GDP.PDF: A Benchmark for Grounded Multimodal Reasoning over Professional PDF Documents

Surge AI Research

Abstract

Professional work in finance, healthcare, law, engineering, construction, insurance, real estate, and scientific research is still overwhelmingly mediated by PDFs. Yet existing evaluations typically isolate one slice of the problem: OCR, layout analysis, chart reasoning, table QA, or document VQA, rather than testing whether a system can answer realistic, domain-grounded questions over visually complex, multi-page professional PDFs. We introduce GDP.PDF*, a multimodal reasoning challenge benchmark of expert-authored document–question pairs spanning ten professional domains. Items are organized by a three-tier, eleven-axis taxonomy covering extraction fidelity, grounding, reading order, tables, charts, cross-references, spatial reasoning, noise robustness, and abstention on unsupported queries. Candidate prompts are retained only when two strong frontier multimodal models both exhibit a major failure, making the benchmark explicitly diagnostic rather than merely representative. Each item includes an expert rubric decomposed into atomic grading criteria, enabling fine-grained scoring as well as a conservative strict-pass metric for model-level comparison. In a pilot evaluation of seven frontier multimodal models, no evaluated model exceeds 15% pass rate, with recurring failures in table alignment, chart reading, footnotes and exclusions, spatial plans, and superseded or noisy documents.* GDP.PDF *is designed for evaluation and benchmarking of document-grounded reliability in knowledge-intensive multimodal reasoning.*

1. Introduction

Recent multimodal models perform strongly on broad visual reasoning suites, but many high-value deployments look nothing like textbook visual QA. They look like reading a benefits packet to compare plan tiers, locating an indemnification clause in a lease, reconciling a packing list against a certificate of conformance, following a clinical flowchart whose meaning changes in a footnote several pages away, or counting fixtures from a floor plan while respecting a legend. These are ordinary professional tasks, and PDFs are the format in which this work is commonly performed.

We argue that current evaluation misses an important conjunction of capabilities. First, professional PDFs are structurally difficult: multi-page tables, sidebars, footnotes, legends, callouts, scanned artifacts, and amendments that supersede earlier content. Second, success is often knowledge-intensive: the right answer depends not only on perceiving the page but also on understanding domain semantics such as policy exclusions, regulatory wording, benefits tiers, or engineering conventions. Third, the dominant failure mode is not always blank inability but confident mis-grounding: a polished answer that cites the wrong clause, wrong row, or wrong diagram element.

Existing benchmark families capture parts of this landscape, but few jointly test realistic workflow questions, original professional PDFs, and grounded reliability in high-stakes domains. This paper introduces GDP.PDF, a benchmark designed around that gap. The current benchmark version contains expert-authored document–question pairs across ten professional domains: *Finance*, *Healthcare*, *Legal*, *STEM/Research*, *Engineering*, *Construction*, *Manufacturing/Supply Chains*, *Insurance*, *Real Estate*, and *Human Resources*. Each item is paired with an expert rubric and capability-axis tags, and candidate items are retained only when two strong frontier models fail materially on the same document-grounded task.

Contributions.

1. We present GDP.PDF, a new multi-domain frontier benchmark for evaluating knowledge-intensive reasoning over domain-specific PDFs, with expert-authored items spanning ten real workflow domains.
2. We formalize the benchmark through a three-tier, eleven-axis taxonomy that decomposes professional PDF understanding into interpretable capabilities including grounding, tables, charts, cross-referencing, spatial parsing, robustness, and abstention.
3. We use a rubric-element based evaluation for open-ended document questions, pairing diagnostic rubric score with a conservative strict-pass metric for model-level comparison.
4. We provide a pilot evaluation of frontier models and a detailed failure analysis showing recurring weaknesses

that are consequential in real professional settings.

