

15 · Teams and operations

30-Day AI Pilot Plan

A four-week AI pilot: measure a baseline, run small trials, review quality, then decide on the numbers. With roles, check-ins and a report template.

For: For team leads and business owners who have chosen two or three tasks and want to trial them in an organised way before rolling out.

Level: Intermediate Version 1.0 · 2026-09-21

Why a limited pilot before rolling out?

Bringing AI to a whole team at once makes it impossible to know what worked and what did not. A 30-day pilot narrows the question: a few specific tasks, named people, measures written in advance, and a decision at the end of the month. The pilot moves from "we tried it and liked it" to "we measured it and found this".

The plan rests on three principles: every week ends with a tangible output you can show, the team works on its real material rather than generic examples, and every output passes a human eye before it reaches another human.

Before you start you need two things: a list of two or three tasks chosen from your AI use map (resource 13), and a tool that passed the privacy conditions you set (resource 14). Without them, week one disappears into debate.

Roles: who does what?

Role	Responsibility	Expected time
Sponsor	The decision-maker. Approves the pilot scope and success criteria and makes the final call on day 30.	Kick-off and decision meetings
Pilot lead	Coordinates the plan, gathers the numbers, runs the weekly check-ins and writes the report.	Two to three hours a week
Participants	Two to five people who actually do the tasks with the tool and log time and notes.	Within their normal work, plus minutes to log
Reviewer	Reviews a sample of outputs against a fixed quality rubric and decides what is fit to use.	One to two hours a week
Data owner	Makes sure what is pasted into the tool respects the agreed rules and follows up any incident.	A short weekly check

In small teams one person may hold two roles, but do not combine participant and reviewer for the same task: whoever wrote the output should not be its only judge.

The four weeks

1. **Week 1: baseline and preparation.** Participants do the chosen tasks the current way, with no tool, and log the time for each instance. The reviewer writes a three-to-five-point quality rubric for each task. The team writes the first instructions (prompts) for each task and stores them in a shared place. The kick-off meeting is held and success criteria are written down. *Output:* a baseline table, a pilot charter signed by the sponsor, and a first prompt library.
2. **Week 2: small pilots.** Participants start using the tool on the chosen tasks only, logging for each instance the time with the tool, the review time, and whether the output needed rework. Mid-week, a short session improves the prompts: which mistake keeps repeating? *Output:* a usage log of at least ten cases per task and an improved second version of the prompts.
3. **Week 3: quality review.** The reviewer takes a random sample of outputs and scores them with the same rubric used on week-one work, ideally without knowing which were made with the tool. Errors are logged by type: wrong information, unsuitable tone, weak Arabic, something important left out. *Output:* a quality comparison table and a list of recurring errors with how to avoid them.
4. **Week 4: measured adoption and decision.** Use continues with the improved prompts, and time is measured again, because the first week with any tool is usually slower. The pilot lead runs the numbers through the time and value calculator (resource 16), writes the report, and the decision meeting is held. *Output:* the pilot report and a written decision: adopt, adjust and extend, or stop.

What do we measure?

- **Net time per instance:** time working with the tool + review time + rework time, compared with the baseline. Never compare tool time alone.
- **Quality:** the reviewer's average score on the same rubric before and during the pilot.
- **Rework rate:** how many outputs out of ten needed a major edit or a rewrite.
- **Errors caught:** their number and type. A high number early on is not failure; it shows the review is working.
- **Actual use:** did participants really use the tool on the agreed tasks, or drift back to the old way? And why?
- **Data incidents:** any time something was pasted that should not have been. The target is zero; any incident is logged without blame and the procedure is adjusted.

Check-ins: four kinds of meeting are enough

Meeting	When	Length	Questions
Kick-off	Day 1	45 minutes	Which tasks? Who takes part? What are the success criteria? What must never be pasted? When do we decide?

Meeting	When	Length	Questions
Weekly check-in	End of weeks 1, 2 and 3	30 minutes	What did we log? What broke? Which prompt do we improve? Was there a data incident?
Mid-point review	Around day 15	Within the check-in	Should we stop a task that is not working and focus on the others?
Decision meeting	Day 30	60 minutes	What do the numbers say? What is the decision? Under what conditions? Who follows up?

Pilot charter template

Pilot charter (written in week 1)

Pilot name: [e.g. draft replies to customers]

Sponsor: [name] | Pilot lead: [name] | Reviewer: [name]

Participants: [names or roles]

Tool: [name and version/plan] | Start date: [] | Decision date: []

Tasks in the pilot:

1. [task] — baseline: [] minutes per instance — frequency: [] per month
2. [task] — baseline: [] minutes per instance — frequency: [] per month

Explicitly out of scope: [tasks or data the tool will not be used for]

Never paste into the tool: [customer names, phone numbers, financial data...]

