SynthDocBench: Controlled Benchmark for Long-Context Visual Document Understanding

Abhigya Verma[1*] , **Khyati Mahajan**[1*] , **Amit Kumar Saha**[1*] , **Shruthan Radhakrishna**[1] , **Sagar Davasam**[1] , **Vikas Yadav**[1] , **Sai Rajeswar**[1, 2, 3]

[1]ServiceNow AI [2]Mila [3]Université de Montréal

Vision language models (VLMs) have achieved strong performance on visual document understanding benchmarks such as DocVQA, ChartQA, and MMLongBench-Doc. However, real-world documents combine multiple factors such as length, layout complexity, modality, and question difficulty, which makes it difficult to attribute model failures to specific causes. We introduce SYNTHDOCBENCH, a fully synthetic benchmark for long-context visual document understanding that systematically controls factors including document length, layout structure, modality composition, and question type. The benchmark is constructed using a combinatorial design, each factor is varied independently across generated documents, enabling controlled analysis of model behavior. Documents are generated end to end using an LLM pipeline across six layout archetypes, with a 40 percent random override to prevent models from exploiting spurious correlations. Additionally, SynthDocBench spans long-context documents with substantially greater length and structural diversity than existing benchmarks. Evaluating seven frontier VLMs, we uncover three failure modes that existing benchmarks cannot surface: sharp degradation with document length, a systematic positional sensitivity in which the middle third of a document is hardest for five of six models and five of six models show a negative Early→Late trend (steepest decline: 8.3 percentage points), and breakdown of chart comprehension in long-document settings. These results suggest that current models may be overfitting to benchmark artifacts rather than achieving robust long-context visual document understanding.

Correspondence: abhigya.verma@servicenow.com
Code: https://github.com/ServiceNow/SynthDocBench
Dataset: https://huggingface.co/datasets/ServiceNow-AI/SynthDocBench

[*]Equal contribution.

servicenow

1 Introduction

Understanding long, visually rich documents is a defining challenge for vision language models (VLMs). Real-world documents interleave text, tables, charts, and complex layouts across dozens or hundreds of pages, demanding both long-range retrieval and cross-modal reasoning. Benchmarks such as DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), and MMLongBench-Doc (Ma et al., 2024) have driven substantial progress, yet this progress conceals a fundamental diagnostic blind spot: *when a model fails on a real document, it is difficult to know why.*

On single-page tasks, evaluation is approaching saturation. Frontier models exceed 95% on DocVQA (Bai et al., 2025a; Wang et al., 2025a) and 89% on ChartQA (Anthropic, 2025; Bai et al., 2025a) — though harder benchmarks such as ChartQAPro (Masry et al., 2025), VisuLogic (Zhang et al., 2025), and ChartMuseum (Xu et al., 2025) show that chart understanding is far from solved, with degradations exceeding 30 percentage points. Yet all evaluate charts in isolation, abstracted from the multi-page contexts in which they naturally occur. Long-context benchmarks like MMLongBench-Doc (Ma et al., 2024) address the document-length dimension, targeting long PDF comprehension with rich visual content, and the strongest model achieves only 57%. LongDocURL (Deng et al., 2025) and M-LongDoc (Chia et al.,

2025) extend to documents spanning 100s of pages, but prioritize breadth of coverage over controlled diagnosis: neither constructs questions requiring joint reasoning over charts and textually distributed evidence across distant pages. Furthermore, these benchmarks draw on real documents, confounding potential sources of difficulty (answer depth, presentation modality, layout density, cross-page evidence integration) which co-vary and cannot be disentangled.

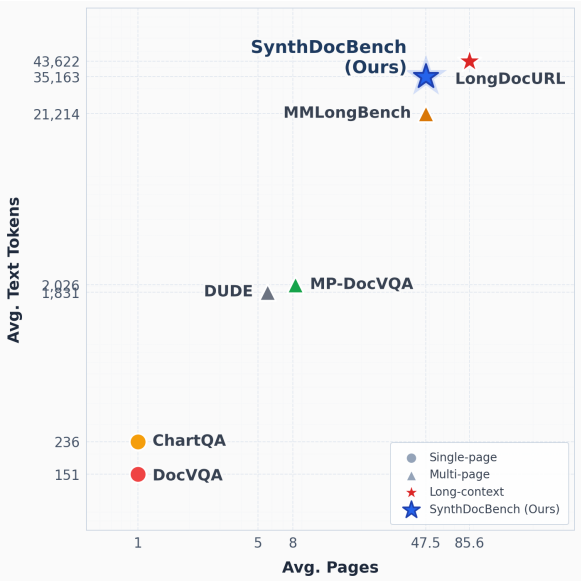
Synthetic benchmarks have a long history of enabling precisely this kind of decomposition. The bAbI tasks (Weston et al., 2016) and SCAN (Lake & Baroni, 2018) used programmatic generation to isolate reasoning primitives in NLP; CLEVR (Johnson et al., 2017) and RAVEN (Zhang et al., 2019) did the same for visual relational and analogical reasoning; and PuzzleVQA (Chia et al., 2024) recently applied synthetic abstract patterns to diagnose multimodal reasoning bottlenecks in VLMs. The common principle is that synthetic control trades ecological validity for interpretability, enabling attribution of failures to specific causes rather than an opaque bundle of confounds. No comparable instrument exists for long-context visual document understanding, an important setting where confounds are arguably most severe.

We address this gap with **SYNTHDOCBENCH**, a fully synthetic, controlled benchmark that enables the first systematic decomposition of VLM failure modes in long-context document understanding by varying document length, page depth, modality composition, and question type as independent axes. We contribute:

1. SYNTHDOCBENCH, a controlled synthetic benchmark with independently variable length, depth, modality, and question-type axes, released as three task-specific subsets, which probes 24 distinct D3.js chart types (including visually ambiguous forms such as dumbbell, lollipop, slope, and sparkline grids) requiring exact numerical reads that cannot be inferred from surrounding text; `cross_modal`, which places supporting charts and corroborating text in separate, non-adjacent sections to test cross-modal grounding over long-range dependencies; and `complex`, which requires combining 2 to 4 evidence units from text and charts across difficulty levels L1 to L5, ranging from direct value lookup to cross-section synthesis.
2. A fully automated LLM-based generation pipeline with a *dual-layer document design*: every chart is generated simultaneously as a rendered D3.js visualisation and as a hidden structured metadata block used only for ground-truth derivation, making answers deterministic by construction across 24 chart types and 6 layout archetypes.
3. A systematic empirical analysis of seven frontier VLMs, validated with a cross-judge robustness check (GPT-5 and Gemini-as-judge agree within 3.5 ACC points, $r \geq 0.94$, across all question types), revealing three concurrent, previously unobservable failure modes: sharp performance degradation with increasing evidence complexity and reasoning depth (L1→L5), systematic positional sensitivity in which the middle section of a document is hardest for five of six models and five of six models exhibit a negative Early→Late trend (steepest decline: 8.3 pp), and collapse of precise chart-reading accuracy in long-document contexts, even for models that perform well on existing benchmarks.

2 Related Work

Charts and Visual Reasoning. Chart understanding has been evaluated primarily in isolation from document context. ChartQA (Masry et al., 2022) established the standard evaluation setup for chart-oriented VQA, now approaching saturation. ChartQAPro (Masry et al., 2025) addresses this through harder, more diverse real-world charts, exposing over 30 percentage points of degradation for models that saturate ChartQA, though it retains the isolated-image setting. ChartGalaxy (Li et al., 2025) demonstrates that programmatic generation can yield million-scale synthetic diversity spanning 75 chart types and 330 stylistic variations, providing a methodological precedent for SYNTHDOCBENCH's synthetic approach. VisuLogic (Zhang et al., 2025) and ChartMuseum (Xu et al., 2025) probe fine-grained logical and perceptual reasoning over isolated charts, while MultiChartQA (Zhu et al., 2025) extends this to simultaneous multi-chart reasoning. Collectively, these benchmarks establish



Benchmark	Scope	Avg. Context	#Docs/Charts	#Questions
Chart Benchmarks				
ChartQA (Masry et al., 2022)	Isolated charts	1 chart	21,953	32,701
ChartQAPro (Masry et al., 2025)	Isolated charts	1 chart	1,341	1,948
ChartGalaxy (Li et al., 2025)	Isolated charts	1 chart	1,763,189	—
VisuLogic (Zhang et al., 2025)	Isolated images	1 image	1,000	1,000
ChartMuseum (Xu et al., 2025)	Isolated charts	1 chart	928	1,162
MultiChartQA (Zhu et al., 2025)	Multi-chart	2–3 charts	655	944
Document VQA Benchmarks				
DocVQA (Mathew et al., 2021)	Single-page docs	1 page	12,767	50,000
MP-DocVQA (Tito et al., 2023)	Multi-page docs	≤20 pages	—	46,000
SlideVQA (Tanaka et al., 2023)	Multi-page docs	∼20 pages	2,619	14,500
MMLongBench-Doc (Ma et al., 2024)	Multi-page docs	47.5 pages	135	1,082
LongDocURL (Deng et al., 2025)	Multi-page docs	∼83 pages	396	2,325
M-LongDoc (Chia et al., 2025)	Multi-page docs	>200 pages	—	851
MMLongBench (Wang et al., 2025b)	Multi-page docs	8K–128K tokens	13,331	—
SYNTHDOCBENCH	**Multi-page docs**	**Avg. 51.1 pages**	**200 / 3,340**	**1,788**

Figure 1: **Landscape of benchmarks** Top: benchmark comparison by average document length (pages) and average textual context (tokens). SynthDocBench occupies a unique region with both long multi-page context and high textual density. Bottom: comparison with existing chart and document VQA benchmarks. Prior benchmarks typically isolate either charts or long documents, whereas **SYNTHDOCBENCH** is designed to study their intersection under long contexts.

that chart understanding is far from solved, yet none evaluate charts within the document contexts where they naturally occur and where interpretation requires engagement with surrounding text, tables, and figures.

Long-Context Multimodal Benchmarks. Early long-context evaluation focused on needle-in-a-haystack retrieval (Wang et al., 2024), which provides insufficient difficulty ceilings for frontier models and correlates poorly with downstream reasoning performance (Wang et al., 2025b). Document-centric benchmarks have progressed from single-page evaluation via DocVQA (Mathew et al., 2021) through multi-page extensions including MP-DocVQA (Tito et al., 2023) and SlideVQA (Tanaka et al., 2023), though both are constrained to roughly twenty pages and principally assess span extraction. MMLongBench-Doc (Ma et al., 2024)

was the first benchmark targeting long PDF comprehension with interleaved visual content, and MMLongBench (Wang et al., 2025b) extended this framework to five task categories spanning up to 128K tokens. LongDocURL (Deng et al., 2025) and M-LongDoc (Chia et al., 2025) advance the state of the art through cross-element localisation and open-ended responses, but both prioritize breadth of coverage over diagnostic decomposition. SYNTHDOCBENCH is distinguished by its explicitly diagnostic design: document length, page depth, modality composition, and question type are varied as independent axes, with charts and figures foregrounded as primary reasoning targets whose interpretation requires engagement with arbitrarily distant contextual evidence.