2. Related Work

Document parsing, OCR, and layout analysis. Document understanding benchmarks have long focused on extraction, layout, and OCR. Representative resources include FUNSD for noisy forms [11], CORD for receipts [23], PubLayNet [34], DocLayNet [25], and DocILE for business-document localization and extraction [30]. More recent robustness- and parsing-oriented evaluations include OCRBench [13], OCRBench v2 [8], OmniDocBench [22], Real5-OmniDocBench [35], RoDLA [3], and MMDocBench [38]. These resources are foundational, but they primarily emphasize page parsing, OCR, or layout robustness rather than end-to-end reasoning over professional workflow questions.

Document QA, long documents, and retrieval-aware systems. Document VQA benchmarks moved the field toward end-to-end question answering: DocVQA [19], InfographicVQA [20], MP-DocVQA [28], DUDE [29], and TAT-DQA [37]. Recent long-document and raw-file evaluations include DocBench [39], MMLongBench-Doc [15], LongDocURL [6], and M-LongDoc [5]. Retrieval-aware systems and benchmarks such as PDFTriage [26] and MMDocRAG [7] highlight the importance of evidence selection across long multimodal documents. GDP.PDF is complementary: it focuses on professional PDFs where the core problem is often not merely finding a page but grounding the answer correctly under domain-specific semantics, footnotes, exclusions, legends, and superseded content.

Charts, tables, and domain-specific reasoning. Chart and table understanding have their own mature benchmark lines, including ChartQA [17], PlotQA [21], CharXiv [31], ChartQAPro [18], HybridQA [2], TAT-QA [36], and FinQA [4]. Domain-specific reasoning benchmarks such as FinanceBench [10], LegalBench [9], and ClinicBench [12] target specialized knowledge. However, many of these evaluations operate over extracted text, isolated tables, or standalone charts rather than messy, multi-page PDFs in which the answer depends on reading the original artifact correctly.

Broad multimodal evaluation. Broad benchmarks such as MMMU [33], MMMU-Pro [32], MathVista [14], and MEGA-Bench [1] are invaluable for tracking general multimodal progress, while knowledge-intensive VQA datasets such as OK-VQA [16] and A-OKVQA [27] test knowledge use on natural images. Their breadth is a strength, but it is not a substitute for a specialized benchmark focused on the reliability of professional PDF reasoning.

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

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

For conservative model-level comparison, we also define a strict item pass 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

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

Rubric score preserves within-item gradations for diagnostic slicing; strict pass rate treats a task as solved only when all required factual obligations are satisfied. In this paper, failure analysis relies on the rubric formalism, while Table 6 reports strict pass rate.

3.2. Design Principles

Document dependence. The answer must genuinely depend on the provided PDF rather than recoverable world knowledge. Tasks are chosen so that generic priors are insufficient and often harmful.

Workflow realism. Prompts are authored from real professional workflows, not synthetic stress tests. Experts write the question in the form they would naturally ask of an assistant while doing their job. In fact, contributors were encouraged to submit prompts corresponding to real tasks that they have recently performed in their occupations.

Adversarial challenge-set construction. GDP.PDF is intentionally a challenge benchmark. A candidate enters the

Benchmark family	Representative work	Evaluates well	Under-measured relative to GDP.PDF
Layout / OCR / parsing	FUNSD, CORD, PubLayNet, DocLayNet, DocILE, OmniDocBench	OCR, layout detection, extraction, document parsing quality	End-to-end answers to realistic professional questions whose correctness depends on domain semantics and grounded interpretation
Document QA / long-doc	DocVQA, MP-DocVQA, DUDE, DocBench, MMLongBench-Doc, M-LongDoc	QA over raw or long documents, often with multi-page evidence	High-stakes professional workflows with exclusions, footnotes, spatial plans, amendments, and adversarial challenge-set construction
Chart / table / domain	ChartQA, CharXiv, TAT-QA, FinQA, FinanceBench, LegalBench	Reasoning over charts, tables, or specialized domains in more controlled settings	Joint perception of the original professional PDF plus grounded reliability under realistic workflow prompts
Broad multimodal	MMMU, MMMU-Pro, MathVista, MEGA-Bench	General multimodal reasoning breadth	Fine-grained diagnosis of document-grounded reliability in professional PDF settings
This work	**GDP.PDF**	**Knowledge-intensive reasoning over professional PDFs with rubric-based grounded scoring**	**—**

Table 1. Positioning GDP.PDF relative to neighboring benchmark families. The main gap is not any single phenomenon in isolation, but the conjunction of realistic documents, domain-specific workflow semantics, and grounded reliability.

pool only when two strong frontier multimodal models both exhibit at least one major failure. As a result, the benchmark is optimized for diagnosis of present capability gaps rather than estimation of average task success in everyday office use.

