QCRCE Accuracy Audit Report

Batch: `samples` (10 documents). Generated 2026-10-31 10:12:16 CST(UTC+0800). Preregistration: `PREREGISTRATION.md` (frozen before run B). **The samples are synthetic** (generated documents with fictional companies); real client layouts will score differently.

> **Read first:** every number below is from an in-sample run on synthetic documents. It shows how the cross-check workflow behaves; it is not a production accuracy guarantee. Accuracy on your documents is PROPOSED until measured on a reviewed sample of them.

Evidence labels. OBSERVED = measured in one run. REPLICATED = two independent extractors (A and B, different methods/vendors) produced the same value on the same document; this is method-level replication, not replication on a new dataset. PROPOSED = not yet tested.

Method

- Run A: QCRCE pipeline (`pipeline.py`): text-layer parsing for digital PDFs, local OCR for scans, rule-based field mapping, arithmetic cross-checks. No LLM.
- Run B: independent vision-LLM extraction (`google/gemini-2.5-flash`, temperature 0) from the page image only; B never sees A's output.
- Every field is compared. A==B → auto-accept candidate. A≠B → human review queue. A document is auto-accepted only if all its fields agree **and** A's arithmetic checks pass.

Headline results

- Field-level A/B agreement (auto-accept candidates): **225/251 (89.6%)** [OBSERVED]
- Documents fully auto-accepted: **7/10** [OBSERVED]; the rest go to review.
- Fields where A and B agree are REPLICATED values. Of these, wrong vs ground truth: **0/225 (0.0%)** [OBSERVED, checked against truth]
- Agreement is judged ignoring whitespace. 18 agreed fields differ only by spaces OCR dropped in A (e.g. 'Coppercable2.5mm100m'); the auto-accepted value in `accepted_fields.csv` uses B's text with spacing. [OBSERVED]
- A's field errors caught by the A/B disagreement queue: **26/26 (100.0%)** [OBSERVED]

Accuracy vs ground truth (field level) [OBSERVED, single run per extractor]

subset	fields	A strict	A ws-insens.	B strict	B ws-insens.	A==B
all	251	207/251 (82.5%)	225/251 (89.6%)	251/251 (100.0%)	251/251 (100.0%)	225/251 (89.6%)
digital	116	116/116 (100.0%)	116/116 (100.0%)	116/116 (100.0%)	116/116 (100.0%)	116/116 (100.0%)
scan	135	91/135 (67.4%)	109/135 (80.7%)	135/135 (100.0%)	135/135 (100.0%)	109/135 (80.7%)

subset	fields	A strict	A ws-insens.	B strict	B ws-insens.	A==B
invoice	127	117/127 (92.1%)	126/127 (99.2%)	127/127 (100.0%)	127/127 (100.0%)	126/127 (99.2%)
delivery_note	52	34/52 (65.4%)	37/52 (71.2%)	52/52 (100.0%)	52/52 (100.0%)	37/52 (71.2%)
packing_list	42	29/42 (69.0%)	32/42 (76.2%)	42/42 (100.0%)	42/42 (100.0%)	32/42 (76.2%)
bill_of_lading	30	27/30 (90.0%)	30/30 (100.0%)	30/30 (100.0%)	30/30 (100.0%)	30/30 (100.0%)

Of 26 disagreeing fields: A correct in 0, B correct in 26 (both can be wrong). [OBSERVED]

Note on B: a perfect score here most likely reflects clean synthetic text that is easy for a modern vision model, not a ceiling for real documents. B was given the field names (the target schema) but no values. Ground truth is produced by construction: the generator writes each document and its truth JSON from the same data, so no human labelling was involved.

Preregistered bars

Verdicts are in-sample and indicative only. The 95% lower bound is a Wilson interval on the field or error count; fields within a document are correlated, so the true uncertainty is larger.

hypothesis	bar	result	95% lower bound	verdict
H1 A digital accuracy	>=99%	116/116 = 100.0%	96.8%	PASS on point estimate only; interval crosses bar
H2 B scan accuracy (ws)	>=90%	135/135 = 100.0%	97.2%	PASS (in-sample)
H3 error rate in agreed fields	<=1%	0/225 = 0.0%	upper 1.7%	PASS on point estimate only; interval crosses bar
H4 A errors flagged by disagreement	>=90%	26/26 = 100.0%	87.1%	PASS on point estimate only; interval crosses bar
H5 auto-accept rate	>=60%	225/251 = 89.6%	85.3%	PASS (in-sample)
H6 audit spend (OpenRouter key usage delta)	<= USD 3	USD 0.0337	n/a	PASS

H1 is in-sample: A's rules were written on these same documents.

Per document

doc	type	kind	A==B fields	A checks pass	decision
01_invoice_digital	invoice	digital	33/33	True	auto-accept
02_invoice_digital_alt	invoice	digital	23/23	True	auto-accept
03_commercial_invoice_scan	invoice	scan	33/33	True	auto-accept
04_delivery_note_digital	delivery_note	digital	22/22	True	auto-accept
05_delivery_order_scan	delivery_note	scan	15/30	False	human review
06_packing_list_digital	packing_list	digital	21/21	True	auto-accept
07_packing_list_scan	packing_list	scan	11/21	False	human review
08_bill_of_lading_digital	bill_of_lading	digital	17/17	True	auto-accept
09_bill_of_lading_scan	bill_of_lading	scan	13/13	True	auto-accept
10_invoice_scan_heavy	invoice	scan	37/38	False	human review

Field-by-field disagreements: `agreement_table.md`. Reproducibility: `MANIFEST.md`.

In this batch, 25 of 26 disagreements were line-item cells, mostly downstream of a single row that A's OCR dropped (row shift), not independent errors.

Limitations

- Small n: results are indicative, not statistically meaningful.
- Agreement is not proof of correctness: two extractors can make the same mistake; on client data without ground truth, the agreed-field error rate is PROPOSED until a reviewed sample measures it.
- Cost and speed on client volumes are PROPOSED; this run's spend covers {len(docs)} documents.
- B is a third-party hosted model; its results can change if the provider updates the model (served model IDs are logged in MANIFEST.md / run_b/).
- Line items are aligned by row index; one dropped row shifts all later rows into disagreement (conservative: more review, not less).
- Synthetic samples; the same team wrote the generator and the rules. Run A was developed on these same documents, so A's numbers are optimistic.