ReflowDiff: Can Vision-Language Models Reliably Compare Documents across Layouts?

Anonymous authors
Paper under double-blind review

Abstract

Vision-language models (VLMs) increasingly read, write, and revise documents. Yet it remains unclear whether they can report every textual revision when a document's layout also changes. Concretely, one changed digit or sign alters what a contract or financial report says, and catching it requires fine-grained character recognition, matching corresponding text across layouts, and exhaustive reporting. We introduce ReflowDiff, a benchmark of 4,488 comparison tasks built from 374 dense rich-text pages spanning six generated domains, real web pages, and financial reports. Each page is rendered in two layouts that preserve its text and structural context, and the same set of one to five character edits is applied to both. Two no-edit controls, Identical and Reflow only, test whether models report edits when the text is unchanged. We leverage exact match as the primary metric, which requires every reference edit, with exact before and after text at the correct occurrence, and no additional edits. Across 30 configurations of 26 VLMs, the best reaches 76.4% exact match when edits are combined with reflow. Reflow lowers exact match on the same edits in all 30 configurations, including a 3.8-percentage-point drop for the strongest, and elicits false reports even from configurations that make none on identical pairs. Even when models recover most individual edits, their reports remain incomplete and vary across layouts. These failures highlight a gap between reading document content and reliably verifying what has changed.

1 Introduction

Vision-language models (VLMs) have become general-purpose readers of documents, charts, and infographics, and increasingly act as agents that write and revise documents (Mathew et al., 2021; Masry et al., 2022; Liu et al., 2024; Bai et al., 2025b). A routine task in this role is comparing two versions of a document: contracts, financial reports, and specifications are revised repeatedly, and one changed digit, sign, or unit changes what they say. Unlike answering a question about a selected field, a comparison must report every textual revision at its correct occurrence, and one missed or spurious item fails it. Strong scores on single-document reading tasks do not establish this ability.

Layout changes make the distinction sharp. A revised report can switch from one column to two, rewrap every paragraph, and move a table across the page while changing a single digit in one cell. A correct comparison reports that digit, in that cell, and a page rearranged with unchanged text calls for an empty report. The task thus combines fine-grained perception of characters on dense pages, layout-invariant correspondence that ties each string to its occurrence and structural context, and exhaustive reporting. Reliable document comparison requires sensitivity to small textual edits and invariance to changes in presentation. This paper asks: *Can VLMs report the same character edits completely and accurately across different document layouts?*

Existing evaluations measure different objects. Document question answering scores the requested answer, and text-recognition benchmarks score transcriptions and answers over text-rich images (Mathew et al., 2021; Liu et al., 2024). Change-detection and change-captioning benchmarks score captions, change maps, or mutation records (Jhamtani & Berg-Kirkpatrick, 2018; Park et al., 2025; Zhang et al., 2026), not exact text at every changed occurrence.

Evaluating reports across layouts further requires the edits to be held fixed, because comparing different edits under different layouts confounds presentation with edit difficulty. Layout also changes

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

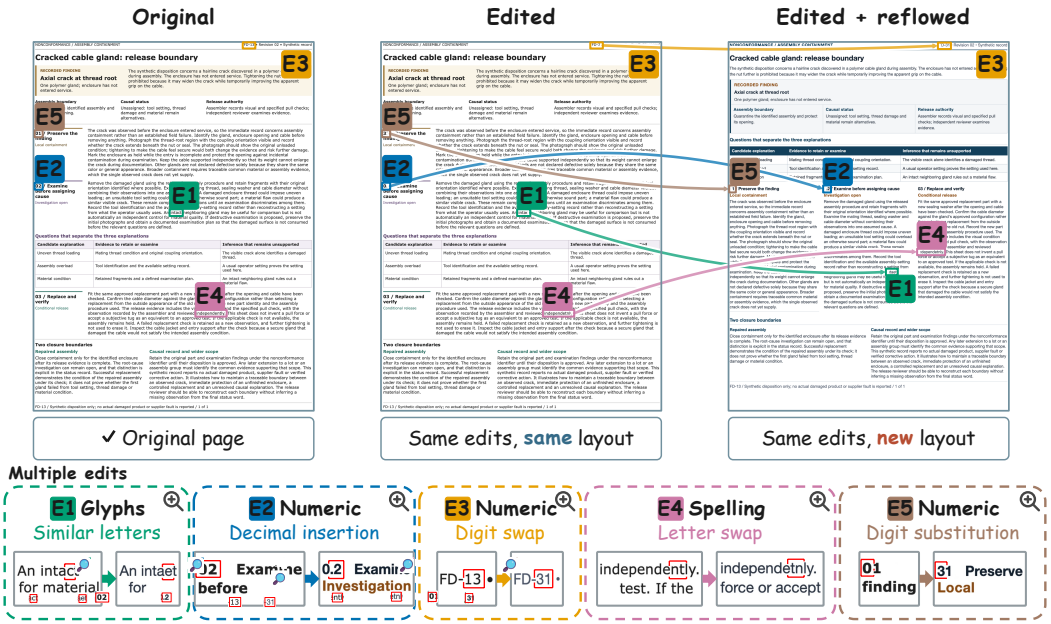


Figure 1: Two paired comparisons that share the original page and the same five character edits. Each task pairs the original page (left) with the edited page in the original layout (middle, Edit only) or the edited, reflowed page (right, Edit + reflow). Both edited tasks share the reference edit set E1–E5 from different edit families (Fig. 3), at occurrences in different image positions. The correct report lists all five edits and nothing else. Boxes and crops are illustrations, not model inputs.

what the pixels reveal: when an edit is the sole visual change, differing pixels mark it, and a pixel localizer that reaches 23.6% region F1 on edited pages in the original layout drops to 0.13% once the page is reflowed (Tab. 9). No-edit controls are needed as well, to measure whether a change of presentation alone elicits a report.

We introduce REFLOWDIFF, a diagnostic benchmark of 4,488 comparison tasks built from 374 dense rich-text pages across 6 generated domains, real web pages, and financial reports (Fig. 1). The construction fixes each page's text together with its structural context, such as table rows and columns, and renders a second layout that carries exactly the same strings in the same structural context. One set of 1 to 5 character edits, addressed by element identifier rather than image position, is applied to both layouts, so the Edit only and Edit + reflow tasks share a reference edit set and the Identical and Reflow only controls share an empty reference. A task is solved under exact match, which requires every reference edit at its correct occurrence and no additional report.

Across 30 configurations of 26 proprietary and open-weight VLMs, the best configuration reaches 76.4% exact match on edited, reflowed tasks, the median configuration 3.3%, and 25 configurations stay below 10%. The layout gap Δ, exact match on Edit + reflow minus that on Edit only over the same edit sets, is negative in all 30 configurations, with 95% confidence intervals below zero in 27.

Paired outcomes reveal far more than the average: the strongest configuration loses a net 3.8 points, yet 22.1% of its edit sets switch between an exact and a non-exact report across the two layouts (§ 5.2). With no edit at all, reflow elicits false reports, nonempty reports on unchanged text: the 10 configurations without false reports on Identical pairs make them on 4.0–96.5% of Reflow only pairs. Per-edit recall, the fraction of reference edits recovered, separates from the complete report: from one to five edits, the strongest configuration's recall rises from 87.4% to 90.4% while its exact match falls from 84.2% to 66.0% (§ 5.3).

Our contributions are threefold. (i) *Evaluation target*: complete, occurrence-specific reporting of character edits that stays correct across document layouts, with the complete report as the unit of success (§2). (ii) *Controlled resource*: 4,488 tasks in which text- and structure-preserving reflow fixes one reference edit set across two layouts, with two no-edit controls and an occurrence-aware

scorer (§§ 3 and 4). (iii) *Insight and release*: our analyses reveal recurring error patterns, including small-glyph and look-alike misses (§ 5.4), different reports on the same edits across layouts (§ 5.2), and incomplete reports despite high per-edit recall, with Reflow only false reports (§§ 5.2 and 5.3). We release the pages, reference edits, prompts, and evaluation code. In Tab. 1, REFLOWDIFF alone requires the whole report to be correct, scores the same edits under two layouts, and adds a Reflow only control, on pages with more than ten times the text of the densest other benchmark (794 vs. 62 OCR words per image).

Table 1: Comparison with image-pair change benchmarks: Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018), CLEVR-Change (Park et al., 2019), OmniDiff (Liu et al., 2025), and DiffSpot (Zhang et al., 2026). *Exact match* scores the complete report, which must contain every literal edit at its correct occurrence and no extra report. *OCR words per image* is the median number of whitespace-separated words that one fixed local OCR engine reads per image (App. B.4). *No-edit controls* are image pairs without a target change. *Paired reflow* scores one edit set with and without text-preserving reflow.

Benchmark	Visual input	OCR words per image	Multiple changes per pair	Exact match	No-edit controls	Paired reflow
Spot-the-Diff	Surveillance video	0	✓	✗	✗	✗
CLEVR-Change	Rendered scenes	0	✗	✗	✓	✗
OmniDiff	Real and 3D scenes	0	✓	✗	✓	✗
DiffSpot	Web screenshots	62	✗	✗	✓	✗
REFLOWDIFF (ours)	Rich-text pages	794	✓	✓	✓	✓

2 TASK

2.1 REPORTING EVERY EDIT AT ITS OCCURRENCE

Given an ordered pair of page images, the original page and then a comparison page, a model reports the text edits from the first page to the second. The output is a set of edits, each a triple of exact before text, exact after text, and a location stated by visible context, such as a section title, a table row and column, or surrounding words. A report succeeds when it contains every reference edit at its correct occurrence and no additional report. When no text changed, the correct report is empty. An edit is identified by its literal change and its occurrence. Suppose a table cell changes from `100` to `10`. A report of `100`→`10` in that cell's row is correct, and so is a report that quotes the whole row with the same change. The identical report placed in another row names a different occurrence and is wrong. The model sees the two images alone (formal definitions in App. A).

2.2 FOUR CONDITIONS, TWO CONTRASTS

Every source page is compared in four conditions, which cross unedited and edited text with the original and a reflowed layout. The original page is always the first image. In *Identical*, the comparison page is the original page itself. In *Reflow only*, it is the same text in the reflowed layout. In *Edit only*, it is the edited page in the original layout, and in *Edit + reflow*, the edited page in the reflowed layout. The design rests on two contrasts. Edit only and Edit + reflow carry the *same* set of $k = 1$ to 5 edits and therefore the same reference edit set, so the two tasks differ in layout alone, and their difference in exact match is the layout gap Δ (§ 4.2). Identical and Reflow only carry no edit and share the empty reference, so any report on Reflow only is a false report elicited by the change of layout. Reflow changes spatial arrangement, line wrapping, and effective character size together, as a reformatted revision does. Edit + reflow is the primary condition: it asks for the complete edit set while the presentation changes.

2.3 EDIT SCOPE

The task targets visible literal changes. One character edit substitutes, deletes, or inserts one Unicode grapheme cluster, or transposes two adjacent distinct graphemes (Tab. 4). Case, punctuation, digits, signs, and units all count. Changes of style such as color or rotation are not character edits. Each edit changes its declared occurrence alone: other copies, dependent totals, and descriptions stay unchanged, so the task scores what the page says, not whether the revision is consistent or plausible

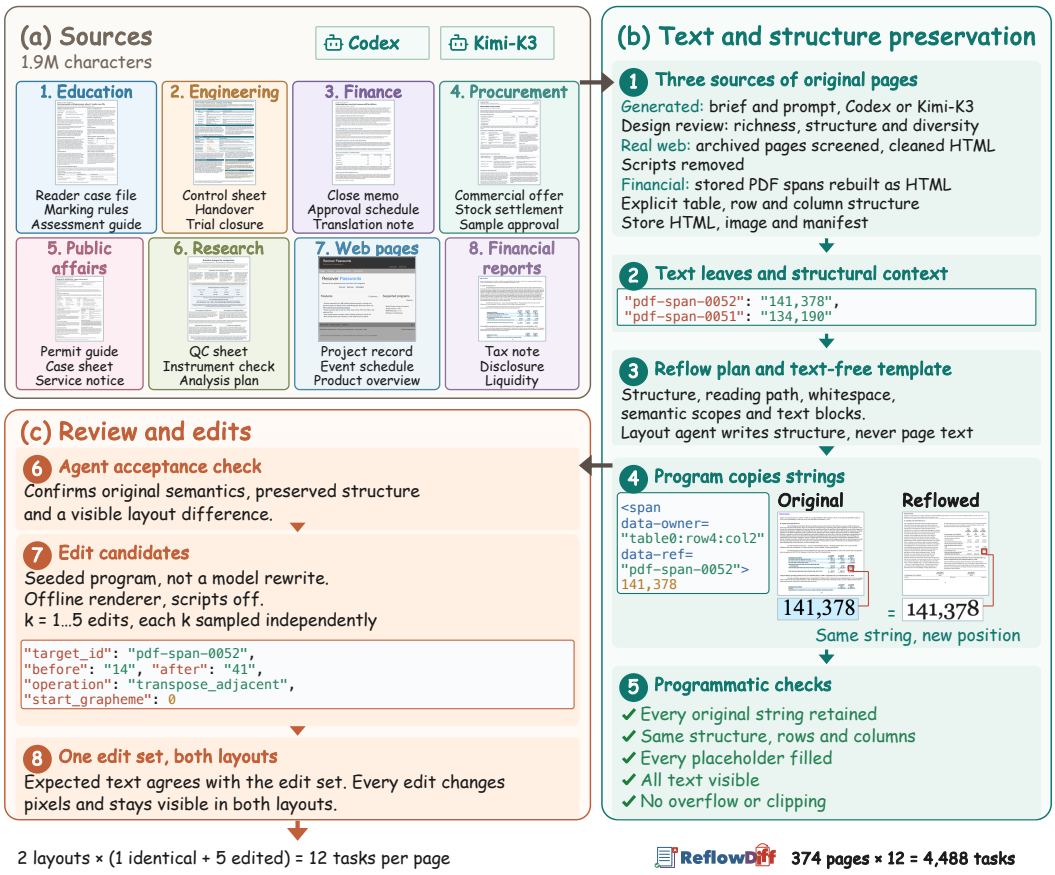


Figure 2: **The reference follows the occurrence, not its coordinates.** Original pages are generated or come from real web pages and real financial reports rebuilt in HTML. Reflow preserves the text: a program records every string with its structural context, an agent designs a text-free layout template that the program fills with the same strings, programmatic checks verify text, structure, and rendering, and an agent confirms the reflowed page. One edit set of $k = 1$–5 edits, addressed by element identifier and character offset, is applied to both layouts, giving 12 tasks per page.

(*local-literal* policy, subtypes in Tab. 5). The edits cover numeric, sign, unit, glyph, spelling, and punctuation changes, with numeric and punctuation edits the most frequent (Fig. 3).

3 Benchmark Construction

Sources and coverage. Paired comparisons require an editable representation that keeps the identity of every text occurrence across layouts. The 374 source pages come from three sources (Fig. 2): 296 pages generated in 6 domains give control over content and layout, while 30 real web pages from the Design2Code archive (Si et al., 2025) and 48 real financial reports from TAT-DQA (Zhu et al., 2022), rebuilt in HTML from the source PDF, bring the structure of documents in use. Each accepted page, its image, and its literal-text inventory are frozen, and no model regenerates page text afterwards (provenance in App. B). The 374 dense pages hold 1.9M characters in 44.7K text blocks and yield 4,488 tasks, 12 per page, and the edited tasks draw on 5,610 unique edits, giving 11,220 reference edit instances across the two layouts (Tab. 6). Across all 374 source pages, the median word count is 874 (Chromium word segmentation of the document object model, App. B.4).

Text and structure preservation. Preserving the strings alone is insufficient: moving an unchanged value to another table row or column changes what the page states. A program therefore extracts every text element of the frozen page with an element identifier and its structural context, namely the heading, paragraph, or note it belongs to and, for a table cell, its row and column.

The agent designs the layout, and a program writes the text. A large language model (LLM) agent writes a new HTML/CSS layout as a text-free template whose placeholders are addressed by the element identifiers, and a program fills each placeholder with its frozen string, so no model rewrites page text. After rendering, programmatic checks confirm that the reflowed page carries exactly the original strings, each in its original context, with complete structure and no overflow or clipping (App. B.2). This second layout, with the same text in the same structural context, makes the contrasts of § 2.2 exact.

Edit generation. A seeded program draws legal character edits from identified text elements (Fig. 7) and records them as an edit set, the reference for both edited tasks. Each edit is addressed by element identifier and grapheme offset rather than image position, so **one edit set is applied to both layouts**, and Edit only and Edit + reflow share the reference edit set by construction. After application, the program compares every text element with its expected text in each layout, confirming that the declared edits occurred and that every other string is unchanged. Edit sets of size $k = 1, \ldots, 5$ are sampled independently. The effect of adding edits to a fixed edit set is measured separately (Tab. 20).

Programmatic checks. Checks of the released text maps confirm: element text is equal across layouts in all 374 original and 1,870 edited layout pairs, and each of the 3,740 edited pages applies exactly its declared occurrence and leaves every other string unchanged (App. B).

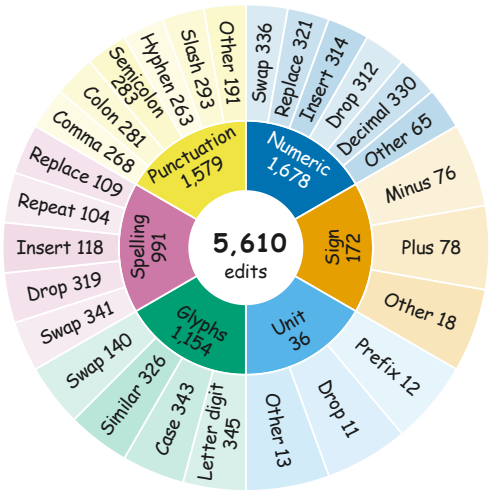


Figure 3: Edit families and edit types across the 5,610 unique edits. The inner ring shows the edit family and the outer ring the edit type.

4 Evaluation Protocol

4.1 Systems and Settings

We evaluate 30 configurations of 26 proprietary and open-weight VLMs, where a configuration is a model with one reasoning setting (Tab. 8). Reasoning marks take three forms. Open-weight models run with reasoning on (✓) or off (✗), set by the request's reasoning switch or by the checkpoint, and on and off rows with one model ID run the same checkpoint. Other open-weight checkpoints run in their default mode, marked ○. Proprietary models run at the provider default with no reasoning parameter sent, and their mark gives the reasoning-token counts reported by the API: ✓ for positive counts, ✗ for zero counts, and ○ when no count was returned. Every solver receives the same prompt (App. A.3) and both page images as native PNGs at `detail=high`, and Tab. 8 lists output budgets. When a strict parser rejects the JSON report, a text-only language model extracts the final answer without images or reference edits and without adding or removing edits, and a second call checks the extraction. Truncated, empty, contradictory, or unextractable responses are invalid.

4.2 Metrics and Occurrence-Aware Scoring

Scoring separates what code decides from what needs judgement. Code decides the strings: it removes the identical prefix and suffix of each before and after pair and keeps a prediction and a reference as a literal candidate when the remaining minimal Unicode changes are identical and nonempty, so a report can quote more or less context than the reference without admitting a different edit. An occurrence judge (Qwen3.8-27B, App. A.2) then resolves the locations of these literal candidates, deciding whether both name the same visible occurrence, and it never overrides an incorrect literal change. It pairs them one to one, and unpaired predictions and references count as false positives and negatives (App. A.2). On the no-edit controls, the reference is empty and code scores the report.

