

Report 2: Baseline After Fixes

Aliyan Faisal · same invoices, same scoring as Report 1

1. Summary

Report 1 measured the invoice checker as it stood and listed six reading problems. For this report I fixed those problems in the extractor and scored the same invoices in the same way, so the numbers can be compared directly. The two sets are the 30 independent invoices (corpus A) and the 40 built-in generator invoices (corpus B). The *Evaluation Harness Guide* explains how the scoring works.

Rules extractor (the real one)	Report 1 (baseline)	Report 2 (after fixes)
Corpus A field accuracy (30 independent invoices)	84.5%	98.1%
Corpus A planted mistakes found	25 of 32	31 of 32
Corpus A false alarms	30	6
Corpus A correct invoices with no findings	3 of 15	12 of 15
Corpus B field accuracy (40 generator invoices)	99.9%	99.9%
Corpus B planted mistakes found	20 of 20	20 of 20
Corpus B false alarms	0	0

Result. Corpus A field accuracy rose from 84.5% to 98.1%, planted mistakes found from 25 to 31 of 32, and false alarms fell from 30 to 6. All six problems from Report 1 improved. Corpus B did not change (99.9% in both reports, one tax rate still missing on invoice-031), which is expected because the generator invoices were already read well.

Read the corpus A gain with care. I tuned the extractor against these same 30 invoices, so the new score is higher than a fresh set of invoices would give. Four problems remain on corpus A (section 6); they are listed as findings and were not fixed in this round.

2. What was run

The same invoices, the same answer keys and the same scoring as Report 1, using the rules extractor only. The oracle and broken controls do not use the extractor, so they were not run again. The rules-ocr extractor (scanned PDFs) was not run again for this report, so its Report 1 figures are not updated here. The raw results for both reports can be downloaded from the Harness Reports page of this site.

3. What was fixed

Problem found in Report 1	What changed in the extractor	Result on corpus A
Tax IDs that are not VAT-style were not recognised	Tax ID lines are now recognised by their local label: ABN, UID, GSTIN, GST/HST, EIN, NTN, CIF, NIF, BN and similar, as well as VAT.	Tax ID read: 16 -> 30 of 30. False "tax ID not detected" warnings: 14 -> 0
Line items of the receipt and quantity-first layouts were not extracted	A line may now have its description between or after the numbers (quantity first). A "10 x 55.75" run is split into quantity and unit price, and the description is taken from the row above when it sits alone.	Line quantity, unit price and amount: 138 -> 168 of 168. Quantity-first layout 37.0% -> 100%. Receipt layout 38.5% -> 100%
Some invoice-number formats and labels were missed	Added Factura and Fattura labels, "n°", "Invoice FA-26-0140" with no "No.", a "#" kept as part of the number, and a number printed under its header. A date is no longer accepted as a number.	Invoice number read: 22 -> 30 of 30
Customer name sometimes taken from a label such as "Date"	When the neighbour beside "Bill to" is another label or a date, the name below it is used. Added Cliente and Kunde.	Customer name read: 22 -> 29 of 30

Problem found in Report 1	What changed in the extractor	Result on corpus A
Currency decided from "\$" when the invoice says AUD or CAD	A statement such as "All amounts in AUD" now outweighs a bare "\$".	Currency read: 25 -> 30 of 30
A line or total row mistaken for shipping	Mostly removed as a side effect of reading the line items correctly.	Spurious shipping amounts: 7 -> 3
(Not in Report 1) Supplier missing on right-aligned letterheads	If no name is found in the left half of the page, the whole page width is searched.	Supplier name read: 27 -> 30 of 30

The checking rules themselves were not changed.

4. Results on the independent invoices

Fields read right, before and after. Green is 95% or better, amber 90 to 95%, red below 90%:

Field	Report 1	Report 2	Change
Line description	121/168	151/168	+30
Line quantity	138/168	168/168	+30
Line unit price	138/168	168/168	+30
Line amount	138/168	168/168	+30
Tax ID	16/30	30/30	+14
Invoice number	22/30	30/30	+8
Customer name	22/30	29/30	+7
Currency	25/30	30/30	+5
Shipping	23/30	27/30	+4
Supplier name	27/30	30/30	+3
Issue date	29/30	29/30	+0
Due date	29/30	29/30	+0
Subtotal	30/30	30/30	+0
Discount	30/30	30/30	+0
Tax rate	29/30	29/30	+0
Tax amount	29/30	29/30	+0
Total	29/30	29/30	+0

Page structure (3 invoices each)	Report 1	Report 2
quantity first	37.0%	100.0%
receipt	38.5%	100.0%
tax summary	86.2%	88.2%
compact	90.1%	95.4%
cards	91.8%	97.4%
letterhead	94.9%	100.0%
sidebar	95.4%	96.5%
banner	97.0%	100.0%
letter	99.4%	100.0%
multi-page	99.5%	100.0%

Planted mistake type	Planted	Found in Report 1	Found in Report 2
duplicate charge	6	4 of 6	6 of 6
grand total	3	2 of 3	3 of 3
line total	9	6 of 9	9 of 9
missing tax ID	2	2 of 2	2 of 2
subtotal	7	7 of 7	7 of 7
tax	5	4 of 5	4 of 5

All planted line-total and duplicate-charge mistakes are now found; the one still missed is a tax mistake on invoice-028, where the tax row is not found (section 6).

