

A good free plugin, and what the owner still can't see

Anthropic's free Small Business plugin for Claude: 44 skills for closing the month, reports, payroll, taxes and invoices, written as instructions, so the model writes its own code each time. Run with Opus 5.5 on synthetic books and three synthetic Shopify stores, three runs per question, every answer kept. An independent check, not affiliated with or endorsed by Anthropic.

PLANTED PROBLEMS FOUND

12/12

in every close, with the plugin and without

REVENUE IN THREE RUNS

3

different figures for the same month; net income the same

What it does well

CLOSING AUGUST, BOOKS WITH 12 PLANTED PROBLEMS	WITH THE PLUGIN	PLAIN CLAUDE
All 12 found, in the answer or the file the run saved	3 of 3 runs	3 of 3 runs
The 14.20 coffee, under its 25.00 receipt rule	not chased in 3	receipt asked for in 3
The weekly 45.00 email plan, not a duplicate	not questioned in 3	questioned in 2
The 0.30 difference, under its 0.50 rule	noted in 3, one also said to correct it	an error in 3

With its skill called by name it was right in 25 of 27 runs of the bench (26 if a refund split the owner's definitions do not settle is accepted). When the owner uses its words, the intended skill ran in 29 of 30 sessions. A saved report definition was followed on every rerun, 6 of 6.

What the owner still cannot see, or did not choose

	RUN 1	RUN 2	RUN 3
Revenue in the corrected P&L, with the plugin	96,459.17	94,396.54	96,680.77
Net income	73,281.22	73,281.22	73,281.22

- **Which revenue.** Every close said Shopify deposits are booked net, then picked its own revenue line. Which one the business reports is the owner's rule, never asked.
- **Which orders count.** One run counted an order tagged "test" (59 orders, 2,999.72), two left it out (58 orders, 2,918.82); neither asked first.
- **Whether it ran at all.** Asked plainly with the export in the folder ("How did August go?"), none of its skills ran in 27 sessions.

None of these is a bug to report. It is the part a general plugin cannot know about one business.

The same questions: the plugin, my skills, plain Claude

NINE OWNER QUESTIONS, 3 RUNS EACH	SMALL BUSINESS	MY SKILLS	PLAIN CLAUDE
Right by the bench's automatic checks	25/27	27/27	26/27
Every order or product the owner should see named	19/27	27/27	21/27
Says it added the figures up by hand	0/27	0/27	9/27

The plugin has no skill for the bridge from orders to payouts; five of its six misses on naming, counting its own workbooks, are on that question.

Closing the gap, on top of the plugin

- **Your rules, asked once:** which revenue line, whether test orders are sales, when a sale counts, how a partial refund is split: answered in a file you can read.
- **Printed under every figure:** so this month and next are counted the same way.
- **Checked on your exports:** recent months and a copy with planted problems, three runs each.
- **Tried on your team's words:** the right skill runs when someone asks "how did the month go?".

Every answer and every grade is handed over.

How to read this

- Synthetic data, Opus 5.5 only, three runs per question: a rate of 1 in 3 is a sign, not a measurement.
- The plugin ran cold, as an owner who skipped onboarding would have it: no stored business context, not in Cowork, in unattended sessions. Onboarding does not ask the counting rules above, so the main finding stands.
- The bench's checks were written for my own skills; where a plugin answer used another defensible basis it was read, and each reading is recorded.

Not affiliated with or endorsed by Anthropic, Shopify or Intuit. The plugin is Anthropic's, under Apache-2.0.