Would you put more money behind this?

A campaign report shows $190 of contribution after advertising. The arithmetic checks pass. The recommendation is to increase spending. It looks like the work is finished, right up until someone reads the transactions the report was supposed to explain.

Use the cost recorded for each sale, include the relevant returns, and that same campaign is down $23. Nothing dramatic happened to the formatting. The business answer changed.

This is a new fictional case, built for this guide. It draws on a lesson from my work connecting operational systems and checking what they produce: a useful review has to challenge the business conclusion, not simply confirm that a program ran. We will work the answer, break a check and decide what can actually be claimed.

Write the sentence the build must earn.

> The promise
> For the September Desk reset sales cohort, show contribution through October 5, 2026, using sale-date costs and returns known by that cutoff. State whether the $100 expansion threshold is met.

The fictional decision owner requires at least $100 before considering more spend. That threshold is a case assumption, not a recommended margin policy. Even a pass would support a review, not automatically authorize a campaign change.

Our measure deducts original product costs, outbound freight, retained fees, return freight, handling and campaign spend from net sales. Resellable returns recover their original product cost. Damaged returns do not. Sales tax, fixed overhead and later returns are outside this defined measure. Call it campaign contribution, because a broader profit claim would promise more than these inputs support.

The join matters as much as the formula.

Evidence	Key	Why it matters
Sales	sale_id	Defines the cohort and original quantity
Cost history	sku + effective date	Prices cost at the sale date
Returns	return_id + sale_id	Connects refunds and recovered stock
Spend	campaign	Includes the cost of acquiring sales

The kit has four sales, four return records, a cost history, a current cost list and one campaign-spend record. Three sales belong to Desk reset. The fourth is an older campaign and stays outside this decision.

Two returns qualify by the cutoff. One arrives later, and another belongs to the older sale. Counting every September refund would mix populations. Counting only September postings would miss a relevant October return. Begin with the sales cohort, then follow its records through the stated cutoff.

Today's cost cannot rewrite last month's sale.

Sale	Sale-date cost	Current cost
S-201: 10 CLIP	10 x $18 = $180	10 x $15 = $150
S-202: 8 CLIP	8 x $22 = $176	8 x $15 = $120
S-203: 6 LAMP	6 x $31 = $186	6 x $27 = $162
Total	$542	$432

The report used the current catalog, reducing cost by $110. Its multiplication is correct. Its basis is wrong for this promise. A checker that recalculates the same quantities against the same current-cost field will confidently reproduce the mistake.

The fixture defines cost by its effective date. For a real business, agree which source governs historical cost and whether adjustments are expected. Do not infer that a field named cost carries the meaning your decision needs. The matching date and governing policy belong in the evidence.

A refund and a stock recovery are different events.

Return R-301 refunds two CLIP units from S-201 for $60. Both can be resold, so the case recovers $36 of original cost. Return freight is $9 and handling is $4. The contribution effect is minus $37: minus $60, plus $36, minus $13.

R-302 arrives on October 2 against S-203. It refunds $55, but the LAMP is damaged, so no inventory value comes back. Add $8 return freight and $3 handling. Its effect is minus $66. It belongs in a September sales-cohort decision through October 5, even though its posting month is October.

> Keep the boundary visible
> R-303 is dated October 7 and is excluded as future evidence. R-304 belongs to the older campaign. Neither is silently deleted.

Every correction has somewhere to land.

Step	Contribution	Change
Published-looking result	$190	Starting claim
Use sale-date cost	$80	-$110
Include R-301	$43	-$37
Include R-302	-$23	-$66

The final view contains $755 of net sales and $506 of net product cost after the resellable return. Deduct $66 outbound freight, $52 fees, $17 return freight, $7 handling and $130 campaign spend. The result is minus $23. The complete bridge explains the $213 difference from the first report.

This is why I want a reconciliation, not another confident paragraph. Each movement points to records and a rule. If someone disagrees with the conclusion, we can identify which fact or treatment they would change instead of debating whether the dashboard feels trustworthy.

Ask what the green check actually checked.

The supplied flawed report balances perfectly. Its net sales equal gross sales less its stated refunds. Its contribution equals its stated revenues less its stated costs. Its verdict agrees with its stated threshold. All three checks pass because the report can be consistently wrong.

Those checks still have value. They catch broken arithmetic and presentation drift. They cannot discover an omitted return or decide which cost basis the business intended. The weakness is treating internal consistency as evidence that the inputs and meaning are right.

> A question for the reviewer
> If the builder and the checker believe the same wrong assumption, what evidence in this review could make either of them change its answer? Write down that evidence before adding more tests.

Let the reviewer reach an answer first.

Give a second reader the business promise and source fixtures before showing the report or the builder's explanation. Ask for the cohort, included returns, cost basis and contribution. A different AI can help, but independence comes from the evidence and method, not from a second logo.

The reviewer should commit a calculation before opening the candidate report. Then compare the two and reproduce any disagreement at record level. Keep the independent answer separate from the generated report so the build cannot quietly refresh its own answer key.

For this exercise, read the sources and write your calculation before opening references/expected.md. That file contains the completed manual answer. The automated audit reconstructs the fixture result without importing the report builder, then checks it against that separate answer.

Watch the wrong report pass, then fail.

The kit runs offline with Python 3.10 or newer, without third-party packages. From the extracted skill folder, use these commands. The first creates a deliberately flawed report in the local outputs folder. The next proves its arithmetic balances. The independent audit must reject it.

```
python scripts/report.py --mode flawed
python scripts/selfcheck.py outputs/flawed.json
python scripts/audit.py outputs/flawed.json
```

Then generate the corrected result and run the audit against that file. Read the errors as evidence, not a score. A failing audit should identify the mismatched amounts or record coverage. It should never replace the source data with whatever makes a green result.

```
python scripts/report.py --mode corrected
python scripts/audit.py outputs/corrected.json
```

Plant a mistake that ought to matter.

