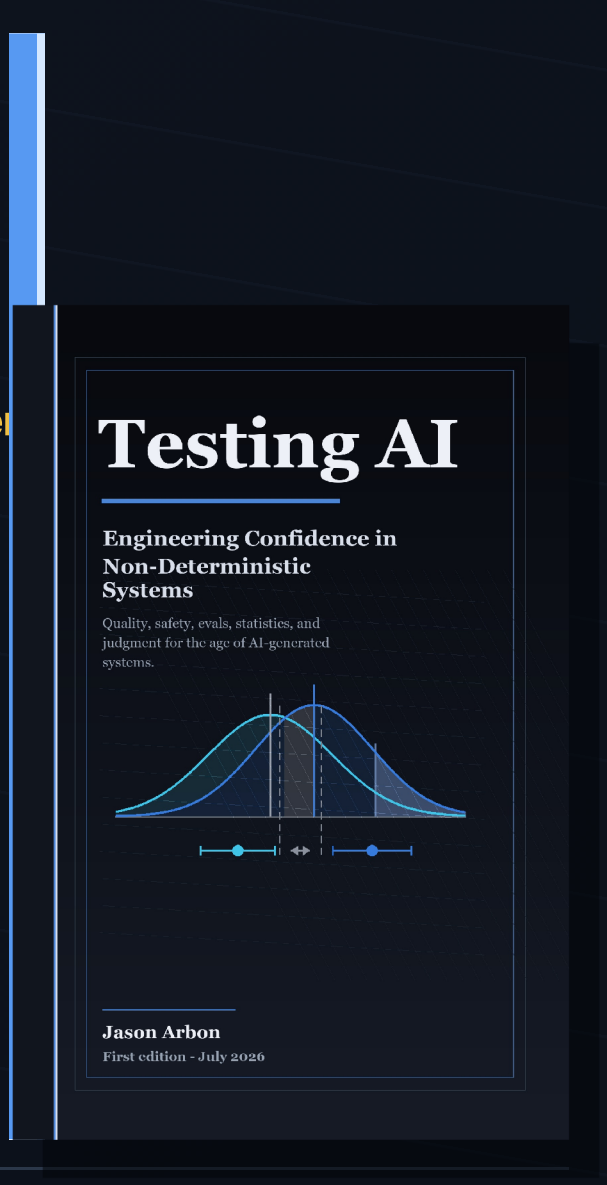
ANTI-PATTERNS THAT CREATE FALSE CONFIDENCE

Old testing habits can produce new lies.

AI quality gets dangerous when a familiar metric, demo, or test ritual is mistaken for enough evidence to ship.

Swipe for the traps, and the confidence-engineering replacer



THE BOOLEAN TRAP

A green light can hide the uncertainty you need to explain.

Pass / fail

PASS

Useful for strict contracts and hard blockers. Too small for ordinary AI behavior.

Risk envelope

EVIDENCE

Frequency, severity, slices, variance, latency, cost, and human-review load.

Keep hard blockers for privacy leaks and unsafe tool actions. Report ordinary quality as a distribution.

PERCENT PASSED IS NOT QUALITY

A 90% pass rate can be a comforting lie.

Flat dashboard

The number says nothing about what was selected, weighted, duplicated, or left out.

- **Easy permutations dominate**
- **Critical failures get averaged away**
- **Risky slices are absent**
- **Test mix changed underneath**

Report severity, blockers, slices, confidence interval, and test composition beside the headline number.

THE GOLDEN ANSWER PROBLEM

One expected sentence is not the whole quality model.

One “correct” answer

Exact-match testing rejects helpful alternatives and can still miss fabricated sources or unsafe next steps.

Measured dimensions

Allow legitimate variation while protecting the properties the product needs.

- correct
- grounded
- safe
- actionable
- appropriate tone

Use exact checks for schemas, policy wording, citations, and irreversible confirmations. Use rubrics and calibrated judgment for open-ended work.

THE BUG-TICKET TRAP

One bad output is evidence of a failure family.

Screenshot bug

"The model said this strange thing."

A vivid example becomes a demand for one surgical patch.

Behavior pattern

Cluster, reproduce, slice, estimate frequency, identify likely layers, and test nearby cases.

Judge the mitigation by distribution movement, not whether the screenshot disappeared.

AI failures can involve prompts, retrieval, tools, policy, labels, context, and model behavior at the same time.

WHACK-A-MOLE TUNING

Fixing the visible failure can create quieter failures nearby.

Patch

Add a prompt rule, new retrieval hint, or fine-tune to hide the embarrassing failure.

Shift

Over-refusal, slower responses, policy confusion, weaker helpfulness, or a different broken slice appears.

Measure

Run the original failure, neighbor cases, benign counterexamples, slices, and holdouts.

> A tuning change needs a blast-radius eval. Celebrate distribution improvement, not one repaired demo.

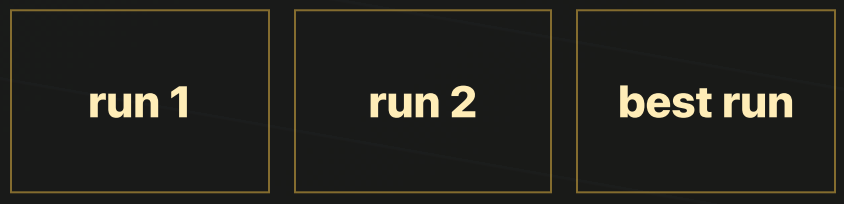
Every fix should explain what improved, what regressed, and what risk still remains.

THE ONE-RUN DEMO FALLACY

A beautiful demo proves what the system can do once.

Demo success

The team retries a prompt, finds the magical output, and shows it to the room.



Release evidence

- Repeated, locked conditions
- Versioned tools and traces
- Representative sampling
- Tail-risk and slice reporting

A demo can inspire investment. It cannot prove reliability, safety, or readiness.

THE AGGREGATE-SCORE TRAP

The average user does not exist.

English support

Improved 4 points.

Spanish recovery

Regressed 11 points.

Low-risk chat

Same score, lower cost.

High-risk account action

More incorrect approvals.

A strong aggregate cannot wash away a failed high-risk slice. Define slice thresholds before you run the comparison.

THE FINAL-ANSWER TRAP

The visible answer is only the end of a larger system.

Retrieve

Was the evidence relevant, allowed, current, and complete?

Plan

Did the agent understand the task and missing information?

Act

Were tool choice, permissions, arguments, and side effects safe?

Explain

Did the final answer faithfully represent what happened?

> Store traces as eval artifacts. A lucky final answer can hide a wrong source, missing confirmation, or unsafe tool trajectory.

Score the path and the ending. Debugging begins when you know which layer actually failed.

THE JUDGE-AS-TRUTH TRAP

An LLM judge is a measurement device, not an oracle.

Uncalibrated verdict

Fluent writing wins. Longer answers win. The score changes with prompt order, rubric wording, and model version.

Calibrated evidence

Compare to humans, inspect disagreements, track bias and drift, and escalate high-risk decisions.

Verify the verifier. A precise measurement system can still be measuring the wrong thing.

IcebergQA

AI-FIRST QUALITY WORK

THE REPLACEMENT

Replace confidence theater with confidence engineering.

AI quality is not more checklists, more dashboards, or more green dots. It is building an evidence system that stays honest about variation, slices, risk, and real-world outcomes. IcebergQA helps teams make that shift.

Talk to IcebergQA

Testing AI by Jason Arbon · Chapter 10: Anti-Patterns That Create False Confidence