Vision-Language Models for Document Understanding. Recent vision-language models have made single-page document QA increasingly saturated: frontier systems such as Qwen3-VL (Bai et al., 2025a) and InternVL3.5 (Wang et al., 2025a) now approach ceiling performance on DocVQA, so these benchmarks provide limited diagnostic separation among leading models. In contrast, long-context visual document reasoning remains substantially harder: the best reported results on MMLongBench-Doc remain far below single-page performance.[1] This gap suggests that failures are driven by properties that emerge in multi-page settings e.g., long-range dependency tracking, retrieval over dispersed evidence, yet current evaluations do not isolate which document factors are most responsible. Figure 1 situates document VQA benchmarks by average document length (pages) and average token count. Single-page benchmarks (DocVQA, ChartQA) lie in the bottom-left, whereas multi-page benchmarks shift toward higher length and token regimes. SYNTHDOCBENCH lies near the frontier along both axes, comparable to LongDocURL in page count and substantially denser in tokens than MMLongBench, while using controlled synthetic documents and requiring retrieval and reasoning over multi-modal evidence across multiple pages.

3 The SYNTHDOCBENCH Benchmark

The benchmark is generated via three coupled stages (Figures 2 and 3): (i) *document generation*, mapping a topic seed to a styled visual report with an aligned structured manifest; (ii) *QA generation*, converting the manifest into difficulty-controlled QA pairs; and (iii) *vision-only evaluation*, measuring model performance on rendered page images with no access to the underlying metadata. Reference answers are derived deterministically from the same structured artifacts used to generate the documents, eliminating the annotation bottleneck of real-document benchmarks (Weston et al., 2016; Johnson et al., 2017).

3.1 Synthetic Visual Document Generation

The document-generation pipeline maps a topic seed τ to a styled visual report $\mathcal{D}$ and a document-level QA manifest $\mathcal{M}$, factorizing synthesis into content generation, visual grammar, visualization synthesis, and assembly (Figure 2).

Topic-grounded content generation: Given τ, the pipeline constructs a semantic backbone from retrieved topic-relevant evidence and reorganizes it into a structured intermediate representation that exposes section boundaries, data-bearing spans, and salient content units. Downstream stages operate over this representation rather than free-form text, enabling precise evidence tracing.
Design and visual grammar: A layout archetype $a \in \mathcal{A}$ is sampled with probability 0.6 from a topic-conditioned distribution and with probability 0.4 uniformly at random (see Appendix L.2), preventing trivial correlations between subject matter and layout. The archetype governs page grammar, chart placement strategy, and auxiliary component types (metric cards, timelines, pull quotes), yielding a realistic and diverse report distribution.
Grounded visualization synthesis: Each visualization is generated in two aligned forms: (i) a visible D3.js rendering that appears in the document, and (ii) a structured metadata object V_k recording chart semantics (axes, data points, derived insights). This *dual-layer* formulation ensures ground truth is available by construction without post-hoc chart parsing

[1] https://huggingface.co/spaces/OpenIXCLab/mmlongbench-doc, accessed March 2026.

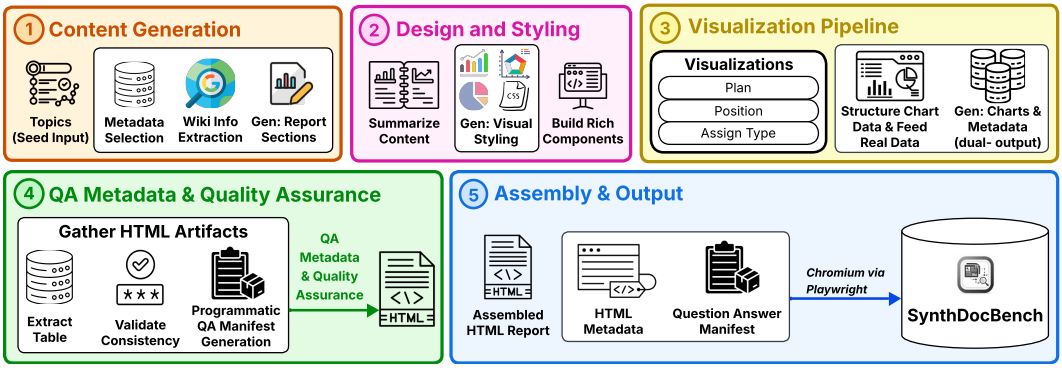


Figure 2: **Synthetic visual document generation pipeline.** From a topic seed, the pipeline generates grounded report content, applies document-level visual styling, synthesizes visualizations, performs metadata and QA validation, and assembles the final HTML/PDF reports with a machine-readable QA manifest.

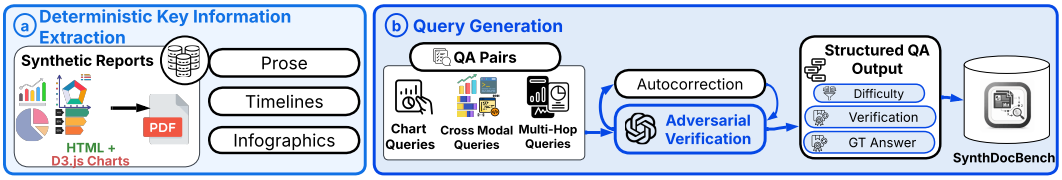


Figure 3: **QA generation pipeline.** The pipeline parses the generated report into structured evidence channels, extracts and synthesizes key information, and generates chart-reading, cross-modal, and multi-hop questions. A verification stage filters weak or malformed items before serializing the final QA output.

or human labeling (full schema in Appendix L.3).
Validation and assembly: Numeric values in both chart and table metadata are recomputed from structured data and corrected before finalization; all validated metadata are aggregated into $\mathcal{M}$. The report is assembled as HTML and rendered to PDF using Playwright,[2] producing an aligned pair $(\mathcal{D}, \mathcal{M})$.

3.2 Question-Answer Generation

The second stage converts a generated report into structured QA items via evidence recovery, synthesis, and generation with validation (Figure 3).

The pipeline parses $\mathcal{D}$ into text, table, and visualization channels, recovering chart semantics directly from embedded metadata rather than pixels. Evidence units are then synthesized into higher-order compositions supporting aggregation, comparison, and cross-source reasoning.

The benchmark produces three question families: *chart-reading* (direct chart or table semantics), *cross-modal* (joint reasoning over textual and visual evidence), and *complex multi-hop* (composition across 2–4 evidence units). Each question carries a difficulty label $L1$–$L5$ (Table 13) and a structured evidence trace. A validation stage filters malformed, weakly supported, or ambiguous items; the serialized output schema is detailed in Appendix L.4.

[2] https://github.com/microsoft/playwright

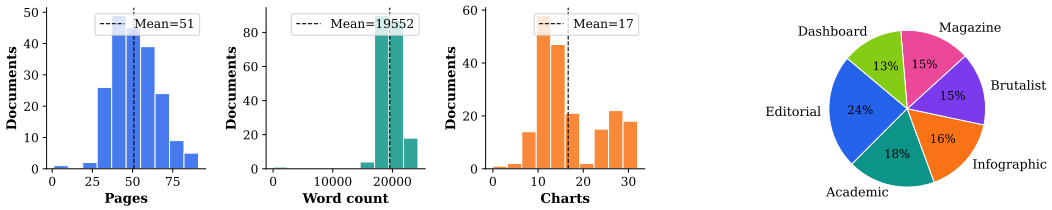
Metric	Mean	Med.	Min	Max
Pages	51.1	49	24	91
Words	20,568	20,450	17,367	25,138
Charts	16.7	14	5	32
PDF (MB)	2.12	2.10	1.20	3.80

Table 1: Per-document statistics across all 200 reports.

Three-layer quality control (numeric recomputation, automated consistency filtering, and 100-sample manual review with >96% acceptance rate) is described in Appendix F.

3.3 Benchmark Statistics

The resulting benchmark comprises **200 synthetic reports** and **1,788 questions** distributed across three subsets: 597 `chart` reading, 597 `complex` multi-hop, and 594 `cross_modal`. Documents average **51.1 pages**, **16.7 charts**, and ≈**20,568 words**, placing SYNTHDOCBENCH firmly in the long-document regime. The benchmark spans **24 distinct chart types** across **6 layout archetypes**, ensuring models cannot exploit narrow visual or structural distributions. Figure 4 shows the distributions of page count, word count, and chart count per document: all three are unimodal and tightly concentrated, reflecting the controlled generation pipeline rather than the heavy-tailed distributions typical of real-world corpora. The tight ranges reflect the controlled generation pipeline: page count, word count, and chart count are all bounded by design, enabling ablations that hold document complexity constant while varying other axes (modality, question type, layout).



(a) Document property distributions (pages, words, charts). *(b) Layout archetypes*

Figure 4: Document composition statistics across 200 reports.

4 Evaluation Setup

Candidate models operate under a strict *vision-only* protocol: they receive only the rendered page-image sequence $\mathcal{I}$ and never access HTML source, embedded metadata, or $\mathcal{M}$. The end-to-end pipeline is illustrated in Figure 8 (Appendix A). Each PDF is rasterized at 144 DPI, capped at 120 pages, and concatenated into single-column 5-page vertical strips (max 7,900 px, 4 MB); we use 5-page strips as the default to satisfy the input constraints of all evaluated models simultaneously, noting that denser strips further improve performance where API limits permit. Full hyperparameters are in Table 19. Candidate models predict $\hat{a} = f_\theta(\mathcal{I}, q)$ at temperature 0 via a fixed system prompt requiring concise, vision-grounded 2 to 4 sentence answers (prompt in Appendix L.7). GPT-5 (Singh et al., 2025) scores each $(\hat{a}, a^*)$ pair at temperature 0, returning a JSON score in $[0, 10]$ (rubric in Table 6). Parse failures receive score -1 and are excluded from all aggregates (Q_{valid}). To validate judge reliability, we re-scored a subset of responses using Gemini-3.1-Pro and Claude-Sonnet-4.5 as alternative judges (Appendix C): GPT-5 and Gemini-as-judge agree to within 3.5 ACC points (Pearson $r \geq 0.94$ across all question types), confirming that model rankings are robust to judge choice (see Table 9).

We report mean judge score and threshold accuracy:

$$\mathrm{ACC}(f_\theta) = \frac{1}{|Q_{\text{valid}}|} \sum_{q \in Q_{\text{valid}}} \mathbf{1}\Big[\mathcal{J}(\hat{a}_q, a_q^*) \geq 6\Big], \tag{1}$$

where τ=6 ("core answer correct") aligns with MMLongBench-Doc (Ma et al., 2024). Both metrics are stratified by question family and difficulty level $L1$–$L5$.

4.1 Experiments and Results

We evaluate the following eight vision-language models on SYNTHDOCBENCH: Gemini-3.1-Pro (Google, 2026), GPT-5.4 (Singh et al., 2025), GPT-4o (OpenAI et al., 2024), Claude-Sonnet-4.5 (Anthropic, 2025), Qwen3.5-VL-122B (Qwen Team, 2025), Qwen3-VL-235B (Bai et al., 2025a), InternVL3-78B (Wang et al., 2025a), and Qwen2.5-VL-7B (Bai et al., 2025b). All models are evaluated with GPT-5 as judge.

