

# Report 1: Baseline

Aliyan Faisal · checker unchanged: this is the baseline

## 1. Summary

I ran the invoice checker on two sets of invoices whose correct values I know, to see how accurate it is before I change anything. This is the baseline; later reports use the same invoices and the same procedure so the numbers can be compared. How the harness works is in the separate *Evaluation Harness Guide*.

| Rules extractor (the real one)    | Built-in generator (40 invoices) | Independent invoices (30)                  |
|-----------------------------------|----------------------------------|--------------------------------------------|
| Field accuracy                    | 99.9%                            | 84.5%                                      |
| Planted mistakes found            | 20 of 20                         | 25 of 32                                   |
| False alarms                      | 0                                | 30 (26 "not found" warnings, 4 arithmetic) |
| Correct invoices with no findings | 20 of 20                         | 3 of 15                                    |

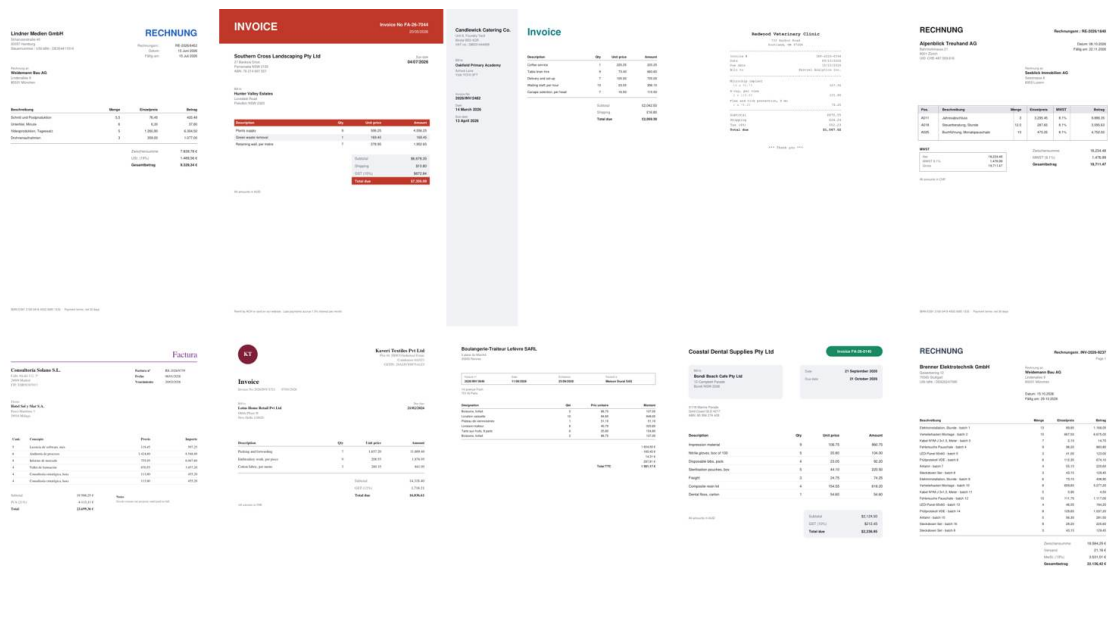
**Result.** The checking rules are correct: when given perfect values (the oracle) they found every planted mistake with no false alarms. The weak part is reading invoices. The extractor does very well on the invoices from the project's own generator, which it was built alongside, but only 84.5% on invoices written independently with different layouts, languages and tax types. It still found 19 of 24 arithmetic mistakes, but it often misses tax IDs and invoice numbers and does not understand two layouts. I did not change the extractor for this report.

## 2. The test invoices

**Independent invoices (30).** I wrote these separately from the extractor: 10 page structures (letter, banner, sidebar, receipt, tax-summary table, quantity-first, letterhead, compact, cards, multi-page), 8 currencies, 4 languages and several tax types (VAT, GST, sales tax, MwSt., TVA, IVA). 15 are correct; 15 contain one to three randomly chosen mistakes, and the same kinds repeat across invoices (32 planted mistakes in total).

