMDocBench [39]. All of these measure how well a model reads the page. None of them measures whether a system can use what it reads to complete a task someone is paid to do, which is the question we wanted to answer.

Document QA, long documents, and retrieval-aware systems. The document VQA line moved evaluation from parsing toward question answering: DocVQA [20], InfographicVQA [21], MP-DocVQA [30], DUDE [31], and TAT-DQA [38]. DocBench [40], MMLongBench-Doc [16], LongDocURL [7], and M-LongDoc [6] extended the setting to long documents and raw files, and the retrieval-oriented papers (PDFTriage [27], MMDocRAG [8]) argued that finding the evidence inside a long multimodal file is a large part of the problem in its own right. GDP.pdf overlaps most with this group. The difference lies in what the questions assume of the reader. Ours concern professional documents where the answer hinges on footnotes, exclusions, legends, or superseded sections, and where a reader is expected to know what those constructs mean.

Charts, tables, and domain-specific reasoning. Charts have dedicated benchmarks (ChartQA [18], PlotQA [22], CharXiv [32], ChartQAPro [19]), as do tables and numerical reasoning (HybridQA [3], TAT-QA [37], FinQA [5]). FinanceBench [11], LegalBench [10], and ClinicBench [13] test financial, legal, and clinical knowledge respectively. What nearly all of these have in common is that the model receives a cleaned input – extracted text, one table, or one chart by itself. By the time the input has been cleaned to that degree, much of what makes a real PDF difficult has already been removed.

Broad multimodal evaluation. The general multimodal suites (MMMU [33], MMMU-Pro [34], MathVista [15], MEGA-Bench [2]) track progress across many task types at once, and the knowledge-focused VQA datasets, OK-VQA [17] and A-OKVQA [28], test world knowledge over natural images. These benchmarks serve a purpose different from ours. None of them was designed to indicate whether a model can be trusted with a benefits packet or a deed. Table 1 summarizes where GDP.pdf sits relative to these families.

3. GDP.pdf: Benchmark Design

3.1. Task Formulation

Each benchmark item is a tuple

$$x_i = (P_i, q_i, R_i, a_i, d_i, \tau_i),$$

where P_i is a PDF document, q_i is a natural-language question, $R_i = \{r_{ij}\}_{j=1}^{K_i}$ is an expert rubric containing K_i atomic criteria, a_i is an expert reference answer used during rubric authoring, d_i is a domain label, and $\tau_i \subseteq \mathcal{T}$ is a set of capability-axis tags drawn from the taxonomy $\mathcal{T}$.

Given a model response $\hat{y}_i = M(P_i, q_i)$, the grader assigns a binary score $g_{ij} \in \{0, 1\}$ to each rubric element r_{ij}. The item-level rubric score is

$$s_i(M) = \frac{1}{K_i} \sum_{j=1}^{K_i} g_{ij},$$

and the benchmark-level mean rubric score is

$$\text{Score}(M) = \frac{1}{N} \sum_{i=1}^{N} s_i(M).$$

We also define a strict pass-rate indicator

$$p_i(M) = \begin{cases} 1 & \text{if } g_{ij} = 1 \text{ for all } j \\ 0 & \text{otherwise,} \end{cases}$$

with benchmark-level strict pass rate

$$\text{PassRate}(M) = \frac{1}{N} \sum_{i=1}^{N} p_i(M).$$

Models are ranked by strict pass rate; all other analysis uses the mean rubric score. An item passes strictly only if every element of its rubric is satisfied; Table 6 reflects how rarely that happens. The rubric score keeps the partial credit and shows the sub-tasks and criteria that the model gets right.

3.2. Design Principles

Prior knowledge. If general prior knowledge suffices to answer a question, the task was rejected: the answer has to depend on the attached PDF.

Realism. The questions came from contributors' actual jobs and workflows. Each contributor was encouraged to submit

Benchmark family	Representative work	Evaluates well
Layout / OCR / parsing	FUNSD, CORD, PubLayNet, DocLayNet, DocILE, OmniDocBench	How well the page is read: OCR quality, layout detection, extraction
Document QA / long-doc	DocVQA, MP-DocVQA, DUDE, DocBench, MMLongBench-Doc, M-LongDoc	Question answering over raw or long documents, often with evidence spread over many pages
Chart / table / domain	ChartQA, CharXiv, TAT-QA, FinQA, FinanceBench, LegalBench	Charts, tables, or one specialized domain, usually with a cleaned input
Broad multimodal	MMMU, MMMU-Pro, MathVista, MEGA-Bench	General multimodal reasoning across many task types
This work	**GDP.pdf**	**Whether answers to professional questions are grounded in the right evidence from the original PDF**

Table 1. GDP.pdf relative to neighboring benchmark families.

the kind of question they would actually have typed to an assistant mid-task at work.

