

Report 5: Fresh Invoice Test, Rules and Vision Model

Aliyan Faisal · a new set of invoices neither reader has seen, and a reader that is a model

This report adds a second kind of extractor. The *rules* extractor reads the text of the PDF with hand-written rules. The new *vision* extractor shows a picture of each page to an AI model and asks it to read the values. Both were scored on a new set of 30 invoices (corpus D) that was written after every fix in Report 4 and that neither has seen. The rules extractor was not changed for this report.

1. Summary

Reports 3 and 4 showed that the rules extractor does well on the layouts it was built beside and fixed against, and that every fix was made on invoices it was then scored on. A fresh set is the honest test. Corpus D has 30 single-page invoices in ten different kinds of document, including some with no table at all. The *Evaluation Harness Guide* explains how the scoring works.

Corpus D (30 fresh invoices)	Rules extractor	Vision, first instruction	Vision, improved instruction
Field accuracy	62.0%	99.4%	99.8%
Planted mistakes found	15 of 26	26 of 26	26 of 26
False alarms	30	0	0
Correct invoices with no findings	5 of 15	15 of 15	15 of 15

Result. On invoices it had not seen, the rules extractor scored 62.0%, found 15 of 26 planted mistakes and raised 30 false alarms. The vision extractor, reading the same invoices as pictures, scored 99.4% on its first instruction and found all 26 mistakes with no false alarms. The layouts without a table (a prose list, an e-mail, cards) are where the rules extractor fails and the model does not. On the older sets, which the rules extractor was fitted to, the two are level on the single-page invoices and the rules extractor leads on the multi-page ones (section 5), so the gain from reading pictures is on layouts nobody wrote rules for.

As in the earlier reports, the oracle control, which returns the answer key itself, scored 100.0% on corpus D with 26 of 26 planted mistakes found and no false alarms. So the answer keys and the checking rules are right on these invoices, and every difference above comes from how the invoices were read.

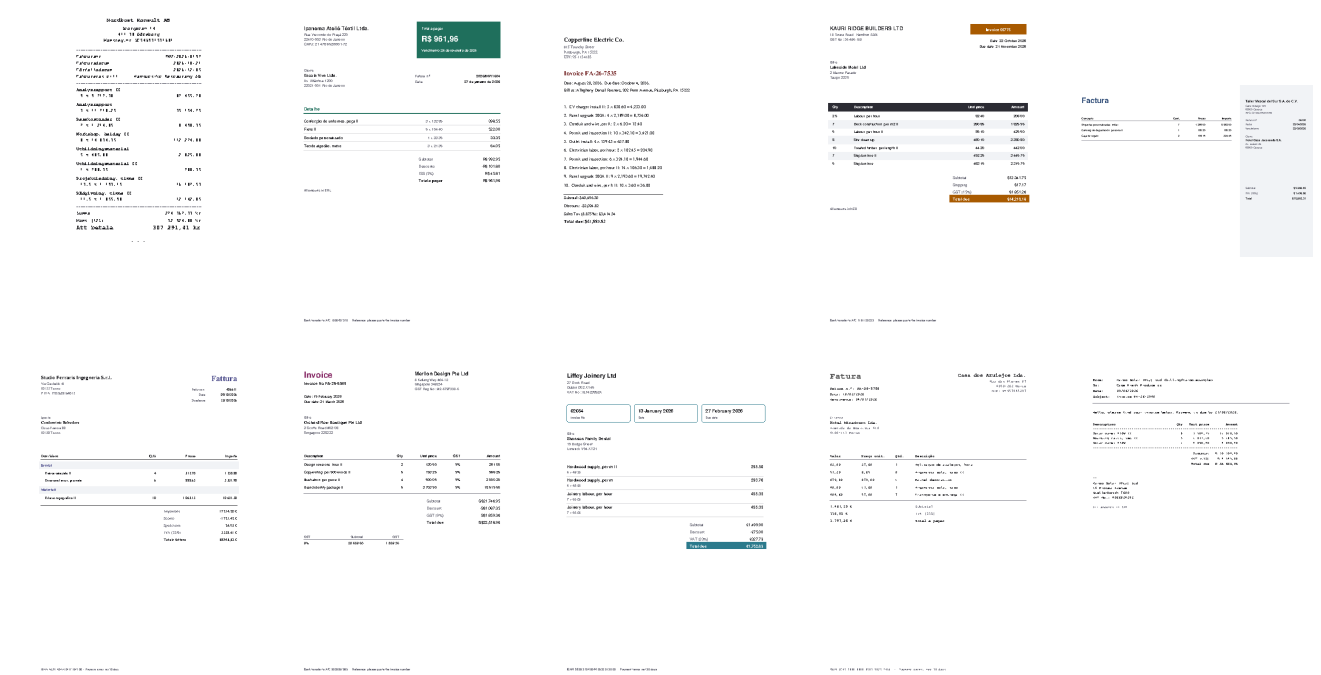
Read the improved-instruction column with care: I wrote that instruction after seeing the model's mistakes on corpus D and on the multi-page set, so corpus D is no longer a fresh test for it (section 3). The clean checks of the improved instruction are the older sets A and B, which were not used to write it.