Knowledge-intensive grounding. Items are selected to expose the interaction between perception and domain semantics: exclusions, footnotes, legends, plan symbols, superseding amendments, and other professional conventions that change what the "right" answer actually is.

Diagnostic granularity. Each item is tagged with one or more capability axes so that performance can be decomposed by failure mode rather than reported only as a single aggregate.

3.3. Benchmark Profile and Domain Coverage

Table 2 summarizes the current benchmark version. The collection is intentionally heterogeneous at both the domain and artifact levels: financial filings, benefits packets, technical datasheets, clinical reviews and guidelines, floor plans, insurance policies, deeds, inspection reports, and scanned amendments all appear in the benchmark. This heterogeneity matters because many models that tolerate clean reports fail once evidence is split across schedules, figures, footnotes, legends, and later superseding sections.

Table 3 expands the domain coverage. Some questions are primarily parsing-heavy, while others mix parsing with constrained downstream reasoning because that is intrinsic to the workflow. We exclude prompts that can be answered mainly from general knowledge without careful document reading.

Property	Details
Domains	10 professional domains; even distribution across domains
Prompt source	Expert-authored real workflow tasks paired with source PDFs and collector notes
Document regimes	Short policies and forms; multi-page tables; charts and infographics; technical datasheets; floor plans; long scanned deeds and guidelines
Screening rule	Candidate retained only if two strong frontier models both exhibit a major failure
Annotations	Reference answer, atomic rubric, domain label, capability-axis tags, and notes on the parsing challenge
Question styles	Lookup, comparison, reconciliation, calculation, summarization, and unsupported-query detection
Reporting metrics	Mean rubric score for fine-grained analysis; strict pass rate for model-level leaderboard reporting

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

Domain	Representative PDFs	Typical task form	Recurring challenge modes
Finance	Earnings releases, investor filings, analyst materials	Extract revenue, margins, or year-over-year changes	Adjacent numeric columns, note fields, cross-page tables, precise numeric grounding
Healthcare	Reviews, dosage tables, clinical guidelines	Map mutation to subtype; follow a guideline branch; extract treatment detail	Long-document navigation, figure footnotes, absent information, high-stakes abstention
Legal	Contracts, leases, policy language, filings	Locate indemnification, liability-cap, or exclusion language	Dense prose, precise clause grounding, cross-references, section scoping
STEM / Research	Climate reports, technical reports, scientific figures	Aggregate evidence from maps, figures, and narrative text	Legend reading, chart extraction, multi-figure synthesis
Engineering	Datasheets and specification sheets	Read thresholds, tolerances, or settings from plots and tables	Log-scale plots, unit sensitivity, merged headers, dense technical notation
Construction	Floor plans, schedules, and county plan sets	Count fixtures or windows; verify plan comments against drawings	Spatial legends, symbol matching, schedule–plan alignment, small visual marks
Manufacturing	Process notes, packing lists, CoCs, technical instructions	Reconcile shipments or select process parameters from instructions	Reconciliation across sections, footnotes, prior-vs-document conflicts
Insurance	Auto policies and endorsements	Determine whether coverages apply under definitions and exclusions	Non-linear policy reading, exclusion logic, term scope, document-grounded fault analysis
Real Estate	Valuation reports, inspection reports, deeds, amendments	Interpret comparable-sales status; summarize urgent repairs; identify current deed terms	Status-column semantics, scanned noise, superseded sections, long-form amendments
HR	Benefits packets, leave-policy tables, personnel handbooks	Compare plans; order announcements; read tenure-banded entitlements	Multi-page tables, tenure bands, footnote dates, chronology from scattered evidence

Table 3. Coverage by domain in GDP.PDF. The same model must handle qualitatively different artifact types and evidence patterns across domains.

