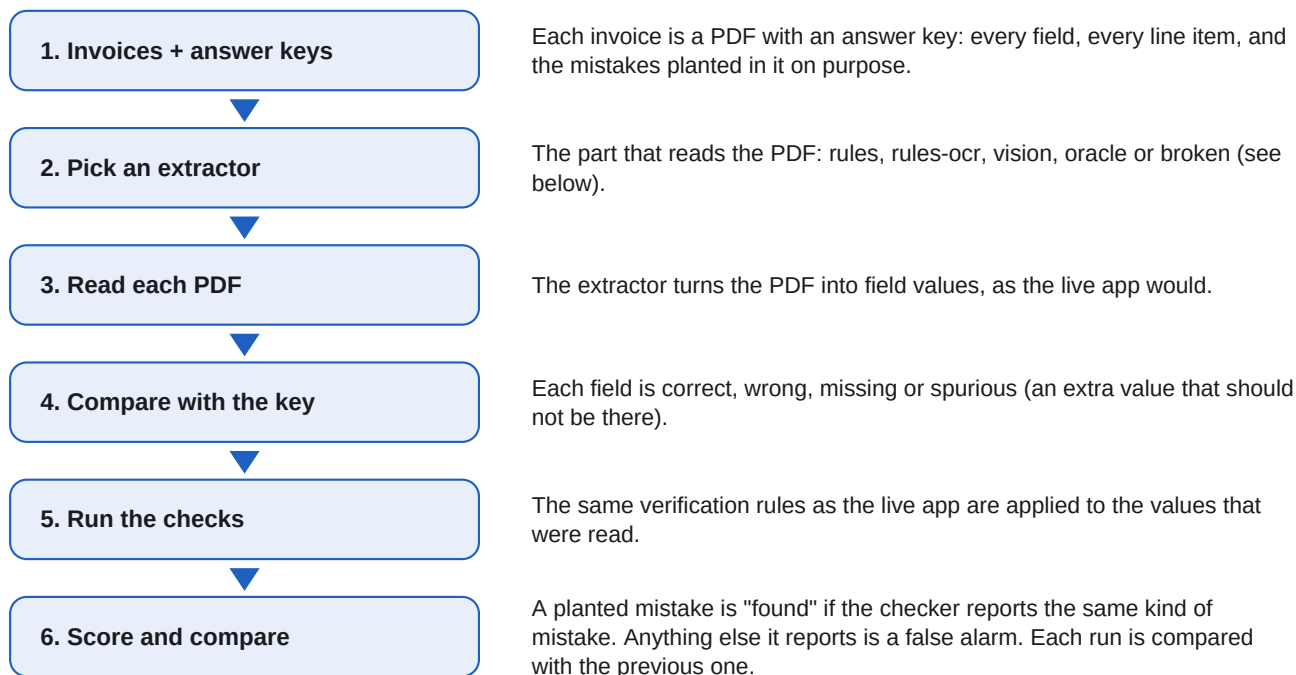


InvoiceInspect: Evaluation Harness Guide

How the accuracy of the invoice checker is measured

What it is. The evaluation harness is a testing tool that answers one question: how often does the invoice checker get it right? It runs the checker on invoices whose correct values are already known, compares what the checker found with those known answers, and produces a score. Because the test invoices stay the same, the harness can be run again after every change to see whether the score went up or down.

How it works



The five extractors

Name	What it does	In the live app?
rules	The real extractor. Reads the text of the PDF and finds each field by pattern, position and arithmetic. This is the checker's own logic, built on a PDF-reading library.	Yes
rules-ocr	The same extractor, run on a scanned picture of the page that text-recognition (OCR) software has read. It shows how the checker copes with scans.	Yes, for scans
vision	Shows a picture of each page to an AI model and asks it to read every value and line item. It uses no text layer and no OCR software, and each answer is saved so a score can always be rebuilt. It reads layouts that have no rules written for them.	Similar to the AI second reading
oracle	Ignores the PDF and returns the answer key itself. A "test oracle" is a source that already knows the right answer. It must score 100%; if it does not, the checking rules or the harness have a fault.	No, test only
broken	A control: returns the answer key with about 30% of invoices damaged on purpose. It must score badly, which shows the harness notices mistakes.	No, test only

How to read the results

Field accuracy counts fields read exactly right; totals and tax weigh more than names. **Found** is the number of planted mistakes that the checker reported. **False alarms** are reports that match no planted mistake. **Recall** is found divided by planted; **precision** is found divided by everything the checker reported. Results are also split by page structure, so a weakness in one layout shows up on its own.

For example, this is how the rules extractor scored on the 30 independent invoices in Report 1 (before any fixes). Green is 95% or better, amber 90 to 95%, red below 90%:

Field read wrong or missing	Correct
Tax ID	16/30
Line description	121/168
Customer name	22/30
Invoice number	22/30
Shipping	23/30
Line quantity	138/168
Line unit price	138/168
Line amount	138/168
Currency	25/30
Supplier name	27/30
Issue date	29/30
Due date	29/30
Tax rate	29/30
Tax amount	29/30
Total	29/30

Page structure	Accuracy
quantity first	37.0%
receipt	38.5%
tax summary	86.2%
compact	90.1%
cards	91.8%
letterhead	94.9%
sidebar	95.4%
banner	97.0%
letter	99.4%
multi-page	99.5%

Overall field accuracy 84.5%. Planted mistakes found: 25 of 32. False alarms: 30. Recall 78%, precision 45%.

What an answer key holds

For each invoice: the supplier and customer names, the supplier's tax ID, the invoice number, the issue and due dates, the currency, every line item (description, quantity, unit price and amount), the subtotal, discount, shipping, tax rate, tax and total as the invoice prints them, and the list of planted mistakes. Each planted mistake records its kind, the correct value and the value printed. Amounts are held in whole cents so that no rounding creeps in.

Limits

- The invoices are made up, not collected from real companies, so the scores describe behaviour on these invoices, not exact real-world accuracy.
- The built-in generator was written alongside the extractor, so it is the easier test; that is why a second set was written independently of it.
- The harness trusts whatever the extractor returns. The live app marks doubtful values "could not verify", which the harness does not model.
- A planted tax mistake that is reported as a total mistake counts as one miss plus one false alarm, even though a user would still be warned.
- The AI reading is not scored.
- With 30 to 40 invoices per set, one invoice moves a percentage visibly, so compare runs on the same invoices.