	Prior Benchmarks		SYNTHDOCBENCH (Ours)							
	Accuracy		Overall		Chart		Complex		Cross-Modal	
Model	**DocVQA** (1 pg)	**MMLong-Bench-Doc** (47.5 pg)	**ACC**	**Score**	**ACC**	**Score**	**ACC**	**Score**	**ACC**	**Score**
Proprietary Models										
Gemini-3.1-Pro (Google, 2026)	93.4	45.1	**0.725**	**7.19**	**0.759**	**7.62**	**0.789**	**7.28**	**0.628**	**6.65**
GPT-5.4 (Singh et al., 2025)	—	—	0.423	4.68	0.425	4.24	0.457	5.21	0.387	4.59
GPT-4o (OpenAI et al., 2024)	92.8	46.3	0.386	4.38	0.457	4.44	0.360	4.66	0.342	4.05
Claude-Sonnet-4.5 (Anthropic, 2025)	92.0	40.1	0.314	3.96	0.353	4.11	0.337	4.28	0.250	3.49
Open-Weight Models										
Qwen3.5-VL-122B (Qwen Team, 2025)	—	—	0.655	6.77	0.713	7.21	0.690	6.89	0.561	6.22
Qwen3-VL-235B (Bai et al., 2025a)	**96.5**	**57.0**	0.586	6.18	0.642	6.60	0.611	6.23	0.503	5.71
InternVL3-78B (Wang et al., 2025a)	95.1	24.3	0.383	4.39	0.456	4.75	0.397	4.73	0.296	3.69
Qwen2.5-VL-7B (Bai et al., 2025b)	93.7	25.1	0.081	1.08	0.162	1.76	0.012	0.37	0.067	1.12

Table 2: Performance on SYNTHDOCBENCH (200 reports, 1,788 questions) alongside two established benchmarks. **DocVQA** (Mathew et al., 2021) tests single-page understanding; **MMLongBench-Doc** (Ma et al., 2024) tests multi-page documents (avg. 47.5 pages). — = not publicly reported at time of writing. Bootstrap 95% CIs (2,000 resamples, seed 42) are $\leq \pm 0.023$ overall and $\leq \pm 0.041$ per subset across all models; all pairwise ACC gaps between adjacent-ranked models exceed their combined CI half-widths. Per-model CI details are reported in Table 10.

Table 2 presents ACC (τ=6) and mean judge score across all eight models, stratified by question subset. Gemini-3.1-Pro leads all models by a substantial margin, achieving an overall ACC of 0.725 and a mean judge score of 7.19. The ranking Gemini > Qwen3.5-VL-122B > Qwen3-VL-235B > GPT-5.4 > GPT-4o $\approx$ InternVL3-78B > Claude-Sonnet-4.5 $\gg$ Qwen2.5-VL-7B holds consistently across all three question subsets. Qwen3.5-VL-122B ranks second overall (ACC 0.655), demonstrating that a 122B MoE open-weight model can approach Gemini-level performance on long-context document understanding. GPT-5.4 ranks fourth overall (ACC 0.423), between Qwen3-VL-235B (0.586) and GPT-4o (0.386). GPT-4o (0.386) and InternVL3-78B (0.383) are statistically indistinguishable despite their different architectures, suggesting that neither parameter scale nor training approach alone determines long-context document performance on this benchmark.

OCR + Text-Only Baseline. Chart-reading is the easiest subset for most models, while cross-modal questions are consistently the hardest, confirming that integrating evidence across text and charts within long documents remains an open challenge for all current VLMs. The gap between Gemini and the next-best model (Qwen3.5-VL-122B, 0.655) is 7.0 ACC points; we discuss a potential distribution-familiarity confound in Section 5. Model

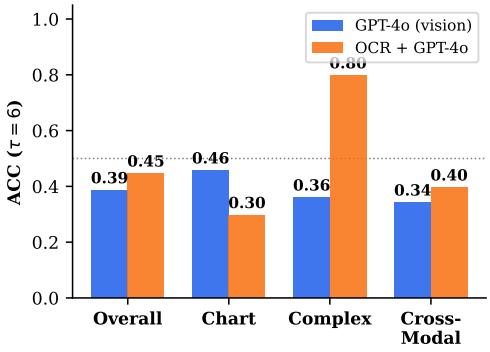


Figure 5: GPT-4o vision vs. OCR+GPT-4o (text-only) ACC by subset. Vision dominates on Chart; OCR dominates on Complex.

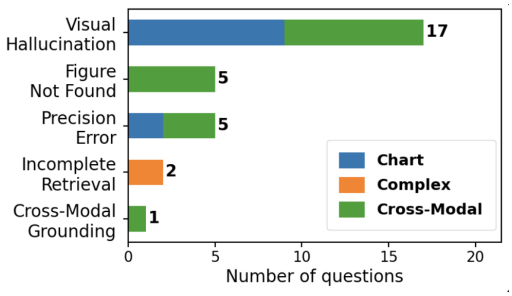


Figure 6: Hard failures: 109 questions where all six models score ≤ 3, broken down by error category and question subset (**Ch.** = chart-reading; **Cx.** = complex; **XM.** = cross-modal).

rankings on SynthDocBench correlate with MMLongBench-Doc (Spearman ρ=0.657, Pearson r=0.683), supporting external validity of the benchmark ordering (Appendix C).

PyMuPDF extracts page text (up to 60,000 chars) and GPT-4o answers without images (n=1,525; 263 parse failures excluded). As Figure 5 shows, performance is highly asymmetric: complex multi-hop ACC is 0.798 for OCR vs. 0.360 for vision, confirming that complex evidence is largely text-recoverable; chart-reading reverses sharply (OCR 0.297 vs. vision 0.457), confirming that chart questions genuinely require pixel-level visual decoding. The 46 pp gap between OCR chart ACC and Gemini's (0.759) isolates visual perception as the primary bottleneck.

5 Analysis and Discussion

	L1		L2		L3		L4		L5	
Model	ACC	Sc.	ACC	Sc.	ACC	Sc.	ACC	Sc.	ACC	Sc.
Proprietary Models										
Gemini-3.1-Pro	**0.784**	**7.84**	**0.731**	**7.53**	**0.675**	**6.91**	**0.766**	**7.19**	**0.670**	**6.48**
GPT-5.4	0.231	2.37	0.422	4.66	0.501	5.29	0.489	5.28	0.259	4.02
GPT-4o	0.271	2.68	0.455	4.78	0.451	4.83	0.400	4.64	0.173	3.54
Claude-Sonnet-4.5	0.382	3.81	0.300	3.96	0.312	3.99	0.360	4.28	0.154	3.18
Open-Weight Models										
Qwen3.5-VL-122B	0.707	7.16	0.673	7.11	0.639	6.66	0.683	6.76	0.528	6.02
Qwen3-VL-235B	0.648	6.48	0.595	6.55	0.568	6.01	0.619	6.25	0.457	5.36
InternVL3-78B	0.437	4.42	0.382	4.35	0.390	4.48	0.427	4.63	0.198	3.62
Qwen2.5-VL-7B	0.015	0.17	0.123	1.48	0.155	1.88	0.033	0.71	0.005	0.26

Table 3: Model performance stratified by difficulty level $L1$–$L5$. **ACC** = fraction of questions with judge score ≥ 6; **Sc.** = mean judge score (0–10). Best per column in **bold**. Bootstrap 95% CIs (2,000 resamples, seed 42) are $\leq \pm 0.070$ per level for all models; all adjacent-ranked model gaps exceed their combined CI half-widths. Per-level CI details are in Table 11.

We analyse model performance across three dimensions: difficulty level, question category, and error type, surfacing failure modes hidden by aggregate scores.

Figure 7: ACC (τ=6) by fine-grained question category. Rows are grouped by subset (Chart Reading, Cross-Modal, Complex Reasoning); columns are models ordered by overall ACC. The colour scale runs from red (0) to green (1). Full numerical values are in Table 12 (Appendix I).

Difficulty-Stratified Results Table 3 shows ACC by level $L1$–$L5$. All models except Gemini-3.1-Pro degrade monotonically toward $L5$; Claude-Sonnet-4.5 drops 23 pp ($L1 \rightarrow L5$) while Gemini stays flat (0.670–0.784). Qwen3.5-VL-122B follows a similar flat-then-drop pattern (0.707 at $L1$, 0.528 at $L5$), tracking closely with Gemini across all levels. GPT-4o's anomalous $L1$ dip (0.271) suggests precise value-extraction is harder for it than compositional reasoning.

Question-Category Breakdown Figure 7 breaks down ACC by fine-grained category. Trend/pattern questions are easiest (perceptually salient direction cues), while value-reading and *integrate-sources* cross-modal questions are hardest — precise axis-label reading and multi-page evidence alignment each bottleneck distinct model families. The widest cross-model gap is in *technical/quantitative* complex questions (Gemini: 0.643 vs. GPT-4o: 0.111, InternVL3: 0.156), isolating quantitative multi-step reasoning as the primary frontier differentiator.

Positional Bias: Evidence Location Within Documents Questions are bucketed by relative chart position $p{=}k/K$ into equal thirds (Table 4; see Appendix G for the bar chart). The **middle third is hardest for 5 of 8 models**, dropping 5–18 pp below Early. Qwen3.5-VL-122B shows the steepest Early$\rightarrow$Middle drop (-18.5 pp) while partially recovering in Late; Claude-Sonnet-4.5 shows the steepest monotonic Early$\rightarrow$Late decline (-11.7 pp); Gemini-3.1-Pro shows a U-shaped pattern echoing the lost-in-the-middle effect (Liu et al., 2023).

Hard Failure and Domain Error Analysis Of 109 questions where all eight models score $\leq$ 3 (Figure 6; Appendix M), cross-modal failures dominate, confirming multi-modal evidence integration as the hardest challenge.

Domain analysis (Figure 11, Appendix H) shows cross-modal ACC lags chart reading by 13-16 pp in every topic domain.

Visual hallucination dominates errors. Models return plausible-looking values absent from the ground-truth chart, concentrated on dense value-reading charts, dumbbell plots, and multi-series comparisons, localising the bottleneck to pixel-level decoding, not reasoning. Figure-not-found and precision errors account for the remainder; both point to precise

Model	Early	Middle	Late	Δ
Proprietary				
Gemini-3.1-Pro	**0.788**	**0.717**	**0.784**	−0.004
GPT-5.4	0.440	0.419	0.468	+0.028
GPT-4o	0.489	0.443	0.443	−0.046
Claude-Sonnet-4.5	0.418	0.342	0.301	−0.117
Open-Weight				
Qwen3.5-VL-122B	0.820	0.635	0.660	−0.160
Qwen3-VL-235B	0.674	0.578	0.693	+0.019
InternVL3-78B	0.451	0.460	0.455	+0.003
Qwen2.5-VL-7B	0.174	0.135	0.188	+0.014
n	184	237	176	—

Table 4: Chart-reading ACC (τ=6) by position bucket. Δ = Late−Early; negative = early-content advantage. See Figure 10 (Appendix G) for the bar chart.

quantitative visual-textual alignment as a second bottleneck that prompting alone does not close (Appendix L.7).