3.4. Capability Taxonomy

We organize the benchmark around a three-tier taxonomy with eleven capability axes. The taxonomy serves two roles: it guides prompt authoring and it supports diagnostic error analysis. Real failures are often compositional, so items may be tagged with multiple axes.

3.5. Collection and Curation Workflow

The benchmark is built through an expert-in-the-loop workflow intended to maximize realism while keeping each item auditable.

(1) Candidate sourcing. A domain expert contributes a real task, the source PDF, an explanation of why the task matters in practice, and an initial reference answer.

(2) Challenge calibration. Each candidate is run against two strong frontier models. A candidate advances only when both incur at least one *major failure*: a materially wrong final answer, omission of decisive evidence, or unsupported invention. Benign stylistic variation does not qualify.

(3) Novelty and ambiguity filtering. We reject items that can be answered from general priors, rely on hidden worker context, or hinge on ambiguous wording instead of document understanding.

(4) Rubric authoring. For retained items, the expert decomposes the target into atomic criteria covering required facts, acceptable equivalences, and disallowed claims. Unsupported-query items also specify what a correct abstention must say.

(5) Axis tagging and worker notes. Items receive one or more capability-axis tags, and the collector records the intended parsing challenge (for example merged cells, footnotes, superseded content, or spatial legend matching) so that later analysis can separate broad trends from one-off artifacts.

(6) Sanity checks. Gold answers must satisfy the rubric, and known failing outputs are checked against it to ensure that the intended error is penalized consistently.

3.6. Frontier Challenge-Set Characterization

Pilot collection surfaced three recurring patterns. First, prompt length is a poor proxy for difficulty. Some of the hardest items are short, single-sentence questions about a single table, figure, or plan; the challenge lies not in linguistic complexity, but in precise grounding to the correct row, column, symbol, or annotation. Second, reconciliation tasks are especially diagnostic. Questions that require comparing two sections, integrating a footnote with a main table, or determining whether a later amendment supersedes an earlier statement reliably induce confident but brittle errors. Third, some items intentionally combine parsing with limited downstream reasoning. We retain such cases when that reasoning is native to the workflow: for example, ordering companies after incorporating footnote-corrected dates, or determining that a requested conclusion is unsupported because the document does not provide the necessary evidence. Accordingly, GDP.PDF should be understood as a benchmark for document-grounded task completion rather than for OCR or layout analysis in isolation. It is a deliberate challenge set designed to expose consequential failure modes, not a population-level estimate of routine office work.

4. Evaluation Protocol

Atomic rubric design. Professional document tasks often require open-ended answers with multiple factual obligations. Surface metrics such as string overlap or ANLS [19, 24] are therefore poorly aligned with correctness: a fluent

Family	Axis	What the axis measures
Tier 1: Foundational extraction & grounding	Textual fidelity & completeness	Whether the model extracts needed text without dropping critical content or altering key facts
	Grounding (zero hallucination)	Whether asserted facts are verifiable in the PDF rather than supplied from prior knowledge or fabrication
	Basic spatial awareness	Whether the model correctly identifies relative page positions when the prompt depends on location
Tier 2: Structural & multimodal comprehension	Semantic reading order	Whether the model follows human reading flow instead of flattening columns, sidebars, or callouts
	Typographical hierarchy	Whether headings, bullets, emphasis, and section boundaries are interpreted correctly
	Standard table parsing	Whether rows, columns, merged headers, and adjacent cells are aligned correctly
	Chart & multimodal interpretation	Whether legends, axes, and visual marks are read correctly in context
Tier 3: Advanced reasoning & robustness	Complex & multi-page tables	Whether context is preserved across page breaks, nested headers, and merged cells
	Relational cross-referencing	Whether footnotes, citations, appendices, and distant definitions are linked correctly
	Artifact & noise resilience	Whether the model ignores scan noise, headers/footers, watermarks, and superseded content
	Unanswerable query detection	Whether the model abstains when information is missing, redacted, or unsupported

Table 4. The eleven-axis capability taxonomy used to author, tag, and analyze GDP.PDF items.