**Built-in generator (40).** The project's own generator in 10 visual designs, half correct and half with one mistake each. This is the easier set.

All invoices are fictitious. Every invoice and its answer key can be opened from the Harness Reports page of this site, and the invoice numbers in section 4 are links to the PDFs.



The first ten independent invoices, one per page structure.

### 3. Results on the independent invoices

| Extractor                | Field accuracy | Mistakes found | False alarms |
|--------------------------|----------------|----------------|--------------|
| oracle (reference)       | 100.0%         | 32 of 32       | 0            |
| rules (real, text PDFs)  | 84.5%          | 25 of 32       | 30           |
| rules-ocr (scanned PDFs) | 66.5%          | 22 of 32       | 55           |
| broken (control)         | 94.9%          | 32 of 32       | 33           |

The oracle at 100% with no false alarms shows the checking rules are right; the broken control scoring badly shows the harness notices bad reading.

| Field read wrong or missing | Correct | Page structure | Accuracy |
|-----------------------------|---------|----------------|----------|
| Tax ID                      | 16/30   | quantity first | 37.0%    |
| Line description            | 121/168 | receipt        | 38.5%    |
| Customer name               | 22/30   | tax summary    | 86.2%    |
| Invoice number              | 22/30   | compact        | 90.1%    |
| Shipping                    | 23/30   | cards          | 91.8%    |
| Line quantity               | 138/168 | letterhead     | 94.9%    |
| Line unit price             | 138/168 | sidebar        | 95.4%    |
| Line amount                 | 138/168 | banner         | 97.0%    |
| Currency                    | 25/30   | letter         | 99.4%    |
| Supplier name               | 27/30   | multi-page     | 99.5%    |
| Issue date                  | 29/30   |                |          |
| Due date                    | 29/30   |                |          |
| Tax rate                    | 29/30   |                |          |
| Tax amount                  | 29/30   |                |          |
| Total                       | 29/30   |                |          |

| Planted mistake type | Planted | Found |
|----------------------|---------|-------|
| duplicate charge     | 6       | 4     |
| grand total          | 3       | 2     |
| line total           | 9       | 6     |
| missing tax ID       | 2       | 2     |
| subtotal             | 7       | 7     |
| tax                  | 5       | 4     |

Subtotal and tax-ID mistakes were found every time; line-total and duplicate-charge mistakes were missed mostly on the receipt and quantity-first layouts, where the line items are not extracted.

### 4. Result for each independent invoice

Found = planted mistakes the tool reported. Extra = reports that match no planted mistake (false alarms).

| Invoice                     | Structure | Planted mistakes       | Found | Extra findings |
|-----------------------------|-----------|------------------------|-------|----------------|
| <a href="#">invoice-001</a> | letter    | none (correct invoice) | -     | -              |
| <a href="#">invoice-002</a> | banner    | tax                    | 1/1   | missing tax ID |
| <a href="#">invoice-003</a> | sidebar   | none (correct invoice) | -     | -              |