Effect of Image Presentation: Pages per Strip and Resolution All ablations use **GPT-5** as judge on the full benchmark (Table 5; full details in Appendix N). Gemini-3.1-Pro ACC increases monotonically from 0.369 (1 page/strip) to 0.792 (10 pages), with cross-modal benefiting most (0.339→0.707), confirming multi-page context is essential. 144 DPI is optimal; higher resolution degrades performance due to JPEG compression under the 4 MB API cap. GPT-4o and Claude-Sonnet-4.5 show model-specific optima at 2 and 5 pages respectively, indicating model-specific optimal context density.

Gemini-3.1-Pro dominance and potential confounds. The 13.9 pp gap between Gemini (0.725) and Qwen3-VL-235B (0.586) warrants caution: SynthDocBench uses web-rendered (HTML/D3.js) charts, which may be stylistically consistent with Gemini's training distribution. We cannot rule out rendering-familiarity as a partial confounder; future work should validate with alternative rendering backends to isolate long-context reasoning effects.

6 Conclusion

We introduce SYNTHDOCBENCH, a long-context visual document understanding benchmark, measuring VLM performance for the ability to locate multiple key information facts over long-contexts from a given document to reason and answer a multi-step question. It is generated synthetically with controlled difficulty axes by independently varying document length, layout complexity, modality composition, and question type across a combinatorial design. Our results reveal three concurrent failure modes in frontier VLMs: sharp performance degradation with increasing evidence complexity and reasoning depth, systematic positional sensitivity in which the middle third of a document is the hardest section for four of six models and three of six models show a negative Early→Late trend, with Claude-Sonnet-4.5 exhibiting the steepest monotonic decline (−11.7 pp) and Gemini-3.1-Pro showing a distinctive U-shaped recovery (§5), and collapse of precise chart-reading accuracy in long-document contexts. SYNTHDOCBENCH reveals a clear disparity between benchmark performance and genuine long-context visual reasoning ability, serving as a diagnostic platform for future model development, providing clean, controlled signals needed to identify, understand, and ultimately improve model capabilities. Code and data will be publicly released upon acceptance.

As future work, we plan to extend SYNTHDOCBENCH to broader multimodal reasoning settings, including richer document types (e.g., tables, forms, and mixed-layout reports),

Table 5: Rendering and prompting ablations on the full 200-report benchmark (judge: GPT-5).

	Overall		Chart		Complex		Cross-Modal	
Setting	ACC	Score	ACC	Score	ACC	Score	ACC	Score
(a) Pages per strip (concat-num) — Gemini-3.1-Pro, DPI fixed at 144								
1 (single page)	0.369	4.04	0.451	4.65	0.318	3.48	0.339	3.99
2	0.474	5.12	0.509	5.27	0.531	5.60	0.381	4.50
5 (default)	0.725	7.19	0.759	7.62	0.789	7.28	0.628	6.65
10	**0.792**	**7.80**	**0.838**	**8.32**	**0.831**	**7.70**	**0.707**	**7.37**
(b) Rasterization resolution (DPI) — Gemini-3.1-Pro, concat-num fixed at 5								
72 DPI	0.686	6.97	0.753	7.50	0.732	7.14	0.572	6.25
144 DPI (default)	**0.725**	**7.19**	**0.759**	**7.62**	**0.789**	**7.28**	**0.628**	**6.65**
216 DPI	0.683	6.92	0.729	7.36	0.747	7.18	0.570	6.23
(c) Prompting strategy — Gemini-3.1-Pro								
Default	**0.725**	7.19	**0.759**	**7.62**	0.789	7.28	**0.628**	**6.65**
Chain-of-thought	0.702	7.15	0.731	7.32	0.783	7.63	0.591	6.50
No system prompt	0.723	**7.25**	0.756	7.52	**0.820**	**7.71**	0.592	6.53
(d) Prompting strategy — GPT-4o								
Default	0.386	4.38	0.457	4.44	0.360	4.66	**0.342**	4.05
Chain-of-thought	0.396	4.61	0.458	4.49	0.393	4.91	0.336	**4.41**
No system prompt	**0.420**	**4.72**	**0.506**	**4.63**	**0.419**	**5.22**	0.335	4.31
(e) Prompting strategy — Claude-Sonnet-4.5								
Default	0.314	3.96	0.353	4.11	0.337	4.28	0.250	3.49
Chain-of-thought	0.369	4.44	**0.403**	**4.36**	0.363	4.49	0.340	4.47
No system prompt	**0.422**	**4.90**	0.392	4.31	**0.515**	**5.73**	**0.357**	**4.64**
(f) Pages per strip — GPT-4o (DPI fixed at 144)								
1 (single page)	0.374	4.17	0.450	4.26	0.337	4.14	0.337	4.10
2	**0.396**	**4.39**	**0.470**	4.41	**0.381**	4.63	0.337	**4.13**
5 (default)	0.386	4.38	0.457	**4.44**	0.360	**4.66**	**0.342**	4.05
10	0.354	4.21	0.438	4.35	0.332	4.51	0.292	3.75
(g) Pages per strip — Claude-Sonnet-4.5 (DPI fixed at 144)								
1 (single page)	0.258	3.15	0.303	3.51	0.236	2.78	0.235	3.15
2	0.288	3.41	**0.361**	3.92	0.276	3.21	0.227	3.11
5 (default)	**0.314**	**3.96**	0.353	**4.11**	**0.337**	**4.28**	**0.250**	**3.49**
10	0.312	3.70	0.378	4.01	0.314	4.06	0.244	3.03

longer contexts, and more diverse reasoning tasks such as multi-hop aggregation and cross-document grounding.

Reproducibility Statement

The authors are committed to aiding researchers in reproducing our benchmark and results. The evaluation source code is publicly available at https://github.com/ServiceNow/SynthDocBench, and the dataset is publicly available at https://huggingface.co/datasets/ServiceNow-AI/SynthDocBench. We also make every effort to disclose experiment hyperparameters wherever necessary throughout the paper, and in the appendix.

Ethics Statement

Synthetic data and content generation. SYNTHDOCBENCH is constructed entirely from programmatically generated synthetic documents. No human subjects were involved, no personal data were collected, and no real-world documents were reproduced. Document content is grounded in broad, publicly available topic seeds (geopolitics, economics, environmental science, technology, and related domains) and generated by large language models under structured constraints. We reviewed generated content to confirm that it does not contain personally identifying information, hate speech, or other harmful material. Because all textual content is machine-generated and factually non-binding, it should not be treated as authoritative.

Model evaluation and API usage. Evaluation is performed via commercial APIs (Gemini, GPT-4o, Claude) and open-weight models under their respective terms of service. No model was fine-tuned on SYNTHDOCBENCH data during the study; all evaluations use frozen model weights. The GPT-5 judge is used solely to score candidate responses against deterministic reference answers; the rubric and prompts are fully disclosed in the appendix to allow independent replication.

Environmental impact. Large-scale model evaluation carries a non-trivial computational and energy cost. We limited redundant evaluation runs through careful ablation design (Section 5) and report all configurations transparently so that future work can build on our results rather than repeating them. Estimated compute was approximately 150 GPU-hours for document rendering and 200 API-hours for model inference and judge scoring.

Benchmark integrity and Goodhart's Law. Releasing a controlled benchmark creates the risk that future models overfit to its specific design choices (chart types, layout archetypes, question templates). We mitigate this through (i) a 40% random layout override that prevents spurious topic–layout correlations and (ii) programmatic generation that admits straightforward extension. We encourage the community to use SYNTHDOCBENCH as a *diagnostic* tool and to contribute extensions that preserve the benchmark's diagnostic validity rather than optimizing against its current instantiation.

Intended use. SYNTHDOCBENCH is intended for research on long-context visual document understanding. It is not intended for deployment in high-stakes decision-making systems. Results on SYNTHDOCBENCH characterize specific, controlled failure modes and should not be generalized uncritically to real-world document understanding performance.

References

Anthropic. Claude 4.5 Sonnet model card addendum. https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd.pdf, 2025.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Jialong Song, Peng Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16259–16273, 2024.

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 super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9244–9261, 2025.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1135–1159, 2025.

Google. Gemini 3.1 technical report. *Technical Report*, 2026.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Brenden M Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2873–2882, 2018. URL https://arxiv.org/abs/1711.00350.

Zhen Li, Duan Li, Yukai Guo, Xinyuan Guo, Bowen Li, Lanxi Xiao, Shenyu Qiao, Jiashu Chen, Zijian Wu, Hui Zhang, Xinhuan Shu, and Shixia Liu. ChartGalaxy: A dataset for infographic chart understanding and generation. *arXiv preprint arXiv:2505.18668*, 2025. URL https://arxiv.org/abs/2505.18668.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.

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*, pp. 2263–2279, 2022.

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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, 2025.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. URL https://www.docvqa.org/.

OpenAI, Aaron Hurst, Adam Lerer, Adam Goucher, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Qwen Team. Qwen3.5-VL technical report. https://huggingface.co/Qwen/Qwen3.5-VL-7B-Instruct, 2025.

Aaditya Singh, Lawrence Chan, Ted Li, Amelia Glaese, et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13636–13645, 2023.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37:20540–20565, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, et al. MMLONGBENCH: Benchmarking long-context vision-language models effectively and thoroughly. *arXiv preprint arXiv:2505.10610*, 2025b.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. In *Proceedings of the International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1502.05698.

Liyan Xu et al. ChartMuseum: A benchmark for fine-grained chart understanding. *arXiv preprint arXiv:2505.13444*, 2025. URL https://arxiv.org/abs/2505.13444.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5317–5327, 2019.

Weiye Zhang et al. VisuLogic: A benchmark for evaluating visual reasoning in multi-step logic problems. *arXiv preprint arXiv:2504.15279*, 2025. URL https://arxiv.org/abs/2504.15279.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. Multichartqa: Benchmarking vision-language models on multi-chart problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11341–11359, 2025.

A Evaluation Pipeline Diagram

Figure 8 details the end-to-end evaluation pipeline used to assess all candidate models on SYNTHDOCBENCH. Rendered PDFs are rasterized to page images at 144 DPI, concatenated into 5-page vertical strips, and passed directly to each candidate vision-language model at temperature 0 (vision-only; no OCR or metadata). Each candidate response is then scored against the reference answer by GPT-5 acting as judge $\mathcal{J}$, using the Table 6 rubric.

Table 6: GPT-5 judge scoring rubric. Scores of -1 indicate parse failures and are excluded from aggregate statistics.

Score	Criterion
10	All key facts present and fully correct
8–9	Mostly correct, with only minor omissions or imprecision
6–7	Core answer correct, but with some missing detail
4–5	Partially correct, with significant gaps or errors
2–3	Mostly incorrect or severely incomplete
0–1	Incorrect, hallucinated, or contradictory to the reference
-1	Judge parse failure (excluded)

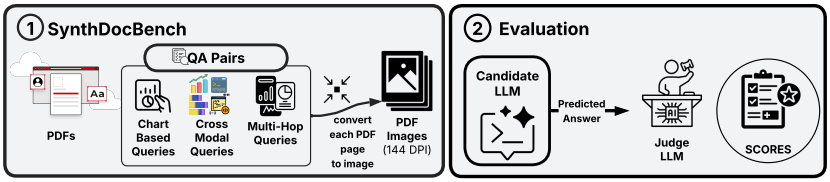


Figure 8: **Evaluation pipeline.** Rendered PDFs are converted to page images at 144 DPI, grouped into concatenated 5-page strips, and supplied directly to candidate models at temperature 0. Candidate answers are then scored against deterministic reference answers by GPT-5 acting as the judge model $\mathcal{J}$.

