

Two Hands: a vision model as the second keyer of self-checking paper forms

Technical report. Every number in this document is generated from the committed results files by `scripts/build_report.py` and checked by `make eval` (the check fails if a number drifts). No model was called to write it.

Abstract

Paper forms that carry their own arithmetic (rows that must sum to a total, a total written in words and in digits) are still keyed by hand, often twice. We ask whether an open vision model, run on AMD GPUs, can act as the second keyer: it reads every field, the form's own identities decide whether the read is accepted, and every disagreement with the keyed record goes to a person with the crop of the paper. On a pre-registered, stratified random sample of 300 handwritten tally sheets (Mexico, 2024; official double-keyed capture as the comparison), Gemma-4-31B-IT read 289 pages and disagreed with the capture on 267 of 6,358 fields (4.2%). Settled on the photo, **52 of those 267 are fields where the official capture does not match the paper** (1 in 5; Wilson 95% 15.2% to 24.6%), against a capture error rate of at least 1 in 122 for a random field. The frozen gate accepted **55 of 300 forms (18.3%) with no person**, with **2 wrong fields in 1,210** (Wilson 95% 0.04% to 0.60%); a secondary two-reader tier (Gemma and Qwen3.8-27B must agree; defined in Amendment 2 after the one-reader result was seen and before the second reader ran; lighter gate, no section checks) accepted 89 forms with 0 wrong fields in 1,958 (full gate, section 7: 22 forms, 0 wrong in 484). Wrong is judged against the official capture; every disagreement was checked on the photo; fields where the model and the capture agree were not individually checked. Every gate variant we tried is reported in one table, pre-registered and post-hoc alike. Reading cost \$0.000144 per form.

1. The problem

A keyed record of a paper form is trusted because two people typed it. When they agree on a wrong number (or one of them marks a readable field "illegible"), nothing downstream notices. Checking keyed numbers against the paper by hand does not pay, so it is rarely done. Many forms, however, check themselves: their rows sum to a printed total, sections must equal each other, and every quantity is written twice, in words and in digits. A reader that transcribes both, independently, lets the paper referee the read.

The tool is clerical: it compares a keyed record with the image it was keyed from. It does not count, total or interpret anything; it reports, field by field, where the record and the paper disagree. The tally sheets are used because they are public, handwritten, self-checking and come with an official two-person capture to grade against, not for what they record. Vote rows are named by printed position only (Option 01 to Option 15).

2. Data and pre-registration

Population and sample. Type-2 tally sheets of the 2024 count with a PREP image (phone photo or scanner), unique id, excluding the 43 pilot forms. Stratified by state x capture type, proportional allocation, seed 20261015: 300 forms (196 phone photos, 104 scans). Manifest sha256 605af9775666... , schema sha256 c9240632c3f0... , frozen prompt sha256 3664253035b9... (checked on every call row by `make eval`).

Order of events (git commit times and the first recorded call, UTC):

Document	Commit	Committed	Governs	First governed call
PROTOCOL.md (sample, privacy, readers, gate, metrics, headline rule)	f85ca1f	2026-10-15 01:09:08Z	the 300	Gemma 2026-10-15 01:38:17Z
results/manifest.json (the 300)	f85ca1f	2026-10-15 01:09:08Z	the 300	Gemma 2026-10-15 01:38:17Z
Amendment 1: the page locator	ddfabe9	2026-10-15 01:29:19Z	preparation of the 300	first successful Gemma read 2026-10-15 05:56:11Z
Amendment 2: secondary analyses (two-reader tiers, adjudication of every disagreement)	56fe2e1	2026-10-15 10:36:46Z	Qwen pass, adjudication	Qwen 2026-10-15 11:53:45Z; first adjudication commit 2026-10-15 10:37:43Z
Amendment 3: gate v2 and a fresh 100-form confirmation draw	693451b	2026-10-15 13:36:57Z	the confirmation 100	2026-10-16 08:51:07Z

PROTOCOL.md is append-only (make `eval` checks that no commit deletes a line of it). Amendment 1 changed only the locator, before any successful read of the 300: the pre-registered locator left 52 of 300 pages unlocated (50 phone photos, 2 scans). The new locator was tuned on the pilot forms and a fresh calibration set (seed 20261016, $n = 150$, disjoint from the 300): 144 of 150 located, 143 correctly aligned by an eye audit of every page (Wilson 95% 90.7% to 97.7%). Amendment 2 was written after seeing the Gemma accept rate and before any Qwen read or adjudication. Amendment 3 defines gate v2 **after** seeing the reads and their adjudication, so every v2 number on the 300 is post-hoc; its test is the fresh 100-form draw, scored in section 4 (`results/confirm100.md`).