Success criteria (written now, not after the results):

- Net time per instance falls by at least []%
- Average quality stays at or above [] out of 5
- Major rework in fewer than [] out of 10 outputs
- Zero unresolved data incidents

Stops the pilot at once: [e.g. a sensitive data incident, a complaint caused by an output]

Sponsor signature: []

Weekly log template

Weekly check-in log

Week: [1 / 2 / 3 / 4] | Date: []

Cases done with the tool: task 1 [] | task 2 []

Average time per instance (tool + review + rework): task 1 [] | task 2 []

Outputs needing major rework: [] of []

Most frequent error this week: []

Change to the prompts: [what changed and why]

Data incidents: [none / short description and what was changed]

Blocker needing a sponsor decision: []

Next week's step: []

Pilot report template

30-day pilot report

1. Summary in three sentences

[what we tried, what we found, what we recommend]

2. Scope

Tasks: [] | Participants: [] | Tool: [] | Period: [from ... to ...]

3. Results against the baseline

Task | Baseline (min) | With tool + review + rework | Difference | Quality before/after

[] | [] | [] | [] | [] / []

4. Monthly estimate (from the time and value calculator)

Net hours saved per month: [] | Estimated monthly value: []

Assumptions: [hourly value used, monthly volume, what was not counted]

5. Quality and errors

[types of error, how they were caught, what the tool is not fit for]

6. What we learned about the way we work

[prompts that worked, how review changed, what participants said]

7. Limits and caveats

[small sample, slower first week, tasks not tried, estimated figures]

8. Recommendation and decision

[adopt / adjust and extend 30 days / stop] — conditions: [] — owner: []

A complete example: a pilot at a distribution company

Illustrative example

"Al-Marfa Distribution" (a fictional company) chose two tasks from its use map: draft replies to repeat customer questions, and summarising the weekly meeting. The sponsor is the general manager, the pilot lead is the operations coordinator, the participants are the two customer service officers and the admin assistant, and the reviewer is the sales manager.

Week 1: the customer service officers logged 60 replies at an average of 6 minutes each. The reviewer wrote a four-point rubric: correct information, complete answer, tone, sound Arabic. A written rule: no customer name or number is pasted; messages are written as "the customer is asking about...".

Week 2: average time with the tool was 2.5 minutes, review 1.5 minutes, and rework about one minute on average, because three replies in ten gave wrong delivery times. Prompt change: paste the official delivery schedule at the start of each conversation.

Week 3: the reviewer scored 20 random replies. Average quality rose from 3.6 to 3.9 out of 4, and delivery-time errors fell to one in ten after the prompt change. Meeting summaries were weaker: they missed two decisions in two meetings, so it was agreed that the meeting owner writes the decisions and the tool only summarises the discussion.

Week 4: net time per reply was 4 minutes instead of 6 (2 + 1.5 + 0.5). At 120 replies a month, that is about 4 net hours a month. Meeting summaries did not meet the quality criterion.

Decision: adopt the tool for draft replies with review kept in place, and stop the meeting-summary pilot in its current form, to be redesigned later. Condition: a monthly review of a sample of ten replies.

Common mistakes

- **Mistake:** starting to use the tool before measuring the baseline. **Fix:** week one is for measuring the current way, even if it feels slow.
- **Mistake:** writing success criteria after seeing the results. **Fix:** they go in the charter and the sponsor signs them on day one.
- **Mistake:** counting tool time alone and ignoring review and rework. **Fix:** net time = tool + review + rework, compared with the baseline.
- **Mistake:** widening the pilot midway to new tasks because enthusiasm is high. **Fix:** note new tasks for the next pilot and finish the current scope.
- **Mistake:** blaming the person after a data incident, so people stop reporting. **Fix:** log the incident without blame and change the procedure or prompt.
- **Mistake:** treating "stop" as failure. **Fix:** learning that a task does not suit the tool is a useful result that saves money and time.

Completion checklist

- ☐ The chosen tasks are specific and there are no more than three.
- ☐ Roles are assigned, and the reviewer is not the person who wrote the output.
- ☐ The baseline was measured the current way before using the tool.
- ☐ The pilot charter is written and the sponsor signed the success criteria.
- ☐ Every participant knows what may and may not be pasted.
- ☐ Weekly check-ins were held and the log was kept.
- ☐ Quality was scored with the same rubric before and during the pilot.
- ☐ The report was written with its numbers and limits, and a written decision was made with conditions and an owner.

About the author — Mohamad Charafeddine is an entrepreneur from Tripoli, Lebanon: founder of the Passion digital marketing agency, president of Takaful Youth Association, and founder of digital platforms. He trains individuals and teams in applied AI and marketing.

Examples in this file are fictional. Free to use in your work and organisation; credit the source if you republish.
mohamadcharafeddine.com/en/resources/ai-30-day-pilot/