B Benchmark Comparison

Table 7 situates SYNTHDOCBENCH relative to the most closely related benchmarks and reports frontier model scores on each. Despite strong numbers on DocVQA, ChartQA, and MMLongBench-Doc, coverage is sparse and inconsistent across models—a direct consequence of benchmarks built from heterogeneous real corpora with no controlled variation. Crucially, high scores on these benchmarks do not transfer: the same models that approach saturation on DocVQA and ChartQA exhibit substantial failure rates on our controlled subsets, demonstrating that SYNTHDOCBENCH exposes failure modes that existing benchmarks cannot surface.

C Judge Validation

To assess the reliability of GPT-5 as judge, we re-scored the full 1,788-question benchmark for four candidate–judge pairs using Gemini-3.1-Pro and Claude-Sonnet-4.5 as alternative judges. The same two candidates (GPT-4o and Qwen3-VL-235B) are evaluated under both alternative judges, enabling a direct comparison of judge behaviour on identical responses. Table 8 reports overall pairwise Pearson correlation, within-1-point agreement rate, and ACC delta between GPT-5 and each alternative judge; Table 9 further stratifies by question type.

GPT-5 and Gemini-as-judge agree to within 3.5 ACC points ($r \geq 0.94$, $w_1 \geq 0.84$) across all tested candidates. Agreement is highest for chart-reading and cross-modal questions ($r \geq 0.95$) and slightly lower for complex multi-hop questions ($r \geq 0.91$), where scoring subjectivity is higher. These gaps are consistent across both candidate models and are well within the benchmark's bootstrap CI half-widths, confirming that model rankings are robust to judge choice regardless of question type. Claude-Sonnet-4.5 is a systematically lenient judge (+11–16 ACC points overall, rising to +24–27 pp on complex questions), consistent with the positivity bias documented for smaller LLM judges, and was excluded on this basis. The GPT-5/Gemini convergence addresses a potential vendor-conflict concern:

Model	DocVQA	ChartQA	MathVista	MMMU	MMLongBench-Doc
Proprietary Models					
Gemini-3.1-Pro	**93.4**	88.5	**74.8**	**81.0**	45.1
GPT-4o	92.8	85.7	63.8	70.7	46.3
Claude-Sonnet-4.5	92.0	**89.0**	72.0	77.8	40.1
Open-weight Models					
Qwen3-VL-235B	**96.5**	**91.2**	**85.8**	71.3	**57.0**
InternVL3-78B	95.1	—	79.0	**72.2**	24.3
Qwen2.5-VL-7B	93.7	83.8	68.1	58.0	25.1

Table 7: Performance of evaluated VLMs on established visual document understanding benchmarks. — = not publicly reported at time of writing.

Table 8: Overall judge agreement statistics on the full SYNTHDOCBENCH (200 reports, 1,788 questions). r: Pearson correlation between score sequences. w_1: fraction of responses where $|\text{score}_{\text{GPT-5}} - \text{score}_{\text{alt}}| \leq 1$. Δ**ACC**: ACC(alt judge) $-$ ACC(GPT-5) at τ=6.

Candidate	Alt. Judge	r	w_1	ΔACC
GPT-4o	Gemini-3.1-Pro	0.942	0.838	-0.035
Qwen3-VL-235B	Gemini-3.1-Pro	0.960	0.893	-0.010
GPT-4o	Claude-Sonnet-4.5	0.881	0.654	$+0.135$
Qwen3-VL-235B	Claude-Sonnet-4.5	0.916	0.723	$+0.156$

Gemini-as-judge independently replicates GPT-5's ranking with a maximum ACC deviation of 3.5 points.

Ranking correlation with MMLongBench-Doc. To assess external validity, we compute the Spearman rank correlation between SynthDocBench overall ACC and published MMLongBench-Doc scores for the six models evaluated on both benchmarks. The correlation is $\rho = 0.657$ ($r = 0.683$), indicating moderate positive agreement: models that perform well on our benchmark tend to perform well on MMLongBench-Doc, but the rankings are not identical. Notably, GPT-4o ranks higher on MMLongBench-Doc (#2) than on SynthDocBench (#3), suggesting that our benchmark's emphasis on precise chart-value extraction and cross-modal alignment surfaces capabilities that MMLongBench-Doc's more heterogeneous question set partially masks. The imperfect correlation ($\rho < 1$) is itself evidence that SynthDocBench provides complementary diagnostic signal beyond existing benchmarks.

Table 9: Judge agreement stratified by question type on the full SYNTHDOCBENCH (200 reports, 1,788 questions). r: Pearson correlation. w_1: within-1-point agreement. Δ**ACC**: ACC(alt) $-$ ACC(GPT-5) at τ=6. Positive ΔACC means the alternative judge is more lenient than GPT-5.

		Overall			Chart			Complex			Cross-Modal		
Candidate	Alt. Judge	r	w_1	ΔACC	r	w_1	ΔACC	r	w_1	ΔACC	r	w_1	ΔACC
Alt. Judge: Gemini-3.1-Pro													
GPT-4o	Gemini	0.942	0.838	-0.035	0.953	0.801	-0.030	0.915	0.869	-0.054	0.950	0.845	-0.020
Qwen3-VL-235B	Gemini	0.960	0.893	-0.010	0.968	0.910	-0.022	0.933	0.896	-0.013	0.967	0.872	$+0.005$
Alt. Judge: Claude-Sonnet-4.5													
GPT-4o	Claude	0.881	0.654	$+0.135$	0.913	0.647	$+0.022$	0.833	0.610	$+0.271$	0.895	0.705	$+0.113$
Qwen3-VL-235B	Claude	0.916	0.723	$+0.156$	0.950	0.824	$+0.069$	0.849	0.618	$+0.241$	0.916	0.727	$+0.159$

Qualitative alignment with real-document failures. Consistent with our benchmark's error taxonomy, the failure modes we identify in the controlled setting correspond to qualitatively similar patterns in MMLongBench-Doc: models that score lowest on our chart-reading subset also show the steepest degradation on MMLongBench-Doc's figure-heavy questions, and the cross-modal integration failures we isolate (where GPT-4o drops to 0.342 ACC) align with the documented difficulty of questions requiring evidence from non-adjacent pages in real documents. The key distinction is attribution: in real documents these factors co-vary and failures cannot be cleanly assigned to a single cause, whereas SynthDocBench's factorial design makes each failure mode individually observable.

D Bootstrap Confidence Intervals

Table 10 reports per-model bootstrap 95% confidence intervals (2,000 resamples, seed 42) for ACC on each question subset; Table 11 breaks these down by difficulty level. All pairwise ACC gaps between adjacent-ranked models exceed their combined CI half-widths, confirming the reliability of the rankings in Table 2.

Table 10: Bootstrap 95% CI half-widths (±) for ACC at τ=6, computed over 2,000 resamples (seed 42) on the full SYNTHDOCBENCH (200 reports, 1,788 questions).

Model	Overall	Chart	Complex	Cross-Modal
Gemini-3.1-Pro	±0.020	±0.036	±0.034	±0.039
GPT-4o	±0.022	±0.040	±0.039	±0.038
Claude-Sonnet-4.5	±0.022	±0.041	±0.037	±0.034
Qwen3-VL-235B	±0.022	±0.038	±0.041	±0.040
InternVL3-78B	±0.023	±0.040	±0.038	±0.038
Qwen2.5-VL-7B	±0.013	±0.029	±0.008	±0.021

Table 11: Bootstrap 95% CI half-widths (±) for ACC at τ=6 stratified by difficulty level, computed over 2,000 resamples (seed 42).

Model	L1	L2	L3	L4	L5
Gemini-3.1-Pro	±0.055	±0.044	±0.040	±0.036	±0.069
GPT-4o	±0.063	±0.048	±0.044	±0.042	±0.053
Claude-Sonnet-4.5	±0.070	±0.044	±0.044	±0.042	±0.051
Qwen3-VL-235B	±0.070	±0.049	±0.044	±0.044	±0.069
InternVL3-78B	±0.070	±0.046	±0.042	±0.043	±0.056
Qwen2.5-VL-7B	±0.018	±0.031	±0.032	±0.015	±0.008

E Score Distribution per Model

Figure 9 shows the distribution of raw judge scores (0–10) for each model. A bimodal distribution (mass near 0–2 and near 8–10) indicates a model that either fully recovers or fully misses an answer. A unimodal distribution centered at 6–7 indicates consistent but imprecise retrieval. Gemini-3.1-Pro is the only model with a strong right-skewed distribution, while Qwen2.5-VL-7B is concentrated near 0.

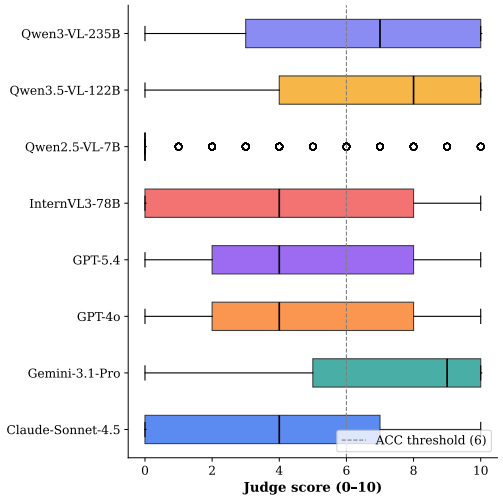


Figure 9: Distribution of judge scores (0–10) per model.

F Question Quality Control

We apply three layers of quality assurance to the generated questions.

Numeric recomputation. For every question whose ground-truth answer is a single numeric value (e.g. a chart data point, percentage, or count), we independently re-derive the answer from the structured metadata $\mathcal{M}$ used during generation, including raw `chart_data` arrays and `key_facts` tables. Any question whose recomputed answer differs from the LLM-generated answer by more than a 5% relative tolerance is flagged and either corrected or dropped. This filter removed fewer than 2% of chart-reading questions.

Automated consistency filtering. All 1,788 questions pass a rule-based filter checking: (1) the question references an element present in $\mathcal{M}$; (2) the ground-truth answer is non-empty and at least six tokens long; (3) no exact-match overlap between question text and answer text exceeding 40% of answer tokens (to screen out trivially answerable questions); and (4) difficulty level L is consistent with the number of inferential steps encoded in the `reasoning` field.

Manual review. A random sample of 100 questions (stratified by group type and difficulty level) was reviewed by two of the authors independently. Reviewers rated each question on three dimensions: *answerability* (is the answer determinable from the document?), *answer correctness* (is the ground-truth answer correct?), and *question clarity* (is the question unambiguous?). Inter-rater agreement was $\kappa = 0.81$. Overall acceptance rate across all three dimensions was 96%. The four rejected questions (two from *complex* and two from *cross-modal*) involved ambiguous referents that could not be resolved from the document alone; all four were removed from the final benchmark.

G Positional Bias: Bar Chart

Figure 10 visualises the ACC per position bucket from Table 4 in the main text.