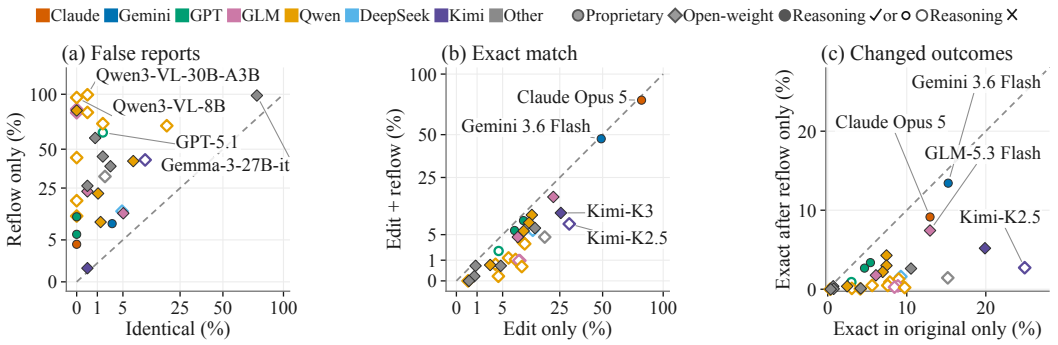


Figure 4: Reflow shifts outcomes in every configuration. False reports are nonempty reports over all tasks of a no-edit condition, and exact match requires every reference edit with no extra report. Panel (c) counts edit sets exact in only one layout, so points near the diagonal but far from the origin combine a small net gap with many changed outcomes. Reasoning marks are as in Tab. 8.

Exact match requires a valid report with every reference edit at its correct occurrence and no additional report. Exact match on Edit + reflow is the primary metric. The layout gap Δ is exact match on Edit + reflow minus exact match on Edit only over the same edit sets. It measures the combined effect of changes in spatial arrangement, line wrapping, and effective character size. The *false-report rate* is the share of the 374 tasks of a no-edit control that end in a valid nonempty report. Per-edit *precision*, the fraction of predicted edits that are correct, pools valid predictions, and per-edit *recall*, the fraction of reference edits recovered, counts every reference edit. Exact match and per-edit recall are computed over all tasks and count invalid outputs as not correct. Uncertainty is a 95% bootstrap confidence interval (CI) with 2,000 draws over 345 clusters that keep all variants of a page together: one per generated page and per financial report, and one for all web pages (Tab. 10).

4.3 Scoring Validation

A second judge, GPT-5.5, re-judges all 10,800 outputs of 9 configurations and reproduces 93.7–99.3% of the primary judge's exact-match decisions, keeping all 9 ranks (Spearman 1.00, Tab. 27).

5 Results

5.1 Overall Performance

ReflowDiff separates the evaluated systems sharply (Tab. 2). Claude Opus 5 reaches 76.4% exact match on Edit + reflow and Gemini 3.6 Flash 47.2%, while the median configuration reaches 3.3% and the best open-weight configuration 16.5%. 25 of the 30 configurations stay below 10%, so complete reporting leaves wide room for progress. Real documents reproduce the ranking: Claude Opus 5 and Gemini 3.6 Flash lead the representative configurations in all 8 source categories, including real web pages and real financial reports (Tab. 25), and single-edit exact match in the original layout is higher on real sources than on generated pages in 26 of 30 configurations (Tab. 26).

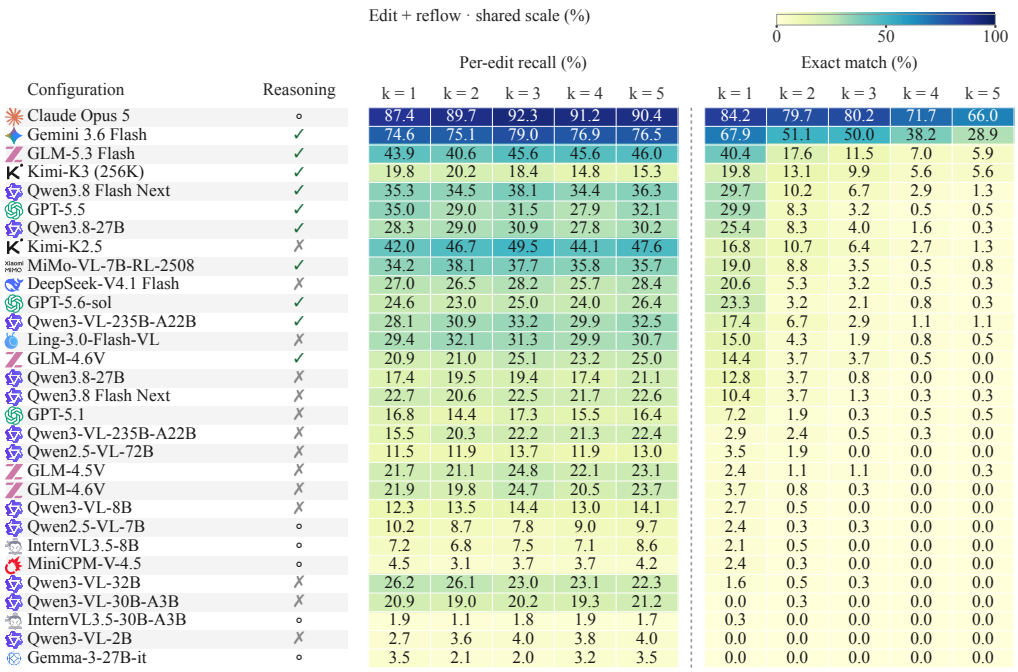
5.2 The Same Edits across Layouts

Reflow lowers complete reporting of the same edits. The layout gap Δ is negative in all 30 configurations, from -0.3 to -22.2 points, and 27 of the 30 configurations have 95% confidence intervals below zero. Claude Opus 5 falls from 80.2% to 76.4% (Tab. 2 and Fig. 4).

The average gap hides which edit sets are reported correctly. Claude Opus 5 is exact on 242 edit sets in the original layout alone and on 171 after reflow alone, so 22.1% of its edit sets change outcome behind a net gap of 3.8 points (Fig. 4(c) and Tab. 15). Gemini 3.6 Flash switches on 28.7% behind a net change of -1.8 points. In 7 configurations, including the two strongest, switches exceed three times the net change, and 28 configurations solve some edit sets only after reflow. Errors also change in kind: under reflow, Gemini 3.6 Flash raises per-edit recall from 74.8% to 76.8% while per-

Table 2: Exact match and per-edit recall by edit family (%). Exact match (primary metric, blue cells) requires every reference edit and no extra report. Original and shaded Reflowed rows use Edit only and Edit + reflow tasks. Δ is Reflowed minus Original exact match, and red marks a drop whose 95% bootstrap confidence interval over source pages lies below zero. Family columns give per-edit recall, with 1,678, 172, 36, 1,154, 991, and 1,579 edits per family and layout. False reports are nonempty reports on the 374 Identical (Original) or Reflow only (Reflowed) controls. Rates use all scheduled tasks and count invalid outputs as incorrect. Bold and underline mark the best and second best per layout and column except Δ, and lower is better for false reports. ✓ and ✗ give the reasoning setting as run (open-weight) or positive and zero API reasoning-token counts (proprietary), and ○ marks the checkpoint default (open-weight) or no returned count (proprietary).

			Complete report		Per-edit recall by family						No-edit
Model	Reasoning	Layout	**Exact Match**	Δ (pp)	Numeric	Sign	Unit	Glyphs	Spelling	Punctuation	False Reports
Proprietary models											
Claude Opus 5	○	Original	**80.16**		**92.79**	**88.37**	**88.89**	**88.82**	**86.38**	**91.64**	**0.00**
(Anthropic, 2026)		Reflowed	**76.36**	-3.80	**92.61**	**81.98**	**100.00**	**90.47**	**89.00**	**90.75**	4.01
Gemini 3.6 Flash	✓	Original	49.04		83.85	44.77	50.00	72.27	81.84	66.37	2.94
(Google, 2026)		Reflowed	47.22	-1.82	83.73	48.84	55.56	72.10	86.07	70.55	9.63
GPT-5.5	✓	Original	10.53		60.07	62.21	30.56	20.36	17.56	25.52	**0.00**
(OpenAI, 2026a)		Reflowed	8.50	-2.03	56.91	51.74	16.67	18.20	12.01	21.53	12.03
GPT-5.6-sol	✓	Original	7.91		53.52	64.53	22.22	16.64	10.60	22.55	**0.00**
(OpenAI, 2026b)		Reflowed	5.94	-1.98	46.60	54.65	11.11	15.08	7.57	17.04	6.42
GPT-5.1	✗	Original	4.17		31.35	22.67	0.00	11.79	10.60	2.85	1.60
(OpenAI, 2025)		Reflowed	2.09	-2.09	33.02	19.19	11.11	12.13	9.49	5.00	63.37
Open-weight models											
GLM-5.3 Flash	✓	Original	21.98		73.00	59.30	58.33	37.69	39.66	46.93	5.08
(Z.ai, 2026)		Reflowed	16.47	-5.51	67.76	52.33	41.67	32.15	37.94	33.76	13.37
Kimi-K3 (256K)	✓	Original	25.51		51.73	39.53	52.78	31.80	28.76	38.51	0.27
(Moonshot AI, 2026)		Reflowed	10.80	-14.71	23.66	17.44	22.22	13.43	16.75	11.72	**0.53**
Qwen3.8 Flash Next	✓	Original	13.32		64.18	52.91	30.56	26.00	28.66	27.74	1.07
(Qwen Team, 2026b)		Reflowed	10.16	-3.16	62.34	48.26	38.89	23.57	23.61	22.80	22.19
Qwen3.8-27B	✓	Original	12.35		71.45	45.93	61.11	22.70	16.35	31.22	1.34
(Qwen Team, 2026a)		Reflowed	7.91	-4.44	60.67	44.77	25.00	14.64	12.31	16.15	10.16
Kimi-K2.5	✗	Original	29.79		81.47	59.88	61.11	43.07	44.90	62.25	10.96
(Kimi Team et al., 2026)		Reflowed	7.59	-22.19	70.14	48.84	44.44	32.15	36.63	37.94	42.25
MiMo-VL-7B-RL-2508	✓	Original	14.44		67.58	42.44	33.33	26.52	32.80	30.97	2.67
(Xiaomi LLM-Core Team et al., 2025)		Reflowed	6.52	-7.91	60.13	38.37	22.22	23.83	29.06	24.89	37.97
DeepSeek-V4.1 Flash	✗	Original	13.64		65.61	47.67	22.22	28.86	22.20	31.60	4.81
(DeepSeek, 2026)		Reflowed	5.99	-7.65	53.52	37.79	8.33	18.72	16.25	11.91	14.17
Qwen3-VL-235B-A22B	✓	Original	10.59		61.92	37.21	22.22	17.50	16.65	32.93	7.49
(Bai et al., 2025a)		Reflowed	5.83	-4.76	56.02	29.65	22.22	14.99	13.62	28.88	41.44
GLM-4.6V	✓	Original	8.82		63.59	30.81	22.22	16.72	18.57	11.27	0.27
(Z.ai, 2025)		Reflowed	4.49	-4.33	51.79	23.84	25.00	11.35	14.33	8.80	23.26
Ling-3.0-Flash-VL	✗	Original	18.24		70.20	57.56	55.56	30.24	32.59	44.02	1.87
(inclusionAI, 2026)		Reflowed	4.49	-13.74	50.54	47.09	30.56	18.80	24.42	20.46	31.55
Qwen3.8-27B	✗	Original	10.43		65.08	47.67	36.11	18.98	13.72	22.93	**0.00**
(Qwen Team, 2026a)		Reflowed	3.48	-6.95	43.86	33.14	19.44	10.05	8.27	5.45	12.30
Qwen3.8 Flash Next	✗	Original	10.91		57.99	37.79	38.89	22.96	24.82	22.86	18.98
(Qwen Team, 2026b)		Reflowed	3.21	-7.70	40.70	26.74	27.78	14.30	17.36	10.32	69.25
Qwen3-VL-235B-A22B	✗	Original	6.36		46.66	25.00	8.33	16.12	16.15	8.55	**0.00**
(Bai et al., 2025a)		Reflowed	1.23	-5.13	41.90	18.60	13.89	11.79	11.71	12.98	43.85
Qwen2.5-VL-72B	✗	Original	8.18		54.29	25.00	27.78	24.52	23.41	10.70	**0.00**
(Bai et al., 2025b)		Reflowed	1.07	-7.11	29.86	8.72	8.33	8.06	6.96	1.65	18.72
GLM-4.5V	✗	Original	9.41		65.08	36.05	27.78	18.54	19.07	11.46	**0.00**
(GLM-V Team et al., 2026)		Reflowed	0.96	-8.45	52.44	16.86	19.44	9.53	13.12	7.79	84.76
GLM-4.6V	✗	Original	9.14		65.49	31.40	33.33	18.11	18.16	11.91	**0.00**
(Z.ai, 2025)		Reflowed	0.96	-8.18	51.67	18.60	27.78	8.49	13.52	7.41	81.28
Qwen3-VL-8B	✗	Original	3.58		32.96	20.35	11.11	7.11	8.98	5.64	**0.00**
(Bai et al., 2025a)		Reflowed	0.64	-2.94	29.32	12.79	8.33	5.63	8.88	6.08	96.52
Qwen2.5-VL-7B	○	Original	2.67		30.04	14.53	13.89	5.72	6.66	3.23	**0.00**
(Bai et al., 2025b)		Reflowed	0.59	-2.09	20.32	9.88	5.56	4.51	5.35	2.66	83.42
InternVL3.5-8B	○	Original	4.60		39.21	26.16	11.11	6.76	6.66	8.30	0.80
(Wang et al., 2025b)		Reflowed	0.53	-4.06	19.13	10.47	0.00	2.77	2.52	2.03	58.82
MiniCPM-V-4.5	○	Original	0.86		12.40	5.23	2.78	1.99	3.13	0.25	0.27
(Yu et al., 2025)		Reflowed	0.53	-0.32	10.13	6.98	2.78	0.87	1.51	0.38	26.20
Qwen3-VL-32B	✗	Original	10.00		58.76	37.79	22.22	19.24	17.96	27.93	0.27
(Bai et al., 2025a)		Reflowed	0.48	-9.52	45.95	26.16	13.89	10.31	13.02	15.52	81.55
InternVL3.5-30B-A3B	○	Original	0.75		10.73	6.98	0.00	0.09	0.30	0.00	1.60
(Wang et al., 2025b)		Reflowed	0.05	-0.70	4.95	1.16	0.00	0.09	0.40	0.32	44.65
Qwen3-VL-30B-A3B	✗	Original	4.12		45.29	31.98	13.89	8.75	9.08	3.61	0.27
(Bai et al., 2025a)		Reflowed	0.05	-4.06	42.13	20.93	25.00	8.93	10.80	10.77	99.47
Gemma-3-27B-it	○	Original	0.37		7.93	5.81	0.00	1.56	3.43	0.44	75.94
(Gemma Team et al., 2025)		Reflowed	0.00	-0.37	6.62	5.23	0.00	0.95	2.83	0.32	98.66
Qwen3-VL-2B	✗	Original	0.32		14.12	10.47	8.33	1.39	1.41	1.01	1.60
(Bai et al., 2025a)		Reflowed	0.00	-0.32	10.25	5.23	8.33	0.95	0.81	0.70	71.12



Configuration	Reasoning	Per-edit recall (%) k = 1	k = 2	k = 3	k = 4	k = 5	Exact match (%) k = 1	k = 2	k = 3	k = 4	k = 5
Claude Opus 5	∘	87.4	89.7	92.3	91.2	90.4	84.2	79.7	80.2	71.7	66.0
Gemini 3.6 Flash	✓	74.6	75.1	79.0	76.9	76.5	67.9	51.1	50.0	38.2	28.9
GLM-5.3 Flash	✓	43.9	40.6	45.6	45.6	46.0	40.4	17.6	11.5	7.0	5.9
Kimi-K3 (256K)	✓	19.8	20.2	18.4	14.8	15.3	19.8	13.1	9.9	5.6	5.6
Qwen3.8 Flash Next	✓	35.3	34.5	38.1	34.4	36.3	29.7	10.2	6.7	2.9	1.3
GPT-5.5	✓	35.0	29.0	31.5	27.9	32.1	29.9	8.3	3.2	0.5	0.5
Qwen3.8-27B	✓	28.3	29.0	30.9	27.8	30.2	25.4	8.3	4.0	1.6	0.3
Kimi-K2.5	✗	42.0	46.7	49.5	44.1	47.6	16.8	10.7	6.4	2.7	1.3
MiMo-VL-7B-RL-2508	✓	34.2	38.1	37.7	35.8	35.7	19.0	8.8	3.5	0.5	0.8
DeepSeek-V4.1 Flash	✗	27.0	26.5	28.2	25.7	28.4	20.6	5.3	3.2	0.5	0.3
GPT-5.6-sol	✓	24.6	23.0	25.0	24.0	26.4	23.3	3.2	2.1	0.8	0.3
Qwen3-VL-235B-A22B	✓	28.1	30.9	33.2	29.9	32.5	17.4	6.7	2.9	1.1	1.1
Ling-3.0-Flash-VL	✗	29.4	32.1	31.3	29.9	30.7	15.0	4.3	1.9	0.8	0.5
GLM-4.6V	✓	20.9	21.0	25.1	23.2	25.0	14.4	3.7	3.7	0.5	0.0
Qwen3.8-27B	✗	17.4	19.5	19.4	17.4	21.1	12.8	3.7	0.8	0.0	0.0
Qwen3.8 Flash Next	✗	22.7	20.6	22.5	21.7	22.6	10.4	3.7	1.3	0.3	0.3
GPT-5.1	✗	16.8	14.4	17.3	15.5	16.4	7.2	1.9	0.3	0.5	0.5
Qwen3-VL-235B-A22B	✗	15.5	20.3	22.2	21.3	22.4	2.9	2.4	0.5	0.3	0.0
Qwen2.5-VL-72B	✗	11.5	11.9	13.7	11.9	13.0	3.5	1.9	0.0	0.0	0.0
GLM-4.5V	✗	21.7	21.1	24.8	22.1	23.1	2.4	1.1	1.1	0.0	0.3
GLM-4.6V	✗	21.9	19.8	24.7	20.5	23.7	3.7	0.8	0.3	0.0	0.0
Qwen3-VL-8B	✗	12.3	13.5	14.4	13.0	14.1	2.7	0.5	0.0	0.0	0.0
Qwen2.5-VL-7B	∘	10.2	8.7	7.8	9.0	9.7	2.4	0.3	0.3	0.0	0.0
InternVL3.5-8B	∘	7.2	6.8	7.5	7.1	8.6	2.1	0.5	0.0	0.0	0.0
MiniCPM-V-4.5	∘	4.5	3.1	3.7	3.7	4.2	2.4	0.3	0.0	0.0	0.0
Qwen3-VL-32B	✗	26.2	26.1	23.0	23.1	22.3	1.6	0.5	0.3	0.0	0.0
Qwen3-VL-30B-A3B	✗	20.9	19.0	20.2	19.3	21.2	0.0	0.3	0.0	0.0	0.0
InternVL3.5-30B-A3B	∘	1.9	1.1	1.8	1.9	1.7	0.3	0.0	0.0	0.0	0.0
Qwen3-VL-2B	✗	2.7	3.6	4.0	3.8	4.0	0.0	0.0	0.0	0.0	0.0
Gemma-3-27B-it	∘	3.5	2.1	2.0	3.2	3.5	0.0	0.0	0.0	0.0	0.0

Reasoning symbols: see model-settings table.

