

Report 7: Fresh Test, 50 Invoices

Aliyan Faisal · a fresh set of test invoices: the rules extractor and the vision model on layouts, countries and languages new to both

This is a fresh test. Nothing in the extractor was changed for these 50 invoices before they were scored, and the instruction given to the vision model was written before they existed. Any improvements made after this report and scored on the same invoices are not a fresh measure.

1. Summary

Corpus E has 50 invoices: 20 from the project's own generator (10 single-page and 10 multi-page) and 30 written by me in ten page structures that appear in none of the earlier sets, in ten countries and languages that are new (Germany, Poland, Czechia, Turkey, Finland, Denmark, Norway, India, China and Switzerland). Half have planted mistakes. The rules extractor was scored on all 50. The vision model, which costs money for every invoice, was run on the 30 written by me.

	Rules extractor	Vision extractor
Field accuracy, 20 generator invoices	100.0%	not run on these; see below
Field accuracy, the 30 new-layout invoices	70.3%	97.9%
Field accuracy, all 50 invoices	92.2%	not run
Planted mistakes found, the 30 new-layout invoices	14 of 27	27 of 27
False alarms, the 30 new-layout invoices	53	8

Result. The rules extractor read the 20 generator invoices without a single wrong field. On the 30 invoices in new layouts and languages it scored 70.3%, found 14 of 27 planted mistakes and raised 53 false alarms. The vision model scored 97.9% on the same 30, found all 27 planted mistakes and raised 8 false alarms. The vision model was not run again on the 20 generator invoices, because it had already been run on 20 invoices of the same kind in Report 5: 98.0% on 10 single-page invoices and 85.9% on 10 multi-page invoices, at the same resolution and with the same instruction.

2. The invoices

Group	Invoices	What they test
Generator, single page	10	Three designs the extractor was built on, with new content. Five correct, five with planted mistakes.
Generator, multi-page	10	Wrapped lines, sections, carried-forward rows. Five correct, five with planted mistakes.
Written by me, ten structures	30	Mixed tax rates, tax over two rows, dated rows, credit lines, a pro forma with a deposit, a wide table with item codes, a tear-off payment slip that repeats the amounts, dot-leader statements, a ruled grid, and totals on page one with the items on page two.

The 30 written by me use number formats the earlier sets did not: lakh grouping (1,23,456.00), the Swiss apostrophe (1'234.50) and amounts followed by the signs for the zloty, koruna, lira or krone (zl, Kc, TL, kr.) or preceded by the yen or rupee sign. I fixed the layouts, countries and formats before running anything. Chinese names and descriptions are not scored, because the comparison ignores non-Latin letters.

3. Rules extractor on the 30 new-layout invoices

Page structure	Rules extractor	Vision extractor
dot leaders, no columns	11.2%	91.4%
wide table, item codes	56.1%	100.0%

Page structure	Rules extractor	Vision extractor
pro forma with deposit	70.0%	99.4%
totals before items	72.7%	95.2%
tax over two rows	76.5%	96.4%
dated rows	78.1%	99.5%
credit lines	83.4%	99.5%
mixed tax rates	84.0%	99.5%
ruled grid	86.7%	100.0%
tear-off payment slip	86.7%	97.7%

Field	Rules extractor	Vision extractor
Tax ID	9/30	26/30
Customer name	9/30	30/30
Tax amount	9/30	30/30
Subtotal	11/30	30/30
Tax rate	11/30	27/30
Total	11/30	30/30
Currency	16/30	28/30
Issue date	17/30	29/30
Due date	17/30	30/30
Invoice number	19/30	30/30
Shipping	22/30	30/30
Line description	121/152	132/152
Line quantity	127/152	152/152
Line unit price	127/152	151/152
Line amount	131/152	151/152
Discount	26/30	30/30
Supplier name	30/30	30/30

4. Problems found

Rules extractor

Problem	Evidence	Effect
Amounts with a currency word or symbol it does not know	14 of the 15 invoices with no total are in currencies printed with the zloty, koruna, lira, krone, yen or rupee sign (the total is missing on 14 of 18 invoices in those currencies).	Subtotal, tax, total and currency missing.
Local words for labels	Tax ID not found on 21 of 30 invoices, customer on 21, invoice number on 11, issue and due dates on 11 each. The labels are in languages and wordings the extractor has no words for.	Fields missing, which also triggers false alarms.
Rows with no columns (dot leaders)	Accuracy 11.2%: description, price and amount are one run of text with dots between them.	Lines and totals missing.

Problem	Evidence	Effect
Totals before the items, and a wide table with extra columns	Totals-before-items 72.7%, wide table 56.1%.	Totals missing, line amounts mixed with item codes.
Amounts taken for the wrong kind of row	On an invoice with the tax over two rows the tax rate and tax amount are wrong; on a pro forma a discount and shipping appear that are not printed; shipping appears on a payment-slip invoice.	Wrong or spurious fields.

Vision extractor

Problem	Evidence	Effect
A note kept in the description	On the three dot-leader invoices (Polish, Czech, Turkish) the quantity and price in brackets, such as "(7 x 1 232,85)", stayed in the description (16 lines). On two Chinese invoices the roman numeral II was read as two bars (4 lines).	20 wrong descriptions of 152.
Tax ID changed	Two Swiss tax IDs returned without the trailing "MWST"; two long IDs had one character dropped or changed.	4 of 30 tax IDs wrong.
Tax rate halved on a split tax	On all three invoices where the tax is printed over two rows the model returned half the rate (for example 12.5 for 25).	3 of 30 tax rates wrong.
Currency symbol returned	The Czech koruna sign was returned instead of the code CZK on two invoices.	2 of 30 currencies wrong.
A date and a digit	One issue date read with the wrong month; one unit price off by ten cents on a seven-digit number.	One date, one line wrong.

5. Cost and speed

The vision run on 30 invoices cost about 15 cents and took several minutes. The rules extractor scored all 50 in about ten seconds at no cost. The page pictures are sent to the model provider.

Conclusion

On layouts and languages that neither reader had seen, the vision model read 97.9% of the fields correctly and found every planted mistake, and the rules extractor read 70.3% and found 14 of 27. The rules extractor was perfect on the invoices from the generator it was built against and weak where the invoice used words, symbols or structures it had no rule for. Most of its failures are vocabulary (currency symbols and label words), not structure, which is the part that can be fixed in a general way.

Read the numbers with care. The invoices are synthetic and I wrote 30 of the 50 myself, as I wrote the extractor, so they are not an independent outside sample. Each layout and country has only three invoices. The vision model was run once on 30 invoices and its answers vary slightly between runs, which is why they were saved. Chinese text is not scored. The harness trusts whatever an extractor returns.

The invoices, their answer keys and the raw results for this report can be downloaded from the Harness Reports page of this site.