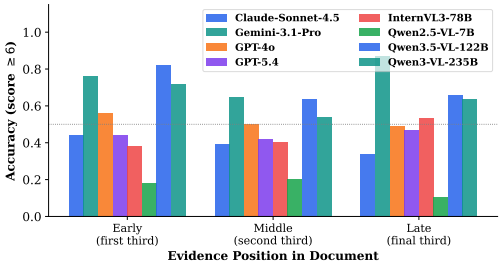


Figure 10: Chart-reading ACC (τ=6, judge: GPT-5) by evidence position bucket (n=597, 200 reports). Questions bucketed by relative chart position $p = k/K$ into equal thirds. Middle third is hardest for 4 of 6 models; Claude-Sonnet-4.5 steepest decline (−11.7 pp).

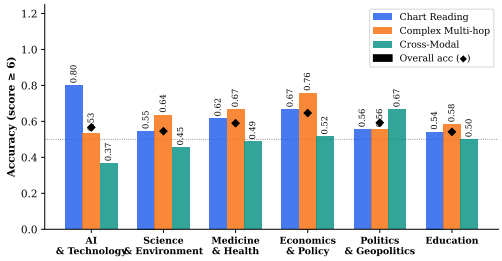


Figure 11: Gemini-3.1-Pro accuracy (ACC, τ=6) per topic domain, grouped by question type. Black diamonds = overall accuracy per domain. Dashed line at 0.5. Based on the 57 domain-annotated reports (513 questions).

H Domain Analysis

Figure 11 shows Gemini-3.1-Pro's accuracy stratified by six topic domains and three question types. Domain annotations cover 57 of the 200 reports (original curated set); the remaining 143 are not grouped by domain. Within the annotated subset, cross-modal accuracy falls 13–16 pp below chart-reading and complex in every domain; the gap is widest in AI & Technology topics.

I Question-Category Breakdown (Full Table)

Table 12 provides the complete numerical values visualised in Figure 7.

Category	Gemini-3.1	GPT-4o	Claude-S-4.5	Qwen3-VL	InternVL3	Qwen2.5-7B
Chart Reading						
Value reading	**0.784**	0.271	0.382	0.648	0.437	0.015
Comparison	**0.759**	0.377	0.266	0.623	0.327	0.075
Trend / pattern	0.734	**0.724**	0.412	0.653	0.603	0.397
Cross-Modal						
Verify with chart	**0.616**	0.399	0.278	0.480	0.318	0.045
Integrate sources	**0.692**	0.298	0.260	0.545	0.293	0.030
Compare repr.	**0.576**	0.328	0.213	0.485	0.278	0.126
Complex Reasoning						
Historical/timeline	**0.910**	0.553	0.508	0.784	0.598	0.015
Technical	**0.643**	0.111	0.137	0.427	0.156	0.000
Impact / reception	**0.814**	0.417	0.364	0.623	0.437	0.020

Table 12: ACC (τ=6) by fine-grained question category (numerical companion to Figure 7). Best per row in **bold**.

J Performance by Question Subset

Figure 12 reports ACC (τ=6) broken down by question subset (chart-reading, complex, cross-modal) for all six models. Cross-modal integration is consistently the hardest subset

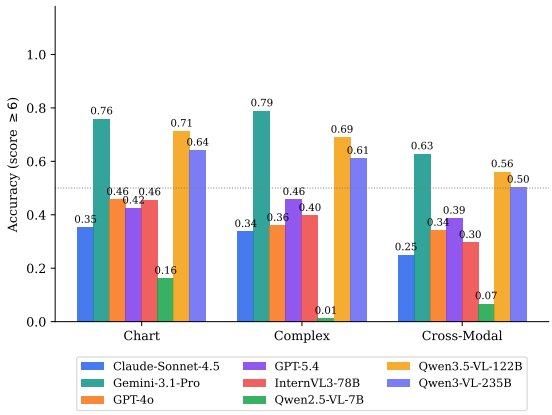


Figure 12: ACC (τ=6) by question subset. Cross-modal questions are consistently hardest across all models, confirming modality alignment as the primary bottleneck.

across all models, confirming modality alignment as the primary bottleneck in long-context document understanding.

K Question Category Taxonomy

Table 14 defines all nine fine-grained question categories used in the category-level breakdown of Table 12. Each category belongs to one of the three question subsets and targets a distinct reasoning capability. The difficulty label *L*1–*L*5 is orthogonal to category: any category can appear at any difficulty level depending on the complexity of the evidence chain.

Level	Modality	Operation	Example
L1: Direct lookup	Chart / Table	Read-off	`What is the value for X in chart Y?`
L2: Comparison	Chart / Table	Compare	`Which is larger, A or B?`
L3: Aggregation	Chart / Table	Compute	`What is the total or average of ...?`
L4: Domain reasoning	Chart + Text	Infer	`Why does X outperform Y given the context?`
L5: Visual interpretation	Chart only	Interpret	`What does colour C encode in chart Y?`

Table 13: Difficulty taxonomy for SYNTHDOCBENCH

Table 14: Fine-grained question category definitions for SYNTHDOCBENCH. Categories are grouped by subset; all nine are mutually exclusive within their subset.

Category	**Definition**
	Chart Reading
Value Reading	Extract one or more exact numerical values from a single chart or table element (e.g., read a bar height, a cell in a table, or a point on a line). No cross-source reasoning is required; the answer is localized to a single visual element.
Comparison	Compare two or more values within the same chart or table to identify the larger, smaller, or closest item, or to compute a difference or ratio. The answer requires reading at least two elements from the same visual.
Trend / Pattern	Describe the directional behavior, distribution shape, or temporal pattern visible in a chart (e.g., "increasing trend," "bimodal distribution," "peaks in Q3"). Exact numerical precision is secondary to correctly characterizing the overall pattern.
	Cross-Modal
Verify with Chart	A specific quantitative claim appears in the document text. The model must locate the corresponding chart and confirm, refute, or quantify the claim using visual evidence.
Integrate Sources	Answer a question whose complete response requires combining a textual claim with a specific numerical value readable only from an associated chart. Neither source alone is sufficient; the model must execute a four-step pipeline: locate claim $\rightarrow$ find chart $\rightarrow$ parse value $\rightarrow$ synthesize.
Compare Representations	The same phenomenon is described both in text (e.g., a percentage trend) and in a chart (e.g., a line plot). The model must reconcile or contrast these two representations, identifying any discrepancy or providing a richer joint characterization.
	Complex Reasoning
Historical / Timeline	Synthesize facts distributed across multiple sections or time periods to construct a causal or chronological narrative (e.g., sequence of policy changes, evolution of a metric over decades). Requires integrating at least two temporally separated evidence units.
Technical / Quantitative	Apply domain-specific knowledge or multi-step arithmetic to evidence recovered from the document (e.g., derive a compound growth rate, convert units, or evaluate a quantitative tradeoff between two technical approaches).
Impact / Reception	Assess the downstream effects, societal implications, or critical reception of a phenomenon discussed in the document. Requires synthesizing evaluative language from multiple sections rather than extracting a single localized fact.

L Additional Implementation Details

This appendix summarizes implementation details that support reproducibility but are not strictly necessary for understanding the main methodological contributions. We include them here to document the design space of document layouts, the structured schema used to encode visualization semantics, the serialized format of generated QA items, and the rendering constraints imposed by the evaluation harness.

L.1 Notation Glossary

For readability, the main text introduces notation only when needed. Table 15 collects the full symbol glossary in one place. The notation spans the three benchmark stages: synthetic document generation, structured question generation, and image-based evaluation.

Table 15: Notation used in the SYNTHDOCBENCH methodology.

Symbol	Type	Definition
τ	string	Topic seed used to initialize document generation
$\mathcal{A}$	set	Layout archetype space
$a \sim P_{\mathcal{A}}(\tau)$	sample	Archetype sampled from the topic-conditioned archetype distribution
$\mathcal{D}$	document	Final rendered document, represented as HTML and exported PDF
$\mathcal{V}_k$	object	Structured metadata object associated with the k-th visualization in $\mathcal{D}$
$\mathcal{T}_k$	object	Structured metadata object associated with the k-th table in $\mathcal{D}$
$\mathcal{M}$	manifest	Document-level QA manifest aggregating all $\mathcal{V}_k$ and $\mathcal{T}_k$
q	string	Natural-language evaluation question
a^*	string	Reference answer derived deterministically from $\mathcal{M}$
$\hat{a}$	string	Model-predicted answer for q from rendered page image(s) of $\mathcal{D}$
f_θ	model	Candidate vision-language model parameterized by θ
$\mathcal{I}$	sequence	Rendered page-image sequence supplied to the evaluated model
ℓ	integer	Difficulty level associated with a question
$\mathcal{J}$	model	Judge model used to score $(\hat{a}, a^*)$

L.2 Layout Archetypes

The report-generation pipeline uses a small but diverse inventory of layout archetypes to vary document structure, rhetorical style, and chart placement strategy. These archetypes control the global page grammar of a report, including whether visualizations are embedded inline, arranged in grids, or surfaced as full-width elements. They also affect the types of auxiliary components that can appear, such as metric strips, pull quotes, or timeline bands. Table 16 lists the archetypes used in SYNTHDOCBENCH together with their distinctive design characteristics.

L.3 Visualization Metadata Schema

A key property of the benchmark is that every generated visualization is accompanied by a structured metadata object that records its semantic content independently of the rendered pixels. This metadata is used downstream for deterministic QA generation and for consistency validation between rendered charts and their underlying values. The schema is designed to capture both data semantics (e.g., chart values, axes, and derived insights) and presentation-level attributes (e.g., legend presence, color encoding, and highlighted items). Table 17 summarizes the fields stored for each chart.

L.4 Question Output Schema

Generated QA items are serialized in a structured format so that evaluation can be performed not only at the answer level but also at the level of question family, difficulty, and supporting

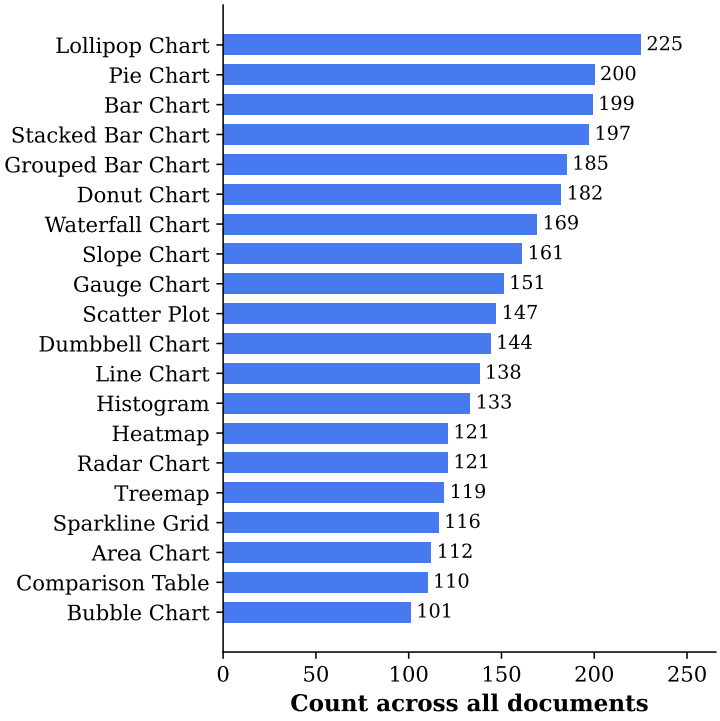


Figure 13: Distribution of chart types (top 20 shown). The corpus covers 24 distinct types spanning common (bar, line, scatter) and specialized (dumbbell, sankey, lollipop) forms.