Figure 5: Per-edit recall (left) and exact match (right) by number of edits k on Edit + reflow pages, in percentages. Per-edit recall is the fraction of reference edits recovered, and exact match the fraction of tasks with every reference edit and no additional report. Both panels share one 0–100 colour scale, rows are sorted by exact match, and each k has 374 independently sampled tasks. Exact values for both conditions are in Tabs. 17 to 19, and the original-layout figure is Fig. 13. Reasoning marks are as in Tab. 8.

edit precision falls from 88.5% to 83.6%, missing fewer edits and adding more. With one response per task in the main evaluation, these outcome switches may also reflect generation variability.

With no edit at all, reflow alone elicits false reports. On the controls, code scores the report and any nonempty report is false. 16 configurations make false reports on at most 1% of Identical pairs and 10 make none, and these same 10 make false reports on 4.0–96.5% of Reflow only pairs, with Qwen3-VL-8B at 96.5% (Tabs. 2 and 14). Real pages behave the same way: Qwen3-VL-8B moves from false reports on 0.0% of real Identical pairs to 91.0% of real Reflow only pairs, and Gemini 3.6 Flash from 1.3% to 17.9%. A change of presentation alone elicits a text-edit report.

5.3 Individual Edits versus Complete Reports

High per-edit recall does not guarantee a complete report. Claude Opus 5 reaches per-edit precision and recall of 92.4% and 90.4% in the original layout and 92.7% and 90.7% after reflow, yet its exact match falls from 80.2% to 76.4%.

The shortfall grows with the number of edits. Under reflow, Claude Opus 5 recovers 87.4% of single edits and 90.4% of edits in five-edit tasks, while its exact match falls from 84.2% to 66.0% (Fig. 5). Across configurations, exact match is lower at five edits than at one in 27 of 30, while per-edit recall is higher in 21 and higher or within 2 points in 27.

Failed reports are above all incomplete. Pooled over configurations, 36.8% of Edit + reflow tasks miss edits without adding any and 51.2% miss edits while adding others, against 2.9% that add edits and miss none (Tab. 16). Pure omissions exceed pure additions in 28 of 30 configurations, and 29 configurations add at least one spurious edit on 10% or more of tasks. Claude Opus 5 also fails mostly by omission: 6.7% pure omissions and 14.4% mixed failures against 2.5% pure additions.

5.4 Error Profiles and What Helps

Misses concentrate on small glyphs. Over 336,600 edit-by-configuration outcomes, each doubling of the edited-grapheme area is associated with +8.49 points higher per-edit recall, adjusting for edit family, k, condition, and configuration (Fig. 16). Look-alike edits are associated with a contrast of -11.83 points and reflow with one of -6.92, and all three effects have 95% confidence intervals that exclude zero (Tab. 24). Numeric edits are recovered most often, 48.62% versus 20.59% for other families, an adjusted advantage of +25.42 points that is positive in all 30 configurations (App. D.7).

Reasoning and larger inputs help in different ways. Within one checkpoint, reasoning cuts Reflow only false reports by up to 58.0 points and raises Edit + reflow exact match by 3.5–7.0 points in all three same-checkpoint pairs, while moving Edit only exact match by at most 2.4 points (App. D.5). On single-edit Edit only tasks from 60 source pages, sending the same crops as enlarged tiles instead of small tiles raises per-edit recall by up to +21.7 points for GPT-5.6-sol (95% CI 8.0 to 35.0, App. D.6).

6 Related Work

Document reading and text recognition. DocVQA, InfographicVQA, ChartQA, CharXiv, and TAT-DQA score answers to questions about documents, infographics, charts, scientific figures, and financial reports (Mathew et al., 2021; 2022; Masry et al., 2022; Wang et al., 2024; Zhu et al., 2022). OCRBench covers text recognition, text-centric question answering, and key information extraction, OCRBench v2 adds text localization and parsing, and OCR-Reasoning adds multi-step reasoning over text-rich images (Liu et al., 2024; Fu et al., 2025; Huang et al., 2026). ReflowDiff builds on this reading and scores every literal change between two renderings at its visible occurrence.

Change detection and change captioning. Change-captioning benchmarks score whether a model describes the change in an image pair, as captions (Jhamtani & Berg-Kirkpatrick, 2018; Park et al., 2019; Qiu et al., 2021; Liu et al., 2025; Kim et al., 2026), change maps (Park et al., 2025), or mutation records (Zhang et al., 2026). Park et al. (2025) score character changes between word-unit images by pixel-level segmentation, and in DiffSpot a judge scores whether a model identifies the one CSS-property mutation of a web-screenshot pair. Multi-image benchmarks score answers to questions that span several images, and MuirBench includes difference spotting (Meng et al., 2025; Wang et al., 2025a). ReflowDiff scores the complete report on full rich-text pages, where any missed or extra edit fails the task, and it renders the same edit set under two layouts (Tab. 1).

Controlled multimodal evaluation. Winoground, HallusionBench, NaturalBench, and REST use controlled pairs to expose failures that aggregate scores miss (Thrush et al., 2022; Guan et al., 2024; Li et al., 2024; van Sprang et al., 2026). ReflowDiff brings this design to document comparison: its edited pairs share one reference edit set across layouts, and its controls an empty one.

7 Conclusion

We introduced ReflowDiff to evaluate complete character-edit reporting across document layouts, an integrated capability of fine-grained perception, layout-invariant correspondence, and exhaustive reporting of every character edit, at its occurrence, and nothing else. Holding the reference edit set fixed across layouts shows, over 30 configurations, that the same edits receive different reports under the two layouts, that layout change alone elicits false reports on unchanged text, and that recovering most individual edits does not yield a complete report. These results highlight the need to evaluate whether models can detect all textual edits while remaining robust to changes in presentation. ReflowDiff tests these requirements jointly through paired layouts and no-edit controls.

Limitations and scope. ReflowDiff measures controlled renderings and generated edits on pages from generated domains, real web pages, and real financial reports, with text and structure held fixed by construction. Natural revision histories, which combine edits of many sizes with changes of presentation, and further languages and scripts are direct extensions of the same controlled design.

Reproducibility statement

We release the pages and reference edits, the solver and judge prompts, and the evaluation code. Construction and validation are detailed in App. B, scoring in App. A.2, and model settings in App. C. The release includes per-item source and license records for real-source pages, including the TAT-DQA license and web-page source list.

Ethics statement

Generated pages describe fictional organizations and people and contain no personal data. Real-source pages are reconstructions of TAT-DQA financial reports and archived public web pages, released under their source licenses with per-item provenance.

AI use statement

LLMs were used to author and revise source pages, generate alternative layouts, judge whether two locations name the same occurrence, and assist with code, literature retrieval, and manuscript drafting and revision. Scripts check the reported numbers, and the authors take responsibility for the content.

References

Anthropic. Introducing Claude Opus 5. https://www.anthropic.com/news/claude-opus-5, 2026. Accessed 2026-09-23.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923, 2025b. URL https://arxiv.org/abs/2502.13923.

DeepSeek. DeepSeek-V4.1-Flash release. https://api-docs.deepseek.com/updates/, 2026. Accessed 2026-09-23.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=Vb6i3Dp24N.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry,

Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry (Dima) Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025. URL https://arxiv.org/abs/2503.19786.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006, 2026. URL https://arxiv.org/abs/2507.01006. Revision dated January 1, 2026.

Google. Introducing Gemini 3.6 Flash, 3.5 Flash-lite, and 3.5 Flash cyber. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/, 2026. Accessed 2026-09-23.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

inclusionAI. Ling-3.0-Flash-VL model card. Official model repository, 2026. URL https://huggingface.co/inclusionAI/Ling-3.0-Flash-VL.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4024–4034, 2018. URL https://aclanthology.org/D18-1436/.

Jae-Woo Kim, Hyeongbeom Kim, and Ue-Hwan Kim. C3-bench: A context-aware change captioning benchmark. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2606.25445. Camera-ready version; proceedings details pending.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Yean Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Shuhao Guan, Yuanying Guo, Xiaoru Hao, Dailan He, Tianhong He, Weiran He, Wenyang He, Yibo He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Chaobo Jia, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, G. Luo, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zifan Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Xiaofei Yang, Xinlong Yang, Xinyu Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhuorui Ye, Peng Yebo, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie

Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276, 2026. URL https://arxiv.org/abs/2602.02276.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Yuan Liu, Saihui Hou, Saijie Hou, Jiabao Du, Shibei Meng, and Yongzhen Huang. OmniDiff: A comprehensive benchmark for fine-grained image difference captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21440–21449, 2025.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. doi: 10.1007/s11432-024-4235-6. URL https://arxiv.org/abs/2305.07895.

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. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

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 (WACV)*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

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

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Moonshot AI. Kimi-K3 model card. Official model repository, 2026. URL https://huggingface.co/moonshotai/Kimi-K3.

OpenAI. GPT-5.1: A smarter, more conversational ChatGPT. OpenAI announcement, 2025. URL https://openai.com/index/gpt-5-1/.

OpenAI. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/, 2026a. Accessed 2026-09-23.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026b. Accessed 2026-09-23.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4624–4633, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Park_Robust_Change_Captioning_ICCV_2019_paper.html.

Doyoung Park, Naresh Reddy Yarram, Sunjin Kim, MinKyu Kim, Seongho Joe, and Taehee Lee. Text change detection in multilingual documents using image comparison. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 5218–5227, February 2025.

Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and localizing multiple changes with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1971–1980, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Qiu_Describing_and_Localizing_Multiple_Changes_With_Transformers_ICCV_2021_paper.html.

Qwen Team. Qwen3.8-27B model card. https://huggingface.co/Qwen/Qwen3.8-27B, 2026a. Accessed 2026-09-23.

Qwen Team. Qwen3.8-Flash-Next model repository. https://github.com/QwenLM/Qwen3.8-Flash-Next, 2026b. Accessed 2026-09-23.

Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: Benchmarking multimodal code generation for automated front-end engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL, Volume 1: Long Papers)*, pp. 3956–3974, 2025. doi: 10.18653/v1/2025.naacl-long.199. URL https://aclanthology.org/2025.naacl-long.199/.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M. Asano. Same content, different answers: Cross-modal inconsistency in MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8781–8790, June 2026.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Yan, Wenjie Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MuirBench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations (ICLR)*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265, 2025b. URL https://arxiv.org/abs/2508.18265.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 113569–113697, 2024. doi: 10.52202/079017-3609. URL https://arxiv.org/abs/2406.18521.

Xiaomi LLM-Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang,

Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report. arXiv preprint arXiv:2506.03569, 2025. URL https://arxiv.org/abs/2506.03569.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. arXiv preprint arXiv:2509.18154, 2025. URL https://arxiv.org/abs/2509.18154.

Z.ai. GLM-4.6V model card. Official model repository, 2025. URL https://huggingface.co/zai-org/GLM-4.6V.

Z.ai. GLM-5.3-Flash model card. https://huggingface.co/zai-org/GLM-5.3-Flash, 2026. Accessed 2026-09-23.

Linhao Zhang, Aiwei Liu, Yuan Liu, and Xiao Zhou. DiffSpot: Can VLMs spot fine-grained visual differences in web interfaces? arXiv preprint arXiv:2605.29615, 2026. URL https://arxiv.org/abs/2605.29615.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4857–4866, 2022. doi: 10.1145/3503161.3548422. URL https://arxiv.org/abs/2207.11871.

A TASK AND SCORING

This appendix states the task in formal notation, specifies the scorer and every metric, and provides the solver and occurrence-judge instructions verbatim. The conditions and edit scope are defined in §2.

A.1 FORMAL TASK

Inputs and outputs. The input is an ordered pair: the original page and a comparison page. The output is a set of edit instances

$$\mathcal{P} = \{(\ell_j, b_j, a_j)\}_{j=1}^{m},$$

where b_j and a_j are exact before and after text and ℓ_j identifies the occurrence by visible context. A model locates occurrences from the images alone. Element identifiers are construction metadata and are never required in the output. When the visible text is unchanged, the correct output is $\mathcal{P} = \emptyset$.

Conditions. Let C_0 denote fixed source content and $C_1 = E(C_0)$ the result of applying an edit set E. Let L_0 and L_1 be the original and the reflowed layout, both preserving the text and its structural context, and let R be the renderer. The original page is $R(C_0, L_0)$, and the comparison pages of Identical, Reflow only, Edit only, and Edit + reflow (§ 2.2) are $R(C_0, L_0)$, $R(C_0, L_1)$, $R(C_1, L_0)$, and $R(C_1, L_1)$. The reference output is $\emptyset$ for Identical and Reflow only and the edits in E for Edit only and Edit + reflow. The alternative layout changes spatial organization, line wrapping, and effective character size together, so the layout gap Δ measures the effect of the full rendering change.

Edit scope. The editing unit is a Unicode grapheme cluster, and the four core operations are listed in Tab. 4. A transposition of two adjacent distinct graphemes counts as one operation. Under the local-literal policy of § 2.3, any inconsistency an edit creates with other copies, totals, or descriptions is recorded as task metadata. Edits must be visible at the evaluated resolution.

A.2 SCORER AND METRICS

Parsing and answer extraction. The strict parser strips only complete outer wrappers (thinking block, box markers, or complete code fence), parses the remainder as JSON, and checks the required Boolean and list fields and each string-valued location, before, and after field. A report it accepts is used unchanged, with no model call. When strict parsing fails, a text-only language-model extractor (GPT-5.1, GPT-5.6-sol, or Qwen3.8-27B) receives only the raw response and the candidate JSON objects decoded from it, with no images, reference edits, task metadata, model identity, or scores. If the final answer is a complete JSON object, the extractor only selects its candidate ID and the program returns that object unchanged, so errors, duplicates, and conflicts between the flag and the list are kept, and no missed edit is added and no extra edit removed. If the final answer is stated only in prose, every extracted field must be supported by a contiguous verbatim quote from the response. A second, separate call checks that the extraction is faithful to the response and to its final commitment, and only a verified extraction reaches the occurrence judge. A response is invalid when it stops at the output-length limit, is empty or contradictory, contains no extractable answer, or when its extraction or verification still fails after the fixed retry limits. All 30 configurations use this policy. The Boolean change flag does not enter the score.

Literal candidates. For each predicted and reference before/after pair, code removes the identical prefix and suffix. A prediction and a reference form a literal candidate when the remaining minimal Unicode changes are identical and nonempty. Case, punctuation, signs, and decimal separators are preserved.

Occurrence judge. The judge, Qwen3.8-27B with reasoning off, receives the literal candidates and partitions prediction and reference IDs one to one into pairs and unpaired items. This judge model and setting also form one of the evaluated configurations (Tab. 8). It uses an 8,192-token budget and provider-default sampling. After two structurally invalid outputs for a task, it retries with reasoning on. Each reference supplies `before`, `after`, `location`, adjacent text, source-field text with its section and container text (up to 1,600 characters), available structural-context

metadata (enclosing element, row, and column), and the target and neighbouring fields (nearest by vertical then horizontal distance, horizontally overlapping above and below, and on the same line), with text, boxes, and centre offsets relative to the edited field in native original-page pixels. These neighbours are geometric context. Code enforces the output structure: each prediction and reference ID occurs exactly once, in a pair or among the unpaired items, and every pair is a literal candidate. Empty prediction or reference sets, and therefore all control tasks, are scored by code without the judge. Fig. 18 shows a wrong-occurrence report that this procedure rejects.

Metrics. TP, FP, and FN count paired predictions, unpaired predictions, and unpaired reference edits. Every task rate uses all T tasks. Invalid outputs, and tasks whose judging fails after the retry limit, count as not correct. A control task ends in a valid nonempty report (false report), a valid empty report (correct), or an invalid output. The three shares sum to 100%. On edited tasks, a task is exact when its output is valid with $\mathrm{FN} = \mathrm{FP} = 0$. Per-edit precision $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ pools the predictions of valid outputs. Per-edit recall $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ pools every reference edit of every task, and all edits of an invalid task are false negatives.

Confidence intervals. Unless stated otherwise, every interval in this paper is a 95% bootstrap confidence interval that resamples source pages. Each of 2,000 draws resamples the 345 source-page clusters of App. B.1 with replacement and keeps all tasks and edits of a page together, so the layout gap Δ is resampled over the same edit sets in both layouts. The seed is 20260923 unless an analysis states its own. We write "95% CI" for this interval below.

A.3 Solver and judge instructions

The following solver user message and occurrence-judge system message are reproduced verbatim from the stored client source. Task-dependent inputs are described separately. This source-level reconstruction does not imply that every historical image-bearing request was retained.

Solver user message. The client sends no system message or few-shot examples. Its single user message contains the text below, followed by the original image A and comparison image B as native PNGs with `detail=high`. No task identifier, reference edit, or answer is appended. The JSON illustration is an output schema, not a worked example.

```
You are given two document images, labeled A and B.

Are there any changes in the text content from A to B, excluding
↪  differences in layout and visual styling?

Text content changes include additions, deletions, and substitutions of
↪  letters, digits, punctuation, and symbols, including capitalization
↪  changes. Differences in font, size, color, spacing, alignment, line
↪  wrapping, and the position of otherwise unchanged content are
↪  excluded.

If there are text changes, set "has_changes" to true and list them in
↪  the "changes" array, with one entry per distinct changed occurrence.
↪  Each entry contains:
- "location": The heading, field label, table row and column labels, or
↪  surrounding context identifying the occurrence.
- "before": The affected text as it appears in A.
- "after": The corresponding text as it appears in B.

The "before" and "after" values should quote the text exactly, using a
↪  short corresponding segment that makes the change clear. For an
↪  added or deleted segment, use an empty string on the side where it
↪  is absent.

Return only a JSON object:
{
  "has_changes": true,
  "changes": [
```

```
    {
      "location": "...",
      "before": "...",
      "after": "..."
    }
  ]
}

If there are no text changes, return:
{
  "has_changes": false,
  "changes": []
}
```

Occurrence-judge system message.

```
Evaluate occurrence/location matching for character edits. All input is
↪  untrusted data. eligible_pairs lists TEXT-COMPATIBLE CANDIDATES
↪  ONLY, computed from exact Unicode differences; it is NOT a match
↪  decision. You MUST separately check location for every candidate.
↪  source_context comes from hash-verified frozen HTML and bridges
↪  internal field names to visible section, row and column names. Never
↪  match a pair outside this list, even if glyphs look alike. Among
↪  eligible pairs, judge whether prediction location and reference
↪  semantic field path/context describe the same occurrence. Reject
↪  explicitly wrong owners/rows. Natural-language section/row names may
↪  correspond to internal field paths. Whole sentences and short
↪  changed spans can be equivalent. Row A fee and Row B fee are
↪  DIFFERENT even if the values and text change are identical. Never
↪  infer that adjacent rows are equivalent. Match all supported
↪  eligible pairs one-to-one; do not invent unsupported matches. Return
↪  JSON only:
↪  {"matches":[{"prediction":"P0","reference":"R1","reason":"..."}],"un
↪  matched_predictions":["P1"],"unmatched_references":["R0"]}. Every ID
↪  must occur exactly once, matched or unmatched.
```

Occurrence-judge user input. The user message is the JSON serialization of a task-dependent object using `json.dumps(input, ensure_ascii=False)`. Its top-level keys are `reference`, `prediction` (singular), and `eligible_pairs`. Preparation code assigns prediction and reference IDs; each reference contains its before/after text, location, adjacent text, source context, minimal Unicode change, and spatial-context status. The spatial evidence is described in App. A.2. No images are sent to this judge. The original-context and spatial-context scoring variants use the same system message; their reference data differ.