Domain	Task sketch	Decisive 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.	Models ignore or fail to integrate the footnote, producing a polished but chronologically wrong list.
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.	Models override the PDF with machining priors, recommend brad-point bits, and add generic process advice 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.	Models present *TNXB* as if it were a marker for hEDS and reverse or omit the key clinical distinction.
Insurance	Decide whether Medical Payments and UMBI apply after a highway collision involving a full-time RV.	Exclusions specify that a vehicle used as a residence is not an uninsured motor vehicle for the relevant coverages.	Models stop at the insuring agreement, miss the exclusion, and incorrectly grant coverage.
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.	Models identify the right figure but extract implausible values from the plot.
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.	Models choose the wrong schedule row or quantity column, then propagate the mistake 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.	Models treat pending registration as missing price evidence and predict a value change that the document does not support.

Table 5. Representative benchmark items. The decisive evidence is often small, local, and easy to miss even when the overall document topic is understood.

answer may be lexically similar to the reference while citing the wrong clause or wrong table row. We therefore evaluate against expert rubrics written as *atomic* criteria. A criterion may require, for example, naming the correct company set, excluding an ineligible company after footnote correction, or refraining from unsupported extrapolation. Partial credit is implemented by decomposition into multiple binary elements rather than by fuzzy scalar judgments.

Inclusion-time failure versus benchmark score. The dataset inclusion rule and the evaluation metrics are inten-

tionally different. A candidate enters the benchmark only when two calibration models both exhibit a *major failure*. At evaluation time, models receive continuous rubric scores together with a derived strict-pass indicator. This distinction matters: an item can be hard enough for calibration yet still allow one model to recover some rubric elements correctly without fully solving the task.

Unsupported queries and abstention. Some items are negative-constraint questions: the requested information is absent, blurred, redacted, or not inferable from the PDF. For these items, full credit requires an explicit statement that the document does not support the requested conclusion and zero invented specifics. Helpful fabrication receives zero on the relevant criteria.

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

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

We use rubric score for these diagnostic slices because it preserves partial progress on open-ended items. For compact model-level comparison, Table 6 instead reports strict pass rate, which is the more conservative summary of whether a system fully solves a professional document task.

Judge inputs and manual verification. The grader receives the original question, the expert rubric, the model response, and the source PDF. Gold answers are used during rubric authoring and sanity checks, but they are not exposed to the grader at evaluation time in order to reduce anchoring. All qualitative failures discussed in Section 5.3 were manually verified against the source documents and expert notes, which is important because many errors are subtle but consequential.

5. Pilot Evaluation

5.1. Evaluated Models

We evaluate seven frontier models available through public model 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 system receives the full question text and the source PDF.

5.2. Main Results

The central quantitative finding is important: no evaluated model exceeds 15% strict pass rate. Even the top-ranked system fails on roughly 85% of the current benchmark, and four of the seven evaluated models remain at 7% or below. On this challenge set, apparently basic tasks still fail once the document contains merged tables, footnotes, plan symbols, buried exclusions, or superseded sections.

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

Qualitatively, the hardest pilot slices involve floor plans and construction drawings, dense technical plots, policies with buried exclusions, and long noisy documents with superseding amendments. These patterns recur across construction, engineering and manufacturing, insurance, and scanned real-estate materials in the current collection. Some HR and STEM/Research items are somewhat more tractable when the task reduces to a clean table or figure lookup, but even these cases remain brittle once footnotes or cross-page dependencies enter the workflow.

5.3. Failure Analysis

1. Table alignment failures. The most common error is pulling values from the wrong row or column once a table contains merged cells, nested headers, or a small inference step. In an HR benefits item, one system chose the wrong tenure band when comparing plan costs after the employee added a dependent; another selected the right band but hallucinated the values. In a construction window-schedule example, multiple models chose the wrong row entirely because they misread which column represented total quantity.