Table 16: Layout archetypes used in SYNTHDOCBENCH.

Archetype	Viz layout	Font mood	Max vizzes	Distinctive features
Magazine	Hero + smaller	Editorial	3	90vh hero, drop caps, gradient overlays, pull quotes
Dashboard	2× grid	Technical	4	KPI strip, metric cards, compact chart grids
Academic	Inline	Scholarly	2	Abstract box, numbered sections, formal tables
Editorial	Breakout column	Narrative	2	Reading progress cues, chapter markers, block quotes
Infographic	Full-width stack	Bold	4	Icon strips, timeline bands, data callouts
Brutalist	Full bleed	Raw	2	Thick borders, monospace, oversized numbers, stamp labels

Table 17: Visualization metadata schema used by the report-generation pipeline.

Field	Type	Contents
`vizId`	string	Unique identifier of the form `viz-{section}-{index}`
`chartType`	enum	Chart type associated with the rendered visualization
`title`	string	Descriptive chart title
`axes.x / axes.y`	object	Axis label, type, optional unit, and optional range
`data`	array	[{label, value, category, metadata}] — one entry per rendered mark
`insights`	array	Typed facts such as maximum, minimum, comparison, gap, trend, outlier, or proportion
`visualProperties`	object	colorEncoding, colorMap, sortOrder, hasLegend, hasGridlines, highlightedItems, annotationCount
`sourceContext`	string	Attribution to the source passage supplying the underlying values

evidence. In addition to the surface question and reference answer, the output schema stores evidence traces such as required facts, chart indices, and required data points, which are useful for debugging, validation, and downstream error analysis. Table 18 describes the fields included in each serialized QA record.

Table 18: Structured output schema for generated question-answer pairs.

Field	Question types	Contents
`question`	All	Natural-language question string
`answer`	All	Reference answer
`question_type`	All	Question family/type label
`difficulty`	All	Difficulty level $L1$–$L5$
`required_facts`	Multi-hop	Indices into the extracted evidence list
`required_facts_text`	Multi-hop	Verbatim supporting evidence
`fact_sources`	Multi-hop	Source section for each supporting fact
`reasoning`	All	Stepwise evidence-to-answer derivation
`chart_index`	Visual, Cross-modal	Index of the referenced chart
`chart_title`	Visual, Cross-modal	Title of the referenced chart
`required_data_points`	Visual, Cross-modal	Specific {label, value} pairs required for the answer
`category`	Visual, Cross-modal	Fine-grained category label

L.5 Rendering Hyperparameters

The evaluation harness operates on rendered page images rather than HTML or PDF source, so rendering choices directly affect the visual evidence available to the model. The hyperparameters listed in Table 19 define the image construction process, including rasterization resolution, maximum page budget, batching of pages into concatenated strips, and compression constraints required to satisfy API limits across providers. These values therefore influence both evaluation efficiency and the effective difficulty of the visual inference problem.

L.6 Inference and Judge Configuration

Table 20 summarizes the fixed inference configuration applied to both candidate models and the judge. All models are queried at temperature 0 to ensure fully deterministic outputs across runs.

Table 19: Rendering and compression hyperparameters.

Parameter	Description	Value
RESOLUTION	PDF rasterization DPI	144
MAX_PAGES	Maximum pages extracted per PDF	120
CONCAT_NUM	Pages per concatenated image strip	5
COLUMN_NUM	Columns per concatenated image	1
MAX_IMG_BYTES	Maximum compressed image size	4 MB
MAX_DIMENSION	Maximum pixel dimension (longest side)	7,900 px
ACC_THRESHOLD	Minimum judge score counted as correct	6

Table 20: Inference configuration for candidate models and the judge.

Parameter	Value	Notes
Judge model	GPT-5	Fixed across all experiments
Temperature	0	Applied to both candidates and judge
Candidate response	2–4 sentences	Enforced via system prompt
Judge format	JSON only	`{"score": <0-10>, ...}`
Parse failure	Score $= -1$	Excluded from Q_{valid}

L.7 Prompts

All prompts are shown verbatim as supplied to the model APIs. Text in *italics* denotes fields populated at runtime.

Candidate model — default (concise) prompt.

System

```
You are an expert analyst reviewing a PDF report.
Answer the question using ONLY what is visible in the provided pages.
Be concise and direct — 2–4 sentences maximum.
Include specific numbers, dates, or names only if directly relevant.
Do not restate the question. Do not explain your reasoning. Just answer.
If the answer is not in the document, say: “Not found in the document.”
```

User

```
[Rendered page images supplied as vision input]
Question: {question}
```

Candidate model — chain-of-thought (CoT) prompt.

System

```
You are an expert analyst reviewing a PDF report.
Answer the question using ONLY what is visible in the provided pages.
Think step by step: first identify the relevant chart or section, then extract
the key facts, then synthesize your final answer.
Include specific numbers, dates, or names where relevant.
If the answer is not in the document, say: “Not found in the document.”
```

User

```
[Rendered page images supplied as vision input]
Question: {question}
```

Candidate model — no system prompt baseline. In the *none* condition no system prompt is supplied. The user turn contains only the rendered page images followed by the question, identical to the user turns above.

Judge model prompt.

System

```
You are a factual accuracy evaluator.
Score the CANDIDATE ANSWER against the REFERENCE ANSWER for a document-based Q&A
task.

Rubric (0-10):
  10 = all key facts present and correct
  8-9 = mostly correct, minor gap or imprecision
  6-7 = core correct, some detail missing
  4-5 = partial, significant gaps or errors
  2-3 = mostly wrong or missing
  0-1 = incorrect or hallucinated

Respond with ONLY this JSON (no markdown, no extra text):
{"score": <0-10>, "judgement": "<2-3 sentences on accuracy>"}
```

User

```
QUESTION: {question}

REFERENCE ANSWER: {gt_answer}

CANDIDATE ANSWER: {candidate_answer}

Score 0-10 and explain. Return JSON only.
```

M Hard Failure Error Category Definitions

Subset	Total	Ch.	Cx.	XM.
Hard failures (all $6 \leq 3$)	**109**	**37**	**14**	**58**

Table 21: Hard failure counts by question subset (all six models score ≤ 3). **Ch.** = chart-reading; **Cx.** = complex; **XM.** = cross-modal.

Table 22 defines the seven mutually exclusive error categories used to classify hard failures in Section 5. Each category was assigned by GPT-5 based on the question, ground-truth answer, and all model responses. The per-category count breakdown is in Table 21.

Table 22: Error category definitions for the hard failure analysis (Section 5). Categories are mutually exclusive; each failure is assigned exactly one.

Category	Definition
Visual Hallucination	The model confidently reads a wrong value from the chart. The figure is located and the relevant element identified, but the pixel-level value extraction is incorrect (e.g., reports 1,000 ppm instead of 30 ppm).
Figure Not Found	The model cannot locate the referenced figure in the document. It either returns "not found in the document" or retrieves a different, unrelated figure.
Precision Error	The model extracts the correct chart element but with the wrong exact value, scale, or unit (e.g., reads 29% instead of 27.5%, or reports billions instead of millions).
Incomplete Retrieval	The model retrieves part of the required evidence but misses one or more key facts distributed across the document, producing an answer that is partially correct but substantively incomplete.
Cross-Modal Grounding	Both the textual claim and the relevant chart are correctly identified in isolation, but the model fails to perform the required quantitative alignment or comparison between the two modalities.
Reasoning Error	The necessary evidence is correctly retrieved from the document, but the model draws a wrong conclusion through a logical or arithmetic mistake in the final inference step.
Question / Annotation Ambiguity	The question or ground-truth answer is ambiguous, underspecified, or requires information that is not recoverable from the visible document pages.

N Ablation Study Details

This section documents the exact experimental configuration used for each ablation study. All ablations use Gemini-3.1-Pro as the candidate model and GPT-5 as judge unless otherwise stated. Each condition is evaluated on a representative full 200-report benchmark (1,788 questions), and the same judge prompt is used throughout.

N.1 Page Concatenation Ablation

The evaluation harness renders each PDF page as a raster image and groups consecutive pages into a single vertical image strip before passing them to the model. The `concat-num` hyperparameter controls how many pages are stacked per strip.

Parameter	Value	Note
`-concat-num`	1, 2, 5, 10	Pages per image strip
`-resolution`	144	DPI fixed
`-prompt-style`	default	Concise system prompt
Model	gemini-3.1-pro	Fixed candidate
Total images/doc	$\lceil$pages / concat-num$\rceil$	Varies

Table 23: Concat-num ablation configuration.

Parameter	Value	Note
`-resolution`	72, 144, 216	PDF rasterization DPI
`-concat-num`	5	Fixed (default)
`-prompt-style`	default	Fixed
Max bytes/image	4 MB	JPEG recompressed if exceeded
Max dimension	7,900 px	Downscaled if exceeded

Table 24: DPI ablation configuration.

At 72 DPI the resulting strip is typically around 1,200 px tall; at 144 DPI around 2,400 px; at 216 DPI around 3,600 px before any compression rescaling. At `concat-num=1` each page is sent individually, maximising resolution per image but fragmenting document context across many turns. At `concat-num=10`, up to ten pages are merged into one tall strip, providing wider context at reduced per-line resolution. The default `concat-num=5` balances these trade-offs.

N.2 Rendering Resolution Ablation

PDF rasterization converts each page to a pixel image at the specified DPI. Higher DPI produces finer text and chart detail but larger file sizes that may trigger per-image byte-limit compression.

N.3 Prompting Strategy Ablation

Across all three models, the no-prompt condition achieves the highest complex-question ACC (Gemini +5.3 pts, GPT-4o +5.9 pts, Claude +17.0 pts over default), with Claude's complex ACC jumping from 0.380 to 0.550, suggesting the concise-answer framing actively constrains multi-step reasoning. CoT yields modest overall gains (Gemini: +1.6 pts; Claude: +3.0 pts) but reduces GPT-4o chart-reading ACC (0.520→0.468), where longer outputs introduce hallucinated intermediate steps. Chart-reading and cross-modal ACC vary by at most 5 points across all strategies, confirming these subsets are driven by visual perception rather than output formatting.

Parameter	Values	Note
Judge models	GPT-5, Gemini, Claude	Same rubric/JSON
Candidate models	GPT-4o, Qwen3, InternVL3	Pre-scored answers
ACC threshold τ	4, 5, 6, 7, 8	GPT-5 judge only
Temperature	0	All judges

Table 25: Judge sensitivity ablation configuration.

Style	System prompt	Description
`default`	Concise (App. L.7)	2–4 sentence answer
`cot`	CoT (App. L.7)	Step-by-step reasoning
`none`	None	No system prompt

Table 26: Prompting ablation configuration.

N.4 Judge Sensitivity Ablation

Judge sensitivity results are in Table L.7. Across all three models, the no-prompt condition achieves the highest complex-question ACC (Gemini +5.3 pts, GPT-4o +5.9 pts, Claude +17.0 pts over default), with Claude's complex ACC jumping from 0.380 to 0.550, suggesting the concise-answer framing actively constrains multi-step reasoning. CoT yields modest overall gains (Gemini: +1.6 pts; Claude: +3.0 pts) but reduces GPT-4o chart-reading ACC (0.520→0.468), where longer outputs introduce hallucinated intermediate steps. Chart-reading and cross-modal ACC vary by at most 5 points across all strategies, confirming these subsets are driven by visual perception rather than output formatting.

