30-day AI pilot plan — fictional example

Scope: three workflows, chosen for lower risk and high repetition

The pilot covers the Chapter 2 weekly report, the Chapter 1 observation briefing and the Chapter 9 closure review. Each uses internal records, with names removed before use, and is produced every week. Thirty days from Thursday 1 April give one manual baseline in the set-up week (Friday 2 April) and three AI-assisted runs of each workflow (Fridays 9, 16 and 23 April). That is enough to find practical problems. It is not enough to establish a stable time saving, and the decision record must say so.

Four workflows are *not* in the pilot. Incident evidence (Chapter 8) contains personal and health information and sits under legal and procedural rules. Subcontractor submissions (Chapter 3) contain contractor documents whose handling needs approval. Multilingual briefings (Chapter 6) need named bilingual reviewers. RAMS review (Chapter 4) depends on a competent person being available for every referral. They can follow when the conditions are met. The toolbox talk (Chapter 5) and the inspection plan (Chapter 7) are left out only to keep the first pilot small.

Controls

- **Tool, account and data.** Only an organisation-approved account, never a personal one. Before the pilot starts, IT and data protection approve the tool and account for this use under your organisation's rules; their questions include the provider's terms, whether inputs are retained or used to train the provider's models, where data is processed, and how long chats are kept. No personal data and no contractor commercial content go in. Replacing names with codes may reduce some identification risks, but it does not make data anonymous, because a record can often be traced back by its context; coded records count as personal data for this rule unless your data-protection adviser says otherwise. This book does not recommend a tool or claim any tool's performance.
- **Review.** Every AI-assisted output is checked against sources and approved by a named person before use. The approval line from each chapter is used.
- **Log.** For every use: the model or tool name and version, the date, the prompt version, the inputs used, the reviewer, and the manual baseline time. Time with AI counts the whole task: preparing and checking inputs, prompting, review, corrections and any rework. One-off set-up and training time is logged separately and reported beside the result. Errors are logged in two groups, each with its severity: **contained** errors, found in review before use, and **released** errors, which reached a user or recipient. A **material** error, defined before the pilot, is one that could change a decision, a reported figure, a status or a safety instruction (an author recommendation; set your own definition).
- **Manual fallback.** If the tool is unavailable, gives an output that cannot be checked, or the inventory of inputs does not reconcile, the task is done by hand that day. The report is never held back because the tool failed.
- **Stop.** Any data-handling breach stops the pilot at once, and the incident is reported under the data-protection procedure. An output that was used without review also stops the pilot, on the day it is found, until the cause is understood and the control is fixed.

Go / no-go criteria, fixed before the pilot starts

Criterion	Go if	Stop or fix if
Data handling	No personal or confidential data entered outside the rules	Any breach, however small
Review	Every output approved by a named person before use	Any output used without review
Accuracy	No material error released to a recipient; every released and contained error logged with its severity	A material error was released, or contained material errors keep recurring
Time	Measured saving of 25% or more on the pilot workflows, whole task included	Saving below 10%, or not measurable
Users	Most users would keep using it	Most users say it creates extra checking

Figure 10.4 — Go / no-go criteria, fixed before the pilot starts. The thresholds are fictional management choices, set before the pilot, not after. *Source: Chapter 10 pilot plan (fictional).*

Measured saving is calculated outside the AI tool: (total manual baseline minutes − total AI-assisted minutes for the whole task) ÷ total manual baseline minutes, over all logged runs, with each AI-assisted run paired with the manual baseline for the same workflow. Set-up and training time are reported beside it, so that the reader can see what the saving cost to achieve.

Between 10% and 25% is a third outcome: **extend and adjust**. The thresholds (25%, 10%) are management choices for this fictional pilot, not industry benchmarks. Set your own with your approver, based on what the time is worth and what an error would cost. Weigh errors by severity, not only by count: one released material error matters more than several minor errors caught in review, although a high rate of contained errors is itself a sign that the review is carrying the risk. What matters is that the criteria are written down now, so that the result cannot be made to fit.