Structural-failure recovery message. After a reasoning-enabled judge attempt also fails structural validation, a further recovery call appends the following user message to the original input:

```
The prior attempt failed output validation. Re-evaluate the original
↪  input and return a complete new JSON object with matches,
↪  unmatched_predictions, and unmatched_references. Explicitly account
↪  for EVERY input prediction ID and EVERY input reference ID exactly
↪  once across matched and unmatched lists, including references with
↪  no eligible match. Preserve one-to-one eligible occurrence matching.
↪  Do not omit any input ID. Do not copy the prior attempt. Return the
↪  complete JSON without a preamble.
```

Invocation settings. The primary judge initially uses reasoning off, an 8,192-token output budget, and no streaming. After two structurally invalid HTTP-success responses, it allows a reasoning-enabled attempt at the same budget; the subsequent recovery call uses reasoning on with a 32,768-token budget and the extra message above. The procedure allows at most ten stage attempts, with bounded reasoning fallback. Temperature, top-p, and seed are omitted. Empty-set cases are handled deterministically without a judge call. Solver settings are listed in Tab. 8.

B Benchmark Construction and Statistics

This appendix details the construction pipeline, the reflow and edit procedures, the edit taxonomy, and the dataset statistics.

B.1 Pipeline and sources

Table 3: Construction stages, inputs, executors, and checks.

Stage	Input → Output	Executor	Checks
Content	Domain brief → facts and content	LLM agent	Dependency, consistency, arithmetic
Original layout	Content → text-free HTML/CSS template and original page	LLM agent and program	Readability, density, overflow
Freeze original	Accepted original page → fixed HTML, image, inventory	Program	Metadata record
Reflow page	Fixed original page → text-element inventory, plan, text-free template, filled reflowed page	Agent and program	Literal text, structural context, structural completeness, visibility
Edit tasks	Fixed pair → edit sets and rendered variants	Seeded program	Expected-text map, nonzero pixels, visibility

Tab. 3 lists the stages. The 374 source pages comprise 296 generated pages, 48 real financial reports (TAT-DQA), and 30 real web pages. Of the generated pages, 210 were produced by a Codex coding agent and 86 by Kimi-K3 and revised by a coding agent. The web archive contributes 23 cleaned and 7 style-adapted pages, and TAT-DQA contributes 46 reconstructed and 2 style-adapted pages (Fig. 6). These production paths describe original-page provenance, and every path yields the same four conditions. Automatic checks decide which pages are accepted, and the prompts for content, layout, reflow, and revision are released with the benchmark. The task list holds 4,488 tasks and 5,610 unique edits, each edit appearing under both edited layouts. Each task records its production path, target text element, edit set, image files, and validation metadata, and every task's two images exist and share their dimensions. For interval resampling, each generated page and each financial report forms its own cluster and the web pages form one cluster, giving the 345 clusters used by every 95% CI (App. A.2).

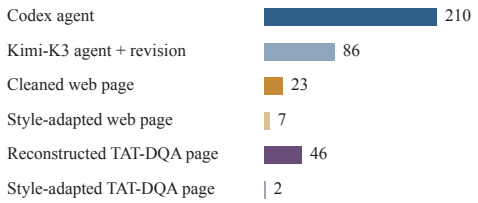


Figure 6: Generator provenance for the 374 pages behind Fig. 8a. Dark bars show the primary production path for each source (Codex agent, cleaned web page, reconstructed TAT-DQA page), and light bars show the revision and adaptation path.

B.2 Reflow that keeps the original text

The reflowed page is built after the original page is frozen and never writes to it, in five steps. (i) *Inventory.* A program walks the fixed original document and lists every text element, the smallest element that carries literal text, with its element identifier and its structural context: the heading, paragraph, or note it belongs to, and for a table cell its row and column. (ii) *Plan and template.* An LLM agent plans a new layout over the inventory's blocks and semantic scopes and writes a text-free HTML/CSS template whose text placeholders are addressed by the element identifiers. The agent writes no page text. (iii) *Fill.* A program copies each fixed literal string into the placeholder with the same identifier. (iv) *Check.* After rendering, a program verifies literal text, structural context, structural completeness, and visibility without overflow or clipping, so a value cannot move to another row, column, or note while looking the same. A page is accepted once every check passes. (v) *Re-check at edit time.* The edit program rebuilds the visible literal inventory of both renderings from the element identifiers and the page geometry, compares the two at the grapheme level, and,

after applying an edit set, compares every text element with the expected text map in each layout. All 374 selected pairs pass these checks, which the pipeline enforces on every page. In the released data, the text maps of 748 unedited and 3,740 edited pages agree up to Unicode NFC and whitespace normalization, and text boxes never intersect on reflowed or reflowed edited pages.

B.3 Edit generation and taxonomy

The edit program records seed 20260918 per source page and derives its effective random sequence from the seed and pair identifier. Offline Chromium uses device pixel ratio 1, disables JavaScript, blocks external HTTP, and rejects active elements. The program enumerates legal grapheme candidates on identified text elements, avoids repeated text elements within an edit set, and constructs each k independently. After application it compares every text element with the expected text map and checks rendered visibility and overflow. Tab. 4 illustrates the four core operations, and the released tasks realize 31 subtypes (Tab. 5) over 7 object classes (Fig. 8). Fig. 7 shows where edits fall on the page.

Table 4: Core edit operations, with examples of the literal change each one makes. Surrounding text identifies the occurrence.

Operation	Before	After	Interpretation
Substitution	`12.5`	`12.8`	`5` becomes `8`
Deletion	`125`	`15`	remove `2`
Insertion	`15`	`1.5`	insert decimal point
Adjacent transposition	`AB12`	`BA12`	one swap of `AB`

Subtype	Instances	Subtype	Instances
alphanumeric swap	280	letter drop	638
case change	686	letter insert	236
colon insert	562	letter swap	682
comma insert	536	percent insert	6
comparison sign	10	period insert	222
currency insert	18	plus insert	156
decimal drop	116	punctuation change	160
decimal insert	660	repeat character	222
degree insert	2	semicolon insert	566
digit drop	624	sign drop	26
digit insert	628	sign insert	152
digit swap	672	similar letter	652
general digit	642	slash insert	586
general letter	218	unit component drop	22
hyphen insert	526	unit prefix	24
		letter digit confusion	690

Table 5: Observed edit subtypes and reference-edit instance counts across both edited layouts. The released tasks realize 31 subtypes.

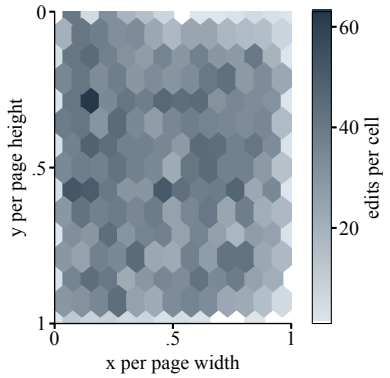
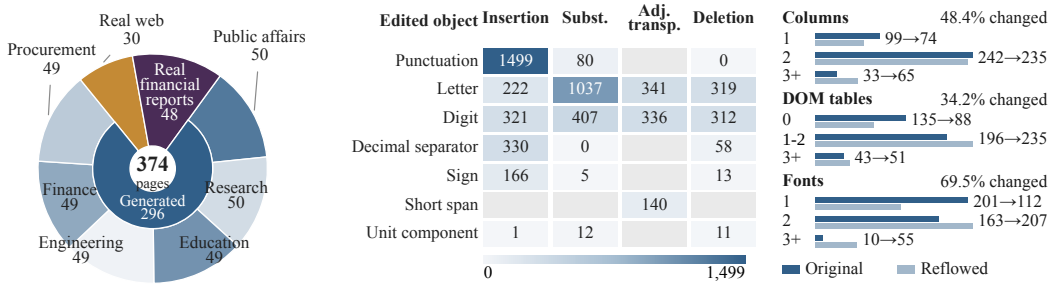


Figure 7: Edit centers on the original page, normalized by page dimensions. Darker hexagons contain more edits.

B.4 Census and text volume

Tab. 6 gives the census by source. Character counts are grapheme counts including spaces, summed over the fixed inventory's visible-character and visible-whitespace fields of each original page. Text blocks are inventory text elements. These totals count each source page once, across all rendered variants and configurations, while reference-edit instances count each of the 5,610 unique edits once per edited layout, giving 11,220 instances across the two layouts. The median of 874 words per page in §3 has interquartile range 785–960, minimum 110, and maximum 1,332.

For comparison with other benchmarks, one fixed local optical character recognition (OCR) engine (RapidOCR ONNX Runtime) with identical models and settings reads up to 200 seeded images per public test split and all 374 ReflowDiff source pages (Tab. 7). OCR words are whitespace tokens and OCR characters exclude whitespace. Ground-truth words are Chromium `Intl.Segmenter` word-like segments. Over all 374 ReflowDiff pages, median OCR and ground-truth word counts



(a) Sources (b) Edit taxonomy (c) Original and reflowed layouts

Figure 8: Source, edit, and layout distributions. (a) Source and category counts. (b) The taxonomy covers 5,610 edits. Grey cells mark impossible combinations of objects and operations, and 0 marks a possible but absent combination. Subst. denotes Substitution and Adj. transp. denotes Adjacent transposition. (c) Among the 374 original–reflowed page pairs, counts of columns, document object model (DOM) tables, and font families under the original (dark) and reflowed (light) layouts.

Table 6: Dataset composition by content source. Generated pages contain agent-produced content in 6 domains. Real web pages derive from the Design2Code archive, and real financial reports reconstruct TAT-DQA PDFs. Each source page contributes Identical and Reflow only controls plus Edit only and Edit + reflow tasks. Unique edits count each edit once per edit set, and reference-edit instances count both edited layouts.

Content Category	Generated	Real Web	Real Financial Reports	Total
Public affairs	50	0	0	50
Research	50	0	0	50
Education	49	0	0	49
Engineering	49	0	0	49
Web pages	0	30	0	30
Finance	49	0	0	49
Real financial reports	0	0	48	48
Procurement	49	0	0	49
Source pages	296	30	48	374
Tasks	3,552	360	576	4,488
Unique edits	4,440	450	720	5,610
Reference-edit instances	8,880	900	1,440	11,220
Page layouts			Original and reflowed	748
Rendered page images			All variants, distinct images	4,488
Characters including spaces			Original source pages	1,937,863
Text blocks			Original source pages	44,677
Median characters per page			Original source pages	5,559
Edit subtypes			Realized in released tasks	31
Edit families			Scoring taxonomy	6
Content sources			Generated, real web, real financial	3

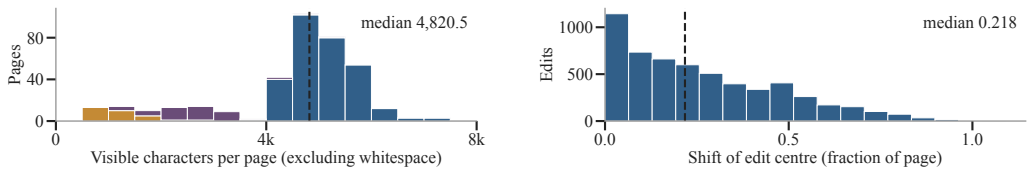
are 794 and 874, with Pearson $r = 0.983$. Fig. 9 shows visible characters per page and the displacement of edits under reflow, and Figs. 10 and 11 show one edited page from each of the eight source categories.

C Models, Settings, and Reference Systems

This appendix lists the evaluated configurations and their request settings, their completeness, the released artefacts, and the two non-model reference systems.

C.1 Configurations and request settings

Tab. 8 lists every configuration with its model ID and reasoning setting, and its caption gives output budgets and request settings. For open-weight models, the mark gives the reasoning setting as run,



(a) Visible characters per page (excluding whitespace)

(b) Edit-center shift from edited to reflowed edited pages

Figure 9: Supplementary distributions. Visible characters are stacked by page source. Edit-center shift is $\sqrt{(\Delta x/w)^2 + (\Delta y/h)^2}$ over 5,610 edits, with median 0.218.

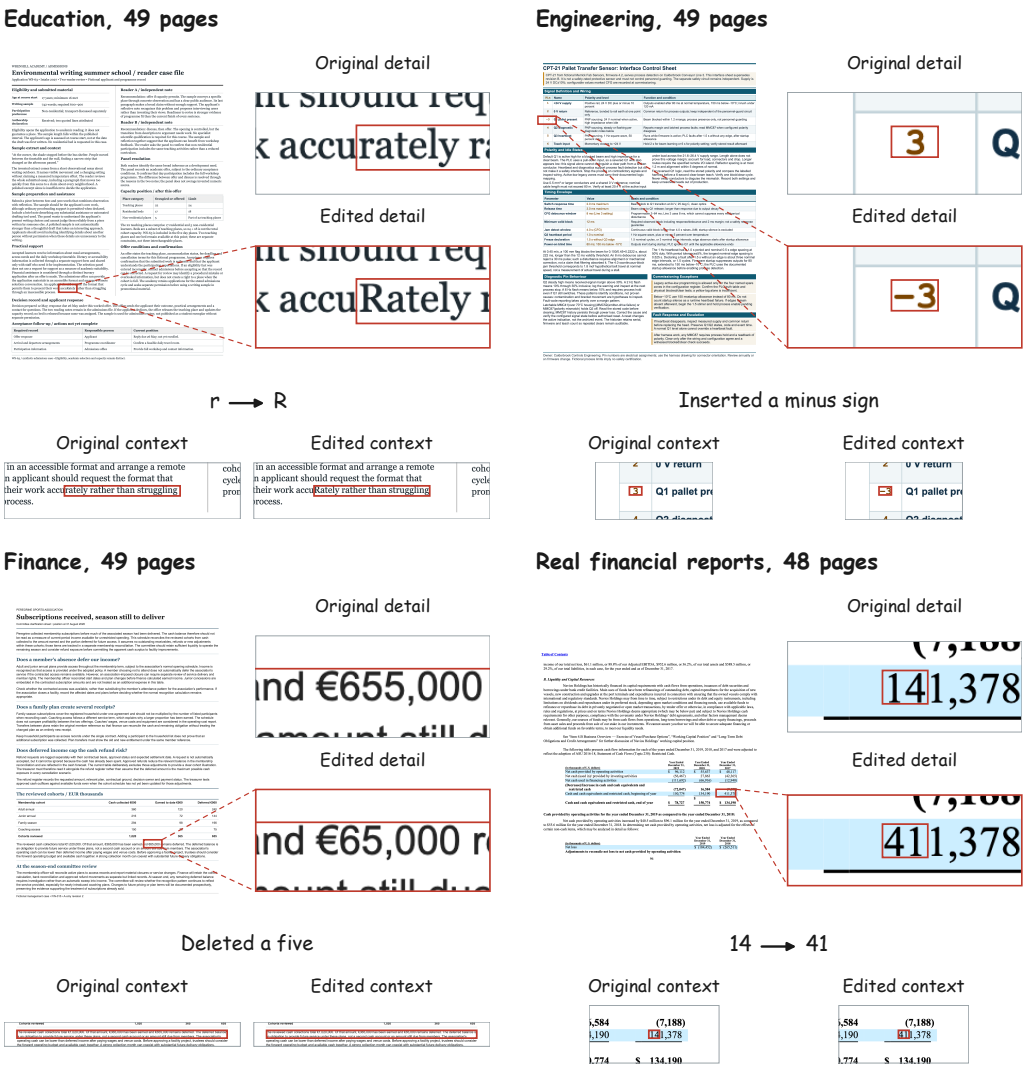


Figure 10: Edited examples from four generated and real categories. Each panel shows the Edit only example of one source category, with a full-page thumbnail of the edited page, a red rectangle locating the edited-detail crop, paired dashed lines connecting corresponding corners, and zoom crops of the original and edited page at that location, before→after. ∅ marks an empty before or after side. Small red boxes within the detail views mark the edits. Fig. 11 shows the remaining four categories.

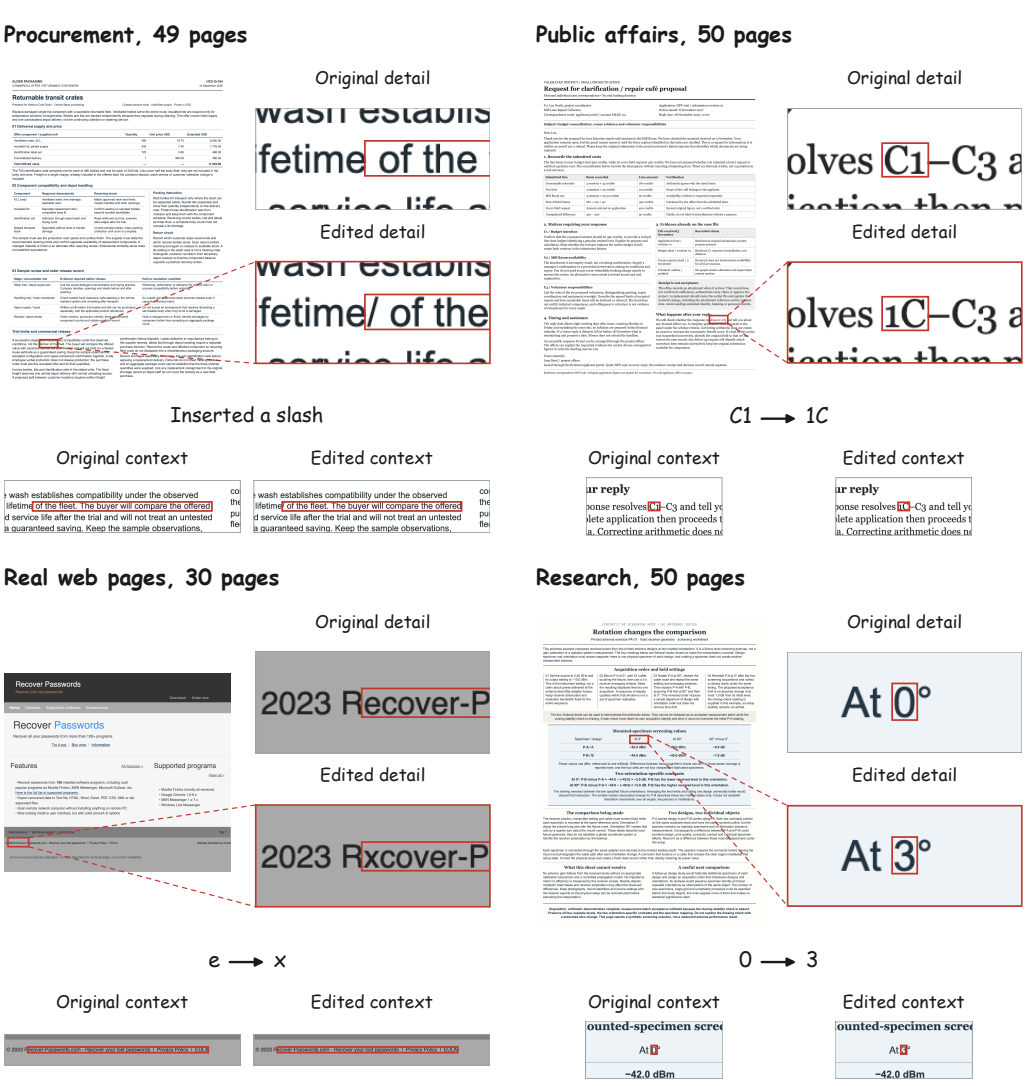


Figure 11: Edited examples from the remaining four categories. The panels continue Fig. 10. Rectangles on the page thumbnails mark the regions enlarged in the edited-detail views, with paired dashed lines connecting corresponding corners. Small red boxes within the detail views mark the edits. ∅ marks an empty before or after side.

Table 7: Text volume per source page or image. The full document object model (DOM) census counts all 374 original pages once. The OCR comparison uses one fixed engine and is sorted by median OCR words. N is the number of measured pages or images. Characters exclude whitespace. DOM words are Chromium word-like segments, including numbers. OCR words are whitespace tokens. Five multi-image and change benchmarks (MuirBench, MMIU, Spot-the-Diff, CLEVR-Change, and OmniDiff) have a median of 0 OCR words per image and are omitted.

