Agent Workflow Evidence Kit

Measure accepted outcomes, expose missing data and make a defensible pilot decision.



Design a small paired experiment for an MCP or WebMCP workflow. Choose the decision and scoring rule before the run. The included dataset is fictional and intentionally imperfect; it demonstrates how to report a missing duration without manufacturing a time-saving claim.

Use this kit to make a concrete plan, rehearse one workflow and capture what you learn. The detailed article and editable templates accompany this compact guide.

Open the full article

https://outcomedriven.xyz/blog/mcp-agent-workflow-evidence/

The implementation sequence

01 / Agree on the decision

Complete the hypothesis sheet before seeing results. Define a reviewed task, a realistic baseline and a primary acceptance condition.

02 / Prepare paired cases

Use the same underlying tasks for both conditions. Freeze or record source versions, vary execution order and define the deadline.

03 / Capture the full run

Keep requests, tool results, artifacts, corrections and review notes. Record human interventions instead of crediting them to the agent.

04 / Score and calculate

Apply the rubric to every eligible case. Keep missing values null and label the denominator and subset used for each calculation.

05 / Write a bounded conclusion

Use the memo to state what the evidence supports, its limitations and the next experiment. Retain failed runs when making improvements.

A completed example

What decision will this experiment inform?

Decide whether to expand the evidence-brief assistant to a supervised team pilot.

What counts as one completed task?

One request becomes an evidence brief reviewed for coverage, grounding, currency and permitted sources.

What are the baseline and candidate?

Baseline: current manual research process. Candidate: MCP-assisted retrieval and drafting using the same approved document snapshot.

What is your primary acceptance rule?

Reviewer accepts the brief by the deadline, with all required evidence and no critical factual or access error.

Review the example as a learning aid. It is not evidence that a customer agreed, a deployment succeeded or a result was measured.

The example, continued

Which cases will you include?

Straightforward lookup, multiple sources, conflicting versions, no answer, inaccessible source, ambiguous request and stale record.

Where does the clock start and stop?

Start at request handoff; stop after final review or the trial deadline. Track review minutes separately as a diagnostic.

How will you handle unresolved or missing data?

Unresolved by deadline is not accepted. Unknown completion time is null. Compare timing only on complete pairs and label that subset.

What will happen after the run?

Investigate each critical failure, retain original results, then repeat affected cases plus fresh cases before expansion.

Review the example as a learning aid. It is not evidence that a customer agreed, a deployment succeeded or a result was measured.

Your working worksheet

01 / What decision will this experiment inform?

Name a specific pilot or product decision, not a vague goal to prove AI works.

02 / What counts as one completed task?

Include the user-visible artifact and required human review.

03 / What are the baseline and candidate?

Record sources, model, prompt, tool version and other differences.

Your working worksheet

04 / What is your primary acceptance rule?

Use an inspectable pass condition. Critical errors should not be offset by formatting points.

05 / Which cases will you include?

Include ordinary work and cases likely to reveal weaknesses.

06 / Where does the clock start and stop?

Use the same boundaries and include review, corrections and interventions.

Your working worksheet

07 / How will you handle unresolved or missing data?

Keep unknown values missing. State which denominator each metric uses.

08 / What will happen after the run?

State the evidence needed, owner, limitations and next review date.

Use the kit, then verify the result

1. Keep a working copy

Extract kit.zip. Open index.html in your browser, or edit the Markdown templates in any text editor. The interactive worksheet keeps notes in page memory only. Download notes before leaving.

2. Make the example your own

Replace every fictional person, source, date and assumption with your own confirmed context. Record unknowns explicitly instead of filling them with plausible guesses.

3. Rehearse a normal and an awkward case

Try one known input, one missing input and one result that requires a reviewer. Check the delivered artifact against its source. Keep the manual path usable.

4. Record a dated decision

State what passed, what failed and what you will change next. Count setup, checking and rework when reviewing effort. A completed worksheet is preparation, not measured business improvement.

Sources and next steps

MCP architecture

https://modelcontextprotocol.io/docs/learn/architecture

WebMCP browser API (experimental)

https://developer.chrome.com/docs/ai/webmcp/imperative-api

Your detailed companion article

https://outcomedriven.xyz/blog/mcp-agent-workflow-evidence/

Independent educational materials. Product names belong to their owners. These templates may be adapted for your own projects and client work.