2. Chart and figure misreads. Models often identify the correct figure but extract the wrong value from it. In an engineering datasheet prompt, both models located the correct threshold-irradiance plot yet read off implausible values from logarithmic axes. In a climate-report item, models failed to recover year-by-year hurricane percentages from a dense infographic even when the task reduced to reading and aggregating document-specific data.

3. Footnotes, cross-references, and exclusions get dropped. Professional documents frequently hide the decisive fact in a footnote or later definition. In an HR prompt about bereavement leave, at least one critical company date appears in a footnote that changes the chronological ranking, yet models ignored it. In an insurance scenario, models

correctly identified the at-fault driver and still failed overall because they stopped after the insuring agreement and missed exclusions that changed coverage applicability.

4. Prior knowledge overrides the document. When the PDF contradicts a model's general prior, the prior often wins. In a manufacturing prompt about drilling Ti-8Al-1Mo-1V pipe, every model recommended brad-point bits even though the PDF explicitly advises against them and instead recommends W+R thinning with a 180° chamfer angle. The responses were helpful-sounding but document-groundedly wrong.

5. Spatial plans are especially brittle. Floor plans, schedules, and symbol legends require matching small visual marks across pages. In a lighting-fixture prompt, models hallucinated fixtures in rooms that had none, missed fixtures that were present, and confused similar symbols. In another construction example, models miscounted windows after failing to connect the schedule on one page with the labeled first-floor plan on another.

6. Noise and supersession cause cascading errors. Long scanned documents with amendments remain difficult. In a real-estate deed example, models cited outdated sections even though later amendments superseded them. In an automated valuation report, all models treated "PENDING" as if it affected the sale value, apparently missing that the term appeared in the registration-date column and therefore should not change the estimated valuation.

6. Discussion

Why standard scores can be misleading. A model can look strong on OCR, chart QA, or broad multimodal tests and still fail the actual workflow because the last mile is document-grounded reliability. GDP.PDF exposes that last mile: not whether a system can say something reasonable about a page, but whether it can stay faithful to the right clause, row, legend, footnote, and revision.

The interaction of perception and domain knowledge. Many errors are neither purely perceptual nor purely reasoning failures. The system must both perceive the artifact correctly and understand what matters in context. The insurance and healthcare examples are instructive: the decisive content is visible, but only a domain-aware reading of exclusions, subtype distinctions, or risk definitions yields the right answer.

A benchmark for the floor, not only the ceiling. Many high-profile multimodal evaluations ask what models can do at their best on difficult but well-posed academic tasks. GDP.PDF instead asks what models should do reliably before they are trusted with routine professional document work. That shift in emphasis is well aligned with evaluation and benchmarking for high-stakes multimodal systems.

The quiet document work behind economic activity. The capability axes emphasized by GDP.PDF are not edge cases. Many economically consequential workflows are still mediated by dense PDFs: contracts, claims policies, medical guidelines, invoices, benefits packets, earnings reports, and construction plan sets. The associated work can often seem mundane: reading a schedule correctly, tracing a footnote, reconciling an amendment, matching a legend to a diagram, or recognizing that a document does not support the requested conclusion. Yet these are precisely the capabilities that broad multimodal benchmarks tend to under-measure. This helps explain why models that appear strong on visually salient or academically convenient tasks can remain brittle on the paperwork that actually governs professional decisions.

Implications for model development. The error patterns suggest that better document understanding will require more than larger context windows. Promising directions include structure-aware page representations, stronger chart and table parsing, explicit handling of footnotes and amendments, retrieval that preserves visual evidence, and better calibration on unsupported questions.

7. Conclusion

We introduced GDP.PDF, a frontier benchmark for evaluating grounded multimodal reasoning over professional PDF documents. The benchmark targets the conjunction of realistic workflows, original PDFs, domain-specific semantics, and grounded reliability. In a pilot evaluation of frontier models, no evaluated model exceeds 15% strict pass rate, and all exhibit recurring failures on tables, charts, footnotes, exclusions, spatial plans, and noisy or superseded documents. We hope GDP.PDF helps shift evaluation from isolated parsing subproblems toward the more practical question of whether frontier models can reliably do the quiet document work that professional settings depend on.

