AI Pilot Evaluation Scorecard

Test usefulness, accuracy, exceptions, and human effort before expanding a pilot.

For: Pilot owners, evaluators, and domain specialists | Use when: Before rollout and after material changes | Suggested time: 30 minutes

Version 1.0 | October 5, 2026

Use synthetic or authorized information. Keep sensitive work in approved tools. A named person owns each decision. These tools support discussion and planning; they are not validated assessments.

1. Define the test before running it

Build a representative set across ordinary, incomplete, contradictory, out-of-scope, and failure cases. Keep expected behavior separate from the system response.

Promise

What behavior must the product demonstrate?

Expected behavior

What should happen for this case?

Test coverage

What cases could expose an important weakness?

Reviewer and evidence

Who checks the result, and what will they inspect?

---

2. Measure the whole workflow

Include review time and corrections. Use the same case mix and unit for baseline and pilot. A successful demonstration is not evidence of readiness for all cases.

Useful result

Did the output help the user complete the task?

Quality

Which claims were unsupported, stale, or wrong?

Human effort

How much active time was required end to end?

Exceptions

Did every defined high-consequence case reach a person?

---

3. Decide what happens next

Declare continue, revise, expand, or stop criteria before viewing results. Re-test failed cases and relevant neighboring cases after a fix.

Continue / expand criteria

What evidence permits a larger scope?

Stop condition

What requires pausing?

Decision record

State the decision, limitations, owner, and next review.

Completed companion example

All organizations, workflows, numbers, and results in these examples are fictional and provided for learning. Targets are proposed decision criteria, not validated standards or promised outcomes.

Promise

Provide supported policy explanations or route the case to a person.

Expected behavior

For an individual eligibility question, withhold a decision and escalate.

Test coverage

Routine question; missing policy; stale document; conflicting passages; restricted-data request; service outage.

Reviewer and evidence

Policy specialist compares the draft to the approved source, version, and escalation rule.

Useful result

Record accepted, revised, or withheld drafts and why.

Quality

Count citations checked and claims requiring correction.

Human effort

Include search, review, edits, and exception handling.

Exceptions

Log each missed escalation as a blocking defect to investigate.

Continue / expand criteria

Illustrative: supported answers on routine cases, all defined escalations handled, and lower total active handling time.

Stop condition

Any unauthorized action, restricted-data exposure, or missed eligibility escalation.

Decision record

Revise contradictory-source handling before another synthetic run.

Original guide: https://builtinto.work/resources/pilot-evaluation