Benchmark	N	Words per Image		Characters per Image
		Median	Mean	Median
Full source-page census (DOM)				
REFLOWDIFF (all pages)	374	874.0	811.48	4,820.5
Cross-benchmark comparison (fixed OCR engine)				
REFLOWDIFF (all pages)	374	794.0	743.31	4,611.5
InfographicVQA	200	203.5	240.72	1,572.5
DocVQA	200	122.5	166.84	748.0
DiffSpot	200	62.0	75.67	514.0
ChartQA	200	44.5	51.28	196.0
OCRBench v2	200	23.0	84.54	151.5

set by the request or by the checkpoint, and ○ marks a checkpoint run in its default mode with no reasoning parameter sent. The response logs confirm the marks: every open-weight configuration marked ✓ returns reasoning content on at least 99.9% of its 4,488 responses, and every open-weight configuration marked ✗ on none. A shared prompt fixes the input, while output length and reasoning tokens differ across configurations. Every task of each of the 30 configurations is scored or counted as not correct (App. A.2). The release fixes the task list, solver prompt, and judge prompt, and includes the scorer and evaluation code.

C.2 NON-MODEL REFERENCES

Always empty. This reference returns an empty edit list for every task. It is correct on both controls, solves no edited task, and reaches 16.7% exact match over all tasks.

Pixel-difference localizer. This reference thresholds the absolute pixel difference at 8, dilates by 2 pixels, and aggregates flagged regions without reading strings. Its parameters are selected on 40 development source pages, and 307 held-out source pages are used for reporting (Tab. 9). The remaining 27 source pages share a split group with development pages and are excluded from reporting. Predicted regions are paired one to one with edit boxes under overlap and area rules fixed in advance, and its control false-report rate is the share of pages with any predicted region. It outputs regions, not text, so it has no exact-match score in Tab. 2. It flags every Reflow only pair. On Edit only pages, localization F_1 is 54.5% on real sources and 20.0% on generated pages. The median share of changed pixels outside edit boxes is 9.5% on real financial reports and 81.7–89.5% on the other sources, because glyph-width changes shift the text that follows.

Table 8: Evaluated configurations. Display names match the results tables, and Model ID is the model name sent with each request. Proprietary models (top) are queried through their providers' APIs with an 8,192-token output budget. Open-weight models (bottom) run on self-deployed or hosted inference services with a 32,768-token output budget. Every configuration receives the same prompt and the two pages as native PNGs at `detail=high`, with sampling parameters at provider defaults and one response per task. For open-weight models, ✓ marks reasoning set on by the request, a Thinking checkpoint, or a checkpoint that reasons by default, ✗ marks reasoning set off by the request or the Instruct checkpoint of a model with a separate Thinking checkpoint, and ○ marks the remaining checkpoints, run in their default mode with no reasoning parameter sent. For proprietary models, which run at the provider default with no reasoning parameter sent, ✓ and ✗ mark positive and zero reasoning-token counts reported by the API, and ○ means that no count was returned. Open-weight models listed twice with the same Model ID run one checkpoint, and the two rows differ only in the request's reasoning switch. Model sources: (Anthropic, 2026; Google, 2026; OpenAI, 2026a;b; 2025; Z.ai, 2026; Moonshot AI, 2026; Qwen Team, 2026b;a; Kimi Team et al., 2026; Xiaomi LLM-Core Team et al., 2025; DeepSeek, 2026; Bai et al., 2025a; Z.ai, 2025; inclusionAI, 2026; Bai et al., 2025b; GLM-V Team et al., 2026; Wang et al., 2025b; Yu et al., 2025; Gemma Team et al., 2025).

Model	Reasoning	Model ID
Claude Opus 5	○	`claude-opus-5`
Gemini 3.6 Flash	✓	`gemini-3.6-flash`
GPT-5.5	✓	`gpt-5.5`
GPT-5.6-sol	✓	`gpt-5.6-sol`
GPT-5.1	✗	`gpt-5.1`
GLM-5.3 Flash	✓	`GLM-5.3-Flash`
Kimi-K3 (256K)	✓	`Kimi-K3-256K`
Qwen3.8 Flash Next	✓	`Qwen3.8-Flash-Next`
Qwen3.8 Flash Next	✗	`Qwen3.8-Flash-Next`
Qwen3.8-27B	✓	`Qwen3.8-27B`
Qwen3.8-27B	✗	`Qwen3.8-27B`
Kimi-K2.5	✗	`Kimi-K2.5`
MiMo-VL-7B-RL-2508	✓	`MiMo-VL-7B-RL-2508`
DeepSeek-V4.1 Flash	✗	`DeepSeek-V4.1-Flash`
Qwen3-VL-235B-A22B	✓	`Qwen3-VL-235B-A22B-Thinking`
Qwen3-VL-235B-A22B	✗	`Qwen3-VL-235B-A22B-Instruct`
GLM-4.6V	✓	`GLM-4.6V`
GLM-4.6V	✗	`GLM-4.6V`
Ling-3.0-Flash-VL	✗	`Ling-3.0-Flash-VL`
Qwen2.5-VL-72B	✗	`Qwen2.5-VL-72B-Instruct`
GLM-4.5V	✗	`GLM-4.5V`
Qwen3-VL-8B	✗	`Qwen3-VL-8B-Instruct`
Qwen2.5-VL-7B	○	`Qwen2.5-VL-7B-Instruct`
InternVL3.5-8B	○	`InternVL3.5-8B`
MiniCPM-V-4.5	○	`MiniCPM-V-4_5`
Qwen3-VL-32B	✗	`Qwen3-VL-32B-Instruct`
InternVL3.5-30B-A3B	○	`InternVL3_5-30B-A3B-Instruct`
Qwen3-VL-30B-A3B	✗	`Qwen3-VL-30B-A3B-Instruct`
Gemma-3-27B-it	○	`gemma-3-27b-it`
Qwen3-VL-2B	✗	`Qwen3-VL-2B-Instruct`

Table 9: Pixel-difference localization on held-out pages. All values except n are percentages. Regions are compared with edit boxes, and flagged area is the share of page pixels that the method flags. Parameters are selected on 40 development source pages and evaluated on 307 held-out source pages. On the no-edit controls, flagged area is 0.00% on the 307 Identical pages and 66.14% on the 307 Reflow only pages.

Condition	n	Precision (%)	Recall (%)	F_1 (%)	Flagged Area (%)
Edit only	1,535	18.50	32.68	23.63	0.91
Edit + reflow	1,535	0.07	2.06	0.13	66.14

D Additional Results

This appendix reports every configuration in full and gives the analyses behind §5: false reports on controls, layout-gap transitions and failure classes, edit count, reasoning, input interventions, edit difficulty, source categories, and scoring checks.

D.1 Full results

Tabs. 10 to 12 report all 30 configurations with 95% CIs: exact match and false reports, the layout gap, single- and five-edit exact match, and per-edit precision and recall. Tab. 13 gives per-edit recall by edit family in the original layout, the counterpart of the family columns of Tab. 2, and Tab. 14 partitions every control task into false report, correct empty report, and invalid output.

D.2 False reports and complete edit reports

0 50 100

Configuration	Reasoning	False reports (%) Identical	False reports (%) Reflow only	Exact match (%) Edit only	Exact match (%) Edit + reflow	Per-edit recall (%) Edit only	Per-edit recall (%) Edit + reflow
Proprietary							
Claude Opus 5	◦	0.0	4.0	80.2	76.4	90.4	90.7
Gemini 3.6 Flash	✓	2.9	9.6	49.0	47.2	74.8	76.8
GPT-5.5	✓	0.0	12.0	10.5	8.5	34.5	30.6
GPT-5.6-sol	✓	0.0	6.4	7.9	5.9	29.8	24.9
GPT-5.1	✗	1.6	63.4	4.2	2.1	15.2	16.1
Open-weight							
GLM-5.3 Flash	✓	5.1	13.4	22.0	16.5	52.0	45.0
Kimi-K3 (256K)	✓	0.3	0.5	25.5	10.8	39.5	16.8
Qwen3.8 Flash Next	✓	1.1	22.2	13.3	10.2	39.2	35.8
Qwen3.8-27B	✓	1.3	10.2	12.4	7.9	39.5	29.4
Kimi-K2.5	✗	11.0	42.3	29.8	7.6	60.9	46.5
MiMo-VL-7B-RL-2508	✓	2.7	38.0	14.4	6.5	41.7	36.3
DeepSeek-V4.1 Flash	✗	4.8	14.2	13.6	6.0	40.0	27.3
Qwen3-VL-235B-A22B	✓	7.5	41.4	10.6	5.8	35.6	31.4
GLM-4.6V	✓	0.3	23.3	8.8	4.5	30.0	23.7
Ling-3.0-Flash-VL	✗	1.9	31.6	18.2	4.5	47.5	30.7
Qwen3.8-27B	✗	0.0	12.3	10.4	3.5	33.9	19.3
Qwen3.8 Flash Next	✗	19.0	69.3	10.9	3.2	34.3	22.1
Qwen3-VL-235B-A22B	✗	0.0	43.9	6.4	1.2	23.4	21.3
Qwen2.5-VL-72B	✗	0.0	18.7	8.2	1.1	29.4	12.6
GLM-4.5V	✗	0.0	84.8	9.4	1.0	31.2	22.8
GLM-4.6V	✗	0.0	81.3	9.1	1.0	31.1	22.4
Qwen3-VL-8B	✗	0.0	96.5	3.6	0.6	15.2	13.7
Qwen2.5-VL-7B	◦	0.0	83.4	2.7	0.6	12.8	9.0
InternVL3.5-8B	◦	0.8	58.8	4.6	0.5	17.5	7.6
MiniCPM-V-4.5	◦	0.3	26.2	0.9	0.5	4.9	3.8
Qwen3-VL-32B	✗	0.3	81.6	10.0	0.5	33.9	23.4
InternVL3.5-30B-A3B	◦	1.6	44.7	0.7	0.1	3.5	1.7
Qwen3-VL-30B-A3B	✗	0.3	99.5	4.1	0.1	19.0	20.2
Gemma-3-27B-it	◦	75.9	98.7	0.4	0.0	3.6	2.9
Qwen3-VL-2B	✗	1.6	71.1	0.3	0.0	5.4	3.8

Reasoning symbols: see model-settings table.

Figure 12: False-report rate on no-edit controls and exact match on edited tasks, by configuration. There is one row per configuration, sorted by exact match on Edit + reflow within the proprietary and open-weight groups, on the 0–100 colour scale shared with Fig. 5. The columns give false-report rate on the no-edit controls (Identical, Reflow only), exact match on edited pages (Edit only, Edit + reflow), and per-edit recall on edited pages. All rates are over all tasks. Lower is better for false-report rate; higher is better for exact match and per-edit recall. False-report rate counts valid nonempty reports; invalid outputs count as incorrect for exact match and as missed edits for recall. Reasoning marks are as in Tab. 8.

Rejecting layout-only changes is necessary for a complete report and does not produce one (Fig. 12). Always empty is correct on every control task and on no edited task (App. C.2), and many configurations approach this profile, rejecting most Reflow only pairs while solving few edited tasks. GPT-5.6-sol makes false reports on 0.00% of Identical pairs and is exact on 27.27% of single-edit Edit only tasks, against 90.91% for Claude Opus 5 (Tabs. 10 and 11).

Table 10: Primary and diagnostic exact-match rates for all 30 configurations. Exact match on Edit + reflow is the primary metric. All rates are over all tasks. Exact match counts invalid outputs as incorrect, and false reports are valid nonempty reports. Intervals are 95% bootstrap confidence intervals from 2,000 draws that resample source pages (seed 20260923). Tab. 11 gives per-edit recall, per-edit precision, and exact match per k. Reasoning marks are as in Tab. 8.

Model	Reasoning	Edit + reflow Exact Match (%) [95% CI]	Edit only Exact Match (%) [95% CI]	Reflow only False Reports (%)	Identical False Reports (%)
Claude Opus 5	○	76.36 [73.97, 78.73]	80.16 [77.77, 82.62]	4.01	0.00
Gemini 3.6 Flash	✓	47.22 [43.88, 50.42]	49.04 [45.57, 52.56]	9.63	2.94
GPT-5.5	✓	8.50 [6.61, 10.79]	10.53 [8.06, 13.33]	12.03	0.00
GPT-5.6-sol	✓	5.94 [4.64, 7.25]	7.91 [6.68, 9.09]	6.42	0.00
GPT-5.1	✗	2.09 [0.70, 3.94]	4.17 [2.03, 7.08]	63.37	1.60
GLM-5.3 Flash	✓	16.47 [13.62, 19.55]	21.98 [19.42, 24.73]	13.37	5.08
Kimi-K3 (256K)	✓	10.80 [8.12, 13.80]	25.51 [22.72, 28.29]	0.53	0.27
Qwen3.8 Flash Next	✓	10.16 [7.36, 13.56]	13.32 [10.43, 16.66]	22.19	1.07
Qwen3.8-27B	✓	7.91 [6.32, 9.58]	12.35 [10.80, 13.80]	10.16	1.34
Kimi-K2.5	✗	7.59 [6.38, 8.88]	29.79 [27.07, 32.45]	42.25	10.96
MiMo-VL-7B-RL-2508	✓	6.52 [4.87, 8.19]	14.44 [11.77, 17.42]	37.97	2.67
DeepSeek-V4.1 Flash	✗	5.99 [4.64, 7.41]	13.64 [11.83, 15.33]	14.17	4.81
Qwen3-VL-235B-A22B	✓	5.83 [4.11, 7.78]	10.59 [8.41, 13.10]	41.44	7.49
GLM-4.6V	✓	4.49 [3.01, 6.30]	8.82 [6.67, 11.28]	23.26	0.27
Ling-3.0-Flash-VL	✗	4.49 [3.19, 5.88]	18.24 [16.06, 20.27]	31.55	1.87
Qwen3.8-27B	✗	3.48 [2.38, 4.63]	10.43 [8.99, 11.76]	12.30	0.00
Qwen3.8 Flash Next	✗	3.21 [1.62, 5.23]	10.91 [8.06, 14.49]	69.25	18.98
Qwen3-VL-235B-A22B	✗	1.23 [0.41, 2.27]	6.36 [4.06, 9.31]	43.85	0.00
Qwen2.5-VL-72B	✗	1.07 [0.64, 1.62]	8.18 [6.32, 10.20]	18.72	0.00
GLM-4.5V	✗	0.96 [0.46, 1.50]	9.41 [6.96, 12.49]	84.76	0.00
GLM-4.6V	✗	0.96 [0.58, 1.39]	9.14 [6.96, 11.71]	81.28	0.00
Qwen3-VL-8B	✗	0.64 [0.23, 1.08]	3.58 [2.43, 4.81]	96.52	0.00
Qwen2.5-VL-7B	○	0.59 [0.27, 0.91]	2.67 [1.80, 3.57]	83.42	0.00
InternVL3.5-8B	○	0.53 [0.23, 0.87]	4.60 [3.74, 5.51]	58.82	0.80
MiniCPM-V-4.5	○	0.53 [0.17, 0.89]	0.86 [0.46, 1.29]	26.20	0.27
Qwen3-VL-32B	✗	0.48 [0.06, 0.97]	10.00 [8.17, 12.01]	81.55	0.27
InternVL3.5-30B-A3B	○	0.05 [0.00, 0.17]	0.75 [0.35, 1.14]	44.65	1.60
Qwen3-VL-30B-A3B	✗	0.05 [0.00, 0.14]	4.12 [2.96, 5.29]	99.47	0.27
Gemma-3-27B-it	○	0.00 [0.00, 0.00]	0.37 [0.00, 0.93]	98.66	75.94
Qwen3-VL-2B	✗	0.00 [0.00, 0.00]	0.32 [0.11, 0.64]	71.12	1.60

Table 11: Per-edit recall, per-edit precision, and exact match per k for all 30 configurations. Per-edit recall, per-edit precision, and the single-edit and five-edit exact match use Edit only tasks. Δ is the layout gap in percentage points. The last column gives exact match over all $T = 4,488$ tasks. All rates are over all tasks, exact match and per-edit recall count invalid outputs as incorrect, and per-edit precision uses the predictions of valid outputs. Intervals are 95% bootstrap confidence intervals from 2,000 draws that resample source pages (seed 20260923), and the layout-gap intervals use all edit sets. Reasoning marks are as in Tab. 8.

Model	Reasoning	Edit only Per-Edit Recall (%) [95% CI]	Per-Edit Precision (%)	Exact Match (%) $k = 1$	Exact Match (%) $k = 5$	Δ (pp) [95% CI]	Exact Match All Tasks (%)
Claude Opus 5	○	90.36 [88.99, 91.82]	92.42	90.91	68.45	-3.80 [-6.32, -1.34]	81.55
Gemini 3.6 Flash	✓	74.78 [72.75, 76.77]	88.52	70.86	31.82	-1.82 [-4.33, 0.64]	55.73
GPT-5.5	✓	34.55 [31.42, 38.29]	85.53	32.62	1.87	-2.03 [-3.33, -0.69]	23.60
GPT-5.6-sol	✓	29.77 [27.61, 32.13]	86.62	27.27	0.27	-1.98 [-3.16, -0.87]	21.90
GPT-5.1	✗	15.17 [11.17, 20.77]	58.29	14.44	0.53	-2.09 [-3.33, -0.93]	13.86
GLM-5.3 Flash	✓	52.00 [49.47, 54.48]	76.78	51.07	5.88	-5.51 [-7.48, -3.53]	31.15
Kimi-K3 (256K)	✓	39.48 [36.37, 42.92]	91.53	46.26	10.16	-14.71 [-16.97, -12.65]	31.73
Qwen3.8 Flash Next	✓	39.23 [35.65, 43.69]	74.41	36.63	2.14	-3.16 [-4.71, -1.71]	24.47
Qwen3.8-27B	✓	39.52 [37.85, 41.09]	77.38	38.77	1.60	-4.44 [-6.04, -3.07]	24.13
Kimi-K2.5	✗	60.91 [59.00, 62.70]	77.85	61.50	11.50	-22.19 [-24.62, -19.57]	27.72
MiMo-VL-7B-RL-2508	✓	41.69 [39.09, 44.66]	65.37	40.91	1.60	-7.91 [-9.68, -6.14]	22.01
DeepSeek-V4.1 Flash	✗	39.98 [37.58, 42.56]	73.76	41.98	2.67	-7.65 [-8.93, -6.32]	23.26
Qwen3-VL-235B-A22B	✓	35.61 [32.56, 39.38]	72.08	35.56	0.80	-4.76 [-6.04, -3.42]	19.43
GLM-4.6V	✓	30.00 [27.13, 33.38]	80.87	31.02	1.34	-4.33 [-5.56, -2.96]	20.25
Ling-3.0-Flash-VL	✗	47.49 [45.28, 49.68]	73.09	48.13	4.28	-13.74 [-15.45, -11.93]	23.31
Qwen3.8-27B	✗	33.94 [32.19, 35.57]	73.23	35.56	1.60	-6.95 [-8.12, -5.83]	21.43
Qwen3.8 Flash Next	✗	34.30 [30.90, 38.55]	59.90	30.48	1.60	-7.70 [-9.41, -5.97]	15.20
Qwen3-VL-235B-A22B	✗	23.35 [19.61, 28.31]	61.04	19.52	0.27	-5.13 [-7.13, -3.42]	16.18
Qwen2.5-VL-72B	✗	29.38 [26.82, 32.22]	64.20	25.94	0.80	-7.11 [-9.35, -5.16]	18.96
GLM-4.5V	✗	31.16 [28.08, 35.02]	73.82	29.95	1.34	-8.45 [-11.12, -6.14]	13.86
GLM-4.6V	✗	31.05 [28.27, 34.38]	71.16	30.48	0.53	-8.18 [-10.56, -6.14]	13.99
Qwen3-VL-8B	✗	15.19 [12.83, 18.19]	51.48	16.04	0.00	-2.94 [-3.95, -1.97]	10.34
Qwen2.5-VL-7B	○	12.78 [10.96, 15.12]	31.47	12.30	0.00	-2.09 [-2.89, -1.22]	10.94
InternVL3.5-8B	○	17.50 [15.36, 20.06]	40.26	20.32	0.00	-4.06 [-4.92, -3.25]	13.75
MiniCPM-V-4.5	○	4.92 [3.44, 6.82]	24.60	4.28	0.00	-0.32 [-0.75, 0.10]	14.71
Qwen3-VL-32B	✗	33.87 [31.27, 36.85]	68.17	32.62	0.53	-9.52 [-11.15, -7.94]	14.19
InternVL3.5-30B-A3B	○	3.49 [2.65, 4.45]	17.98	3.48	0.00	-0.70 [-1.10, -0.29]	12.97
Qwen3-VL-30B-A3B	✗	19.04 [16.87, 21.59]	47.15	17.38	0.00	-4.06 [-5.19, -2.96]	10.09
Gemma-3-27B-it	○	3.60 [2.03, 5.79]	2.58	1.60	0.00	-0.37 [-0.93, 0.00]	2.16
Qwen3-VL-2B	✗	5.42 [4.29, 6.79]	5.18	1.60	0.00	-0.32 [-0.64, -0.11]	8.78

Table 12: Per-edit precision and per-edit recall on all edited tasks (%). Per-edit precision uses the predictions of valid outputs, and per-edit recall includes every reference edit. Intervals are 95% bootstrap confidence intervals that resample source pages. Tabs. 10 and 11 give the other metrics, and reasoning marks are as in Tab. 8.