Adversarial construction. Two strong frontier multimodal models attempted every candidate task, and only candidates on which both made a major, meaningful error were included.

Knowledge-intensive grounding. Tasks in which perception and domain knowledge interact were favored. Professional documents can put a surprising amount of weight on fine print: exclusions, footnotes, legends, plan symbols, amendments.

Diagnostic granularity. Capability tags were added to every item, so the results can be sliced by failure type.

3.3. Benchmark Profile and Coverage

Table 2 summarizes the details of the current version of GDP.pdf. We considered heterogeneity in both domain and in artifact type, so GDP.pdf includes financial filings and benefits packets, datasheets and clinical guidelines, floor plans, insurance policies, deeds, inspection reports and scanned amendments.

Table 3 breaks the coverage down by domain. As previously mentioned, any task that was answerable without reading the document was rejected.

Property	Details
Domains	10 professional domains, evenly distributed
Prompt curation	Questions written by domain experts from their own work, with the original PDFs and the collectors' notes
Document types	Policies and forms; multi-page tables; charts and infographics; datasheets; floor plans; long scanned deeds and guidelines
Screening rule	An item stays only if two strong frontier models both commit a major failure on it
Annotations	Reference answer, atomic rubric, domain label, capability tags, notes on the parsing challenge
Question styles	Lookup, comparison, reconciliation, calculation, summarization; some queries are deliberately unsupported
Reporting metrics	Mean rubric score for analysis; strict pass rate for comparing models
Public release	50 of the 100 examples on Hugging Face, released under Apache 2.0; third-party PDFs retain their original rights

Table 2. Profile of the current version of GDP.pdf.

3.4. Capability Taxonomy

Table 4 lists the eleven capability axes, organized into three tiers (extraction and grounding; structural and multimodal comprehension; advanced reasoning). We solicited prompts against this taxonomy, and the error analysis in Section 5.3 is written in its vocabulary. Most items carry more than one tag.

3.5. Collection and Curation Workflow

1. **Candidate sourcing.** A domain expert submits a task from their own work: the source PDF, a note on why the task matters in that line of work, and a first-pass reference answer.
2. **Challenge calibration.** Two strong frontier models (the screening models) attempt the candidate, and both have to commit at least one *major failure* for it to advance. Major means a materially wrong final answer, a dropped piece of decisive evidence, or a fabricated claim. Superficial differences in phrasing do not count as a failure.
3. **Novelty and ambiguity filtering.** A candidate is rejected at this stage if it is answerable from general priors, rests on context only the original contributor had, or is worded

Domain	Representative PDFs	How models typically fail
Finance	Earnings releases, investor filings, analyst materials	Values come from the adjacent column or the wrong note field; a table which crosses pages tends to lose its alignment.
Healthcare	Reviews, dosage tables, clinical guidelines	The model loses its place in a long review, skips a figure footnote, or answers when the information is simply not there.
Legal	Contracts, leases, policy language, filings	The cited clause is usually real. It is just not the governing one, and cross-references between sections get dropped.
STEM / Research	Climate reports, technical reports, scientific figures	Misread legends, wrong values off the charts, and evidence from several figures that never gets combined.
Engineering	Datasheets and specification sheets	Log-scale plots defeat the value reading; units get mixed up; merged headers scramble the lookup.
Construction	Floor plans and schedules	Symbols never get matched to the legend, so fixture and window counts come out wrong, and the schedule is never reconciled with the drawing.
Manufacturing	Process notes, packing lists, certificates of conformance, technical instructions	Footnotes get skipped, and where the instructions disagree with the model's priors, the priors usually win.
Insurance	Auto policies and endorsements	Reading stops at the insuring agreement, short of the exclusions; defined terms get taken at their everyday meaning.
Real Estate	Valuation reports, inspection reports, deeds, amendments	Status columns are misread, and superseded sections get quoted as if they were still in force.
HR	Benefits packets, leave-policy tables, personnel handbooks	The wrong tenure band gets read, and a date hidden in a footnote silently breaks the chronology.

Table 3. Coverage by domain in GDP.pdf.