2. The new test invoices

Corpus D is 30 invoices with 30 answer keys, three of each of ten kinds of document. They were written with a script of its own, separate from both the extractors and the built-in invoice generators, with different names, suppliers, currencies and line items. They use 11 currencies and 6 languages. 15 are correct and 15 carry between one and three planted mistakes (26 in all). All invoices are fictitious, and every invoice and its answer key can be opened from the Harness Reports page of this site.

Kind of document (3 invoices each)	What it looks like
Receipt	A narrow till-roll receipt in a typewriter font, each item on two lines.
Summary first	The amount due is in a coloured panel at the top; each charge shows "quantity × price" in one column.
Prose list	No table: each item is one sentence, "name: 5 × 12.00 = 60.00".
Quantity first	Quantity is the first column, under a dark header band.

Kind of document (3 invoices each)	What it looks like
Sidebar	A wide page with the parties and the totals in a side panel.
Grouped	Items sit under headings such as Services and Materials.
Tax column	A tax-rate column on every row and a small tax summary table.
Cards	Number and dates shown as three cards, each item on two lines.
Amount first	The columns run backwards: amount, price, quantity, description.
E-mail	A printed e-mail with a fixed-width text table and a signature.



One invoice of each kind.

3. The two extractors

Rules extractor. Reads the text layer of the PDF and finds each value by label, position and arithmetic, using rules written by hand. It needs no AI and no internet connection. It was fixed in Report 4 and is unchanged here.

Vision extractor. Turns every page into a picture at 200 dots per inch and sends all the pages of one invoice together to an AI model (Gemini 2.5 Flash), asking for every value and every line item in a fixed form. It uses no text layer and no character recognition software: the model reads the picture itself. Each answer is saved, so a published number can always be rebuilt and a rerun makes no new request.

Two instructions. The first instruction only said what to return: print only what is on the page, never calculate, money in whole cents, dates as year-month-day. The improved instruction adds plain definitions that are true of any invoice: the currency is the ISO code, a discount is a positive reduction of the price, a deposit or payment already received is not a discount, the total is the invoice total before a deposit is taken off, rows such as carried forward, brought forward and section totals are not line items, a wrapped description is one item, and a label such as "VAT No:" is not part of a tax ID. A small clean-up step also turns a printed currency symbol into its code and writes a discount as a positive amount.

How the improved instruction was made. I wrote it after reading the model's mistakes on corpus D and on the multi-page set (corpus C). Those two sets therefore no longer measure it fairly, and its corpus D score shows that the mistakes were fixed, not how it reads other invoices. Corpora A and B were not used to write it. The model was never fitted to any corpus.

4. Results on corpus D

Kind of document	Rules extractor	Vision, first instruction	Vision, improved instruction
prose list	12.4%	100.0%	100.0%
e-mail	15.5%	99.5%	99.5%
cards	30.3%	100.0%	99.5%
receipt	60.7%	99.0%	99.5%
amount first	65.2%	98.0%	100.0%
tax column	86.7%	100.0%	100.0%
summary first	88.9%	100.0%	100.0%
grouped	89.7%	99.0%	100.0%
quantity first	94.5%	99.5%	100.0%
sidebar	94.8%	99.4%	99.4%

Fields read right (out of 30 invoices, or all line items). Green is 95% or better, amber 90 to 95%, red below 90%:

Field	Rules extractor	Vision, first	Vision, improved
Line quantity	91/177	177/177	177/177
Line unit price	91/177	177/177	177/177
Supplier name	16/30	29/30	29/30
Invoice number	16/30	29/30	30/30
Customer name	17/30	29/30	29/30
Tax rate	17/30	30/30	30/30
Subtotal	18/30	30/30	30/30
Line description	109/177	176/177	177/177
Line amount	109/177	177/177	177/177
Tax amount	19/30	30/30	30/30
Total	19/30	30/30	30/30
Tax ID	23/30	30/30	29/30
Shipping	24/30	30/30	30/30
Currency	26/30	26/30	29/30
Discount	26/30	29/30	30/30
Issue date	27/30	30/30	30/30
Due date	27/30	30/30	30/30

Planted mistake type	Planted	Found: rules	Found: vision, first	Found: vision, improved
duplicate charge	1	0 of 1	1 of 1	1 of 1
grand total	3	2 of 3	3 of 3	3 of 3
line total	11	8 of 11	11 of 11	11 of 11
missing tax ID	1	0 of 1	1 of 1	1 of 1
subtotal	7	3 of 7	7 of 7	7 of 7
tax	3	2 of 3	3 of 3	3 of 3

The rules extractor does well on the kinds that look like a table with labelled totals (quantity first, sidebar, grouped, summary first, tax column) and badly on the ones that do not. The vision extractor reads all ten kinds about equally well.

