

THE CONFIDENCE ENGINEER

The next role is not a job title. It is a discipline.

When AI builds faster than people can inspect, the scarce resource is justified confidence about what should ship.

From Chapter 11: The Confidence Engineer

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 CONVERGENCE

Old roles are converging around one question: **should we trust this?**

Build

Code, prompts, models, and tools.

Test

Evals, regression, and failure discovery.

Product

Value, usability, and taste.

Release

Evidence, gates, canaries, rollback.

Operate

Traces, cost, reliability, incidents.

Explain

Risk, uncertainty, and decisions.

One product. One evidence system.

The title matters less than the work: connect behavior, risk, and consequences to the next decision.

A PRODUCTIVE DIVISION OF LABOR

Use AI for **scale**. Keep people for **judgment**.

AI can multiply the work

- Draft evals and rubrics
- Generate edge cases
- Run repeatable checks
- Cluster failures
- Summarize traces
- Build harness code

Humans decide what matters

- Name the real risk
- Notice the missing slice
- Challenge a seductive metric
- Calibrate the judge
- Apply taste and context
- Accept residual risk

PROBABILISTIC EVALUATION

For AI, quality is a **distribution**, not one winning example.

How often?

Repeated runs reveal the rate of a behavior, not its best moment.

For whom?

Slices expose who benefits, who regresses, and who carries the risk.

How bad?

Severity and tail failures matter more than a tidy average.

Ask what the evidence lets you believe, not only whether one run passed.

THE DELIVERABLE

A dashboard full of activity is not the product. A defensible decision is.

Test theater

Lots of numbers. Green checks. No clear owner. No articulated residual risk.

- What does the score mean?
- Who is missing?
- What would block release?



Decision evidence

Known risks, meaningful slices, calibrated measures, and a clear next action.

- Ship?
- Canary?
- Hold?
- Rollback?
- Collect more evidence?

RISK OVER VERIFICATION

Passing tests can coexist with a product that is still **unsafe, useless, or** untrustworthy.

Safe?

Critical boundaries, abuse paths, and irreversible actions are controlled.

Useful?

The product helps real users complete the job they came to do.

Reliable?

Errors, latency, cost, retries, and fallbacks hold under production pressure.

Fair enough?

Important users and contexts are not averaged out of the release decision.

Evidence is only valuable when it changes a choice: release, constrain, improve, or stop.

SYSTEMS THINKING

Each component can be reasonable while the whole system is **wrong**.

Prompt	Model	Retrieval	Tools	Product
Wrong framing or missing constraints.	Variable output or weak reasoning.	Stale, missing, or unauthorized evidence.	Unsafe calls, retries, permissions, side effects.	Confusing experience, bad action, broken trust.

Trace the path, not only the visible output. System failures live in the joins.

THE SKILLS THAT ENDURE

Confidence engineers are **broad by necessity** and **concrete by habit**.

Automation

Harnesses, regression, telemetry, and repeatable checks.

Statistics

Sampling, variance, calibration, and uncertainty.

AI fluency

Models, judges, tools, traces, and failure modes.

Creativity

Strange but plausible users and future workflows.

Communication

Explain risk without hiding behind jargon.

Judgment

Know what the system cannot decide for itself.

The strongest quality reports are decision artifacts: clear, honest, and useful to the people who must act.

WHAT CHANGES NOW

Generating the first draft is becoming cheaper.
Selecting the survivor is becoming the work.

AI can create overnight

Ten prompts. Twenty UI variants. Hundreds of edge cases. Several product directions. More code than a team can inspect by hand.

Confidence decides what remains

Evaluation, monitoring, cost, safety, value, and real-user evidence determine which candidate deserves real power.

The constraint is no longer imagination. It is trusted evidence about what should survive.

A PRACTICAL OPERATING LOOP

Build the **measurement layer** alongside the product.

Define

Quality, risk, and who matters.

Generate

Cases, variants, and test harnesses.

Measure

Runs, slices, traces, and judges.

Decide

Ship, constrain, canary, or hold.

Learn

Promote failures into future evidence.

AI can accelerate every stage. Accountability for the release decision still needs a clear owner.

THE CONSTRUCTIVE MOVE

Everyone is becoming a confidence engineer.

AI changes who can build. Confidence engineering determines what deserves to ship. IcebergQA helps teams design the evidence, evaluation, and release systems that make AI-first quality practical.

[**Talk to IcebergQA**](#)