CASE STUDY · SEPTEMBER 2026

Numbers Claude gives about your business, the same every month

Claude skills for the numbers a business asks every month, from Shopify, Stripe, Amazon or another export, and a check of the report someone else sends. Tested on nine synthetic datasets, six of them built by separate Claude sessions that never saw the skills. Opus 5.5.

ASKED PLAINLY

3–4

of 11 lines on the owner's definitions, Stripe test account

DEFINITIONS WRITTEN ONCE

10

of 11 lines, in every run

Your definitions, written down once

Asked a month-end question plainly, Claude answers on readings of its own. The arithmetic is not the problem: with the definitions in the prompt it is accurate. The point is that nobody writes them out every month.

	STRIPE	AMAZON	SAAS ON STRIPE BILLING
The question alone Claude on its own readings: the month in local time, payouts by the day they reached the bank, the reserve inside what is owed	3–4 of 11	6–7 of 10	2–4 of 12
With a definitions file the owner's definitions written down once, the question points to them	10 of 11, every run	10 of 10, every run	5–6 of 12
With a pinned calculation the same figure every run (on Stripe the file's tax line moved by cents); turns 6–9 against 11–15 with the file on Stripe	10 of 11, every run	10 of 10, every run	7 of 12

On the subscription business (MRR, churn, billings, recognised revenue) a file was not enough: only the pin rebuilt each subscription's status for the month's end, since the export gives it for the day it was taken. The pin's 7 of 12 is in every run that answered; one of six stopped and asked for a file. No run reached MRR to the cent. On Stripe and Amazon the pin's own check passed 10 of 12 answers and flagged two.

The Shopify month-end skills

THREE TEST STORES, 27 ANSWERS EACH WAY	WITH THE SKILLS	WITHOUT
On the skill's stated definitions the usual answers on two stores, the owner's on the third	27/27	13/27
Everything the owner should see, named every order left out, disputed, not paid out	27/27	21/27

Store D (30,251 orders, built apart, used to develop): the scripts' blind run matched 18 of 50 figures, 55 of 56 after. **Store E, built apart and blind: the skills did not win.** Plain Opus got 9 of 9 on neutral graders, the skills 6 of 9.

Checking a report you already get

An agency's report on the Stripe account, with 7 mistakes planted by a separate session. The skill gives the same verdicts as plain Opus in every run: 16 of 22 right, 6 of 7 mistakes caught; neither caught two swapped digits. It is not more accurate. What it adds is an answer where every figure of the report is accounted for, at about 1.6 times the turns.

Does the right skill run

ELEVEN DEFECTS PLANTED IN A SKILL, ONE PER COPY OF A PLUGIN	FLAGGED
claude plugin validate	2 of 11
My linter including disable-model-invocation, which passes validation and switched the skill off	11 of 11

Three of the eleven stopped the skill cold on its own question; the validator passed one of those three.

It ran, and the answer is wrong

24 MISTAKES PLANTED IN CORRECT ANSWERS	STOPPED
The number check: every figure traces to the script's results	6
The coverage check: everything listed for the owner is named	5
A reviewer model on the other 13 (Sonnet 5 and Opus 5.5, 3 runs each)	13 of 13
Nothing	0

The export changes

Eight changes to an export the owner had confirmed (a new gateway, a pending payment, a second currency, a renamed column, an export taken on the 20th). The earlier version let 6 through silently; now each is named or stops the script.

What you get

- Your definitions asked once, in plain words, and printed under every answer.
- Every figure from a script, checked against the script's results; anything left out named.
- The skills for Claude Code, their tests, and every model answer and grade from the testing.
- The first step is always your own question on your own export.

What this does not claim

- Not real businesses: every store and account is synthetic. Not yet run on a real business's raw export.
- Not a blind test overall: the same person wrote the skills and the graders; every answer the tables count is published with its grades.
- Tested in Claude Code. The Shopify skills were also tried in claude.ai chat; team accounts and Cowork not yet.
- Sizes tested: about 15,000 Shopify orders, 4,000 Stripe payments or 3,200 Amazon orders a month.