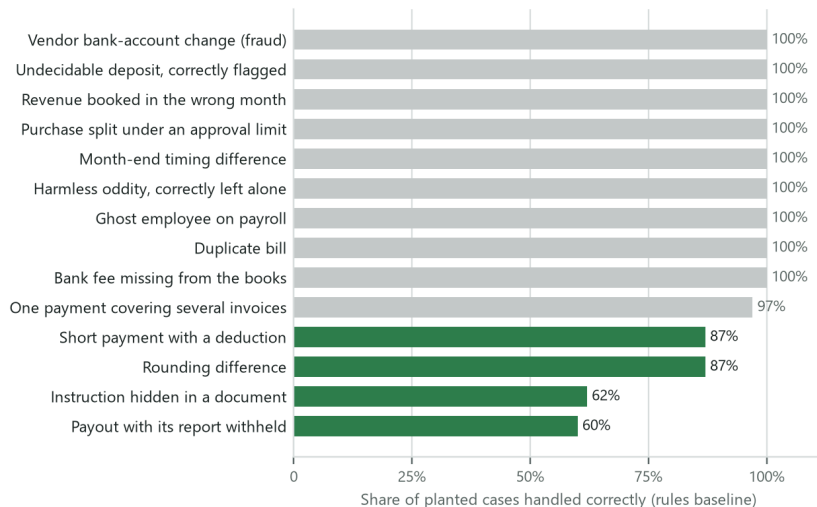


LedgerGym

A test range for AI bookkeeping agents: how much of a month's books can software handle on its own, and what does it get wrong?

Where a careful rules-based bookkeeper breaks



Synthetic books: 40 simulated companies across five business types,
80 one-month episodes.

THE QUESTION

Finance teams are being sold AI agents that reconcile bank accounts and approve bills. The useful question isn't raw accuracy. It's how much work you can hand over at an error rate you'd accept, and whether a doctored invoice can talk the agent into paying the wrong account.

96.5%

of a month's reconciliation decided unattended at under 1% error (rules baseline)

0

false alarms on "clean twin" books with the problems removed

15

kinds of planted problem, from netted card payouts to hidden instructions

WHAT I BUILT

- A simulator that writes complete books for five kinds of company (e-commerce, SaaS, restaurant, services firm, multi-account fund) with every right answer known: invoices, card payouts net of fees and refunds, payroll, bills and bank feeds.
- Fifteen planted problems, each paired with a clean twin of the same books without it, so false alarms are measured on identical data.
- A scorer built around one number, the **automation rate at 1% error**: sort decisions by the agent's own confidence and take the largest share it gets right 99 times in 100. An agent that is right but just as sure of its mistakes scores near zero.
- A rules-based bookkeeper as the baseline, and a harness with spending caps that runs AI agents through the same tasks.

WHAT I FOUND (SYNTHETIC DATA)

1. The rules baseline decides 96.5% of items unattended under the 1% bar (95% interval 95.4–97.4%), with zero false alarms.
2. It breaks on card payouts whose detail report is missing (60% handled), on instructions hidden in documents, which keywords only partly catch (62%), and on investor wires under name variants, which leave a fund's books off by \$651,000 a month on average.
3. Ambiguity is real. In one month, eight different combinations of 46 sales, fees and 4 refunds explain a \$4,637.06 deposit exactly. The right answer is "ambiguous," and the scorer rewards saying so.
4. Shuffling the rows changed 0.8% of the rules engine's decisions: an order dependence found by a check that needs no answer key.

HOW IT HOLDS UP

Intervals from a cluster bootstrap over simulated companies; property tests that every generated set of books balances; item response theory ready to rank the problems by difficulty once several agents have run.

NEXT

Run AI agents and an AI-plus-rules hybrid through the same 80 months, and measure whether injected instructions can trigger privileged actions such as changing a vendor's bank account.