

An AI's sales report, every number checked

A Shopify orders export given to an AI assistant, and its report checked against the same export

A shop owner exports the quarter's orders from Shopify and asks an AI assistant for the key numbers, "exact, for an investor presentation". The assistant writes code, cleans the file and returns a tidy report. We recounted every number of that report from the same export with separate code that never saw the report.

NUMBERS IN THE REPORT

113

CHECKABLE, AND MATCH

27 of 35

WRONG

8

All eight wrong figures come from one reasonable-looking line of the assistant's code.

```
q['refund_prod'] = q['Refunded Amount'] * q['Subtotal'] / q['Total']
```

It takes a "shipping share" out of every refund. For full refunds that is right. The partial refunds in this export are goods only, so shipping that was never refunded was taken out of them: refunds came out €505.35 too low and net sales €505.35 too high.

Figure	In the AI's report	Recounted	Difference
Refunds of goods, Q1	83,002.07	83,507.42	-505.35
Net sales, January	529,138.18	528,930.40	+207.78
Net sales, February	458,990.22	458,841.03	+149.19
Net sales, March	487,268.58	487,120.20	+148.38
Net sales, Q1	1,475,396.98	1,474,891.63	+505.35
Online store, net sales	1,325,008.08	1,324,502.73	+505.35

Everything else that could be checked matched to the cent: gross sales, discounts, orders, refunded orders, returning customers, all ten top-product figures. The assumption is written in a code comment, never in the report. A reader cannot see it; only a recount finds it.

Two figures in the table are printed twice in the report, which makes eight. 78 numbers could not be checked, each with its reason given (a different definition, a split the checking code does not compute, or a note on cleaning).

Right arithmetic, a different answer

A second assistant, the most capable model available at the time, got the same file and request. Its arithmetic holds, and its headline is still €221 thousand away, because it chose the definitions itself:

	Stronger assistant	Checked count
Net sales, Q1	€1,253,965.73	€1,474,891.63
VAT	excluded	included
Orders not yet paid	counted in	left out
Staff purchases	counted in	left out

Both can be defended. Nobody asked the owner which one the investor should see.

Why a stronger model does not replace the check

- **The choice of definitions is the owner's.** A better model still has to pick one; a check that starts by asking makes the choice on purpose.
- **An error that looks reasonable survives its own review.** A model re-reading its work shares its assumptions; a recount by separate code does not.
- **An investor cannot check a model's word.** They can check a document that puts every figure beside an independent recount.

What you send
Your report or deck, the export it was written from, and the list of figures to check. A short written list of definitions follows: what counts as a sale, VAT, shipping, which date.

What you receive
Every figure beside its recount: it matches; it does not, with both values and the difference; or it cannot be checked, and why. And a receipt: file fingerprints that show the document was not changed after the check.

The export is synthetic, generated in the real Shopify format, so the right answers are known. One export, one request, one run of each assistant: this shows what a check finds, not how often assistants are wrong. In one order's notes the export also carried an instruction addressed to "the AI analyst"; neither assistant followed it, and neither mentioned it.