| Invoice                     | Structure      | Planted mistakes                                          | Found | Extra findings                                       |
|-----------------------------|----------------|-----------------------------------------------------------|-------|------------------------------------------------------|
| <a href="#">invoice-004</a> | receipt        | missing tax ID, grand total, line total (line 2)          | 1/3   | -                                                    |
| <a href="#">invoice-005</a> | tax summary    | none (correct invoice)                                    | -     | grand total, missing tax ID                          |
| <a href="#">invoice-006</a> | quantity first | duplicate charge (line 6), line total (line 4)            | 0/2   | missing invoice number                               |
| <a href="#">invoice-007</a> | letterhead     | none (correct invoice)                                    | -     | missing supplier, missing tax ID                     |
| <a href="#">invoice-008</a> | compact        | duplicate charge (line 6), missing tax ID, grand total    | 3/3   | missing invoice number                               |
| <a href="#">invoice-009</a> | cards          | none (correct invoice)                                    | -     | missing invoice number, missing tax ID               |
| <a href="#">invoice-010</a> | multi-page     | duplicate charge (line 17)                                | 1/1   | -                                                    |
| <a href="#">invoice-011</a> | letter         | subtotal, tax, line total (line 4), line total (line 5)   | 4/4   | -                                                    |
| <a href="#">invoice-012</a> | banner         | none (correct invoice)                                    | -     | -                                                    |
| <a href="#">invoice-013</a> | sidebar        | subtotal                                                  | 1/1   | missing tax ID                                       |
| <a href="#">invoice-014</a> | receipt        | none (correct invoice)                                    | -     | missing tax ID                                       |
| <a href="#">invoice-015</a> | tax summary    | subtotal, line total (line 5), line total (line 2)        | 3/3   | grand total                                          |
| <a href="#">invoice-016</a> | quantity first | none (correct invoice)                                    | -     | missing tax ID                                       |
| <a href="#">invoice-017</a> | letterhead     | grand total, tax, subtotal                                | 3/3   | missing supplier                                     |
| <a href="#">invoice-018</a> | compact        | none (correct invoice)                                    | -     | missing invoice number, missing tax ID               |
| <a href="#">invoice-019</a> | cards          | duplicate charge (line 4), subtotal                       | 2/2   | missing invoice number                               |
| <a href="#">invoice-020</a> | multi-page     | none (correct invoice)                                    | -     | missing tax ID                                       |
| <a href="#">invoice-021</a> | letter         | none (correct invoice)                                    | -     | missing tax ID                                       |
| <a href="#">invoice-022</a> | banner         | tax, subtotal, line total (line 4)                        | 3/3   | missing invoice number                               |
| <a href="#">invoice-023</a> | sidebar        | none (correct invoice)                                    | -     | missing tax ID                                       |
| <a href="#">invoice-024</a> | receipt        | duplicate charge (line 7)                                 | 0/1   | -                                                    |
| <a href="#">invoice-025</a> | tax summary    | none (correct invoice)                                    | -     | grand total, missing tax ID                          |
| <a href="#">invoice-026</a> | quantity first | line total (line 2)                                       | 0/1   | -                                                    |
| <a href="#">invoice-027</a> | letterhead     | none (correct invoice)                                    | -     | missing supplier                                     |
| <a href="#">invoice-028</a> | compact        | tax                                                       | 0/1   | grand total, missing invoice number                  |
| <a href="#">invoice-029</a> | cards          | none (correct invoice)                                    | -     | missing issue date, missing due date, missing tax ID |
| <a href="#">invoice-030</a> | multi-page     | duplicate charge (line 17), line total (line 4), subtotal | 3/3   | missing tax ID                                       |

3 of 15 correct invoices produced no findings; 25 of 32 planted mistakes were found.

## 5. Problems found

| Problem                                                                            | Evidence (independent invoices)                                                   |
|------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Tax IDs that are not VAT-style (ABN, EIN, GSTIN, NTN, UID, CIF) are not recognised | Tax ID read on 16 of 30; 14 false "tax ID not detected" warnings                  |
| Line items of the receipt and quantity-first layouts are not extracted             | Six of the seven missed mistakes are on these two layouts; they score 37% and 39% |

| Problem                                                                         | Evidence (independent invoices)                        |
|---------------------------------------------------------------------------------|--------------------------------------------------------|
| Some invoice-number formats and labels are missed ("Factura n°", 2026/INV/3846) | Invoice number read on 22 of 30; 7 false warnings      |
| Customer name sometimes taken from a label such as "Date"                       | Customer name read on 22 of 30                         |
| Currency decided from "\$" when the invoice says AUD or CAD                     | Currency read on 25 of 30; all 5 errors are AUD or CAD |
| A line or total row mistaken for shipping                                       | 7 invoices with a spurious shipping amount             |

## Conclusion

The checking rules are right, and reading the invoices is the weak part, mostly on layouts the extractor was not built beside.

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. Once the extractor is tuned on these 30 invoices its score on them will look better than it is on new invoices, so a final check should use a fresh set of invoices. Scanned PDFs (OCR) score lower and with many false alarms; in the app, confidence checks and the "please confirm" step soften this, but the harness does not measure that.

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