5. Result for each independent invoice

Found = planted mistakes the tool reported. Extra = reports that match no planted mistake (false alarms). Report 1 extras are shown for comparison.

Invoice	Structure	Planted mistakes	Found	Extra (Report 2)	Extra (Report 1)
invoice-001	letter	none (correct invoice)	-	-	-
invoice-002	banner	tax	1/1	-	missing tax ID
invoice-003	sidebar	none (correct invoice)	-	-	-
invoice-004	receipt	missing tax ID, grand total, line total (line 2)	3/3	-	-
invoice-005	tax summary	none (correct invoice)	-	grand total	grand total, missing tax ID
invoice-006	quantity first	duplicate charge (line 6), line total (line 4)	2/2	-	missing invoice number
invoice-007	letterhead	none (correct invoice)	-	-	missing supplier, missing tax ID
invoice-008	compact	duplicate charge (line 6), missing tax ID, grand total	3/3	-	missing invoice number
invoice-009	cards	none (correct invoice)	-	-	missing invoice number, missing tax ID
invoice-010	multi-page	duplicate charge (line 17)	1/1	-	-
invoice-011	letter	subtotal, tax, line total (line 4), line total (line 5)	4/4	-	-
invoice-012	banner	none (correct invoice)	-	-	-
invoice-013	sidebar	subtotal	1/1	-	missing tax ID
invoice-014	receipt	none (correct invoice)	-	-	missing tax ID
invoice-015	tax summary	subtotal, line total (line 5), line total (line 2)	3/3	grand total	grand total
invoice-016	quantity first	none (correct invoice)	-	-	missing tax ID
invoice-017	letterhead	grand total, tax, subtotal	3/3	-	missing supplier
invoice-018	compact	none (correct invoice)	-	-	missing invoice number, missing tax ID
invoice-019	cards	duplicate charge (line 4), subtotal	2/2	-	missing invoice number
invoice-020	multi-page	none (correct invoice)	-	-	missing tax ID
invoice-021	letter	none (correct invoice)	-	-	missing tax ID

Invoice	Structure	Planted mistakes	Found	Extra (Report 2)	Extra (Report 1)
invoice-022	banner	tax, subtotal, line total (line 4)	3/3	-	missing invoice number
invoice-023	sidebar	none (correct invoice)	-	-	missing tax ID
invoice-024	receipt	duplicate charge (line 7)	1/1	-	-
invoice-025	tax summary	none (correct invoice)	-	grand total	grand total, missing tax ID
invoice-026	quantity first	line total (line 2)	1/1	-	-
invoice-027	letterhead	none (correct invoice)	-	-	missing supplier
invoice-028	compact	tax	0/1	grand total	grand total, missing invoice number
invoice-029	cards	none (correct invoice)	-	missing issue date, missing due date	missing issue date, missing due date, missing tax ID
invoice-030	multi-page	duplicate charge (line 17), line total (line 4), subtotal	3/3	-	missing tax ID

12 of 15 correct invoices now produce no findings (3 in Report 1); 31 of 32 planted mistakes were found (25 in Report 1).

6. New problems found in this run

These are the 26 fields still read wrongly or missing on corpus A, grouped by cause. They were found by this run and are not fixed in this report.

Problem	Evidence (independent invoices)	Effect
Line description picks up text from another column: an item code, or a customer name or address printed beside the line	17 wrong descriptions. Tax-summary layout (invoices 005, 015, 025): the item code is prefixed, e.g. "A011 Jahresabschluss" instead of "Jahresabschluss". Sidebar layout (003, 013, 023): sidebar text is prefixed, e.g. "Kestrel Analytics Inc. Implementation services, hour".	Description wrong; amounts and quantities are right, so no false alarm.
Tax-summary table read as a shipping amount; once also as the total	A spurious shipping amount on invoices 005 (1,476.99) and 025 (79.63). On invoice-015 the grand total is read as 20,265.08, expected 20,331.05.	A false "grand total" mismatch on 005, 015 and 025.
Compact layout: the tax row is not found	Invoice-028: tax rate and tax amount missing, and the tax amount (261.48) is read as shipping.	The planted tax mistake is missed, and a false "grand total" mismatch appears.
Cards layout: a date read as the customer name, and both dates lost	Invoice-029: customer read as "April 7, 2026"; issue and due dates missing.	False "missing issue date" and "missing due date" warnings.

False alarms on corpus A went from 30 to 6: four false "grand total" mismatches (invoices 005, 015, 025, 028) and the two missing-date warnings on invoice-029. All six come from the problems above, so fixing them should remove these false alarms too.

Conclusion

Fixing the six reading problems raised accuracy on the independent invoices from 84.5% to 98.1%. Four problems remain and are listed in section 6.

The invoices are synthetic and few (30 and 40), so read the numbers as a comparison and a list of things to fix, not exact real-world accuracy. The extractor was tuned against the 30 independent invoices, so the corpus A score is higher than it would be on invoices the extractor has not seen; a final accuracy figure needs a fresh set. Corpus B did not change. Scanned PDFs (OCR) were not re-measured in this report.