References

[1] 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 *ICLR*, 2025. 2

[2] Wenhu Chen, Hanwen Zha, et al. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of EMNLP*, 2020. 2

[3] Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, and Rainer Stiefelhagen. Rodla: Benchmarking the robustness of document layout analysis models. *arXiv preprint arXiv:2403.14442*, 2024. 2

[4] 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 *EMNLP*, pages 3697–3711, 2021. 2

[5] Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *EMNLP*, 2025. 2

[6] Chao Deng, Jiale Yuan, et al. LongDocURL: A comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*, 2024. 2

[7] Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimodal generation for document question answering. *arXiv preprint arXiv:2505.16470*, 2025. 2

[8] Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2025. 2

[9] Neel Guha, Julian Nyarko, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *NeurIPS*, 2023. 2

[10] Pranab Islam et al. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023. 2

[11] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDAR Workshops*, 2019. 2

[12] Fenglin Liu et al. Large language models in the clinic: A comprehensive benchmark. In *EMNLP*, 2024. 2

[13] Yuliang Liu, Zhang Li, Biao Yang, Wenwen Yu, Chenggang Shi, Xu Zhao, Lin Yuan, Danmeng Wang, Mengchao Li, et al. Ocrbench: On the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 2024. 2

[14] Pan Lu et al. MathVista: Evaluating mathematical reasoning in visual contexts. In *ICLR*, 2024. 2

[15] Yubo Ma, Yuhang Zang, Liangyu Chen, et al. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. In *NeurIPS*, 2024. 2

[16] Kenneth Marino et al. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 2

[17] Ahmed Masry, Do Xuan Long, et al. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL*, 2022. 2

[18] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025. 2

[19] Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. DocVQA: A dataset for VQA on document images. In *WACV*, 2021. 2, 4

[20] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. InfographicVQA. In *WACV*, 2022. 2

[21] Nitesh Methani et al. PlotQA: Reasoning over scientific plots. In *WACV*, 2020. 2

[22] Linke Ouyang, Yuan Qu, et al. OmniDocBench: Benchmarking diverse PDF document parsing with a multi-dimensional framework. *arXiv preprint arXiv:2412.07626*, 2024. 2

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

[24] David Peer, Philemon Schöpf, et al. ANLS*: A universal document processing metric for generative LLMs. *arXiv preprint arXiv:2402.03848*, 2024. 4

[25] 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 *KDD*, 2022. 2

[26] Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Ryan A. Rossi, and Franck Dernoncourt. PDFTriage: Question answering over long, structured documents. In *EMNLP Industry Track*, 2024. 2

[27] Dustin Schwenk et al. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022. 2

[28] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page DocVQA. *Pattern Recognition*, 2023. 2

[29] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, et al. DUDE: Document understanding dataset and evaluation. In *ICCV*, 2023. 2

[30] 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 *ICDAR*, 2023. 2

[31] 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. *Advances in Neural Information Processing Systems*, 2024. 2

[32] 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 *ICLR*, 2025. 2

[33] Xiang Yue et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024. 2

[34] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: Largest dataset ever for document layout analysis. In *ICDAR*, 2019. 2

[35] Changda Zhou et al. Real5-OmniDocBench: Full-scale physical reconstruction benchmark for document parsing. *arXiv preprint arXiv:2603.04205*, 2026. 2

[36] Fengbin Zhu, Wenqiang Lei, et al. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *ACL*, 2021. 2

[37] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *ACM Multimedia*, pages 4857–4866, 2022. 2

[38] Fengbin Zhu et al. MMDocBench: Benchmarking fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*, 2024. 2

[39] Anni Zou et al. DocBench: A benchmark for evaluating LLM-based document reading systems. *arXiv preprint arXiv:2407.10701*, 2024. 2