A DAY-BY-DAY GUIDE FOR THE PILOT YOU SCOPED IN EDITION 01

The 30-Day AI Pilot

From a stated claim to a written decision in one working month. What to do on each day, what to watch for, and when to stop.

AUTHOR
Ted Milton
Founder, Expedient AI

EDITION
July 2026
Version 1.0

FORMAT
A4 print edition
12 pages, measure sheet included

Expedient AI Ltd

INDEPENDENT · BRISTOL · UK

1 THE CORE DISTINCTION

A pilot is a test. It is not a quiet launch.

The difference between learning something and committing to something.

A pilot is a time-limited test of a stated claim. Not a soft launch dressed up as a trial. Not a procurement with extra steps. The point is to find out whether the claim holds under real conditions, with a real reviewer, before any money changes hands for production use.

You must state the claim before you start, in writing, in one sentence. The sentence will not survive contact with the test. That is the point. A claim that survives is worth pursuing. A claim that does not is worth knowing about on day 12, not in month seven.

The difference between a pilot and a procurement

A PILOT	A PROCUREMENT
Tests a claim, with a deadline and a stop rule.	Commits to a deployment and writes the contract.
Costs days of staff time, not licence fees.	Costs a yearly subscription and integration effort.
Ends with a written decision memo.	Ends with an invoice and a renewal date.
Reviews output against an agreed checklist.	Reviews output against vendor promises.
Has an explicit stop condition before it starts.	Has a renewal clause, which is rarely the same thing.
Limits volume, scope and permissions.	Goes firm on the system of record.

The single most common mistake

Treating the pilot as a polite way to start using the tool. Once staff have changed their habits, you cannot run a clean test. The pilot exists to preserve the option of saying "no, this is not for us" on day 30 without breaking anything.

> **One stated claim.** One deadline. One written decision. Everything else is optional.

2 THE READINESS GATE

Before day one. Get the room in order.

You cannot test a process you cannot describe.

A pilot is meant to test AI. It will not do that if you are also testing your own data, your own instructions, or your own willingness to review. Those have to be settled before day one. Otherwise the test becomes about everything except the claim.

The readiness gate

You are ready to begin day one if you can answer "yes" to every line below. Edition 01's readiness worksheet covers the same ground in printable form.

- You have a one-sentence written claim that names the input, the output and the benefit.
- You have a named human owner of the result, not just a project manager of the pilot.
- You have at least ten representative examples of the input, with the correct output alongside them.
- You have data classified to the four tiers in Edition 03, and the tier that will enter the system is known.
- You have instructions written down in the form described in Edition 04, not a sentence in someone's head.
- You have an agreed review checklist and a named reviewer who has read it.
- You have a stop rule. You know what the pilot result must beat to continue past day 30.
- You have written permission, or a written decision that permission is not needed.

Who to invite to the kick-off

Four people, no more. The cost of a longer room is that nobody owns the result.

The process owner. The person who does the work today. They will recognise good output faster than anyone.

The accountable manager. The person who signs off the result. They will recognise the cost of a bad output faster than anyone.

The reviewer. The person who will compare each draft to the agreed checklist. Different from the process owner, ideally.

The pilot sponsor. One person senior enough to extend the deadline, change the scope or stop the test. Quiet authority is fine.

> COMMON TRAP Starting the pilot on day one because the calendar says so. If three of the eight readiness items are unclear, push day one by a week. The lost week costs nothing compared to a polluted test.

3 MEASURE WHAT ALREADY HAPPENS

Days 1 to 5. Measure the current process.

You cannot improve what you have not recorded.

The first five days are not about AI. They are about the process the AI is supposed to help. Without a baseline, the day-30 decision has nothing to compare against. A claim such as "save four hours a week" is unverifiable unless you can show what the week currently takes.

What to measure

- The time taken per item, end to end. Not a self-report. An actual sample, timed.
- The number of items handled per week. Count from a system, not from memory.
- The error rate and the rework rate. What proportion needed a second pass?
- The hidden work. Chasing missing information, fixing templates, re-doing things because the input was unclear.
- The reviewer's decision. Did they accept the work, edit it, or reject it?