		GPT-4o	Qwen3-VL	InternVL3
Judge model	GPT-5 (default)	0.434	0.535	0.354
	Gemini-3.1-Pro	0.417	0.522	0.363
	Claude-Sonnet-4.5	0.607	0.705	0.573
Threshold τ	$\tau = 4$	0.627	0.697	0.568
	$\tau = 5$	0.483	0.575	0.398
	$\tau = 6$ (default)	0.434	0.535	0.354
	$\tau = 7$	0.360	0.451	0.311
	$\tau = 8$	0.299	0.413	0.269

Table 27: Judge sensitivity results. *Top:* ACC for three candidates under three judge models. *Bottom:* ACC under GPT-5 at thresholds $\tau \in \{4, 5, 6, 7, 8\}$. Rankings are fully preserved across all conditions.

O Qualitative Model Response Examples

We present four representative examples from SYNTHDOCBENCH spanning the three question subsets and dominant failure modes from Section 5. Each example shows (i) a zoomed-

in crop of the relevant evidence page, (ii) the question and ground-truth answer, and (iii) the annotated response from a representative subset of evaluated models, where **green** marks correct facts and **red** marks hallucinated or wrong values.

Example 1 — Chart Reading L1 *Mixed: Value Reading from Horizontal Bar*

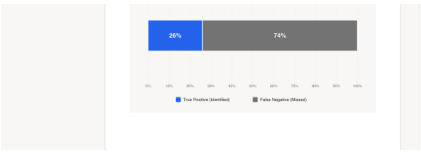


*Figure 14 from the document (p.19) — OpenAI AI Text Classifier Performance on AI Text. The bar shows **26% True Positive** (Identified) and **74% False Negative** (Missed).*

Question *(Chart Reading, L1 — AI in the Classroom)*

According to Figure 14 ('OpenAI AI Text Classifier Performance on AI Text (%)'), what percentage of AI-generated text resulted in a **False Negative (Missed)**?

Ground Truth

74% of AI-generated text resulted in a False Negative (Missed).

Gemini-3.1-Pro `10/10`:

According to Figure 14, **74%** of AI-generated text resulted in a False Negative (Missed).

GPT-4o `0/10`:

According to Figure 14, **43%** of AI-generated text resulted in a False Negative (Missed).

Claude-Sonnet-4.5 `10/10`:

Figure 14 shows that **74%** of AI-generated text was classified as a False Negative (Missed), meaning the classifier failed to identify it as AI-written.

Qwen3-VL-235B `0/10`:

Based on Figure 14, **26%** of AI-generated text resulted in a False Negative (Missed).

Qwen2.5-VL-7B `0/10`:

According to Figure 14, **26%** of AI-generated text resulted in a False Negative (Missed).

InternVL3-78B `0/10`:

According to Figure 14, the percentage of AI-generated text that resulted in a False Negative (Missed) is **16%**.

The chart clearly labels both segments: 26% (True Positive, blue) and 74% (False Negative, gray). Only Gemini and Claude read the correct gray segment. GPT-4o hallucinates 43%—a value absent from the chart. Both Qwen models confuse the two segments, reading the True Positive bar (26%) as the answer. InternVL3 hallucinates 16%.

Example 2 — Cross-Modal L2 *Failure: Visual Hallucination*

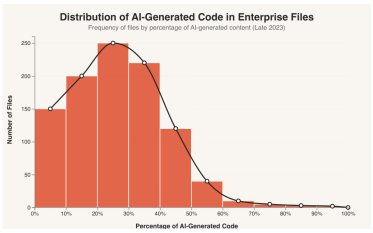


*Figure 1 from the document (p.5) — Distribution of AI-Generated Code in Enterprise Files. The 40–50% bin has **120 files**; the 50–60% bin has **40 files**.*

Question *(Cross-Modal, L2 — AI-Native Software Dev.)*

The text describes the distribution dropping 'sharply after the 40–50% mark.' How does Figure 1 quantify this as the **exact file frequency drop** between the 40–50% and 50–60% bins?

Ground Truth

40–50% bin: **120 files** → 50–60% bin: **40 files**. Drop of **80 files** (66.7% reduction).

Gemini-3.1-Pro `0/10`:

In Figure 1, the 40–50% bin shows a frequency of **200 files**. The subsequent 50–60% bin drops to a frequency of **50 files**. This represents an exact numerical drop of **150 files** between the two bins.

GPT-4o `0/10`:

Figure 1 indicates that the file frequency drops from approximately **17%** in the 40–50% bin to approximately **8%** in the 50–60% bin, representing a numerical decline of about **9 percentage points**.

Claude-Sonnet-4.5 `1/10`:

Not found in the document. While Figure 1 shows the distribution and the text describes a sharp drop after the 40–50% mark, the specific numerical values for file frequency in the two bins are not provided in the visible portions of the document.

Qwen3-VL-235B `2/10`:

Figure 1 shows the frequency of files for the 40–50% bin is approximately **175**. For the next bin, 50–60%, the frequency drops sharply to approximately **50**. This represents a numerical drop of about **125 files**.

Qwen2.5-VL-7B `0/10`:

The figure shows that the number of files drops from approximately **250** in the 40–50% bin to around **150** in the 50–60% bin.

InternVL3-78B `0/10`:

Not found in the document.

Every model hallucinates a different plausible-looking but wrong pair of bar heights. GPT-4o conflates file counts with axis percentages. Claude and InternVL3 decline to answer despite the chart being visible on p.5. No model reads the correct 120→40 counts.

Example 3 — Cross-Modal L3 ***Failure: Figure Not Found***

Figure 3 from the document (p.8) — GitHub Copilot Productivity Study. ***Dev 6*** *(control) reaches* ***180 min****; Dev 6 (treatment) is at* ***85 min****.*

Question *(Cross-Modal, L3 — AI-Native Software Dev.)*

The text states the longest task time in the GitHub Copilot study was 180 min by a developer in the control group. Using Figure 3, which **specific developer** recorded this time, and what was their time in the **Treatment Group**?

Ground Truth

Dev 6 (Control: 180 min → Treatment: **85 min**).

Gemini-3.1-Pro `0/10`:

Not found in the document.

GPT-4o `0/10`:

The document does not specifically identify which developer recorded the 180-minute maximum time nor their completion time in the Treatment Group.

Claude-Sonnet-4.5 `0/10`:

Not found in the document. While the document mentions the 180-minute maximum completion time in the control group, it does not identify which specific developer recorded this time or provide their individual completion time in the treatment group.

Qwen3-VL-235B `0/10`:

Not found in the document.

Qwen2.5-VL-7B `0/10`:

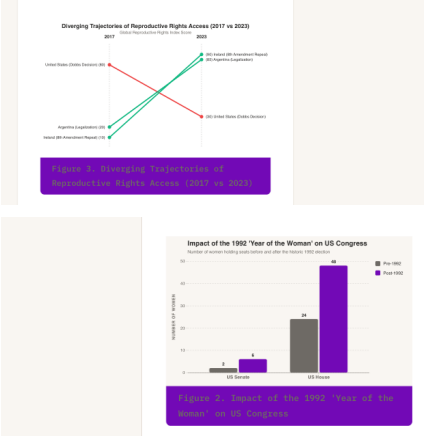
Not found in the document.

InternVL3-78B `0/10`:

Not found in the document.

Figure 3 is clearly visible on p.8 with per-developer labels on the x-axis (Dev 1 through Dev 10). The highest gray dot at 180 min sits at Dev 6. All six evaluated models fail to localise and resolve the individual data point label, returning a uniform "not found" despite the chart being visually present.

Example 4 — Complex Multi-Hop L5 *Failure: Incomplete Retrieval*



Two of three evidence figures from p.5 of the gender equality document: Figure 3 (Reproductive Rights Trajectories, top) and the Congress representation figure (bottom). A third chart on the same page shows child marriage rates 2013–2023.

Question *(Complex Multi-Hop, L5 — Gender Equality)*

How did women's representation in US Congress change after the 1992 'Year of the Woman'? What was the exact decline in global child marriage rates 2013–2023? How does the US reproductive rights 50-point drop compare to changes in **Ireland** and **Argentina** over 2017–2023?

Ground Truth

House: **24**→**48** (doubled). Senate: **2**→**6** (tripled).
Child marriage: **25**→**20** (drop of 5).
US: **80**→**30**. Ireland: **10**→**90**. Argentina: **20**→**85**.

Gemini-3.1-Pro `4/10`:

Following the 1992 election, voters elected 4 new female senators (bringing the Senate total to **6**) and **24** new women to the House of Representatives. Between 2013 and 2023, global child marriage rates declined from **25%** to **19%**. During the 2017–2023 timeframe, the 50-point drop in the US Reproductive Rights Index (from **75** to **25**) contrasted with a 60-point increase in Ireland (from **20** to **80**) and a 45-point increase in Argentina (from **30** to **75**).

GPT-4o `2/10`:

Following the 1992 'Year of the Woman' election, the number of women in the U.S. Congress **doubled**. The global child marriage rate **declined by 15%** between 2013 and 2023. While the U.S. reproductive rights index dropped by **50 points** from 2017 to 2023, during the same timeframe, Ireland's index rose by **25 points** and Argentina's by **30 points**.

Claude-Sonnet-4.5 `0/10`:

Not found in the document.

Qwen3-VL-235B `3/10`:

Following the 1992 'Year of the Woman', the number of women in the US Congress increased from **4** to **6**. Global child marriage rates declined from **25%** in 2013 to **20%** in 2023, a **5 percentage point** drop. The US reproductive rights index dropped **50 points** from 2017 to 2023, while Ireland's index rose from **70** to **90** and Argentina's rose from **60** to **85** during the same period.

Qwen2.5-VL-7B `2/10`:

The number of women in the US Congress increased from **4** to **44** after the 1992 'Year of the Woman' election. The global child marriage rate declined by approximately **25 percentage points** between 2013 and 2023. The US reproductive rights index dropped by **50 points**, while Ireland saw a **10-point increase** and Argentina experienced a **10-point decrease** during the same period.

InternVL3-78B `4/10`:

Following the 1992 'Year of the Woman' election, the number of women elected to the U.S. House of Representatives increased by **24**, marking a then-record high. The global child marriage rate declined by **12 percentage points** between 2013 and 2023. The U.S. reproductive rights index experienced a **50-point drop**, contrasting with Ireland and Argentina, which saw **significant improvements** in their index scores during the 2017–2023 period.

The question requires six independent statistics from three charts on the same page. Gemini retrieves the Senate total and House count but uses wrong base scores for all three countries. Qwen3 correctly reads the child marriage drop and US delta but has wrong starting values for Ireland and Argentina. GPT-4o correctly identifies the Congress doubling and US 50-point drop but reports wrong deltas for the other countries. InternVL3 correctly retrieves the House count and US drop but vaguely describes Ireland/Argentina without numbers. Claude fails to retrieve any evidence. The Ireland (10→90) and Argentina (20→85) base scores are consistently missed.