

F09 / PILOT DESIGN AND EVALUATION

Make the pilot answer a decision

Predeclare the comparison, evidence and guardrails so the result can justify a clear next scope.

SESSION	TEAM	OUTPUT
85 min plan + 35 min review	Ops + evaluator + risk owner	A documented next decision

01 WRITE THE PROTOCOL BEFORE THE RUN

Question + population + unit + thresholds + guardrails

Current practice

A competent comparator

Proposed intervention

A fixed version and scope



Compare quality, effort, recovery and cost.

Keep failures and abstentions. Report denominators and uncertainty.

NEXT DECISION

Stop / revise / extend / limit the release.

A faster result cannot compensate for an unmet consequential guardrail.

A negative or inconclusive result is a legitimate outcome. Sources: page 5.

PILOT PROTOCOL

Write the decision before the run

Predeclare the comparator, measures, cases and stopping conditions. Keep development material separate from the reported evaluation.

POPULATION / ELIGIBLE UNIT / SUBGROUP

INTERVENTION / MODEL AND DATA VERSION

HYPOTHESIS / NEXT DECISION

COMPARATOR / SIMPLER ALTERNATIVE

PRIMARY MEASURE / UNIT / THRESHOLD

QUALITY / SECURITY / HARM GUARDRAILS

CASE PROVENANCE / PERMISSION / DEVELOPMENT SEPARATION / SUBGROUP COVERAGE

CONTAINMENT / STOPPING EVENT / ACCOUNTABLE OWNER

COMPARISON DESIGN / ORDER EFFECTS / CONFOUNDERS / PREDECLARED EXCLUSIONS

Run length and sample adequacy depend on the decision and consequences.

RESULTS AND NEXT SCOPE

Keep the complete evaluation record

Record every eligible case, including failures, abstentions and manual recovery. Use the CSV log for a larger evaluation set.

Case / subgroup / arm	Result / critical issue	Review + recovery effort	Exclusion / disagreement / change

RESULT / DENOMINATOR / UNCERTAINTY / SUBGROUP LIMITS / GUARDRAILS MET?

NEXT DECISION

SPECIFICALLY AUTHORISED NEXT SCOPE / OPEN CONDITION

DECISION OWNER / DATE

FALLBACK / OPERATING READINESS / NEXT REVIEW

ILLUSTRATIVE EXAMPLE

Time improves while quality blocks release

An invented pilot uses 100 eligible evaluation packages. Another 20 packages are kept in a separate development set.

12**Manual median**

minutes of active review

8**Assisted median**

minutes of active review

3**Critical omissions**

assisted outputs / 100

Review time includes reading, checking and correction. The local predeclared target is at least 20% less time with no increase in critical errors and no unauthorised data exposure.

NEXT DECISION / REVISE

Investigate the critical omissions before any release decision.

A revised version must face a fresh evaluation set. Retain the original failures and examine extraction, presentation and reviewer behaviour rather than reporting time alone.

Even zero observed failures would not establish zero underlying risk. Offline performance does not prove adoption, sustained economics or production safety.

FACILITATION AND SOURCES

Run the session

Prepare the inputs in the research companion. Retain evidence, unknowns and counterexamples with the decision record.

15 MIN	State the decision question Specify the intervention, users, conditions and consequence.
15 MIN	Select the comparator Use competent current practice and a simpler alternative.
20 MIN	Predeclare measures Choose the primary outcome and consequential guardrails.
20 MIN	Prepare representative cases Separate development material and record relevant subgroups.
15 MIN	Choose a comparison design Account for order effects, confounding and correlated cases.
RUN	Operate within containment Keep failures, exclusions, version changes and stopping events.
20 MIN	Analyse the complete result Report denominators, uncertainty and total review effort.
15 MIN	Record the next scope Confirm ownership and fallback before operational expansion.

SELECTED SOURCES / ACCESSED 16 SEP 2026

NIST / AI RMF 1.0 / January 2023

Osterwalder / Test Card / Strategyzer, 2015

NIST/SEMATECH / Confidence intervals for proportions

Original pilot and release record informed by established experimentation. Not field validated. A small pilot cannot establish rare-event safety, long-run economics or sustained adoption. Zero observed failures does not mean zero risk. No Test Card layout is reproduced.

Full ledger and access limits: [companion.html](#). Original dotSuper artwork. Reference links identify the underlying sources, not endorsement.