3. Method

Locate and mask. Each image is rectified onto the canonical page (corner markers, or a SIFT + RANSAC match against a masked reference when markers are cut off; ECC refinement); a warp is accepted only if 8 printed landmarks land within 1% of the page width. Otherwise the page is not located: nothing is stored or sent, and the form goes to a person. The right-hand column (names and signatures of officials and representatives) is filled black before the page is stored or sent; the raw download is deleted.

Read. One request per form: the masked page plus 6 zoomed tiles cut from it (tiles never reach into the mask). The prompt is built from the schema and frozen (sha256 3664253035b9...). For each of the 23 fields the model returns the printed row label it sees, the digits and the words, read independently. Gemma-4-31B-IT is the primary reader (temperature 0, max 2000 tokens); Qwen3.8-27B the second (reasoning effort low, max 6000 tokens). Both on AMD Token Factory (OpenAI-compatible endpoint, shared rate limiter).

Gate (frozen). A read is accepted only if: 23 values parse; every printed row label fits its position (a logo row fails if its label fits a neighbouring row better, the trace of a skipped or doubled row); every digit field is read; words = digits wherever words were read; and the four identities hold (vote rows sum to TOTAL; voters + representatives = section 5; section 5 = section 7; section 7 = TOTAL). Anything else is routed to a person with the failing fields named.

Adjudication. Every field where Gemma's digits differ from the official capture (267 fields) was settled on the crop of the masked page, with one verdict: capture error (paper = model), model misread (paper = capture), neither, paper ambiguous, unreadable. One adjudicator: the AI adjudicator (the builder's model, which could see both values; not a native speaker); crops are published so anyone can re-check. Truth = the capture unless the paper says otherwise.

4. Results

4.1 The primary: one reader, frozen gate

Measure	Value
Pages located and read	289 of 300; 11 not located go to a person before any model call
Fields compared with the official capture (22 per form)	6,358
Agreement	6,091 (95.8%)
Disputed fields	267 on 104 forms
Verdicts on the disputed	capture error 52, model misread 156, neither 17, paper ambiguous 34, unreadable 8
Capture errors found	52 on 26 forms
Yield of a disputed field	52 of 267 = 1 in 5 (19.5%; Wilson 95% 15.2% to 24.6%)
Capture error rate, lower bound	52 of 6,358 = at least 1 in 122 (Wilson 95% 0.62% to 1.07%)
Model wrong, alone	173 of 6,358 (1 in 37; Wilson 95% 2.35% to 3.15%)
Accepted with no person	55 of 300 (18.3%; Wilson 95% 14.4% to 23.1%): 27 phone photos, 28 scans
Accepted fields wrong on the paper	2 of 1,210 (Wilson 95% 0.04% to 0.60%)

The headline rule (pre-registered): single-reader Gemma is the headline unless its accepted-field error rate exceeds 0.5%. It is 0.17%, so the single reader stays the primary. On the accepted forms, 3 fields differ from the capture: 1 is a capture error, 2 are model misreads.

4.2 Every gate variant, in one table

Accepted forms out of 300 (unlocated pages and failed reads count in the denominator). Wrong = accepted fields where the accepted value differs from the paper (truth: the verdict on the photo where one exists, else the capture). Undecided = an accepted field the photo does not settle. **Post-hoc rows were designed after seeing these reads and test nothing**; they are shown so that no variant is hidden.

