How to choose your first supervised agent workflow

A decision framework built on four tests: volume, rule density, tolerance for error, and who owns the outcome.

Why the first one matters more than the best one

The first workflow you automate is not primarily an investment in that workflow. It is the thing your organisation will use to decide whether any of this works here. That decision gets made on the first result, by people who were not in the scoping meeting, and it is very difficult to reopen.

Which is why the candidate with the highest theoretical return is usually the wrong place to start. Large returns come from large processes, large processes have many stakeholders, many stakeholders produce long approval chains and contested requirements, and by the time the first result arrives nobody agrees what it was supposed to be. Choose instead for speed of learning: the candidate that will give you a defensible answer soonest.

Four tests, applied in order. Each one has a question you can ask this week, a way to score the answer, and a condition that disqualifies the candidate outright.

Test one: volume

The question. How many times does this happen a month, and where does that number come from?

Volume matters twice. It decides whether the arithmetic closes, because the cost of handling exceptions is roughly fixed per exception while the saving scales with throughput. And it decides whether you can read the result at all. A six week pilot on a workflow that runs fifteen times a month gives you about twenty completed instances, which is not enough to distinguish a genuine improvement from a quiet month.

A rule of thumb that costs nothing to apply: if the workflow will not run at least a hundred times inside your measurement window, lengthen the window rather than strengthen the claim.

SCORE	WHAT IT LOOKS LIKE
3	You can pull the count from a system today, by month, for the last year.
2	You can estimate it within twenty per cent and defend the estimate.
1	Somebody experienced can guess, and the guesses differ.
0	The work is not counted anywhere.

Disqualifier. Too few occurrences inside a measurement window to distinguish signal from variation. A zero here is not fatal to the workflow, but it is fatal to it being first.

Test two: rule density

The question. If you wrote down the rules this work follows, how many would there be, and how many people would agree with the list?

This test fails in two opposite directions. Too few rules and the work is judgement, which means there is nothing to encode and, worse, no cheap way to check the output. Too many rules that nobody agrees on and you have not found an automation project, you have found an undocumented policy dispute, and the project will become a policy project whether or not you planned for one.

The workable middle is work with enough rules to be worth encoding, stable enough to be written down in an afternoon, and boring enough that two experienced staff independently produce the same answer.

> **Run this before you brief a vendor.** Take ten recent instances. Give the same ten to two experienced people separately, without letting them confer. Compare the answers.
>
> The disagreement rate between your own experts is the ceiling on what any agent can be measured against. If two people who have done this for years differ on three of ten, then "correct" is not a well defined term in this workflow, and your acceptance criteria have to say so in writing before anything is built.

Disqualifier. The rules exist but two departments hold different versions of them. Resolve that first. It is cheaper as a meeting than as a deployment.

Test three: tolerance for error

The question. When this is wrong, who finds out, how quickly, and what does it cost to put right?

Put your candidates on two axes: cost to detect an error, and cost to correct one. The quadrants behave very differently and only one of them is a good first project.

	CHEAP TO CORRECT	EXPENSIVE TO CORRECT
CHEAP TO DETECT	Ideal first project. Supervision is light and errors are recoverable.	Workable, with a hard approval gate before anything commits. Cost the reviewed path, not the unreviewed one.
EXPENSIVE TO DETECT	Avoid. Errors accumulate quietly and you learn about them from a customer.	Do not start here. This is what a mature deployment earns the right to.

This is the test that supervision exists to answer. A supervised agent proposes and a person approves, which converts an unbounded error risk into a bounded review cost. That is a trade worth making, but it is a trade: the review cost is real and belongs in your business case. Any ROI figure calculated on the unreviewed path is describing a system you are not going to run.

Ask explicitly whether a wrong output has a regulatory or contractual consequence. If it does, the answer is not no. The answer is a hard gate, a complete audit trail, and a return calculated on the reviewed path.

Disqualifier. Nobody on the team can describe what a wrong answer looks like. If they cannot describe one, they cannot review one, and supervision is theatre.

Test four: who owns the outcome

The question. Name the person whose weekly or monthly number changes if this works.

If the answer is a committee, the project has no owner. If the answer is the IT department, the project has a sponsor but still no owner. The owner is the person who will be asked, in three months, why the figure did or did not move, and who would find that question fair.

Then two follow ups that reveal more than the first question does.

- **Does that person already see the number?** If they have to ask somebody to pull it, you have a measurement project sitting in front of your automation project, and it has to be done first regardless.
- **Does the team doing the work report to that owner?** Automation that saves time for one team while a different team absorbs the review burden fails for organisational reasons that have nothing to do with the technology, and no amount of accuracy fixes it.

Disqualifier. The owner is not in the room when scope is agreed. If they cannot spare an hour to define the metric, they will not defend the result.

Scoring your candidates

Score each candidate zero to three on all four tests, then apply two rules that matter more than the totals.

1. **Any zero is a veto.** Do not average it away. A zero on ownership is not compensated by a three on volume; it is a different kind of problem.
2. **Break ties on time to a readable result.** Between two candidates with the same total, take the one where you will know the answer sooner. You are buying information first and hours second.

A candidate scoring nine or better with no zeros is a reasonable first project. A candidate scoring twelve is rare, and if you have one, start there this quarter.

The three most common wrong answers

In scoping conversations the same three candidates come up, and all three are chosen by a criterion that is not on the list above.

- **The biggest cost centre.** Chosen because the saving would be largest. It is also usually the most political, the most heterogeneous, and the most likely to have four different ways of doing the same task in four offices.
- **The thing an executive mentioned.** Sponsorship is genuinely useful and worth having. It is not evidence that the workflow scores on any of the four tests, and it is a poor substitute for the volume number.
- **The most technically interesting one.** The workflow with the most interesting problem in it is the one most likely to produce an interesting failure. Save it for second.

An honourable mention goes to the workflow that happened to match a vendor's demonstration. A demo is a statement about what the vendor has already built, not about where your money is.

What to do with the winner

Write the baseline before anything is built. Not after the pilot starts and not from memory: the current number, how it is measured, who agrees that is the right measure, and the condition under which you would stop. A baseline agreed afterwards is an argument, and it is an argument the vendor usually wins, which is not in your interest either.

The baseline worksheet in this library is the same document we fill in during week one of a pilot. It takes about twenty minutes with the right person in the room, and it is worth doing even if you never engage anybody, because a workflow whose baseline cannot be written down is a workflow whose improvement cannot be claimed.

NorthCape Technology, Victoria, Australia. northcape.tech