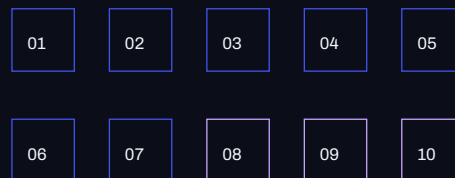


FIELD BRIEFING

Good enough needs a test

Good enough needs a test.



Cases × acceptance rules

What evidence should make us trust a workflow?

A convincing demo answers one question: can the system produce an attractive result on that example? A useful evaluation asks whether it meets the requirements across representative inputs, including difficult cases. The evaluation should describe the work you intend to do, rather than borrow a leaderboard as a substitute.

Explanatory schematics and worked examples are illustrative. No live model, observed client results or measured market performance are presented.

Read the mechanism.

Define acceptance before comparing models or prompts. An invoice extraction task might require the correct total, currency and due date, valid output structure and explicit treatment of missing values. A summary may need all stated decisions, no invented owners and a source reference for each action. Style and factual support deserve separate judgments.

Build a case set from authorised or synthetic material. Include ordinary inputs, edge cases, ambiguous material and deliberate gaps. Preserve expected outputs or a reviewer rubric. Keep development examples separate from a held-out check so repeated tuning does not turn the test into a memorised target.

In an invented ten-case exercise, seven outputs meet the strict rules. An eighth has supported facts but a different structure; it passes only when the formatting rule permits that variation. Two outputs invent commitments and fail under both rules. These counts describe this exercise, not future accuracy. Count failures by type and consequence as well as reporting a percentage.

Use the simplest suitable grading method. Code can check required fields, calculations or exact answers. Humans can review meaning, omissions and practical usability. A model-based judge can help with scale, but needs its own calibration against human judgments and can inherit biases or miss subtle failures.

Repeat enough runs to see instability when outputs vary. Record the model identifier, prompt version, retrieval settings and tool configuration. Comparing two runs while changing all those variables makes the result hard to interpret.

Measure review time and cost alongside quality. A high-scoring draft that requires excessive checking may not meet the workflow's purpose. Re-evaluate after meaningful model, data or configuration changes; a previous pass applies to the system that was tested.

Technical references: Anthropic · Building Evals; Microsoft · Retrieval Augmented Generation; OWASP · Misinformation (2026). Links and source scope appear in the final section.

Try the idea.

INVENTED TEACHING EXAMPLE

Ten cases and two acceptance rules

Seven fictional outputs meet strict structure and support rules. One has supported facts but a different structure. Two invent commitments. Strict acceptance gives 7/10. Permitting the structure variation gives 8/10; both unsupported commitments still fail. These exercise rates are not production accuracy estimates.

Go deeper.

An observed rate needs a denominator

Report which cases were attempted, accepted and excluded. Small synthetic sets explain a method; they do not establish population accuracy or the probability of rare failures.

A judge needs evaluation too

Code graders can be precise for structural or exact-answer checks. Model graders need calibration against trusted judgments, especially when evaluating another model's plausible language.

Change one thing when comparing

Keep prompt, dataset, retrieval, tool permissions and settings recorded. Repeated runs reveal variability. Re-check the system when an important component changes.

Keep the evidence close.

Held-out case

An example kept separate from development to check performance on unseen material.

Rubric

Explicit conditions a reviewer uses to grade an output.

Omission

Important information missing from an answer, even if what remains is correct.

Calibration

Checking that a scoring method agrees with reliable judgments on the task.

Sources · accessed 4 October 2026

[Anthropic · Building Evals](#)

[Microsoft · Retrieval Augmented Generation](#)

[OWASP · Misinformation \(2026\)](#)

Technical and security references support the mechanisms, not our fictional results. KAF's worksheets and examples are original editorial recommendations. No current provider pricing, comparative ranking or legal compliance is asserted. Verify the exact release, configuration and policy for your use.

Practical AI Insights · First editorial edition · 4 October 2026

Good enough needs a test · [Web edition](#)