What to classify

Sort the inputs into the data tiers described in Edition 03. The tier that enters the system on day 6 must be agreed in writing. If any of the inputs are Confidential or Restricted, the pilot does not start until the data handling is settled. There is no useful version of "we'll figure that out later."

What to write

Write the instructions as if a new starter is going to do the work tomorrow. Six parts, in the form described in Edition 04: role, input, output, examples, prohibitions, flags. A draft at the end of week one is normal. A finished version by day 5 is required.

What to agree

The review checklist. The two-minute rule from Edition 04 applies: if reviewing takes longer than doing the work manually, the instructions are not yet good enough. Fix them now, not on day 20.

END OF DAY 5 CHECKLIST

Baseline recorded for ten real items. Data tier agreed. Instructions written. Review checklist signed.

4 TEST WITHOUT TOUCHING LIVE WORK

Days 6 to 10. Run the system on safe data.

Compare outputs to baselines before anything reaches a customer.

The offline test uses synthetic, public or historical data only. No live work, no customer names, no live inbox. The point is to find out whether the system can produce output that passes the review checklist at all. If it cannot on safe data, it will not on live data.

How to run the offline test

Prepare ten test inputs. Take them from the baseline sample, redact any tier-3 or tier-4 detail, and keep the structure.

Run the system. Use the instructions as written. Do not improvise mid-test. If the instructions are unclear, fix the instructions, not the test.

Capture every output. Save them in a folder dated by day. Resist the urge to keep only the good ones.

Review against the checklist. The reviewer marks each item pass, edit or reject. Time the review.

Document failure modes. When something fails, write down what failed and why in one sentence. Patterns matter more than individual errors.

Revise the instructions. If a failure is caused by the instructions being vague, fix the instructions and re-run the affected items.

What you should have by day 10

A pass rate. The proportion of outputs that pass review without edit. Under 50% means the instructions need more work. Above 80% means the system is doing what was claimed.

A failure list. Specific, repeatable categories. "Vague tone" is not a failure mode. "Confuses dates in d/MM/yyyy with MM/dd/yyyy" is.

> **STOP AND RESET** If the system cannot pass 50% of items on safe data after two revisions of the instructions, the claim itself is wrong. Stop the pilot. The instructions are fine. The use case is not.

5 TOUCH LIVE WORK, IN SMALL DOSES

Days 11 to 20. A small portion of eligible work.

Every output reviewed, every output recorded.

Days 11 to 20 are when the system meets real inputs. Not all of them. A small, well-chosen portion. Every draft is reviewed by the named reviewer before it leaves the building. Every review decision is recorded. Twice a week, the four of you sit down and look at what happened.

The eligible work

Decide in advance which items the system sees. The rule is simple: items the reviewer would be comfortable sending without a second look. Anything sensitive, anything contentious, anything with a customer in a difficult moment: out of scope for these ten days. The point is not to test courage. The point is to test the claim.

What to record per item

FIELD	WHY IT MATTERS
Item reference	Traceability.
Input type	Patterns in failures often hide here.
Reviewer decision	Pass, edit or reject. The whole pilot depends on this column.
Edit type	Why it was edited. If "tone" dominates, instructions need rewriting.
Time saved	Estimate per item. Do not try to be precise. Consistent estimates across the team matter more.
Side effect	Anything unexpected. The customer replying faster than usual. The data not being where it should be.

The twice-weekly review

Forty-five minutes on a Tuesday and a Friday. The process owner, the reviewer, the manager. The agenda is fixed: failure modes, edit patterns, time saved, anything unusual. Three questions: are we continuing, are we adjusting the instructions, are we expanding the eligible work?

> **Three signals to keep watching.** Reviewer fatigue. Hidden work growing. Instructions producing inconsistent output across similar inputs. Any one of these on its own is fixable. All three together usually means the claim is wrong.

6 PUSH THE EDGES OF THE ELIGIBLE WORK

Days 21 to 27. Expand only if the measures hold.

Push the edges. Test the difficult cases.