Start with the corrected result. Remove the October return, adjusting refunds, return freight and handling together so the report still balances. Contribution rises from minus $23 to plus $43. The internal check still passes. The independent audit must fail because the report has omitted R-302 and changed the amounts.

The included check script performs that experiment, then restores the valid result. It also rejects the original current-cost report and checks that the audit refuses a missing input file. These are targeted challenges to the promise in this guide, not evidence that every possible accounting error is covered.

```
python scripts/check.py
```

If your test passes both the good answer and the consequentially wrong answer, the test is not yet useful for that distinction.

The consumer does not read your test log.

A correct calculation can still reach the recipient with the wrong label, cutoff or emphasis. Open the artifact they will receive. Check that the headline uses the verified result, the decision threshold is visible, and the coverage note survives the export. Inspect the exact downloaded file or deployed page, not a similar development view.

For our exercise, the JSON file is the delivered artifact. Verify minus $23, HOLD, three cohort sales and two included returns in that file. A future visual report would need its own reading check, because this kit does not include or test that presentation.

For a live action, follow the result through the receiving system. A submission receipt establishes that a request was accepted. The record the recipient actually sees establishes whether the intended change arrived.

The recommendation should fit in one breath.

> The completed case brief
> Hold the expansion. Desk reset contribution through October 5 is -$23 against a $100 review threshold. The prior $190 figure used current costs and omitted two returns. The $213 correction is fully reconciled to the supplied records.

Then name the limit: later returns may change the cohort result, and the measure excludes fixed overhead. The case supports withholding the proposed expansion now. It does not predict what a redesigned campaign would earn or establish the final lifetime economics of these customers.

That is enough for the decision in front of us. The owner can request a fresh cohort view when the return window closes or investigate which part of the offer needs to change. Neither requires pretending this number says more than it does.

Spend the review where a wrong answer would hurt.

A typo fix does not need a finance investigation. A report that will decide inventory, spend or customer commitments deserves an explicit promise and an independent check on the numbers that drive that decision. Match the review to the consequence, then stop when the agreed evidence is complete.

Choose cases that can change the answer: an effective-date boundary, a linked refund, an omitted source, an empty population. Avoid inflating a test count with dozens of cosmetic variations while the business assumption remains untouched. One well-chosen contradiction can be more useful than a page of passes.

Keep a list of the important behaviors you have not tested. It gives the next reviewer a useful starting point and prevents a narrow successful exercise from turning into a universal reliability claim.

Make the disagreement productive.

Assign one person the decision, one the build and one the independent read. Start with the promise and source files. Let the reviewer calculate before opening the report. Compare answers, reproduce the difference, and explain why the internal checks missed it. Then run the planted-fault experiment.

Finish by taking one real report your team uses and writing its equivalent promise. Identify one consequential field, the source that governs it and a mistake your current check would miss. You do not need to audit the entire business to make that next review materially better.

This kit supplies a reproducible case, not evidence that your team has completed it. The next four pages are the complete AI instructions. Use the worksheet to record a real trial and the result you actually observed.

Define the claim and the consequence.

Pressure-test the specific business result I provide. First identify the decision it supports, who will rely on it and what would happen if it were wrong. Write the result as a testable promise: population, period, cutoff, units, calculation basis, required sources and the action threshold where relevant. Separate confirmed facts from assumptions.

Read the supplied evidence before asking for more. If a missing source or unclear rule could change the decision, identify that gap and withhold only the affected conclusion. Keep the review proportional to the requested result; do not expand a focused check into an unrelated universal audit.

Preserve the candidate output and the relevant source inputs so we can reproduce the starting claim. Label fictional practice data and keep it separate from any actual company records.

Build an independent answer.

Reconstruct the consequential result from the governing records and agreed rules. Verify identifiers, coverage, time basis and units before trusting totals. For joined records, show which rows are included, which are excluded and why. Explain how returns, reversals or corrections affect the business measure when those are relevant.

Keep this calculation separate from the candidate output. Do not import the builder's derived totals as your expected answer. If another reviewer is available, give them the promise and primary inputs first, then compare their committed answer with the report. Otherwise describe the actual independence limit.

Reconcile each difference to a record or rule. Challenge your own proposed finding with contrary evidence. A disagreement becomes a defect only when you can reproduce why it violates the stated promise.

Prove the check distinguishes wrong from right.

Select a plausible fault that would change the decision or mislead its recipient. Introduce it into a disposable copy, leaving original inputs intact. Show that the relevant check rejects the faulty result, and explain the reason. Restore the correct result and run the same check again. Do not change the answer key to accommodate the fault.

Use meaningful boundaries rather than a large number of repetitive tests. Record what passed, what failed and what remains outside coverage. If the available environment cannot execute the check, say so and provide a runnable procedure without claiming an observed result.

Inspect the exact final artifact on the consumer's path. Confirm that verified numbers, dates, limitations and labels survived rendering, export or transfer. Distinguish acceptance of a request from confirmation of its eventual effect.

Deliver the decision and its evidence.

Return a concise verdict on the requested result, the evidence behind it, the differences reproduced and the limits that remain. Include the corrected calculation only where the sources support it. Give a clear action recommendation tied to the original decision, and state what evidence would change that recommendation.

Package the source references, reproduction steps, independent check, planted-fault result and final artifact so another person can repeat the review. Use a short issue record with the claim, observed fact, affected conclusion and resolution. Avoid unsupported guarantees, decorative test counts or claims of business impact inferred from a successful command.

Do not send, publish or change live records merely because this review is complete. Finish the requested deliverable and identify any separate action that still requires authorization or evidence.