Review and Results Scorecard

Check the work and decide whether the workflow is useful

Moxie free guide

A fast draft can still create a slow afternoon. This scorecard helps you check the actual result, count the human effort around it, and decide whether an AI workflow deserves another run.

Use one copy for one workflow and a defined test period. Keep quality, effort, and business outcomes separate. A polished summary does not prove accuracy, and a few minutes recovered do not automatically become cash savings.

1 Define the comparison

Workflow and human owner: _____

Baseline dates and number of comparable tasks: _____

AI trial dates and number of comparable tasks: _____

Unit being compared, such as one weekly report: _____

Required output and intended user: _____

Source systems, filters, and time zone: _____

Known differences between the two periods: _____

Use the same task definition and similar complexity. Note changes in volume, staffing, demand, or the offer. A small pilot can reveal practical problems, but it rarely proves what caused a business-wide change.

2 Define a useful result before testing

Choose a few checks that matter to the work. Do not average a serious error away with a lovely tone score.

Accuracy: What facts, calculations, or records must reconcile? _____

Completeness: Which items or sections must be present? _____

Usability: What must the recipient be able to do with the output? _____

Permission: Which sources, actions, and destinations are allowed? _____

Stop conditions: Which errors or boundary violations require an immediate pause?

_____

Evidence location: Where will the reviewer inspect the saved result? _____

3 Review each completed task

Copy this record for every run:

Run ID and date: _____

Input version and output link: _____

Reviewer: _____

Accuracy: pass / fix / unable to verify

Completeness: pass / fix / unable to verify

Usability: pass / fix / unable to verify

Permission boundaries: pass / fail / unable to verify

Errors found and their consequence: _____

Corrections made and rechecked: _____

Final decision: accepted / held / rejected

Open the actual saved output. Recalculate important numbers from the source and check names, dates, links, and destinations. For a dashboard, confirm metric definitions, filters, reporting periods, and freshness. If sources count different things, explain the difference rather than forcing the totals to match.

4 Record the effort and costs

For each comparable task, capture:

Baseline human minutes, including checking and rework: _____

AI trial human minutes for preparation and prompting: _____

AI trial human minutes for review and correction: _____

AI trial human minutes for handoff and troubleshooting: _____

Elapsed delivery time, recorded separately: _____

One-time setup and training hours: _____

Recurring software and usage costs: _____

Human minutes recovered per task = baseline human minutes minus all trial human minutes.

Period capacity estimate = minutes recovered per task × comparable tasks in the period ÷ 60.

Use negative results when the trial takes longer. Keep one-time setup visible rather than hiding it in a favorable run. If you estimate a dollar value, state the hourly-rate assumption and label it a capacity estimate. Call it cash savings only when spending actually falls, with evidence.

5 Work through a fictional example

A fictional service company tests five comparable weekly summaries. Its baseline averages 40 human minutes per summary. The trial averages 8 minutes preparing inputs, 12 reviewing and correcting, and 5 handing off: 25 minutes total.

That is 15 minutes recovered per summary, or 75 minutes across five runs. Setup took three hours, so those five runs have not yet offset setup time. Allocated software cost was $20 for the trial. One draft had an incorrect total that the reviewer caught and fixed before use. The team keeps review in place and tests the calculation step again. None of these illustrative numbers represents a Moxie client result.

6 Connect the results to a decision

Accepted outputs divided by outputs reviewed: _____ / _____

Most frequent error and likely cause: _____

Net human time change over the trial: _____

Extra setup or recurring costs: _____

Evidence the recipient used the work: _____

Next decision: continue / revise / stop

One change to test next: _____

Owner and next review date: _____

For marketing work, track relevant inquiries or bookings alongside visibility measures. Search impressions, clicks, and AI mentions measure different things. They do not by themselves establish sales impact or prove that AI caused an improvement.

Copy this evaluation prompt

> Review my completed scorecard using only the supplied records.
>
> Check the arithmetic, comparable task definitions, sample sizes, missing costs, and unresolved quality failures. Separate observed results from estimates and possible explanations. Do not treat time recovered as cash savings or claim the trial caused a business outcome without supporting evidence.
>
> Return a concise continue, revise, or stop recommendation, its evidence, unresolved questions, and one bounded next test. If a serious error or permission failure is recorded, make it visible rather than averaging it away.
>
> My scorecard:
>
> [PASTE HERE]

Want help deciding what to measure or where a workflow needs work? Talk with Moxie about your process, review effort, and the result your team actually needs.