By day 21, you have a baseline, an offline pass rate, and ten days of live data. The temptation is to declare victory and expand everywhere. Resist it. Days 21 to 27 are about pushing the system into the cases you kept out earlier: the harder inputs, the edge cases, the moments where the answer is not obvious.

What to expand into

- The items you excluded at day 11. The awkward inputs. The incomplete ones.
- The items that came back edited most often. What is it about them?
- The items the reviewer is least confident about. Run the system, see what happens, compare.
- Inputs the process owner would normally handle personally. The cases where judgement matters.

What to calculate

```
Monthly benefit = items handled × time saved per item × loaded hourly cost × realisation rate
```

The realisation rate is the discount for what does not actually happen. If the system saves 12 minutes per item but the reviewer spends 6 minutes checking it, the realisation rate is 50%. Be honest. The number on the day-30 memo has to survive a sceptical reader.

The range, not the point

You do not need a single number. You need a low, expected and high case. The low case assumes the worst pass rate you have seen holds. The high case assumes the best pass rate holds and review time halves. The expected case is somewhere in the middle. The reader of the day-30 memo will trust a range more than a single figure, and they are right to.

> **WATCH THE EDGES** The edge cases are not there to be defeated. They are there to set the realistic upper limit on the system's role. If they fail more than 30% of the time, the eligible work should not include them at scale.

7 END WITH A WRITTEN DECISION

Days 28 to 30. The decision meeting.

Four outcomes. One memo. No ambiguity.

The decision meeting is forty-five minutes. The four people in the room. The agenda is fixed and short: walk through the measure sheet, agree an outcome, write the memo. The memo is one page. It goes to the pilot sponsor the same day.

The four outcomes

STOP

The claim did not hold. Pass rate too low, review time too high, hidden work too large. Stop is a useful outcome. Most pilots that do not stop should not have started.

EXTEND

The claim looks promising but the data is not yet conclusive. Extend for one more month with a tighter scope or a revised claim. Set a new day-30.

REDESIGN

The system can produce useful output, but the instructions, the eligible work or the review process need a substantial rewrite before a full deployment. Run another pilot on the redesign.

ADOPT

The claim held. Pass rate is acceptable, review time is within the agreed limit, hidden work did not grow. Move to a bounded production deployment with the conditions in the memo.

The memo

One page, four sections. The claim, restated. The result, with the measure sheet attached. The decision, with the conditions. The owner of the next step.

WHAT TO BRING

The baseline measurements from week 1. The pass rate from days 6 to 10. The reviewer decisions from days 11 to 20. The edge-case findings from days 21 to 27. The low, expected and high monthly benefit range. The one-page memo draft.

8 WHAT USUALLY GOES WRONG

The week-2 crisis. It is normal.

Three things go wrong between day 10 and day 14. Plan for them.

Every AI pilot we have seen has a difficult moment somewhere between day 10 and day 14. The system that looked promising in the offline test now produces inconsistent output on live work. The reviewer, who was enthusiastic at the kick-off, has stopped volunteering. The hidden work, which the baseline did not capture, is now appearing in the calendar of the person who does the work. None of this means the pilot has failed. All of it means the next decision matters.

Three things that go wrong, and what to do about each

1. Instructions produce inconsistent output

Two similar inputs, two different outputs. The cause is almost always that the instructions left a choice the system could go either way on. Fix it by tightening the instruction, not by running the system again. Rewrite the relevant section in Edition 04's six-part form. Re-run the affected items. If inconsistency persists across three rewrites, the use case is the wrong shape. Stop the pilot.

2. Reviewer fatigue

Reviewers who were alert at day 6 are tired by day 14. They start missing things. They approve drafts they would have edited last week. This is the most dangerous moment, because the reviewer is the only safety net. Either reduce the volume the system is asked to handle, or rotate the reviewer every week. Do not let one person carry the whole load. The pilot is not the reviewer's full-time job.

3. Hidden work increases

The system saved four minutes per item, but it added three minutes of context-setting per item upstream. The process owner has to find the right input, prepare it, feed it in, then deal with the output. That work did not exist before the pilot. It is real work. Capture it in the day-14 review. Adjust the realisation rate in the benefit calculation. If hidden work is over 30% of the claimed time saving, the value case is fragile.

