

FROM TESTS TO RELEASE EVIDENCE

Testing is not a scoreboard.

It is the evidence system that tells a team whether an AI behavior deserves to reach real people.

Swipe for the practical machinery behind a ship decision.

Testing AI

Engineering Confidence in Non-Deterministic Systems

Quality, safety, evals, statistics, and
judgment for the age of AI-generated
systems.

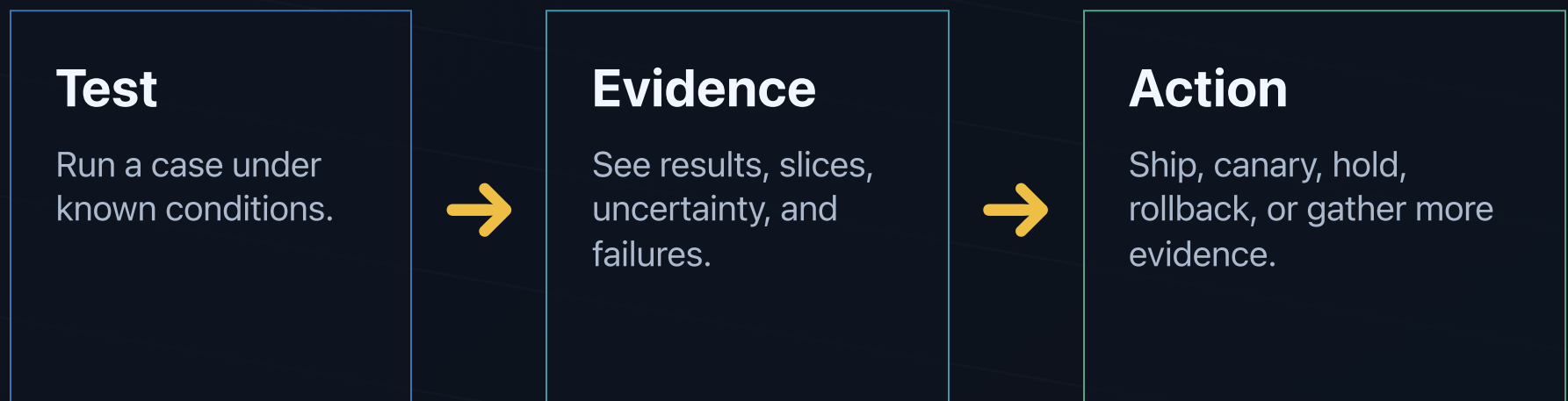


Jason Arbon

First edition - July 2026

THE JOB OF A TEST

A test becomes useful when it drives a **decision.**



A metric without a decision boundary is dashboard decoration.

METAMORPHIC TESTING

When wording changes, the **important property** should not.

Original request

“Can I return shoes after 45 days?”



Same intent, new words

“I bought shoes a month and a half ago. Can I send them back?”

Must remain true: the answer preserves the 30-day policy boundary, even when the sentence changes.

When there is no single right output, test the relationship that must stay true.

MEMORY PLUS REALITY

Keep your **known failures**. Keep meeting new ones.

Golden set

Team memory.

Past regressions, policy boundaries, and the cases that once hurt.



Live sampling

Current reality.

Fresh traffic, new user behavior, changing data, and emerging attacks.

A meaningful new failure becomes tomorrow's regression case.

Golden sets catch regressions. Live samples catch drift.

RISK-BASED SAMPLING

Spend more evidence where being wrong costs **more.**

Low risk

Creative phrasing or a harmless UI suggestion.

Light sample. Fast feedback.

Medium risk

Search relevance, support tone, product recommendations.

Slices and manual review.

High risk

Money, permissions, account actions, private data.

Larger sample. Hard gates.

Irreversible risk

Medical, physical, legal, or autonomous action.

Redundant evidence. Safe fallback.

Equal sampling feels tidy. Users experience consequences, not symmetry.

STRATIFIED REPORTING

An overall average can hide the slice that breaks trust.

Overall

Every evaluated case combined.

8.4 / 10

Common requests

Normal product behavior is strong.

8.9 / 10

Policy edge cases

Small slice. High consequence.

5.8 / 10

Break results down by the categories that matter before you average them away.

RARE FAILURE HUNTING

97% pass can still mean do not ship.

20 evaluated outcomes

19 normal



One privacy leak, unsafe action, or false reassurance can outweigh a great average. Random sampling estimates normal behavior. Targeted tests hunt the failures you cannot afford to miss.

Keep normal-production estimates separate from deliberately risk-weighted probes.

PAIRWISE COMPARISON

When an absolute score is fuzzy, ask a simpler question: **which is better?**

Answer A

Clear and polite. It cites the policy but leaves out the next action.

vs

Answer B

Clear and polite. It cites the policy and gives the user a safe next step.

A reviewer can reliably choose **B** without pretending they can assign a perfectly objective 8.3 versus 8.7.

REPRODUCIBILITY

If you cannot replay the failure, you cannot learn from it.

Input

Prompt, user state, and task.

Model

Version, route, seed, settings.

Context

Retrieved docs, memory, freshness.

Tools

Calls, responses, permissions.

Output

Answer, score, rubric, decision.

Time

When it ran and what changed.

Save the evidence around the failure, not only the failure sentence.

RELEASE GATES

One number is never enough to approve an **AI release.**

Quality

Average, lower confidence bound, and useful coverage are acceptable.

Critical slices

High-risk populations and conditions do not regress.

Hard blockers

No privacy leak, unsafe action, or disallowed policy behavior.

Operations

Latency, cost, monitoring, rollback, and owner are ready.

Release with a canary, keep watching, and keep rollback ready.

A gate is an operating rule for acceptable evidence and risk.

AI-FIRST QUALITY WORK

Tests are inputs. Evidence makes the call.

AI quality is nuanced. IcebergQA helps teams turn testing practice into release evidence for AI-powered products.

[Talk to IcebergQA](#)