Family	Axis	What the axis measures
Tier 1: Extraction & grounding	Correctness & completeness	The text the answer needs is extracted in full, with no dropped content and no altered facts
	Grounding	Every asserted fact can be checked against the PDF; nothing is supplied from prior knowledge or hallucinated
	Spatial awareness	When the prompt depends on location, the model knows where pieces of content lie on the page relative to one another
Tier 2: Comprehension (structural & multimodal)	Semantic reading flow	Columns, sidebars, and callouts are read in the order a person would read them
	Typographic hierarchy	Headings, bullets, emphasis, and section boundaries mean what they should
	Standard table parsing	Rows and columns stay aligned, merged headers and adjacent cells included
	Chart & multimodal interpretation	Legends, axes, and visual marks are interpreted correctly in context
Tier 3: Advanced reasoning	Complex & multi-page tables	Context carries over page breaks, nested headers, and merged cells
	Cross-referencing	Footnotes, citations, appendices, and distant definitions are connected to the text they modify
	Artifact & noise	Scan noise, running headers and footers, watermarks, and superseded content are correctly identified as such
	Unsupported queries	Where information is missing, redacted, or unsupported, the model states this instead of answering incorrectly

Table 4. The capability taxonomy used to author, tag, and analyze GDP.pdf tasks.

so ambiguously that the semantics of the question, rather than the document, becomes the bottleneck.

4. **Rubric authoring.** The expert writes atomic criteria covering the facts the answer must contain, the formulations that count as equivalent, and the claims the answer must not make. For abstention items the rubric states what the abstention has to say.
5. **Axis tagging and worker notes.** Capability tags are added with a note describing the parsing challenge in plain terms: merged cells, a footnote, a superseded section, a legend.
6. **Sanity checks.** The gold answer must pass its own rubric; a deliberately bad output must fail on the criterion the item was built around.

3.6. Data Curation Insights

Assembling the benchmark produced several observations worth recording. Length was a poor predictor of difficulty; several of the hardest items are one-line questions about a single table or figure. Comparisons across two sections, and amendments overriding an earlier line, produced confident hallucination in the best models far more often than expected. Table 5 gives a sample of items, the evidence deciding each, and the usual way most frontier models went wrong.

4. Evaluation Protocol

Atomic rubric design. GDP.pdf deliberately does not adopt ANLS-style overlap scoring [1, 20, 25]. A response can name nearly the same companies as the reference, in

Domain	Task sketch	Document evidence	Typical failure mode
HR	Order the first five companies to offer paid bereavement leave.	A multi-page table must be sorted by announcement date, but a footnote changes Mastercard's date and removes it from the top five.	The footnote is ignored, so the list comes back fluent but chronologically wrong.
Manufacturing	Recommend drill-profile features for cutting holes in Ti-8Al-1Mo-1V pipe.	The relevant section advises against brad-point bits and recommends W+R thinning with a 180° chamfer angle.	General machining priors override the document; models recommend brad-point bits and volunteer process advice that was not requested.
Healthcare	Identify the EDS subtype(s) associated with *TNXB* mutations and distinguish them.	The review supports clEDS, states that hEDS lacks a diagnostic genetic marker, and does not provide enough evidence for the requested hEDS comparison.	*TNXB* gets presented as if it were a marker for hEDS, and the key clinical distinction is reversed or dropped.
Insurance	Decide whether Medical Payments and UMBI apply after a highway collision involving a full-time RV.	Exclusions remove a vehicle used as a residence from the "uninsured motor vehicle" definition and bar medical-payments coverage for injuries it causes.	Reading stops at the insuring agreement; the exclusions are missed and coverage is granted incorrectly.
Engineering	Read threshold irradiance for a specified ambient brightness and wavelength from a datasheet.	The answer requires reading a precise point from a plotted curve with logarithmic scales and the correct source wavelength.	The right figure is identified, but the values read off the plot are implausible.
Construction	Determine how many of the most prominent window type appear on the first floor.	The schedule must be read correctly to identify the dominant type, then matched to plan labels on the first-floor drawing.	A wrong schedule row or quantity column is chosen, and the error propagates into the plan count.
Real Estate	Decide whether three "PENDING" comparable sales would change the automated valuation.	"PENDING" appears in the registration-date column; the sale prices are already reflected, and the PDF does not justify recalculating the estimate.	Pending registration is treated as missing price evidence, leading to a predicted value change the document does not support.

Table 5. Representative items. In most cases the decisive evidence is small and local, and is easy to miss even when the overall topic of the document is understood.

nearly the same words, and still be completely wrong, because the one company a footnote disqualifies is in the list, and an overlap metric will readily reward it. Every item therefore carries a rubric of atomic yes/no criteria.

Unsupported queries and abstention. Some tasks ask for what the document does not contain: the figure is redacted, the value never reported, the conclusion does not follow. To get credit the model has to plainly state this instead of hallucinating. Most models instead produce a fluent, helpful-sounding fabrication, which is scored zero.