> **The honest test.** If, knowing what you now know at day 14, you would not start this pilot at all, stop. The sunk time is not a reason to continue.

9 WORKSHEET

Pilot measure sheet.

One row per measure. Fill it in during the pilot, not at the end.

This is the sheet that goes to the day-30 decision meeting, attached to the memo. Each row is one claim, with the baseline, the target, the actual result, the evidence and the pass or fail decision. Six rows is usually enough. More than ten means the pilot was too ambitious.

MEASURE	BASELINE	TARGET	PILOT RESULT	EVIDENCE	PASS?
Time per item				Sample of 20 timed items	
Pass rate (no edit)				Review checklist log	
Review time per item				Reviewer diary	
Hidden work added				Process owner estimate	
Error rate				Customer reply log	
Reviewer fatigue				Reviewer self-report	
Monthly benefit range				Arithmetic below	

Benefit calculation

```
Low = items × low time saved × low realisation × hourly cost
Expected = items × median time saved × expected realisation × hourly cost
High = items × high time saved × high realisation × hourly cost
```

ITEMS PER MONTH

LOADED HOURLY COST

REALISATION RATE USED

MONTHLY LOW / EXPECTED / HIGH

10 PRACTICAL NOTES

Practical notes for the next pilot.

Lessons that recurred across the pilots this guide is built on.

These notes are not a substitute for running the pilot yourself. They are patterns observed across pilots run with UK small and medium-sized businesses during 2025 and 2026. Read them, then make your own decisions about what applies to your situation.

What tends to work

- Picking a process the process owner is mildly frustrated by, not a process the manager wishes would go away.
- Writing the instructions as a six-part block before day one, even if they change later.
- Recording the baseline in week one even when it feels slow. The memo at day 30 needs it.
- Running the offline test on data that looks like real work, not on toy examples.
- Keeping the reviewer separate from the process owner for the first two weeks of live work.

What tends to fail

- Starting with the tool the vendor sold hardest, instead of the process the business already hates.
- Treating a 30-day pilot as a polite way to begin using the tool.
- Asking the system to handle the worst items first, to "get them out of the way".
- Letting one person carry the whole review load for a month.
- Promising a single number to the board before the benefit range has been agreed.

If you only do five things

1. Write the claim in one sentence before day one.
2. Measure the baseline, even if the answer is "about 30 minutes per item".
3. Run the offline test on data that resembles the real thing.
4. Review every output for the first ten live days.
5. Write the day-30 memo the same day as the decision meeting.

> **If you want help.** Expedient AI Ltd runs pilots like this with UK SMEs on a fixed-scope basis. If you would value an independent reviewer for the day-30 memo, get in touch.

11 COLOPHON

The series and how to reach us.

Six field guides for UK SMEs. One Bristol-based company behind them.

The series

Six field guides for UK small and medium-sized businesses. Independent, plain-English, written by a practitioner.

01 · Where AI Actually Helps
Choose one worthwhile use case, scope a 30-day pilot and estimate the return before buying software.

02 · The 30-Day AI Pilot
A day-by-day guide to running the pilot you scoped in edition 01.

03 · Your Data, Your Rules
AI and privacy in plain English for UK businesses handling personal data.

04 · Writing AI Instructions That Actually Work
The six-part structure that turns vague prompts into reliable outputs.

05 · The AI Vendor Interrogation Kit
Eight questions to ask before signing a contract with any AI vendor.

06 · Your First AI Employee
A non-technical guide to autonomous AI agents for small businesses.

Colophon

PUBLICATION

The 30-Day AI Pilot

Edition 02 · July 2026 · Version 1.0

Author: Ted Milton, Expedient AI Ltd

CONTACT

Expedient AI Ltd

expedientai.co.uk

> You can use this guide without contacting Expedient AI. If you would value an independent reviewer of your day-30 memo or pilot design, Expedient AI Ltd offers practical AI implementation support for UK SMEs.