5. The older sets, same invoices

So that the two extractors can be compared on identical invoices, the vision extractor was run on a spread-out sample of each older set (5 of 30 from corpus A, 10 of 40 from corpus B, 10 of 40 from corpus C), and the rules extractor was scored on the same invoices. The rules extractor was fitted to corpora A and C, and corpus B comes from the generator it was built beside, so these sets favour it.

Same invoices	Rules extractor	Vision, first instruction	Vision, improved instruction
Corpus A: 5 independent invoices	98.4%, found 3 of 3, 2 false alarms	not run	100.0%, found 3 of 3, 0 false alarms
Corpus B: 10 generator invoices	100.0%, found 4 of 4, 0 false alarms	not run	98.0%, found 3 of 4, 1 false alarms
Corpus C: 10 multi-page invoices	100.0%, found 4 of 4, 0 false alarms	79.8%, found 4 of 4, 30 false alarms	85.9%, found 4 of 4, 31 false alarms

The samples are small (5 and 10 invoices), so read them as indications, not rates. On the single-page sets the model is level with the rules extractor. On the multi-page set it scores lower than the rules extractor, which was fitted to those very invoices, and the problems behind that are in section 6.

6. Problems found

Problem	Evidence	Effect
Rules extractor: layouts with no table or no labelled totals	Prose list 12.4%, e-mail 15.5%, cards 30.3%. 11 of 26 planted mistakes missed.	Line items, totals and names missing or wrong; 30 false alarms.
Vision, first instruction: carried-forward and brought-forward rows listed as items (multi-page)	Three of the 10 multi-page invoices returned these rows as items. The harness compares lines by position, so one extra row moved every later line: 500 of the 538 wrong fields came from those three invoices.	Line items wrong; 30 false alarms on the sample.
Vision, first instruction: a deposit read as a discount, and the balance due returned as the total	Two invoices each on the multi-page sample.	Subtotal and total no longer add up.
Vision, first instruction: a printed currency symbol returned instead of its code	The symbol "kr" returned instead of SEK on 3 of 30 invoices; a bare "\$" with no clue left blank on 1.	Currency wrong.
Vision, both instructions: digits misread in small print	Years read as 2020 or 2003 instead of 2026 on three multi-page invoices, tax-ID digits misread on three, and the last digit of unit prices and amounts wrong on the two "stacked" multi-page invoices (34% accuracy, unchanged by the improved instruction).	Wrong dates, tax IDs and line amounts.
Vision, improved instruction: a few reading slips	On corpus D: a Mexican-peso invoice read as US dollars, "S.r.l." read as "S.r.1.", a customer name cut short, and a tax ID with its space removed. On the sample of corpus B: a day and month swapped, and a tax ID without its prefix.	4 wrong fields in all on corpus D.
Vision: amounts that are not labelled or not printed are left blank	On the generator sample the model left the tax rate and a shipping or subtotal amount blank where the invoice shows a bare row of amounts, because it is told never to work out a value.	Missing fields; the rules extractor fills these in by arithmetic.

Resolution matters. The vision results here use page pictures at 200 dots per inch. Most of the digit misreads above are small print on dense multi-page invoices. In a small check on four of those invoices, rendering the pages at a higher resolution removed most of them (for example 154 wrong fields on one invoice fell to 20). That check is not part of the published results. Picture resolution is a setting of the checker, and raising it is the first thing to try for better reading.

On the multi-page sample the improved instruction fixed the carried-forward, deposit and total problems: the sample went from 80% to 86%, the invoice totals from 8 of 10 to 10 of 10, and the layout with item codes from 74% to 100%. The digit misreads remained.

7. Cost and speed

The vision extractor makes one request per invoice, with all pages together. A multi-page invoice took about a minute to read and a single-page invoice less. Every model run behind this report, including the test requests made while setting it up, cost about 65 cents in total. The rules extractor takes a fraction of a second per invoice and costs nothing. The page pictures are sent to the model provider, so using this extractor means sharing the invoice images with that provider.

Conclusion

On 30 invoices that were new to both, the rules extractor scored 62.0% and the vision extractor 99.4% on its first instruction, with all 26 planted mistakes found and no false alarms. The rules extractor needs a rule for each kind of layout and fails on the ones nobody wrote a rule for. The vision extractor reads layouts it was not built for, and its mistakes are mostly in digits of small print and in the meaning of rows such as carried forward and deposits.

Read the numbers with care. The invoices are synthetic and few (30 for the main comparison, 5 to 10 for each older set), they were written by the same person who built the tool, and the model's answers vary slightly from one run to the next, which is why every answer was saved. The improved instruction was written after seeing mistakes on corpus D and on the multi-page set, so its scores there are not a fresh measure; its clean checks are the samples of corpora A and B. One model was tested at one resolution. 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.