Case Study: Receipt Data Extraction with Make.com and Gemini

By Olusegun Marcus · Built and tested on public data (SROIE receipts dataset)

Summary

I built an automated workflow that reads photos of receipts and writes the company name, date and total into a Google Sheet. I then scored its output against the dataset's ground-truth labels, so every accuracy figure below comes from a comparison with known answers.

Headline result: on 15 receipts the workflow had never seen, dates and totals were correct every time (15/15 each). Company name was correct on only 3/15, because on one vendor's receipts the model repeatedly returned the customer's name instead of the business that issued the receipt. I report that failure as a finding, not a footnote.

Problem

Businesses spend time retyping receipt and invoice details into spreadsheets. I wanted to test how far a no-code workflow with a vision model could automate that, and, more importantly, how to measure whether it is accurate enough to trust.

Data

- **Source:** the public SROIE receipts dataset (ICDAR 2019), via a Kaggle copy. Each receipt image comes with a text file giving the correct company, date, address and total.
- **What the receipts are:** mostly Malaysian retail receipts.
- **Test sets:** 3 practice receipts used while building (excluded from all scores); 19 receipts for the first two tests; and 15 "unseen" receipts that neither prompt version had been tested on.
- Answer keys were copied exactly from the dataset's text files and were not edited after seeing results.

Workflow

Built in Make.com on free tiers:

1. **Google Drive:** search a folder of receipt images.
2. **Google Drive:** download each file.
3. **Gemini:** upload the image.
4. **Gemini (extract structured data):** return company, date, address and total as separate fields, using a fixed response schema.
5. **Google Sheets:** add one row per receipt.

Scoring happens in the sheet: answer keys sit on a separate tab and are looked up by file name, and formulas mark each field 1 (match) or 0 (miss).

Prompt v1: read the receipt, copy each field exactly as printed, return only the final amount paid for total, leave blank if not visible.

Prompt v2: v1 plus an instruction to return only the registered business name for company, without registration numbers, branch names or taglines.

How I measured

- **Company:** match if identical after trimming spaces and ignoring upper/lower case.
- **Date:** match only if the text is identical to the key. A date printed as 27 Mar 2018 does not match 27/03/2018 even though it is the same day.
- **Total:** compared as numbers, ignoring a currency symbol or thousands comma.
- This is deliberately strict. A miss is counted as a miss.

Results

Test	Receipts	Company	Date	Total	Overall
Prompt v1	19	8/19	19/19	19/19	46/57 (80.7%)
Prompt v2, same 19 receipts (re-test)	19	15/19	19/19	18/19	52/57 (91.2%)
Prompt v2, unseen receipts	15	3/15	15/15	15/15	33/45 (73.3%)

The v2 re-test run covered 20 receipts (the 19 plus one new one) and scored 54/60 (90.0%).

What I found

1. **Dates and totals were reliable.** Across all three tests, only one date or total was wrong (a total of 101.00 against a key of 100.90).
2. **Company name is the weak point.** Prompt v1 often added registration numbers, branch names or a second business name. Prompt v2 fixed most of those on the receipts it was designed against (8/19 to 15/19).
3. **The fix did not generalise.** On unseen receipts v2 scored 3/15 for company. In 12 cases it returned MAKASSAR FRESH MARKET SDN BHD. On the two receipts I opened, that name was the customer, printed below the issuer, Gardenia Bakeries. Thirteen of the 15 unseen receipts came from that one vendor, so a single repeated confusion drives most of the misses.
4. **Testing on unseen data mattered.** The 90% re-test looked strong, but it used receipts I had already studied. Only the unseen test exposed the buyer-versus-seller problem.
5. **Results varied between runs.** The same receipt sometimes came back with different company text on different runs.

Limitations

- Small samples: 19 and 15 receipts. The unseen set has only 2 vendors, so the 73.3% is not a general accuracy figure.
- The v2 prompt was written after I saw v1's errors on the same receipts, so the re-test shows the fix works on known problems, not on new ones.
- Strict matching counts formatting differences and a few key typos as misses.
- Two date cells were converted by Google Sheets into real dates; I scored them by rebuilding the original text, and the AI's text matched the key.
- Tested on receipts only. I did not test invoices, handwritten documents, or non-Malaysian formats.
- Free-tier limits (daily request quotas and "high demand" errors) forced small batches and retries. Some early duplicate rows from reruns were removed from scoring.
- Not production-ready. Output would need human review, especially for company name.

Next steps

- Test a prompt that says to return the business that issued the receipt, not the customer or "deliver to" party. This is a hypothesis; I have not tested it yet.
- Run it on fresh receipts from many different vendors, then re-score with the same rules.
- Measure run-to-run variation by processing the same receipts several times.
- Add a review step for low-confidence or mismatched company names.

Method

I used an AI-assisted workflow design approach: I used Claude to plan the pipeline, debug errors and design the scoring checks, and I built, ran and verified every step myself against the dataset's ground truth.