Gate	Status	Accepted (/300)	Phone / scan	Accepted fields	Wrong	Wrong, Wilson 95%	Undecided
Gemma, frozen gate (v1)	pre-registered, primary	55 (18.3%)	27 / 28	1,210	2	0.04% to 0.60%	0
Qwen, frozen gate	pre-registered (section 7)	31 (10.3%)	8 / 23	682	1	0.03% to 0.83%	2
Both readers, both frozen gates, 23 digits agree	pre-registered (section 7)	22 (7.3%)	6 / 16	484	0	0.00% to 0.79%	0
Gemma, capture gate (words = digits, rows = TOTAL)	pre-registered secondary (Amendment 2 c)	143 (47.7%)	87 / 56	3,146	14	0.27% to 0.75%	6
Two readers, strict mode	pre-registered secondary (Amendment 2 d)	89 (29.7%)	51 / 38	1,958	0	0.00% to 0.20%	0
Gate v2	POST-HOC (Amendment 3)	47 (15.7%)	24 / 23	1,034	1	0.02% to 0.55%	0
Gate v2b (relaxed logo-row label)	POST-HOC (Amendment 3, secondary)	62 (20.7%)	39 / 23	1,364	1	0.01% to 0.41%	0
Gate v2-2R (unused ballots confirmed by words or Qwen)	POST-HOC (Amendment 3)	54 (18.0%)	27 / 27	1,188	1	0.01% to 0.47%	0
Gate v2-2R strict (unused ballots only if Qwen agrees)	POST-HOC (in no amendment)	45 (15.0%)	19 / 26	990	0	0.00% to 0.39%	0

Confirmation on fresh data (Amendment 3, pre-registered before any read): 100 forms drawn after the fix was designed, Gemma on the AMD Token Factory. The fixed gate (v2) accepted 22 forms; 0 of their 484 fields were wrong against the official capture, with 3 disagreements adjudicated from crops and checked by a person. The pre-registered criterion was an error rate under 0.5%; the point estimate meets it, and the sample is too small to exclude 0.5% at 95% confidence (Wilson upper bound 0.79%). Evidence for v2 across both sets: 1 wrong field in 1,518 accepted, of which 1,034 are a post-hoc re-score of the 300 and 484 are this pre-registered run. (results/confirm100.md)

On the same 100 forms the original gate would have accepted one wrong field: the model read 3114 where the paper says 314, a digit inserted. The v2 rule "a field in no identity needs its digits and words both read and equal" (Amendment 3, condition 2(a); here the words line was blank) blocked it. That is the held-out test of a rule designed after the first run, and it passed.

Gate v3. v3 on the 100: 0 wrong of 504 accepted; with no second reader in this run every SOBRANTES went to a person, so the full v3 test waits for the 3,000 with Qwen.

The two-reader strict mode drops the row-label check and the section identities and relies on the readers' agreement instead, which is why it accepts more forms than either reader alone. The two readers agree on all 23 digits on 147 forms; on the strict tier's accepted fields the agreed read differs from the capture on 3, and 3 of them are capture errors on the photo. The capture gate (Amendment 2 c) is Gemma alone with only words = digits and rows = TOTAL; it is first tabulated in this report. It accepts 143 forms but with 14 wrong fields (0.45%), close to the 0.5% line: without the section identities and labels one reader is not enough. Its accepted fields include 0 where the reader differs from the capture with no verdict (never counted as right). Gate v2 adds one rule to the frozen gate: the field that sits in no identity (unused ballots) must have words and digits

both read and equal; v2-2R also accepts it when Qwen's digits agree; v2-2R strict accepts it only when Qwen agrees.

Route causes of the 234 read-but-routed forms (a form can have several):

Cause	Forms (any cause)	Forms (only cause)
an identity fails	166	64
row label mismatch	86	15
a digit field not read	76	25
words differ from digits	63	7
wrong number of values	1	0

4.3 The capture errors

Type	Fields
Wrong number typed	17
Readable field recorded as illegible (23 on two faint pencil forms)	27
Number typed where the paper is blank	6
Other (readable field captured as "no data")	2
Total	52

A wrong number is typed in about 1 field in 374. **Certified confirmation.** The later certified count covers 31 of the 52 fields; it equals the paper (and the model) on 22, and differs on 9, all 9 of which were recounted (a recount counts ballots, not the sheet, so it need not match the sheet). Example: on form 050815B0100, Option 07, the paper says 175 in digits and words, the first capture keyed 75, the model read 175 and the certified count (by Cotejo) is 175.

