AI Pilot Scorecard

One task, four weeks, a baseline and a stop rule.

Task: _____

Owner (can switch it off): _____

Non-AI alternative: _____

Stop rule, written before starting: _____

Week 0: decide and measure

- [ ] Name the task, the owner and who can switch it off
- [ ] Write down how it works today and the non-AI alternative
- [ ] List what information the tool may see, and what it must never see
- [ ] Choose one main outcome and two to four guardrails
- [ ] Check the vendor's data handling, permissions, retention and emergency stop
- [ ] Measure the baseline over a normal stretch of work

Week 1: a narrow test

- [ ] Use made-up or properly de-identified examples
- [ ] Test normal, odd, badly formatted and deliberately tricky cases
- [ ] A person reviews every output
- [ ] Record wrong statements, missing information and unexpected actions

Weeks 2 and 3: a limited live pilot

- [ ] Go live on a small slice of real work only after privacy and integration checks
- [ ] Review a planned sample, and every output that could cause harm
- [ ] Measure staff time including checking and fixing
- [ ] Track qualified outcomes, complaints and requests for a person

Week 4: decide

- [ ] Compare with the baseline: staff time, cash and customer experience, kept separate
- [ ] Decide: stop, fix the basics, continue the pilot, or expand

☐ Write down why, and the next review date

Scorecard: count every measure out of something

Completion: Eligible requests handled correctly. Count it out of: All eligible requests received. Avoid: Counting 100 calls answered as 100 successes.

Base	Wk 1	Wk 2	Wk 3	Wk 4

Correctness: Outputs that pass your written checks. Count it out of: Outputs you reviewed. Avoid: Treating a confident answer as a correct one.

Base	Wk 1	Wk 2	Wk 3	Wk 4

Escalation: Cases handed to a person that needed a person, and were completed. Count it out of: Cases that needed a person. Avoid: Treating every handoff as a failure.

Base	Wk 1	Wk 2	Wk 3	Wk 4

Review effort: Staff minutes spent checking and fixing outputs. Count it out of: Per task or per week. Avoid: Counting only the AI's processing time.

Base	Wk 1	Wk 2	Wk 3	Wk 4

Lead quality: Leads a person confirms as qualified. Count it out of: Leads captured. Avoid: Treating every contact as a qualified lead.

Base	Wk 1	Wk 2	Wk 3	Wk 4

Extra customers: Customers you can reasonably attribute to the change. Count it out of: Comparable opportunities before and during. Avoid: Counting customers who would have booked anyway.

Base	Wk 1	Wk 2	Wk 3	Wk 4

Complaints and harm: Confirmed wrong, unwanted or harmful interactions. Count it out of: All relevant interactions. Avoid: Hiding one serious error inside a good average.

Base	Wk 1	Wk 2	Wk 3	Wk 4

Running cost: Vendor, phone, usage, staff oversight and upkeep. Count it out of: Per month. Avoid: Leaving out phone charges, review time and retries.

Base	Wk 1	Wk 2	Wk 3	Wk 4

Reliability: Outages and how quickly you recovered. Count it out of: Hours you needed it, or incidents. Avoid: Quoting the vendor's uptime promise as your measured uptime.

Base	Wk 1	Wk 2	Wk 3	Wk 4

Decision

Circle one and write why below

Stop | Fix | Cont. | Expand

A one-month before-and-after comparison is affected by season, chance and other changes. It shows whether a tool is promising and workable, not proven ROI. Vortio method v1.0.0, reviewed 2026-11-23. Full guide: vortio.com/resources/ai-toolkit/pilot-guide