

Agent evaluation record

Record expected behaviour, observed traces and release decisions.

A test-case record for actions and evidence

A runnable example of deterministic evidence checks

A completed conflicting-source test

A free worksheet from ailevrg.com/guides/ai-agent-evaluation

Write the expected behaviour before running a test. Use the sample code for evidence checks, then test the actual agent separately.

01 Test case

Job and test identifier

Real, anonymised or synthetic example

Inputs and authorised evidence

Expected result and stopping rule

02 Run

Model, prompt, tools and retrieval versions

Observed steps and tool actions

Selected sources and validity dates

Final result and human intervention

03 Review

Task, evidence and permission findings

Cost, time and retry findings

Failure consequence and severity

Reviewer and required correction

04 Release

Tests covered and remaining gaps

Conditions blocking rollout

Owner and decision

Changes that trigger another run

A conflicting budget test

Invented teaching example. No client outcome is claimed.

Expected

Client A has two authorised current budget records with opposite values.

Return unknown and request a human decision.

Observation in the sample

The retrieval rule detects different values and returns unknown.

No optional model call is needed to resolve the conflict.

Limit

This demonstrates one deterministic rule on synthetic facts. It is not a test of a live client agent.