4.4 Recall: one state, stated plainly

We only look where the model disagrees with the capture, so an error the model shares with the capture is not found. **Recall on the capture errors the certified count reveals, Colima (pre-registered, a person's blind check): 27 of 28 caught, 96% (Wilson 95% 82-99%). One state, certified-detectable errors only; not total recall.** (PROTOCOL.md Amendment 9, results/colima/recall.md: every field where the certified count differs from the capture and the model agreed with the capture was read blind by a person.)

Earlier, on the random 300: to estimate recall we used the 100 read forms the certified count settled by comparison (no recount): there the certified count differs from the first capture on 23 fields. Judged on the paper, 16 of them are council corrections or blank rows (the paper shows the capture's value; a faithful read cannot dispute them). **The model disputed 7 of the remaining 7:** measured recall 100%, Wilson 95% 65% to 100%. **n = 7**, and only 1 of them is a confirmed capture error (the rest are ambiguous on the paper): too few to measure recall on confirmed capture errors. Counting the corrections in the denominator gives the loose bound 7 of 23 (16% to 51%).

4.5 Gemma vs Qwen

Same 300 forms, same masked page and tiles, same frozen prompt and gate.

	Gemma-4-31B-IT	Qwen3.8-27B
Sent / parsed	289 / 289	289 / 239
finish_reason	stop 289	length 50, stop 239
Accepted, frozen gate	55 (18.3%)	31 (10.3%)
Accepted fields wrong	2 of 1,210 (0.04% to 0.60%)	1 of 682 (0.03% to 0.83%); 2 undecided
Mean latency per form	87.0 s	241.0 s
Tokens in / out (mean)	2,609 / 804	6,612 / 3,340
Cost per form read	\$0.000144	\$0.000597
Spend per form, all attempts	\$0.000139	\$0.000641

Gemma costs about 24% of Qwen per read and is faster; Qwen's 50 replies cut at the token cap are never accepted. Neither reader is accurate enough alone to skip the gate: the value of the second reader is agreement, not accuracy.

4.6 Cost and latency

Gemma: \$0.000144 per form read (mean over 289 reads), about \$0.14 per 1,000 forms; mean latency 87.0 s per form (wall clock on a shared endpoint, including queueing); 2,609 tokens in and 804 out per form; all 289 replies finished with stop. The run needed re-reads on 51 forms after an endpoint outage (355 call rows in all; the last row per form counts).

5. Failure analysis

One call can harmonise its own words and digits. On form 010469C0400, unused ballots, Gemma read 239 in digits and 239 in words; the paper says 234 in both. The field sits in no identity, so words = digits was the only check, and one call satisfied it with itself: the form was accepted with a wrong field (one of the 2). Words and digits from one call are not two independent reads. Gate v2 (words required) does not remove it; only a second reader does (v2-2R strict, 0 wrong). Both readers sharing a misread remains possible and is not counted.

The row-label gate's false alarms. The label check routed 86 forms; on 68 of them the read matched the paper on every graded field, including all 15 of the 15 forms routed for a label alone. Both models misname small logos while reading the values right. Each false alarm costs a person one look; a false accept of a shifted row would cost a wrong record, so the check stays in the frozen gate (v2b relaxes it, post-hoc).

Pages not located. 11 of 300 pages (11 phone photos, 0 scans) could not be located even with the Amendment 1 locator; they go to a person before any model call and count as routed. The pre-registered locator would have left 52.

What the paper's arithmetic can and cannot catch

The guarantee. The tally sheet is redundant twice over: each value is written in digits and in words, and the values are tied by 4 identities in the schema (the 17 vote rows sum to TOTAL; voters + representatives = section 5 = section 7 = TOTAL). **On a paper whose own arithmetic holds, any single misread is caught:** a value misread in its digits fails words = digits when both are written, and a misread value inside an identity breaks that sum. Either way the form routes to a person; nothing is accepted on a broken identity.

Only an error that preserves every identity passes. Three ways:

1. **Compensating errors.** Two misreads that keep a sum true together: +k and -k on two terms of one sum (two vote rows, or voters and representatives), with no words or words misread the same way; or

one misread carried through every side of a chain of identities. A shifted row is a case of the first: values moved inside one sum.

