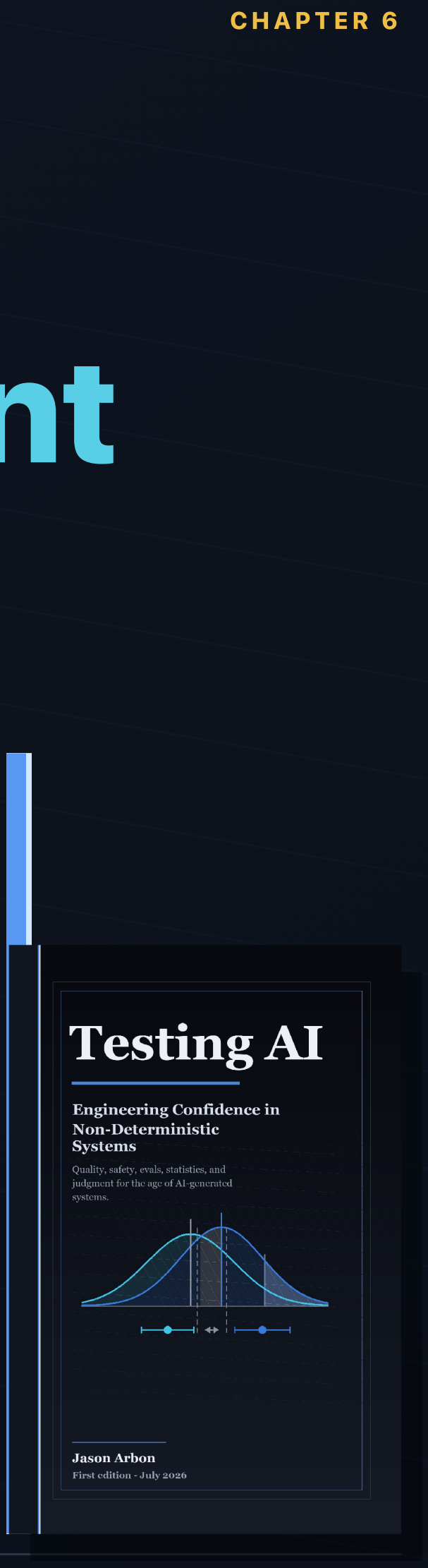
BUILDING EVALS THAT MATTER

An eval is a measurement with a job.

Its task, inputs, judgment method, and score should answer a decision the team actually needs to make.

Swipe for the difference between a benchmark number and product evidence.



THE ANATOMY OF AN EVAL

A score means nothing until you know what created it.

Task

What work is the system trying to do?

Case and context

What input, data, tools, and constraints were available?

Oracle or rubric

What counts as a good, bad, safe, or useful outcome?

Aggregation

How do individual results become a release-relevant signal?

An eval is not “did it say something plausible?” It is a planned measurement of behavior.

PUBLIC EVALS IN ONE GLANCE

Famous benchmarks measure different jobs.

MMLU

Multiple-choice academic and professional questions.

Measures: broad knowledge accuracy

HumanEval

Write a small function, then run executable tests.

Measures: code task pass rate

SWE-bench

Patch a real repository issue against issue-related tests.

Measures: issue resolution

Chatbot Arena

Humans choose between anonymous answers.

Measures: pairwise preference

> The name “eval” covers exams, patches, preference votes, tool workflows, safety probes, and much more.

DIFFERENT TASKS, DIFFERENT ORACLES

There is no single AI score.

A

Exact answer

Useful for a constrained knowledge question with a curated answer key.

</>

Executable test

Useful when a function or patch can be run against behavior checks.

A/B

Human preference

Useful for subjective qualities that resist one exact response.

Accuracy, pass@k, issue-resolution rate, and preference rating are not interchangeable quantities.

BENCHMARKS VERSUS PRODUCT EVIDENCE

A public benchmark is a reference, not a release gate.

Public benchmark

- Shared task shape
- Comparable numbers
- Broad capability signal
- Known blind spots

Your product eval

- Real user intents
- Current tools and context
- Policy and safety boundaries
- Cost, latency, and rollback risk

Borrow the measurement idea. Do not borrow the confidence the leaderboard claims to provide.

CONTAMINATION AND OVERFITTING

A perfect score can mean the system read the exam.

Train against the test set long enough and it stops being a test.

A	B	C	D	100%

The caution is not theoretical. A team can tune prompts, models, and countermeasures against the same held-out cases until the dashboard turns green. **Fresh holdouts protect the meaning of the score.**

Repeated optimization against the eval can turn generalization evidence into exposure evidence.

READ A SCORE LIKE A TESTER

Ask six questions before you believe the number.

What is one unit?

An answer, a conversation, a patch, a ranked list, or a trajectory?

Who are the cases?

Does the population resemble real traffic and real risk?

Who or what judges?

Exact answer, test, rubric, expert, human preference, simulator?

What got averaged away?

Rare failures, slices, cost, latency, disagreement, or churn?

Also ask: could the model have seen the benchmark, and would a better score change a real decision?

FROM BENCHMARK TO ORGANIZATION

A patch can pass a repo and still fail the company.

Benchmark world

- One repo issue
- Known environment
- Tests decide the outcome
- Bounded tool access

Production world

- Private services and credentials
- Canary tenants and rollback
- Permissions and side effects
- Hidden dependencies and humans

A product eval should score the patch and the agent’s investigation, evidence, permissions, and stopping behavior.

BUILD THE EVAL YOUR PRODUCT DESERVES

The same benchmark idea becomes a different eval in each product.

TunedSearch

Queries, ranked evidence, freshness, source authority, and position-aware metrics.

CartCare

Policy checks, tool traces, escalation rules, privacy, and customer risk.

BugPilot

Repo state, permissions, tests, security review, patch scope, and reversibility.

RoseyBot

Perception, authorization, safe motion, recovery, and human override.

A high public score is background evidence. A product-specific eval is a precondition for trust.

LAYERED EVIDENCE

One eval is never the whole story.

Public	Broad capability signals and common measurement patterns.
Product	Private cases tied to real users, tools, data, and decision boundaries.
Adversarial	Tail risks, abuse paths, rare failures, and critical slices.
Live	Monitoring for drift, operational failures, and changing reality after release.

The durable answer is a portfolio of evidence, not a leaderboard screenshot.

Build for the decision. Keep measuring after it ships.

PUT IT INTO PRACTICE

Turn model scores into release evidence.

The hard work is not finding a benchmark. It is building the case population, measurement method, and decision logic your product actually needs. IcebergQA helps teams do that work.

Talk to IcebergQA