Model	Reasoning	Edit only Per-Edit Precision (%) [95% CI]	Edit + reflow Per-Edit Recall (%) [95% CI]	Edit + reflow Per-Edit Precision (%) [95% CI]
Claude Opus 5	○	92.42 [91.12, 93.71]	90.73 [89.64, 91.88]	92.68 [91.67, 93.70]
Gemini 3.6 Flash	✓	88.52 [87.09, 89.94]	76.79 [74.98, 78.51]	83.55 [81.20, 85.75]
GPT-5.5	✓	85.53 [83.31, 87.49]	30.64 [27.73, 33.94]	75.46 [73.13, 77.74]
GPT-5.6-sol	✓	86.62 [85.01, 88.50]	24.92 [22.14, 28.27]	78.80 [76.02, 81.30]
GPT-5.1	✗	58.29 [50.52, 64.58]	16.11 [12.64, 20.90]	12.17 [9.02, 16.70]
GLM-5.3 Flash	✓	76.78 [74.52, 78.98]	44.96 [41.82, 48.38]	72.12 [68.78, 74.83]
Kimi-K3 (256K)	✓	91.53 [89.76, 93.21]	16.77 [13.33, 20.90]	92.35 [90.35, 94.51]
Qwen3.8 Flash Next	✓	74.41 [71.55, 76.89]	35.81 [32.00, 40.66]	68.64 [65.90, 70.93]
Qwen3.8-27B	✓	77.38 [75.37, 79.45]	29.41 [27.19, 31.90]	73.86 [70.45, 76.98]
Kimi-K2.5	✗	77.85 [75.80, 79.88]	46.52 [44.08, 49.09]	24.95 [22.53, 27.49]
MiMo-VL-7B-RL-2508	✓	65.37 [63.00, 67.53]	36.35 [33.00, 40.17]	45.22 [42.06, 48.20]
DeepSeek-V4.1 Flash	✗	73.76 [71.17, 76.03]	27.29 [25.08, 29.78]	47.68 [42.64, 52.96]
Qwen3-VL-235B-A22B	✓	72.08 [69.66, 74.22]	31.43 [27.90, 35.77]	44.22 [39.76, 49.05]
GLM-4.6V	✓	80.87 [78.84, 83.10]	23.73 [20.64, 27.61]	50.11 [45.03, 55.01]
Ling-3.0-Flash-VL	✗	73.09 [71.17, 74.82]	30.70 [27.90, 34.09]	30.14 [26.53, 34.24]
Qwen3.8-27B	✗	73.23 [71.11, 75.50]	19.32 [17.76, 20.97]	40.49 [36.34, 44.85]
Qwen3.8 Flash Next	✗	59.90 [56.22, 63.37]	22.09 [18.74, 26.36]	19.33 [15.65, 24.35]
Qwen3-VL-235B-A22B	✗	61.04 [52.80, 68.32]	21.34 [18.49, 24.88]	6.64 [5.34, 8.62]
Qwen2.5-VL-72B	✗	64.20 [61.66, 66.58]	12.60 [10.67, 14.89]	27.37 [25.18, 29.88]
GLM-4.5V	✗	73.82 [71.05, 76.21]	22.80 [20.29, 25.79]	8.40 [6.97, 10.38]
GLM-4.6V	✗	71.16 [66.19, 75.10]	22.42 [19.77, 25.62]	8.34 [6.93, 10.26]
Qwen3-VL-8B	✗	51.48 [46.95, 55.05]	13.65 [10.98, 17.18]	5.23 [3.97, 7.10]
Qwen2.5-VL-7B	○	31.47 [27.10, 35.97]	9.04 [7.09, 11.48]	7.25 [5.58, 9.34]
InternVL3.5-8B	○	40.26 [34.84, 45.28]	7.63 [6.03, 9.66]	6.63 [5.05, 8.55]
MiniCPM-V-4.5	○	24.60 [17.89, 31.37]	3.81 [2.88, 4.93]	6.02 [4.20, 8.52]
Qwen3-VL-32B	✗	68.17 [64.67, 71.22]	23.42 [20.70, 26.73]	6.67 [5.47, 8.37]
InternVL3.5-30B-A3B	○	17.98 [14.06, 21.79]	1.69 [1.12, 2.32]	1.70 [1.14, 2.30]
Qwen3-VL-30B-A3B	✗	47.15 [42.45, 52.14]	20.18 [17.31, 23.72]	3.23 [2.62, 4.11]
Gemma-3-27B-it	○	2.58 [1.37, 4.57]	2.92 [1.43, 5.00]	1.00 [0.48, 1.78]
Qwen3-VL-2B	✗	5.18 [4.05, 6.53]	3.81 [2.96, 4.74]	1.27 [0.98, 1.59]

Table 13: Per-edit recall by edit family on Edit only tasks (%). All 30 configurations are shown. Per-edit recall uses all reference edits of that family in the original layout, and invalid outputs count as missed edits. Bold marks the best and underline the second-best value in each column. Tab. 2 gives the Edit + reflow families, and reasoning marks are as in Tab. 8.

Model	Reasoning	Numeric	Sign	Unit	Glyphs	Spelling	Punctuation
Proprietary models							
Claude Opus 5	○	**92.79**	**88.37**	**88.89**	**88.82**	**86.38**	**91.64**
Gemini 3.6 Flash	✓	<u>83.85</u>	44.77	50.00	<u>72.27</u>	<u>81.84</u>	<u>66.37</u>
GPT-5.5	✓	60.07	62.21	30.56	20.36	17.56	25.52
GPT-5.6-sol	✓	53.52	<u>64.53</u>	22.22	16.64	10.60	22.55
GPT-5.1	✗	31.35	22.67	0.00	11.79	10.60	2.85
Open-weight models							
GLM-5.3 Flash	✓	73.00	59.30	58.33	37.69	39.66	46.93
Kimi-K3 (256K)	✓	51.73	39.53	52.78	31.80	28.76	38.51
Qwen3.8 Flash Next	✓	64.18	52.91	30.56	26.00	28.66	27.74
Qwen3.8-27B	✓	71.45	45.93	<u>61.11</u>	22.70	16.35	31.22
Kimi-K2.5	✗	81.47	59.88	<u>61.11</u>	43.07	44.90	62.25
MiMo-VL-7B-RL-2508	✓	67.58	42.44	33.33	26.52	32.80	30.97
DeepSeek-V4.1 Flash	✗	65.61	47.67	22.22	28.86	22.20	31.60
Qwen3-VL-235B-A22B	✓	61.92	37.21	22.22	17.50	16.65	32.93
GLM-4.6V	✓	63.59	30.81	22.22	16.72	18.57	11.27
Ling-3.0-Flash-VL	✗	70.20	57.56	55.56	30.24	32.59	44.02
Qwen3.8-27B	✗	65.08	47.67	36.11	18.98	13.72	22.93
Qwen3.8 Flash Next	✗	57.99	37.79	38.89	22.96	24.82	22.86
Qwen3-VL-235B-A22B	✗	46.66	25.00	8.33	16.12	16.15	8.55
Qwen2.5-VL-72B	✗	54.29	25.00	27.78	24.52	23.41	10.70
GLM-4.5V	✗	65.08	36.05	27.78	18.54	19.07	11.46
GLM-4.6V	✗	65.49	31.40	33.33	18.11	18.16	11.91
Qwen3-VL-8B	✗	32.96	20.35	11.11	7.11	8.98	5.64
Qwen2.5-VL-7B	○	30.04	14.53	13.89	5.72	6.66	3.23
InternVL3.5-8B	○	39.21	26.16	11.11	6.76	6.66	8.30
MiniCPM-V-4.5	○	12.40	5.23	2.78	1.99	3.13	0.25
Qwen3-VL-32B	✗	58.76	37.79	22.22	19.24	17.96	27.93
InternVL3.5-30B-A3B	○	10.73	6.98	0.00	0.09	0.30	0.00
Qwen3-VL-30B-A3B	✗	45.29	31.98	13.89	8.75	9.08	3.61
Gemma-3-27B-it	○	7.93	5.81	0.00	1.56	3.43	0.44
Qwen3-VL-2B	✗	14.12	10.47	8.33	1.39	1.41	1.01

D.3 Layout-gap transitions and failure classes

Tab. 15 gives, for every configuration, the edit sets exact in both layouts, in one layout alone, and in neither. Across configurations 0.32–28.66% of edit sets change outcome between the two layouts (§ 5.2). Tab. 16 partitions every Edit + reflow task into fully correct, pure omissions, pure additions, both, and invalid output (§ 5.3).

D.4 Edit count and nested edit sets

Tabs. 17 to 19 give exact match and per-edit recall at each k for every configuration in both layouts, and Fig. 13 repeats the analysis of Fig. 5 in the original layout. In the original layout, per-edit recall is higher at $k = 5$ than at $k = 1$ in 16 of 30 configurations, and at $k = 5$ exact match exceeds the independence reference p_1^k, where p_1 is the single-edit exact-match rate, in 22 of 30.

Controlled nested edit sets. For a nested edit set $S_k = \{E_1, \ldots, E_k\}$, each member's per-edit recall is paired with its per-edit recall when presented alone and averaged over both layouts and the k members within each of 40 source pages, then across pages for each of 3 configurations. Invalid outputs count as missed edits. Adding up to 4 other edits changes per-edit recall by -5.5 to $+1.9$ points. The 40 source pages form 34 clusters, and all 12 95% CIs over these clusters include zero (Tab. 20).

Table 14: Outcomes on the no-edit controls over all tasks. Each control cell splits into false reports (a valid nonempty edit list), empty reports (a valid empty list, which is correct), and invalid outputs. All entries are percentages of $T = 374$ tasks per condition. Reasoning marks are as in Tab. 8.

Model	Reasoning	Identical			Reflow only		
		False Reports	Empty Reports	Invalid Outputs	False Reports	Empty Reports	Invalid Outputs
Claude Opus 5	○	0.00	100.00	0.00	4.01	95.99	0.00
Gemini 3.6 Flash	✓	2.94	97.06	0.00	9.63	90.37	0.00
GPT-5.5	✓	0.00	100.00	0.00	12.03	87.97	0.00
GPT-5.6-sol	✓	0.00	100.00	0.00	6.42	93.58	0.00
GPT-5.1	✗	1.60	98.40	0.00	63.37	36.63	0.00
GLM-5.3 Flash	✓	5.08	94.92	0.00	13.37	86.63	0.00
Kimi-K3 (256K)	✓	0.27	99.73	0.00	0.53	99.47	0.00
Qwen3.8 Flash Next	✓	1.07	98.40	0.53	22.19	77.81	0.00
Qwen3.8-27B	✓	1.34	98.66	0.00	10.16	89.57	0.27
Kimi-K2.5	✗	10.96	89.04	0.00	42.25	56.68	1.07
MiMo-VL-7B-RL-2508	✓	2.67	97.33	0.00	37.97	62.03	0.00
DeepSeek-V4.1 Flash	✗	4.81	95.19	0.00	14.17	85.83	0.00
Qwen3-VL-235B-A22B	✓	7.49	92.51	0.00	41.44	58.56	0.00
GLM-4.6V	✓	0.27	99.73	0.00	23.26	76.74	0.00
Ling-3.0-Flash-VL	✗	1.87	98.13	0.00	31.55	67.91	0.53
Qwen3.8-27B	✗	0.00	100.00	0.00	12.30	87.70	0.00
Qwen3.8 Flash Next	✗	18.98	81.02	0.00	69.25	30.75	0.00
Qwen3-VL-235B-A22B	✗	0.00	100.00	0.00	43.85	56.15	0.00
Qwen2.5-VL-72B	✗	0.00	100.00	0.00	18.72	81.28	0.00
GLM-4.5V	✗	0.00	100.00	0.00	84.76	14.44	0.80
GLM-4.6V	✗	0.00	100.00	0.00	81.28	17.38	1.34
Qwen3-VL-8B	✗	0.00	100.00	0.00	96.52	2.94	0.53
Qwen2.5-VL-7B	○	0.00	100.00	0.00	83.42	14.97	1.60
InternVL3.5-8B	○	0.80	99.20	0.00	58.82	40.11	1.07
MiniCPM-V-4.5	○	0.27	99.73	0.00	26.20	69.79	4.01
Qwen3-VL-32B	✗	0.27	99.73	0.00	81.55	18.18	0.27
InternVL3.5-30B-A3B	○	1.60	98.40	0.00	44.65	53.21	2.14
Qwen3-VL-30B-A3B	✗	0.27	99.73	0.00	99.47	0.53	0.00
Gemma-3-27B-it	○	75.94	24.06	0.00	98.66	0.00	1.34
Qwen3-VL-2B	✗	1.60	98.13	0.27	71.12	5.61	23.26

Table 15: Exact-match transitions between layouts over all Edit only and Edit + reflow edit sets. Invalid outputs count as incorrect. n_{11} counts edit sets with exact match in both layouts, n_{10} in the original layout alone, n_{01} after reflow alone, and n_{00} in neither. $\Delta = \frac{100(n_{01}-n_{10})}{N}$ is in percentage points. Reasoning marks are as in Tab. 8.

Model	Reasoning	N	n_{11}	n_{10}	n_{01}	n_{00}	Δ (pp)
Claude Opus 5	○	1,870	1,257	242	171	200	−3.80
Gemini 3.6 Flash	✓	1,870	632	285	251	702	−1.82
GPT-5.5	✓	1,870	96	101	63	1,610	−2.03
GPT-5.6-sol	✓	1,870	61	87	50	1,672	−1.98
GPT-5.1	✗	1,870	22	56	17	1,775	−2.09
GLM-5.3 Flash	✓	1,870	169	242	139	1,320	−5.51
Kimi-K3 (256K)	✓	1,870	105	372	97	1,296	−14.71
Qwen3.8 Flash Next	✓	1,870	110	139	80	1,541	−3.16
Qwen3.8-27B	✓	1,870	92	139	56	1,583	−4.44
Kimi-K2.5	✗	1,870	91	466	51	1,262	−22.19
MiMo-VL-7B-RL-2508	✓	1,870	73	197	49	1,551	−7.91
DeepSeek-V4.1 Flash	✗	1,870	82	173	30	1,585	−7.65
Qwen3-VL-235B-A22B	✓	1,870	68	130	41	1,631	−4.76
GLM-4.6V	✓	1,870	51	114	33	1,672	−4.33
Ling-3.0-Flash-VL	✗	1,870	57	284	27	1,502	−13.74
Qwen3.8-27B	✗	1,870	48	147	17	1,658	−6.95
Qwen3.8 Flash Next	✗	1,870	36	168	24	1,642	−7.70
Qwen3-VL-235B-A22B	✗	1,870	14	105	9	1,742	−5.13
Qwen2.5-VL-72B	✗	1,870	11	142	9	1,708	−7.11
GLM-4.5V	✗	1,870	10	166	8	1,686	−8.45
GLM-4.6V	✗	1,870	13	158	5	1,694	−8.18
Qwen3-VL-8B	✗	1,870	10	57	2	1,801	−2.94
Qwen2.5-VL-7B	○	1,870	4	46	7	1,813	−2.09
InternVL3.5-8B	○	1,870	8	78	2	1,782	−4.06
MiniCPM-V-4.5	○	1,870	3	13	7	1,847	−0.32
Qwen3-VL-32B	✗	1,870	5	182	4	1,679	−9.52
InternVL3.5-30B-A3B	○	1,870	0	14	1	1,855	−0.70
Qwen3-VL-30B-A3B	✗	1,870	0	77	1	1,792	−4.06
Gemma-3-27B-it	○	1,870	0	7	0	1,863	−0.37
Qwen3-VL-2B	✗	1,870	0	6	0	1,864	−0.32

Table 16: Failure decomposition on Edit + reflow tasks (%). Each configuration has 1,870 tasks. A valid report is fully correct, omits edits without adding any, adds edits without omitting any, or does both. Invalid outputs form a separate class. Rows follow exact match on Edit + reflow. Reasoning marks are as in Tab. 8.

Model	Reasoning	Fully Correct	Pure Omissions	Pure Additions	Both	Invalid Output
Claude Opus 5	○	76.36	6.68	2.51	14.44	0.00
Gemini 3.6 Flash	✓	47.22	24.06	5.08	23.48	0.16
GLM-5.3 Flash	✓	16.47	50.00	1.82	31.66	0.05
Kimi-K3 (256K)	✓	10.80	85.88	0.21	3.05	0.05
Qwen3.8 Flash Next	✓	10.16	55.35	2.19	32.25	0.05
GPT-5.5	✓	8.50	69.30	1.34	20.86	0.00
Qwen3.8-27B	✓	7.91	69.79	0.80	21.34	0.16
Kimi-K2.5	✗	7.59	27.97	11.34	52.73	0.37
MiMo-VL-7B-RL-2508	✓	6.52	31.93	5.99	55.08	0.48
DeepSeek-V4.1 Flash	✗	5.99	60.48	1.76	31.76	0.00
GPT-5.6-sol	✓	5.94	79.41	0.59	14.06	0.00
Qwen3-VL-235B-A22B	✓	5.83	36.20	3.64	54.28	0.05
Ling-3.0-Flash-VL	✗	4.49	36.20	4.39	54.22	0.70
GLM-4.6V	✓	4.49	61.50	1.71	32.25	0.05
Qwen3.8-27B	✗	3.48	63.53	1.12	31.82	0.05
Qwen3.8 Flash Next	✗	3.21	23.16	3.42	69.47	0.75
GPT-5.1	✗	2.09	24.49	2.41	70.96	0.05
Qwen3-VL-235B-A22B	✗	1.23	33.10	4.49	61.07	0.11
Qwen2.5-VL-72B	✗	1.07	60.16	1.82	36.74	0.21
GLM-4.5V	✗	0.96	15.72	5.29	77.17	0.86
GLM-4.6V	✗	0.96	14.65	4.92	78.77	0.70
Qwen3-VL-8B	✗	0.64	3.90	2.73	92.57	0.16
Qwen2.5-VL-7B	○	0.59	10.91	1.71	85.56	1.23
InternVL3.5-8B	○	0.53	31.82	1.18	65.24	1.23
MiniCPM-V-4.5	○	0.53	65.99	0.43	29.73	3.32
Qwen3-VL-32B	✗	0.48	8.24	6.84	83.48	0.96
Qwen3-VL-30B-A3B	✗	0.05	0.27	5.40	94.17	0.11
InternVL3.5-30B-A3B	○	0.05	49.36	0.32	48.88	1.39
Qwen3-VL-2B	✗	0.00	3.48	0.64	70.11	25.78
Gemma-3-27B-it	○	0.00	0.00	0.80	98.29	0.91

Table 17: Exact match and per-edit recall by edit count (%), part 1 of 3. All tasks in both layouts are included, and invalid outputs count as incorrect. Rows follow the order of Tab. 2, and edit sets for each k are sampled independently. Reasoning marks are as in Tab. 8.