Slice reporting. For any subset of items $\mathcal{I}$ defined by domain, tier, or capability axis, we report

$$\mathrm{Score}_{\mathcal{I}}(M) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} s_i(M).$$

For slices we report the rubric score, which retains the partial progress open-ended items produce. For ranking models against each other we use the strict pass rate (Table 6); that number credits fully solved items and nothing else.

Judge inputs and manual verification. The grader is given the question, the rubric, the model's response, and the PDF. The gold answer is withheld at grading time: it is used while authoring and sanity-checking the rubric but nowhere else, so that it cannot anchor the judgment. Every failure discussed in Section 5.3 was also checked by hand against the source document and the collector's notes.

Table 6. Pilot leaderboard on GDP.pdf. We report strict pass rate: the percentage of items on which a model satisfies *all* required atomic rubric criteria. No evaluated model exceeds 15%.

Model	Pass rate (%)
Gemini 3.1 Pro	15
Claude Opus 4.7	14
GPT-5.4	11
Grok 4.20 Beta	7
Kimi K2.5	6
Mistral Large 3	3
Nova 2 Pro	1

5. Pilot Evaluation

5.1. Evaluated Models

We evaluated seven frontier models, all through their public interfaces: Gemini 3.1 Pro, Claude Opus 4.7, GPT-5.4, Grok 4.20 Beta, Kimi K2.5, Mistral Large 3, and Nova 2 Pro. Each model was given the question and the source PDF for all 100 items. No tools and no additional context were provided.

5.2. Main Results

Table 6 reports the model scores.[2] No frontier model scores higher than a 15% strict pass rate. Even the best frontier

[2]The public GDP.pdf leaderboard at https://surgehq.ai/benchmarks/gdp-pdf is re-run periodically as new models and inference configurations become available, so the numbers reported there evolve over time and can differ from the frozen pilot snapshot in Table 6.

model fails roughly five items out of every six, and four of the seven models score 7% or below.

The hardest task slices were the spatial ones (such as floor plans and construction drawings), dense technical plots, and long noisy documents carrying amendments. These show up across the construction, engineering, manufacturing, insurance, and real-estate tasks. On HR and STEM/Research tasks, models scored higher when the task was a lookup in a single table or figure, but a footnote or a cross-page dependency was usually enough to break those as well.

5.3. Failure Analysis

Table alignment failures. The most common error was reading the wrong cell, wrong row or wrong column, and it concentrated in tables with merged cells or stacked headers. For example, an HR benefits item asks how plan costs change after an employee adds a dependent. One model picked the wrong tenure band. A second model picked the right band and then reported numbers which are not in it.

Chart and figure misreads. The model finds the right figure but misreads it. On an engineering datasheet, both screening models located the correct threshold-irradiance plot and returned values that fall well outside the plausible range of the plotted curves.

Footnotes, cross-references, and exclusions get dropped. In professional documents the controlling fact often sits in a footnote, or in a definition pages away from where the question gets answered. The bereavement-leave item contains a footnote which moves one company's date and reorders the ranking; most models never used it. In an insurance item, the models identified the at-fault driver correctly, then stopped at the insuring agreement and never reached the exclusions which decided whether the coverage applied.

Prior knowledge overrides the document. Where the document and the training prior disagree, the prior tended to win. As an example, every model when asked about drilling Ti-8Al-1Mo-1V pipe recommended brad-point bits. As general machining advice this is reasonable. The document in question, however, advises against brad-point bits for this application and calls for W+R-type thinning with a 180° chamfer angle, so the answers sounded helpful and were wrong for the specific document at hand.

Spatial problems are especially challenging. A floor-plan question forces the model to match small symbols to a legend and keep track of them from page to page, and the models could rarely do this without errors. On the lighting question they reported fixtures in rooms which have none and missed fixtures which exist. On the window question they miscounted, because the schedule on one page never got connected to the labeled first-floor plan on the next.

Noise and supersession can cause cascading errors. Long scanned files with amendments were challenging from the first round of curation to the last. Models quoted deed sections which an amendment had already replaced. In one automated valuation report, all seven models treated a "PENDING" label as though it undermined the associated sale prices. The label sits in the registration-date column and should have had no bearing on the estimate.

6. Discussion

Standard scores can be misleading. The models in Table 6 sit at or near the top of most public multimodal leaderboards, yet the best two passed only 15% and 14% of GDP.pdf items. The gap has a simple explanation: GDP.pdf checks something broader suites can skip, namely where the answer came from.

