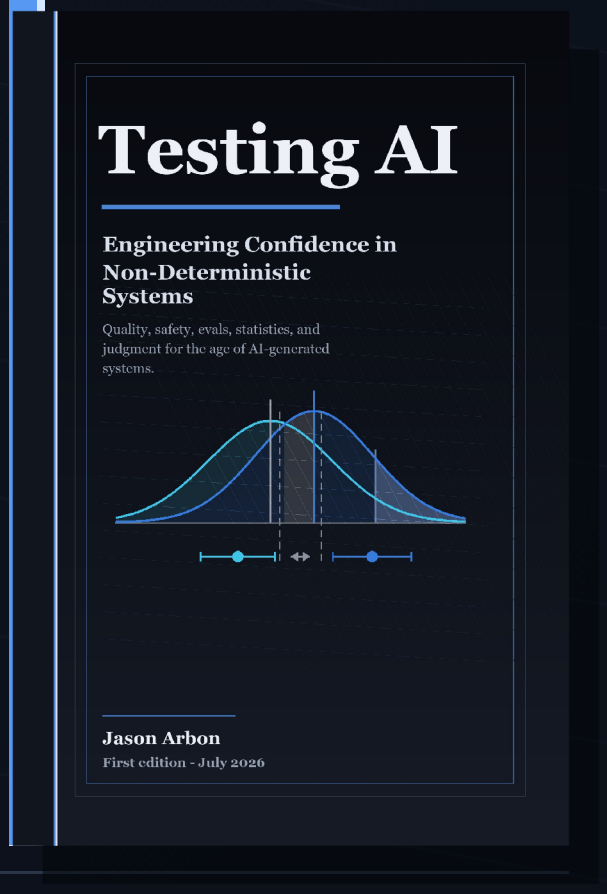
SAMPLING AND UNCERTAINTY

One run is not the truth.

A sample is a small window into future behavior. Confidence comes from looking through it honestly.

Swipe for the evidence behind a responsible AI decision.



ONE RUN VERSUS EVIDENCE

A demo proves possibility.

A representative sample begins to estimate reliability.

ONE RUN

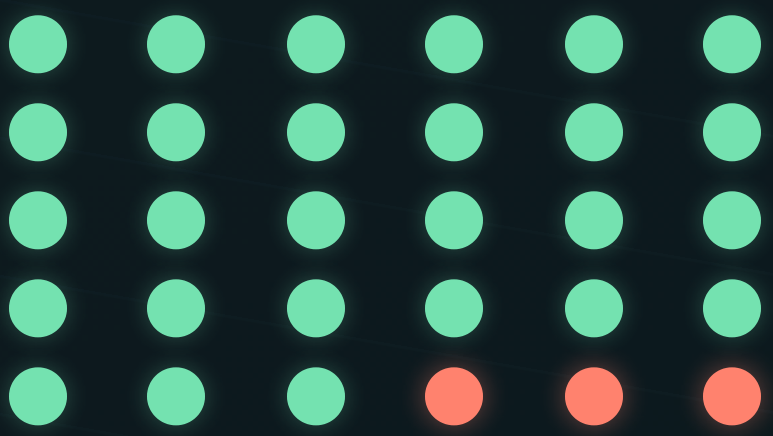
N = 1

One good outcome.
Nothing more.

SAMPLE

REPRESENTATIVE SAMPLE

N = 30



3 high-risk failures.
Now there is a pattern to investigate.

The tail can matter more than a nice-looking demo.

SAMPLE SHAPE MATTERS

More examples do not fix a biased sample.

100 easy cases

Same clean prompt style. Same language. Same happy path.

Measures only what it included.

100 useful cases

Different users, conditions, stakes, phrasing, and failure modes.

Closer to the product's real world.

A thousand easy prompts still measure easy prompts.

EVIDENCE SCALES WITH RISK

There is no magic sample size.

10	**Harness check** Does the system obviously fail?	Fast smoke signal
100	**Rough quality read** Is the direction promising?	Early product decision
1K	**Release evidence** Are important slices stable enough?	Higher-impact change
∞	**Rare-risk hunting** Target the severe failures random sampling may miss.	Privacy, safety, money, autonomy

Match the evidence to the decision and the cost of being wrong.

VARIATION IS THE PRODUCT

Sample the ways users actually show up.

Repeat
Same prompt, many runs.

Phrasing
Typos, slang, incomplete intent.

Context
Fresh, stale, missing, or conflicting data.

People
Languages, accessibility, real needs.

Risk
Rare but unacceptable outcomes.

Time
Production drift and changing reality.

One topic is not one slice. One user is not the population.

DO NOT GRADE WITH ONE NUMBER

The average can hide the user's worst day.

8.4

Mean score

How good was it on average?

8.7

Median score

What did a typical case feel like?

0

Minimum score

What was the worst observed outcome?

4%

Failure rate

How often did it cross a hard line?

Mean, spread, tail, and failure rate answer different questions.

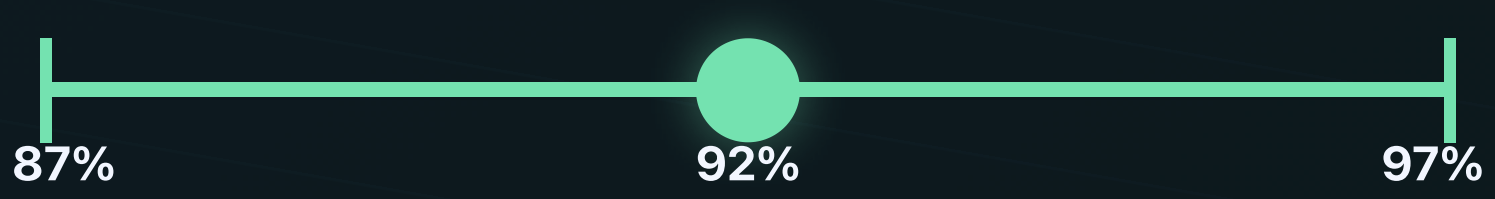
CONFIDENCE INTERVALS

Use about, not false precision.

A sample gives an estimate and a band of plausible values, not the exact quality of the entire product.

92 of 100 cases passed

95% interval



Plain English: **the observed pass rate is 92%, and the estimated underlying rate is roughly 87% to 97% under this method and sample.**

The responsible report names the estimate, method, uncertainty, and population.

COMPARING VERSIONS

82% versus 84% is not automatically a win.

Version A

82%

Its uncertainty may be several points wide.

Version B

84%

Two points higher can still be mostly noise.

> **Better analysis:** run both versions on the same cases, calculate each case's B minus A difference, then put a confidence interval around those paired differences.

Compare the difference, not just two attractive dashboard numbers.

TWO MEANINGS OF CONFIDENCE

A model's certainty is not measured evidence.

Model-reported confidence

“I am 95% sure.”

A self-assessment about one answer. Useful for routing review only after calibration.

Statistical confidence

“Across 100 cases, 92% passed, with an interval of 87% to 97%.”

A calculation from observed outcomes, variation, and sample size.

A confidence label earns trust only when it predicts real accuracy by slice.

THE PRACTICAL OPERATING MANUAL

Do not ask whether one run looked good.

Ask how the system behaves across the distribution of users, contexts, risks, and time that matters.