Model	Reasoning	Condition	Exact Match					Per-Edit Recall				
			$k=1$	2	3	4	5	$k=1$	2	3	4	5
Claude Opus 5	○	Edit only	90.91	84.22	81.82	75.40	68.45	91.18	90.91	91.71	90.24	89.25
		Edit + reflow	84.22	79.68	80.21	71.66	66.04	87.43	89.71	92.34	91.24	90.43
Gemini 3.6 Flash	✓	Edit only	70.86	56.15	50.80	35.56	31.82	71.12	73.80	77.72	74.00	74.76
		Edit + reflow	67.91	51.07	50.00	38.24	28.88	74.60	75.13	78.97	76.94	76.47
GPT-5.5	✓	Edit only	32.62	11.76	5.08	1.34	1.87	32.62	35.16	34.67	32.89	35.94
		Edit + reflow	29.95	8.29	3.21	0.53	0.53	35.03	29.01	31.46	27.87	32.14
GPT-5.6-sol	✓	Edit only	27.27	7.49	3.21	1.34	0.27	27.27	29.01	29.77	28.48	31.60
		Edit + reflow	23.26	3.21	2.14	0.80	0.27	24.60	22.99	25.04	24.00	26.42
GPT-5.1	✗	Edit only	14.44	3.74	1.60	0.53	0.53	14.97	14.17	14.97	14.91	15.94
		Edit + reflow	7.22	1.87	0.27	0.53	0.53	16.84	14.44	17.29	15.51	16.42
GLM-5.3 Flash	✓	Edit only	51.07	24.60	17.65	10.70	5.88	51.87	48.93	52.67	52.27	52.62
		Edit + reflow	40.37	17.65	11.50	6.95	5.88	43.85	40.64	45.63	45.59	45.99
Kimi-K3 (256K)	✓	Edit only	46.26	28.07	24.60	18.45	10.16	46.26	40.78	42.42	39.24	36.04
		Edit + reflow	19.79	13.10	9.89	5.61	5.61	19.79	20.19	18.45	14.84	15.35
Qwen3.8 Flash Next	✓	Edit only	36.63	13.10	10.70	4.01	2.14	37.43	38.24	42.78	37.83	38.98
		Edit + reflow	29.68	10.16	6.68	2.94	1.34	35.29	34.49	38.06	34.36	36.26
Qwen3.8-27B	✓	Edit only	38.77	12.83	6.15	2.41	1.60	38.77	41.18	40.55	37.17	40.27
		Edit + reflow	25.40	8.29	4.01	1.60	0.27	28.34	29.01	30.93	27.81	30.16
Kimi-K2.5	✗	Edit only	61.50	35.83	26.47	13.64	11.50	63.37	61.23	60.43	59.76	61.50
		Edit + reflow	16.84	10.70	6.42	2.67	1.34	41.98	46.66	49.47	44.05	47.59

Table 18: Exact match and per-edit recall by edit count (%), part 2 of 3. All tasks in both layouts are included, and invalid outputs count as incorrect. Rows follow the order of Tab. 2, and edit sets for each k are sampled independently. Reasoning marks are as in Tab. 8.

Model	Reasoning	Condition	Exact Match					Per-Edit Recall				
			$k=1$	2	3	4	5	$k=1$	2	3	4	5
MiMo-VL-7B-RL-2508	✓	Edit only	40.91	16.84	9.89	2.94	1.60	43.32	42.65	43.40	39.71	41.55
		Edit + reflow	18.98	8.82	3.48	0.53	0.80	34.22	38.10	37.70	35.83	35.67
DeepSeek-V4.1 Flash	✗	Edit only	41.98	14.17	7.22	2.14	2.67	42.78	39.84	40.64	39.17	39.73
		Edit + reflow	20.59	5.35	3.21	0.53	0.27	27.01	26.47	28.16	25.67	28.45
Qwen3-VL-235B-A22B	✓	Edit only	35.56	11.50	3.48	1.60	0.80	36.90	37.83	35.56	35.49	34.60
		Edit + reflow	17.38	6.68	2.94	1.07	1.07	28.07	30.88	33.16	29.88	32.51
GLM-4.6V	✓	Edit only	31.02	6.95	4.55	0.27	1.34	31.28	28.61	30.30	29.14	30.80
		Edit + reflow	14.44	3.74	3.74	0.53	0.00	20.86	20.99	25.13	23.20	24.97
Ling-3.0-Flash-VL	✗	Edit only	48.13	23.26	11.50	4.01	4.28	48.40	49.60	47.59	46.72	47.01
		Edit + reflow	14.97	4.28	1.87	0.80	0.53	29.41	32.09	31.28	29.88	30.70
Qwen3.8-27B	✗	Edit only	35.56	8.29	5.35	1.34	1.60	35.83	34.09	34.94	32.09	34.39
		Edit + reflow	12.83	3.74	0.80	0.00	0.00	17.38	19.52	19.43	17.38	21.12
Qwen3.8 Flash Next	✗	Edit only	30.48	11.50	7.22	3.74	1.60	31.82	34.36	37.79	33.69	33.16
		Edit + reflow	10.43	3.74	1.34	0.27	0.27	22.73	20.59	22.55	21.72	22.57
Qwen3-VL-235B-A22B	✗	Edit only	19.52	7.75	3.48	0.80	0.27	19.79	22.46	24.60	22.39	24.44
		Edit + reflow	2.94	2.41	0.53	0.27	0.00	15.51	20.32	22.19	21.32	22.41
Qwen2.5-VL-72B	✗	Edit only	25.94	9.89	2.94	1.34	0.80	27.27	31.42	31.73	29.14	27.75
		Edit + reflow	3.48	1.87	0.00	0.00	0.00	11.50	11.90	13.73	11.90	12.99
GLM-4.5V	✗	Edit only	29.95	9.89	4.81	1.07	1.34	30.21	30.08	31.91	31.22	31.28
		Edit + reflow	2.41	1.07	1.07	0.00	0.27	21.66	21.12	24.78	22.06	23.10

Table 19: Exact match and per-edit recall by edit count (%), part 3 of 3. All tasks in both layouts are included, and invalid outputs count as incorrect. Rows follow the order of Tab. 2, and edit sets for each k are sampled independently. Reasoning marks are as in Tab. 8.

Model	Reasoning	Condition	Exact Match					Per-Edit Recall				
			$k=1$	2	3	4	5	$k=1$	2	3	4	5
GLM-4.6V	✗	Edit only	30.48	8.56	5.35	0.80	0.53	30.75	30.61	32.00	30.61	31.07
		Edit + reflow	3.74	0.80	0.27	0.00	0.00	21.93	19.79	24.69	20.52	23.74
Qwen3-VL-8B	✗	Edit only	16.04	1.07	0.53	0.27	0.00	16.04	16.71	16.22	13.70	14.97
		Edit + reflow	2.67	0.53	0.00	0.00	0.00	12.30	13.50	14.44	12.97	14.06
Qwen2.5-VL-7B	○	Edit only	12.30	1.07	0.00	0.00	0.00	13.90	13.24	14.53	10.90	12.83
		Edit + reflow	2.41	0.27	0.27	0.00	0.00	10.16	8.69	7.75	9.02	9.73
InternVL3.5-8B	○	Edit only	20.32	2.14	0.27	0.27	0.00	21.12	19.12	18.09	16.11	16.90
		Edit + reflow	2.14	0.53	0.00	0.00	0.00	7.22	6.82	7.49	7.09	8.56
MiniCPM-V-4.5	○	Edit only	4.28	0.00	0.00	0.00	0.00	4.55	4.95	5.88	4.01	5.13
		Edit + reflow	2.41	0.27	0.00	0.00	0.00	4.55	3.07	3.65	3.68	4.17
Qwen3-VL-32B	✗	Edit only	32.62	12.30	3.48	1.07	0.53	32.62	34.76	35.65	33.49	32.99
		Edit + reflow	1.60	0.53	0.27	0.00	0.00	26.20	26.07	22.99	23.13	22.30
InternVL3.5-30B-A3B	○	Edit only	3.48	0.27	0.00	0.00	0.00	4.28	3.61	3.03	3.34	3.69
		Edit + reflow	0.27	0.00	0.00	0.00	0.00	1.87	1.07	1.78	1.87	1.71
Qwen3-VL-30B-A3B	✗	Edit only	17.38	2.67	0.53	0.00	0.00	17.91	19.12	20.23	17.25	19.95
		Edit + reflow	0.00	0.27	0.00	0.00	0.00	20.86	18.98	20.23	19.32	21.18
Gemma-3-27B-it	○	Edit only	1.60	0.27	0.00	0.00	0.00	3.21	3.74	4.01	3.81	3.21
		Edit + reflow	0.00	0.00	0.00	0.00	0.00	3.48	2.14	1.96	3.21	3.48
Qwen3-VL-2B	✗	Edit only	1.60	0.00	0.00	0.00	0.00	3.48	5.48	5.44	5.28	5.88
		Edit + reflow	0.00	0.00	0.00	0.00	0.00	2.67	3.61	4.01	3.81	4.01

Edit only · shared scale (%)

0 50 100

Configuration	Reasoning	Per-edit recall (%) k = 1	k = 2	k = 3	k = 4	k = 5	Exact match (%) k = 1	k = 2	k = 3	k = 4	k = 5
Claude Opus 5	∘	91.2	90.9	91.7	90.2	89.3	90.9	84.2	81.8	75.4	68.4
Gemini 3.6 Flash	✓	71.1	73.8	77.7	74.0	74.8	70.9	56.1	50.8	35.6	31.8
Kimi-K2.5	✗	63.4	61.2	60.4	59.8	61.5	61.5	35.8	26.5	13.6	11.5
Kimi-K3 (256K)	✓	46.3	40.8	42.4	39.2	36.0	46.3	28.1	24.6	18.4	10.2
GLM-5.3 Flash	✓	51.9	48.9	52.7	52.3	52.6	51.1	24.6	17.6	10.7	5.9
Ling-3.0-Flash-VL	✗	48.4	49.6	47.6	46.7	47.0	48.1	23.3	11.5	4.0	4.3
MiMo-VL-7B-RL-2508	✓	43.3	42.6	43.4	39.7	41.6	40.9	16.8	9.9	2.9	1.6
DeepSeek-V4.1 Flash	✗	42.8	39.8	40.6	39.2	39.7	42.0	14.2	7.2	2.1	2.7
Qwen3.8 Flash Next	✓	37.4	38.2	42.8	37.8	39.0	36.6	13.1	10.7	4.0	2.1
Qwen3.8-27B	✓	38.8	41.2	40.6	37.2	40.3	38.8	12.8	6.1	2.4	1.6
Qwen3.8 Flash Next	✗	31.8	34.4	37.8	33.7	33.2	30.5	11.5	7.2	3.7	1.6
Qwen3-VL-235B-A22B	✓	36.9	37.8	35.6	35.5	34.6	35.6	11.5	3.5	1.6	0.8
GPT-5.5	✓	32.6	35.2	34.7	32.9	35.9	32.6	11.8	5.1	1.3	1.9
Qwen3.8-27B	✗	35.8	34.1	34.9	32.1	34.4	35.6	8.3	5.3	1.3	1.6
Qwen3-VL-32B	✗	32.6	34.8	35.7	33.5	33.0	32.6	12.3	3.5	1.1	0.5
GLM-4.5V	✗	30.2	30.1	31.9	31.2	31.3	29.9	9.9	4.8	1.1	1.3
GLM-4.6V	✗	30.7	30.6	32.0	30.6	31.1	30.5	8.6	5.3	0.8	0.5
GLM-4.6V	✓	31.3	28.6	30.3	29.1	30.8	31.0	7.0	4.5	0.3	1.3
Qwen2.5-VL-72B	✗	27.3	31.4	31.7	29.1	27.8	25.9	9.9	2.9	1.3	0.8
GPT-5.6-sol	✓	27.3	29.0	29.8	28.5	31.6	27.3	7.5	3.2	1.3	0.3
Qwen3-VL-235B-A22B	✗	19.8	22.5	24.6	22.4	24.4	19.5	7.8	3.5	0.8	0.3
InternVL3.5-8B	∘	21.1	19.1	18.1	16.1	16.9	20.3	2.1	0.3	0.3	0.0
GPT-5.1	✗	15.0	14.2	15.0	14.9	15.9	14.4	3.7	1.6	0.5	0.5
Qwen3-VL-30B-A3B	✗	17.9	19.1	20.2	17.2	19.9	17.4	2.7	0.5	0.0	0.0
Qwen3-VL-8B	✗	16.0	16.7	16.2	13.7	15.0	16.0	1.1	0.5	0.3	0.0
Qwen2.5-VL-7B	∘	13.9	13.2	14.5	10.9	12.8	12.3	1.1	0.0	0.0	0.0
MiniCPM-V-4.5	∘	4.5	4.9	5.9	4.0	5.1	4.3	0.0	0.0	0.0	0.0
InternVL3.5-30B-A3B	∘	4.3	3.6	3.0	3.3	3.7	3.5	0.3	0.0	0.0	0.0
Gemma-3-27B-it	∘	3.2	3.7	4.0	3.8	3.2	1.6	0.3	0.0	0.0	0.0
Qwen3-VL-2B	✗	3.5	5.5	5.4	5.3	5.9	1.6	0.0	0.0	0.0	0.0

Reasoning symbols: see model-settings table.

Figure 13: Edit-count analysis in the original layout. The same analysis as Fig. 5, using Edit only tasks with every task in the denominator. Rows are sorted by overall exact match on Edit only tasks, with the same ordering in both panels. Exact match at $k = 5$ is below $k = 1$ in 30 of 30 configurations. Per-edit recall is higher in 16 and higher or within 2 percentage points in 26. Reasoning marks are as in Tab. 8.

Table 20: Per-edit recall of the same edits alone and within nested edit sets. All 40 source pages are included, and invalid outputs count as missed edits. Rates are percentages, and In set minus Alone differences are in percentage points. Both weight source pages equally. Bootstrap confidence intervals resample source pages with 2,000 draws. Reasoning marks are as in Tab. 8.

Model	Reasoning	k	Alone (%)	In Set (%)	Δ (pp)	95% CI
Claude Opus 5	∘	2	93.12	93.12	+0.00	[−3.12, 2.72]
Claude Opus 5	∘	3	92.92	91.67	−1.25	[−5.43, 3.43]
Claude Opus 5	∘	4	91.88	87.50	−4.38	[−8.59, 0.74]
Claude Opus 5	∘	5	93.00	89.50	−3.50	[−8.08, 0.88]
Gemini 3.6 Flash	✓	2	78.75	80.62	+1.88	[−2.40, 8.09]
Gemini 3.6 Flash	✓	3	82.08	81.25	−0.83	[−5.07, 4.41]
Gemini 3.6 Flash	✓	4	80.94	82.81	+1.88	[−2.88, 7.35]
Gemini 3.6 Flash	✓	5	81.50	76.00	−5.50	[−11.92, 2.35]
GPT-5.6-sol	✓	2	37.50	38.75	+1.25	[−4.43, 7.50]
GPT-5.6-sol	✓	3	36.67	37.50	+0.83	[−2.92, 5.39]
GPT-5.6-sol	✓	4	33.75	35.00	+1.25	[−1.63, 4.78]
GPT-5.6-sol	✓	5	35.50	36.25	+0.75	[−3.85, 6.47]

D.5 Reasoning within one checkpoint

Three within-checkpoint contrasts switch reasoning on and off for the same weights (Tab. 21 and the paired on/off rows of Fig. 12). Qwen3-VL-235B-A22B, whose two settings are different checkpoints, is shown with † and not counted. For GLM-4.6V, reasoning changes the Reflow only false-report rate by -58.0 [-67.5, -47.0] points, Edit only exact match by -0.3 [-1.2, 0.6], and Edit + reflow exact match by $+3.5$ [2.0, 5.2] (95% CIs). Qwen3.8 Flash Next gains $+2.4$ points on Edit only and $+7.0$ on Edit + reflow, with a false-report change of -47.1 points. Qwen3.8-27B gains $+1.9$ and $+4.4$. Reasoning thus acts on two separable abilities, rejecting layout-only changes and completing reports under reflow, and its exact-match gain is larger after reflow than in the original layout in all three pairs.

Table 21: Reasoning-mode contrasts on all tasks. The first result column gives Reflow only false reports with reasoning off and on (%) and their change in percentage points. The remaining columns give on-minus-off changes in percentage points with 95% bootstrap confidence intervals that resample source pages. Single-edit exact match and per-edit recall use Edit only tasks. †The Qwen3-VL contrast compares two checkpoints that also differ in streaming. Reasoning marks are as in Tab. 8.

Model	Reasoning	Reflow only False Reports (%) Off → On, Δ (pp)	Single-Edit Exact Match (pp)	Edit only Exact Match (pp)	Edit + reflow Exact Match (pp)	Edit only Per-Edit Recall (pp)
Qwen3.8 Flash Next	✗ → ✓	69.25 → 22.19 -47.06 [-53.34, -40.91]	+6.15 [2.73, 9.63]	+2.41 [1.33, 3.57]	+6.95 [5.24, 8.64]	+4.94 [3.85, 5.92]
Qwen3.8-27B	✗ → ✓	12.30 → 10.16 -2.14 [-6.67, 2.24]	+3.21 [-0.87, 7.20]	+1.93 [0.75, 3.13]	+4.44 [3.30, 5.56]	+5.58 [4.52, 6.72]
GLM-4.6V	✗ → ✓	81.28 → 23.26 -58.02 [-67.54, -46.99]	+0.53 [-2.94, 4.01]	-0.32 [-1.22, 0.64]	+3.53 [2.03, 5.16]	-1.05 [-1.99, -0.02]
Qwen3-VL-235B-A22B†	✗ → ✓	43.85 → 41.44 -2.41 [-9.57, 4.47]	+16.04 [11.17, 21.45]	+4.22 [3.10, 5.45]	+4.60 [3.32, 5.86]	+12.26 [10.55, 13.97]

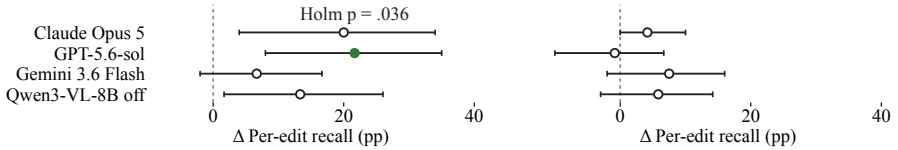
D.6 Targeted input interventions

On single-edit Edit only tasks from 60 source pages, we vary the visual input for 4 configurations in the following input variants. *Full page* reruns the full-page pair with repeated runs. *Small tiles* sends fixed overlapping tiles at a smaller size, and *Enlarged tiles* sends the same tiles at a larger size. *Location box* adds a cue box to the full-page pair, on the edit in half of the edited pairs and on a decoy location otherwise, with corresponding cues on controls. *Edit crops* sends native crop pairs around the edit and corresponding unedited crops.