The interaction of perception and domain knowledge. Most failures could not be cleanly classified into perception errors and reasoning errors. Consider the RV insurance item: the models read the insuring agreement correctly, so perception was fine, and they granted coverage anyway, because they never went looking for the exclusion a policy reader would know to check.

A benchmark for the floor, not just the ceiling. Broader benchmarks mostly ask a ceiling question: how hard a task can this model accomplish in a controlled academic setting? Deploying models for document automation requires asking the floor question instead: what can the model be counted on to get right every single time? The 1–15% range in Table 6 is our answer for the present moment, and it is not a number that supports unsupervised use.

The document work behind economic activity. The tasks in GDP.pdf are not niche edge cases. An insurance adjuster reading exclusions, an HR analyst sorting leave announcements, an estimator counting windows: each task appears in this paper because a contributor actually does that work. Current broader benchmarks sample these skills thinly, and we suspect that explains much of the gap between leaderboard capability and practical reliability.

Implications for model development. Longer context windows alone might not fix this gap in model capabilities. The failures we saw would respond better to page representations that keep structure intact, to parsers that get charts and tables right, and to retrieval that does not strip away the visual evidence. Footnotes and amendments need to be

treated as content rather than noise. And a model with better uncertainty calibration would have scored considerably better on our abstention items.

7. Conclusion

We presented GDP.pdf, a benchmark for multimodal reasoning grounded in professional PDF documents, built from expert-authored tasks with the original files and their domain semantics intact. Seven frontier models took the pilot evaluation, and none exceeded a 15% strict pass rate. The failures concentrated where professional documents are hardest to read: tables, charts, footnotes, exclusions, noisy scans, amendments. We hope the benchmark proves useful for measuring whether models can handle the routine document work that a significant share of professional activity depends on.

References

[1] Ali Furkan Biten, Rubèn Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 4

[2] Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, Dongfu Jiang, Xuan He, Yuan Liu, Hexiang Hu, Xiang Yue, and Wenhu Chen. MEGA-Bench: Scaling multimodal evaluation to over 500 real-world tasks. In *International Conference on Learning Representations*, 2025. 2

[3] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. 2

[4] Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, and Rainer Stiefelhagen. RoDLA: Benchmarking the robustness of document layout analysis models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[5] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021. 2

[6] Yew Ken Chia, Liying Cheng, Hou Pong Chan, Maojia Song, Chaoqun Liu, Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-LongDoc: A benchmark for multimodal superlong document understanding and a retrieval-aware tuning framework. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2025. 2

[7] Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025. 2

[8] Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimodal generation for document question answering. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 2

[9] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 2

[10] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2

[11] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023. 2

[12] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *ICDAR Workshop on Open Services and Tools for Document Analysis*, 2019. 2

[13] Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024. 2

[14] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. 2

[15] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024. 2

[16] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 2

[17] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[18] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. 2

[19] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. 2

[20] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021. 2, 4

[21] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 2

[22] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 2

[23] 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. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[24] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. CORD: A consolidated receipt dataset for post-OCR parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. 2

[25] David Peer, Philemon Schöpf, Volckmar Nebendahl, Alexander Rietzler, and Sebastian Stabinger. ANLS* – a universal document processing metric for generative large language models. *arXiv preprint arXiv:2402.03848*, 2024. 4

[26] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter W. J. Staar. DocLayNet: A large human-annotated dataset for document-layout analysis. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. 2

[27] Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Seunghyun Yoon, Ryan A. Rossi, and Franck Dernoncourt. PDFTriage: Question answering over long, structured documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2024. 2

[28] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 2022. 2

[29] Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalický, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. DocILE benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition*, 2023. 2

[30] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for Multipage DocVQA. *Pattern Recognition*, 144:109834, 2023. 2

[31] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanisławek. Document understanding dataset and evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[32] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 2

[33] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[34] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025. 2

[35] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. PubLayNet: Largest dataset ever for document layout analysis. In *International Conference on Document Analysis and Recognition*, 2019. 2

[36] Changda Zhou, Ziyue Gao, Xueqing Wang, Tingquan Gao, Cheng Cui, Jing Tang, and Yi Liu. Real5-OmniDocBench: A full-scale physical reconstruction benchmark for robust document parsing in the wild. *arXiv preprint arXiv:2603.04205*, 2026. 2

[37] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021. 2

[38] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the ACM International Conference on Multimedia*, 2022. 2

[39] Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. MMDocBench: Benchmarking large vision-language models for fine-grained visual document understanding and grounding. In *Proceedings of the International Conference on Multimedia Modeling*, 2026. 2

[40] Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. DOCBENCH: A benchmark for evaluating LLM-based document reading systems. *arXiv preprint arXiv:2407.10701*, 2024. 2