2. **A field outside every identity.** Computed from the schema (`twohands.gate_v2.unchecked_fields` over `schemas/mx_presidencia_2024.yaml`): SOBRANTES (unused ballots), the same in the Senate schema. Its only check is words = digits, and there is none when the words are not read.
3. **A harmonised read.** Words and digits come from the same model call, so a call that misreads the digits can write words that match. One call is one read, not two.

The guarantee assumes the paper's own sums hold: on a sheet whose arithmetic is off, a misread that happens to repair it passes too.

Measured blind spots. Every wrong field on a form the frozen gate accepted, for both readers (Gemma 2 of 1,210 accepted fields, Qwen 1 of 682); 3 of the 3 sit in the field outside every identity:

Reader	Form	Field	Digits	Words	Paper	Way through
Gemma	010469C0400	SOBRANTES	239	239	234	words = digits, harmonised in one call
Gemma	212668C0300	SOBRANTES	300	not read	379	words not read: no check at all
Qwen	090955C0200	SOBRANTES	200	not read	206	words not read: no check at all

Of the 3: 1 harmonised in one call (010469C0400), 2 with no words read at all (212668C0300, 090955C0200). No compensating error was found among the accepted forms: that describes this sample, not a bound.

Held out on the fresh 100 (pre-registered, Amendment 3): the frozen gate let one more through, a digit insertion (3114 for 314) on 070212B0100, unused ballots with the words line blank. The v2 rule blocked it. The measured kinds so far: harmonised in one call (239/239 for 234); blank words (300 for 379); digit insertion (3114 for 314).

What closes each one (post-hoc on the 300, Amendment 3; the fresh draw is the test):

- **The v2 rule:** a field outside every identity must have digits and words both read and equal. It removes the no-words cases: 1 wrong of 1,034 accepted fields. It cannot remove a harmonised read.
- **A second reader:** v2-2R strict accepts SOBRANTES only when Qwen's digits agree with Gemma's: 0 wrong of
- v2-2R, which also accepts Gemma's own words, keeps the harmonised read: 1 wrong.
- **Separate reads:** digits and words read in separate calls would make words = digits two reads instead of one; not built, not measured.
- **Compensating pairs and shifts:** the row-label check and the tiles make a shift visible; nothing makes a compensating pair visible except a second reader that does not make the same pair. Two readers sharing one misread remain possible and are not counted.

Stated as a guarantee: a single misread cannot pass the gate; an error passes only through the three named ways, and on this sample every one that did went through the field outside every identity.

6. Limits

- One sample of 300 forms of one form type; the gate v2 confirmation is 100 forms read by one reader (no Qwen pass).
- One adjudicator, the AI adjudicator (the builder's model, which could see both values); crops are published for re-checking.
- Recall is measured on one state's certified-detectable capture errors (27 of 28, Amendment 9); errors the certified count cannot reveal are not measured, so this is not total recall.

- The two-reader strict tier checks words = digits and row sums only, without the section identities or labels.
- Latency and throughput were measured on a shared, rate-limited endpoint.
- The capture compared against is the preliminary first capture; the certified count is the official record.

7. Ethics and privacy

- **Masking first.** The column with names and signatures is blacked out before a page is stored or sent; raw downloads are deleted; tiles never reach the mask (tested). A page that cannot be located is neither stored nor sent.
- **Clerical framing.** The tool compares a keyed record with its paper. It produces no totals, no aggregation across forms and no comparison between options; vote rows are shown as Option 01 to Option 15, never by name. The app's policy tests enforce this in code.
- **No claim beyond the paper.** A disputed field is a lead, not a finding, until a person confirms it on the photo; differences with the certified count are reported as recounts, not as errors of anyone.

8. Reproducibility

```
make eval      # python -m twohands.eval; python scripts/gemma_vs_qwen.py; python
scripts/build_report.py --check
make test      # python -m pytest -q
python scripts/build_report.py --pdf      # this report, docs/REPORT.md and report/two-hands-
report.pdf
```

`make eval` recomputes every number on the site from the committed calls files, verdicts, manifests and git, compares each with what the app shows (0 mismatches required), checks the manifest, schema and prompt hashes against PROTOCOL.md and the commit order above, and rebuilds this report to check it is current. Integrity checks (stored gate flag = frozen gate re-run, verdicts cover exactly the disputed fields): all hold. No model call, no network, no raw images.