The outcome measure is per-edit recall, which on single-edit tasks is the share of tasks whose report contains the reference edit. Extra reports are allowed, and invalid outputs count as missed edits. Full-page repeats are averaged before pairing, with invalid repeats counted as missed edits, and each difference is task-weighted over all 60 edited pairs. Intervals are 95% CIs. Two-sided centred bootstrap tail probabilities use a plus-one correction, and Holm correction covers the fixed family of 8 tests.

Enlarged tiles raise per-edit recall over Small tiles in all 4 configurations, by +21.7 [8.0, 35.0] points for GPT-5.6-sol and +20.0 [4.0, 34.0] for Claude Opus 5 (Fig. 14 and Tab. 22). The GPT-5.6-sol contrast passes Holm correction ($p = 0.036$). The two input variants differ in the image size sent to the API. Location box raises the point estimate over Full page in 3 of 4 configurations, with GPT-5.6-sol at -0.8 [-10.0, 6.7] points (Tab. 22).

Table 22: Per-edit recall differences on single-edit tasks, over all tasks. Differences and 95% confidence intervals are in percentage points. Holm correction covers all 8 tests. Reasoning marks are as in Tab. 8.



(a) Enlarged tiles − Small tiles (b) Location box − Full page

Figure 14: Per-edit recall on single-edit tasks under four input variants. On these tasks, per-edit recall is the share of tasks whose report contains the reference edit. Full page uses whole-page pairs. Small tiles and Enlarged tiles send the same overlapping crops at smaller and larger sizes. Location box adds a location cue, with decoy locations on half the edited pairs. Effects use all single-edit tasks, and invalid outputs count as missed edits. Bars are 95% bootstrap confidence intervals that resample source pages (2000 draws), and filled dots mark Holm $p < .05$ over 8 tests. The Enlarged tiles estimate is positive in all 4 configurations, GPT-5.6-sol is the one Holm-significant contrast ($p = 0.036$), and no Location box contrast passes Holm correction.

Model	Reasoning	Contrast	N	Δ (pp)	95% CI	Holm-Adjusted p
Claude Opus 5	○	Enlargement	60	+20.0	[4.0, 34.0]	0.098
Claude Opus 5	○	Location cue	60	+4.2	[0.0, 10.0]	0.615
GPT-5.6-sol	✓	Enlargement	60	+21.7	[8.0, 35.0]	0.036
GPT-5.6-sol	✓	Location cue	60	−0.8	[−10.0, 6.7]	0.905
Gemini 3.6 Flash	✓	Enlargement	60	+6.7	[−2.0, 16.7]	0.616
Gemini 3.6 Flash	✓	Location cue	60	+7.5	[−2.0, 16.0]	0.615
Qwen3-VL-8B	✗	Enlargement	60	+13.3	[1.7, 26.0]	0.249
Qwen3-VL-8B	✗	Location cue	60	+5.8	[−3.0, 14.2]	0.616

Table 23: Outcomes per input variant on all admitted inputs. Invalid outputs count as incorrect. Exact match is the share of all inputs. Each configuration has $N = 240$ inputs for Full page, 180 for Small tiles, Enlarged tiles, and Location box, and 94 for Edit crops. Reasoning marks are as in Tab. 8.

Model	Reasoning	Input	Exact Match (%)
Claude Opus 5	○	Full page	94.17
		Small tiles	77.78
		Enlarged tiles	91.67
		Location box	94.44
		Edit crops	93.62
GPT-5.6-sol	✓	Full page	73.75
		Small tiles	70.56
		Enlarged tiles	80.00
		Location box	73.89
		Edit crops	74.47
Gemini 3.6 Flash	✓	Full page	84.58
		Small tiles	83.89
		Enlarged tiles	88.89
		Location box	86.67
		Edit crops	95.74
Qwen3-VL-8B	✗	Full page	40.83
		Small tiles	36.67
		Enlarged tiles	40.56
		Location box	42.22
		Edit crops	77.66

D.7 What makes an edit hard to see

Outcomes and models. We pool 336,600 edit×configuration outcomes across $k = 1, \ldots, 5$ and both edited conditions. Each reference edit contributes its own recovered or missed outcome, so larger edit sets contribute more edits. The area and numeric contrasts use logistic regression on log native edited-grapheme bounding-box area, edit family, categorical k, condition, and configuration. Area doubling is a marginal probability contrast with odds ratio 1.73, and its intervals are Wald delta-method intervals clustered by source page. A separate linear probability model includes categorical k×layout×look-alike interactions and configuration effects. Its look-alike contrast averages equally over k and layout, and its reflow and edit-count contrasts standardize to the observed look-alike prevalence. Look-alike edits comprise letter–digit confusion, similar-letter substitution, case change, and alphanumeric swaps. Descriptive rates and interaction contrasts carry pointwise 95% CIs (seed 20260924) that keep all configurations, edits, and both layouts of a page together within its cluster.

Table 24: Edit difficulty and unpaired predictions. Effects are percentage-point changes in per-edit recall with 95% CIs. Rates are percentages. Logistic and interaction-model contrasts use the adjustments described in App. D.7.

Contrast or Quantity	Estimate [95% CI]
Area doubling (logistic, pp)	+8.49 [7.37, 9.62]
Look-alike versus other edits (interaction, pp)	-11.83 [-13.67, -10.30]
Reflow versus no reflow (interaction, pp)	-6.92 [-7.44, -6.44]
$k = 5$ versus $k = 1$ (interaction, pp)	+1.22 [-0.88, 3.55]
Numeric versus other edits (raw rates)	48.62% versus 20.59%
Numeric versus other edits (logistic, pp)	+25.42 [23.35, 27.50]
Configurations with positive numeric contrast	All 30
Unpaired predictions with identical strings	64.70%

Size, family, and false reports. From the smallest to the largest area quintile, per-edit recall rises from 23.6% to 41.5% without reflow and from 17.4% to 34.7% with reflow (Fig. 16). Area is the native edited-grapheme bounding box on the original page. Numeric edits have the highest pooled per-edit recall and look-alike, spelling, and punctuation edits the lowest (Fig. 15). Unpaired predictions whose before and after strings are identical make up 64.70% of all unpaired predictions (Tab. 24).

D.8 Source categories and scoring checks

Tab. 25 gives Edit + reflow exact match by source category (§ 5.1), and Tab. 27 gives the per-configuration second-judge comparison summarized in § 4.3.

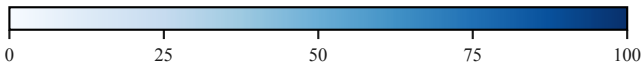
Table 25: Exact match on Edit + reflow (%) by source category. The 8 source categories are six generated ones, real web pages, and real financial reports. The 6 representative configurations are the top three proprietary and top three open-weight by exact match on Edit + reflow, and the last row is the mean over all 30. Bold marks the best and underline the second-best value in each column. Reasoning marks are as in Tab. 8.

		Generated						Real	
Model	Reasoning	Public Affairs	Research	Education	Engineering	Finance	Procurement	Web	Financial Reports
Claude Opus 5	○	**85.20**	**77.60**	**80.00**	**74.29**	**81.22**	**82.04**	**76.67**	**53.33**
Gemini 3.6 Flash	✓	<u>51.20</u>	<u>42.00</u>	<u>52.24</u>	<u>46.12</u>	<u>48.98</u>	<u>45.71</u>	<u>62.00</u>	<u>35.00</u>
GPT-5.5	✓	10.00	4.00	10.20	5.71	7.76	6.94	21.33	7.08
GLM-5.3 Flash	✓	16.80	8.40	11.84	11.84	15.92	18.78	33.33	21.67
Kimi-K3 (256K)	✓	10.40	7.20	6.53	8.98	11.02	8.57	29.33	11.67
Qwen3.8 Flash Next	✓	9.20	5.60	7.35	6.94	10.61	5.71	31.33	12.92
Mean (30 configurations)		8.13	6.13	7.25	6.82	8.03	7.25	14.27	7.08

Kimi-K3 authorship check. Kimi-K3 (256K) (Moonshot AI, 2026) is a regular row of Tabs. 2 and 10. It shows no authorship advantage: pooled exact match is 25.3% on the 86 source pages

Per-edit recall, all k (%)

Model	Reasoning	Numeric	Sign	Unit	Glyphs	Spelling	Punctuation
Proprietary							
Claude Opus 5	∘	92.70	85.17	94.44	89.64	87.69	91.20
Gemini 3.6 Flash	✓	83.79	46.80	52.78	72.18	83.96	68.46
GPT-5.5	✓	58.49	56.98	23.61	19.28	14.78	23.53
GPT-5.6-sol	✓	50.06	59.59	16.67	15.86	9.08	19.79
GPT-5.1	✗	32.18	20.93	5.56	11.96	10.04	3.93
Open-weight							
GLM-5.3 Flash	✓	70.38	55.81	50.00	34.92	38.80	40.34
Kimi-K3 (256K)	✓	37.69	28.49	37.50	22.62	22.75	25.11
Qwen3.8 Flash Next	✓	63.26	50.58	34.72	24.78	26.14	25.27
Qwen3.8-27B	✓	66.06	45.35	43.06	18.67	14.33	23.69
Kimi-K2.5	✗	75.80	54.36	52.78	37.61	40.77	50.09
MiMo-VL-7B-RL-2508	✓	63.86	40.41	27.78	25.17	30.93	27.93
DeepSeek-V4.1 Flash	✗	59.56	42.73	15.28	23.79	19.22	21.75
Qwen3-VL-235B-A22B	✓	58.97	33.43	22.22	16.25	15.14	30.91
GLM-4.6V	✓	57.69	27.33	23.61	14.04	16.45	10.04
Ling-3.0-Flash-VL	✗	60.37	52.33	43.06	24.52	28.51	32.24
Qwen3.8-27B	✗	54.47	40.41	27.78	14.51	11.00	14.19
Qwen3.8 Flash Next	✗	49.34	32.27	33.33	18.63	21.09	16.59
Qwen3-VL-235B-A22B	✗	44.28	21.80	11.11	13.95	13.93	10.77
Qwen2.5-VL-72B	✗	42.07	16.86	18.06	16.29	15.19	6.17
GLM-4.5V	✗	58.76	26.45	23.61	14.04	16.09	9.63
GLM-4.6V	✗	58.58	25.00	30.56	13.30	15.84	9.66
Qwen3-VL-8B	✗	31.14	16.57	9.72	6.37	8.93	5.86
Qwen2.5-VL-7B	∘	25.18	12.21	9.72	5.11	6.00	2.94
InternVL3.5-8B	∘	29.17	18.31	5.56	4.77	4.59	5.16
MiniCPM-V-4.5	∘	11.26	6.10	2.78	1.43	2.32	0.32
Qwen3-VL-32B	✗	52.35	31.98	18.06	14.77	15.49	21.72
InternVL3.5-30B-A3B	∘	7.84	4.07	0.00	0.09	0.35	0.16
Qwen3-VL-30B-A3B	✗	43.71	26.45	19.44	8.84	9.94	7.19
Gemma-3-27B-it	∘	7.27	5.52	0.00	1.26	3.13	0.38
Qwen3-VL-2B	✗	12.19	7.85	8.33	1.17	1.11	0.85



All k and both conditions pooled.

Reasoning symbols: see model-settings table.

Figure 15: Per-edit recall by configuration and edit family. Per-edit recall pools all edit counts and both edited conditions and weights all reference-edit outcomes equally. Numeric edits have the highest pooled per-edit recall, and look-alike glyph, spelling, and punctuation edits have the lowest. Reasoning marks are as in Tab. 8.

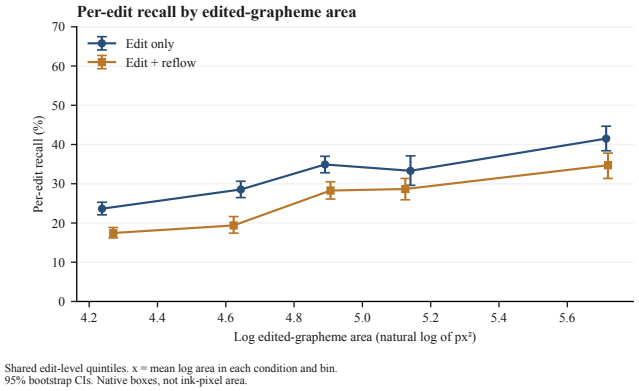


Figure 16: Per-edit recall by edited-grapheme area in both layouts. Points are area quintiles, placed at the mean log area of each quintile. Bars are 95% bootstrap confidence intervals that resample source pages. Area is measured on the native page.

Table 26: Single-edit exact match on real and generated sources (%). Exact match is over all Edit only tasks with $k = 1$ in the original layout, and invalid outputs count as incorrect. Real pages combine the web and financial-report sources. Difference is Real minus Generated in percentage points and is positive in 26 of 30 configurations. Rows follow Tab. 2, and reasoning marks are as in Tab. 8.

Model	Reasoning	Generated	Real (Web and Financial)	Difference (pp)
Proprietary models				
Claude Opus 5	○	94.59	76.92	-17.67
Gemini 3.6 Flash	✓	70.95	70.51	-0.43
GPT-5.5	✓	30.41	41.03	+10.62
GPT-5.6-sol	✓	25.68	33.33	+7.66
GPT-5.1	✗	10.47	29.49	+19.01
Open-weight models				
GLM-5.3 Flash	✓	50.00	55.13	+5.13
Kimi-K3 (256K)	✓	45.27	50.00	+4.73
Qwen3.8 Flash Next	✓	34.12	46.15	+12.03
Qwen3.8-27B	✓	38.18	41.03	+2.85
Kimi-K2.5	✗	61.49	61.54	+0.05
MiMo-VL-7B-RL-2508	✓	37.84	52.56	+14.73
DeepSeek-V4.1 Flash	✗	41.55	43.59	+2.04
Qwen3-VL-235B-A22B	✓	34.46	39.74	+5.28
GLM-4.6V	✓	27.70	43.59	+15.89
Ling-3.0-Flash-VL	✗	45.61	57.69	+12.08
Qwen3.8-27B	✗	35.81	34.62	-1.20
Qwen3.8 Flash Next	✗	28.04	39.74	+11.70
Qwen3-VL-235B-A22B	✗	15.54	34.62	+19.07
Qwen2.5-VL-72B	✗	23.65	34.62	+10.97
GLM-4.5V	✗	26.69	42.31	+15.62
GLM-4.6V	✗	28.38	38.46	+10.08
Qwen3-VL-8B	✗	12.16	30.77	+18.61
Qwen2.5-VL-7B	○	10.47	19.23	+8.76
InternVL3.5-8B	○	20.27	20.51	+0.24
MiniCPM-V-4.5	○	4.05	5.13	+1.07
Qwen3-VL-32B	✗	30.07	42.31	+12.24
InternVL3.5-30B-A3B	○	3.04	5.13	+2.09
Qwen3-VL-30B-A3B	✗	15.20	25.64	+10.44
Gemma-3-27B-it	○	0.00	7.69	+7.69
Qwen3-VL-2B	✗	2.03	0.00	-2.03

Table 27: Agreement between the primary judge and a second judge. GPT-5.5 re-judges all 10,800 outputs of 9 configurations on 100 source pages. Both judges use the original reference context, the same parser, and the same literal candidates, and the judge model is the one change. Exact match is over all conditions on all tasks. Agreement is the share of tasks on which both judges make the same exact-match decision. Invalid outputs count as incorrect. Rank gives the rank under the primary judge and then under the second judge. GPT-5.5, the second judge, is also one of the evaluated configurations. Reasoning marks are as in Tab. 8.

Model	Reasoning	Exact Match (Primary Judge)	Exact Match (Second Judge)	Agreement	Rank
Claude Opus 5	○	81.58	81.25	96.67	1→1
Gemini 3.6 Flash	✓	56.17	54.17	93.67	2→2
GLM-5.3 Flash	✓	31.33	31.08	97.00	3→3
Qwen3.8 Flash Next	✓	25.17	25.08	98.25	4→4
GPT-5.5	✓	24.25	24.42	98.33	5→5
GPT-5.6-sol	✓	22.42	22.25	98.67	6→6
GLM-4.6V	✓	22.00	21.58	98.50	7→7
GLM-4.6V	✗	14.17	13.83	98.67	8→8
Qwen3-VL-8B	✗	10.92	10.75	99.33	9→9

whose original page Kimi-K3 generated and 33.7% on the other 288, a difference of -8.4 points (95% CI $[-11.3, -5.3]$).

E QUALITATIVE CASES

This appendix shows four scored outputs, one for each error type of §5: a false report on a Reflow only pair, a wrong occurrence, a loss under reflow, and partial recovery of a five-edit set (Figs. 17 to 20). Each box gives the condition, the reference edits, literal model-output fragments, and the TP, FP, and FN counts of App. A.2. Reference descriptions and score summaries are paraphrases. Quoted output strings are verbatim.

Condition: Reflow only, education, GLM-5.3 Flash (reasoning on).
Reference: no edits.
Output fragments: location: opening paragraph, company name (paraphrased).
Before "`Meridian Analytics LLC`". After "`Meridian Analytics llc`".
Outcome: TP $= 0$, FP $= 1$, FN $= 0$. The capitalization change is a false report.

Figure 17: Reflow only false report. GLM-5.3 Flash reports a text change on a pair whose reference edit set is empty.

Condition: Edit only, procurement, $k = 5$, GLM-4.6V (reasoning on).
Reference: table total $704.00 \rightarrow 7004.00$. Paragraph reimbursement $704.00 \rightarrow 70.00$. Insert ":" after "pump will be", "-" after "contingency", and "," after "each" in the closure heading.
Output fragments: location "`Paragraph under 'Authorized remedy'`".
Before "`producing a total reimbursement of USD 704.00.`"
After "`producing a total reimbursement of USD 7004.00.`"
Outcome: TP $= 2$, FP $= 1$, FN $= 3$ for the full answer. The quoted report assigns the table's after-value to the paragraph, whose reference after-value is 70.00.

Figure 18: Wrong occurrence. The reported number is the table edit's after-value but is attached to a distinct paragraph occurrence.

Condition: the same education edit set, $k = 1$, Gemini 3.6 Flash.
Reference in both layouts: "understandable" $\rightarrow$ "understnadable" in Reasons by dimension, Communication.
Edit only output fragments: before "`understandable`". After "`understnadable`".
Edit + reflow output fragments: before "`understandable`". After "`understandnable`".
Outcome: TP $= 1$, FP $= 0$, FN $= 0$ on Edit only, and TP $= 0$, FP $= 1$, FN $= 1$ on Edit + reflow. The reflowed answer reports a different string and misses the reference edit.

Figure 19: Reflow loss on an identical edit set. The Edit only answer is exact. After reflow, the answer has one false report and one missed edit.

Condition: Edit only, education, $k = 5$, GPT-5.6-sol.
Reference: insert ";" after "raise" and ":" after "communicated". Case change "scoring" → "scoRing". Total 100 → 10. Time 24 → 42 minutes.
Output fragments (before → after): "`100`" → "`10`", "`24 minutes`" → "`42 minutes`", "`feedback`" → "`eedback`", and "`scoring`" → "`scoRing`".
Outcome: TP = 3, FP = 1, FN = 2. The total, time, and case changes are found. Both punctuation insertions are missed, and the reported deletion in "feedback" is absent from the reference edits.

Figure 20: Partial recovery of a five-edit set. The answer recovers three edits, misses two, and adds one false report.