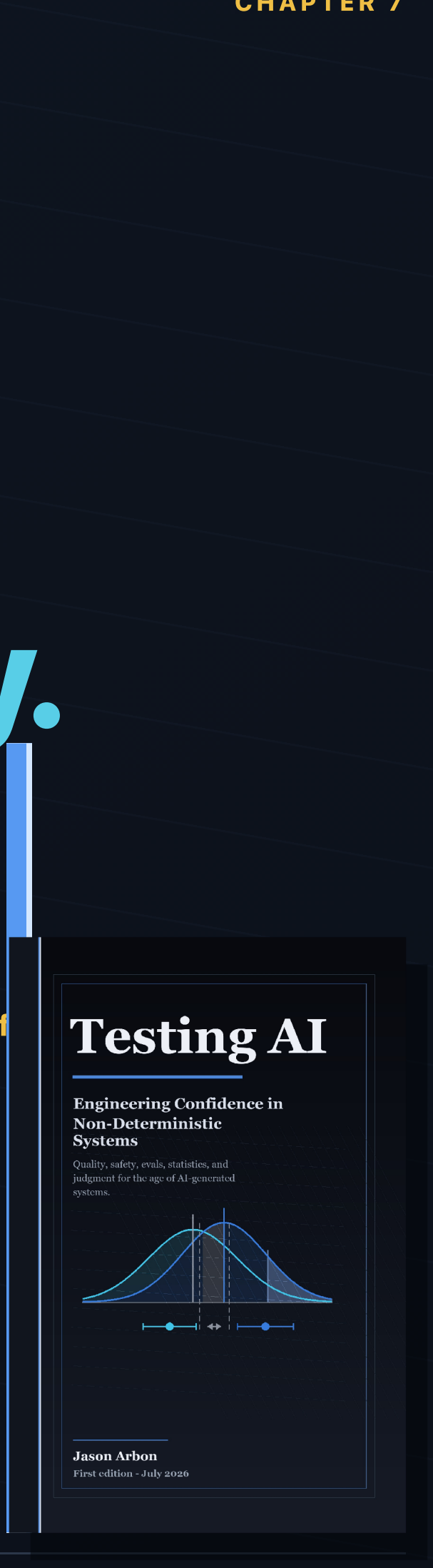
RELEASE READINESS FOR AI SYSTEMS

Release is where testing meets reality.

AI quality does not become fixed at launch. Users, models, tools, policies, and data keep moving.

Swipe for the evidence a system needs before, during, and af release.



OPERATIONAL QUALITY

A no response is a zero-quality experience.

User need

A real task arrives with urgency, context, and expectations.

→

AI product

Model, retrieval, tools, policies, and infrastructure work together.

→

Useful response

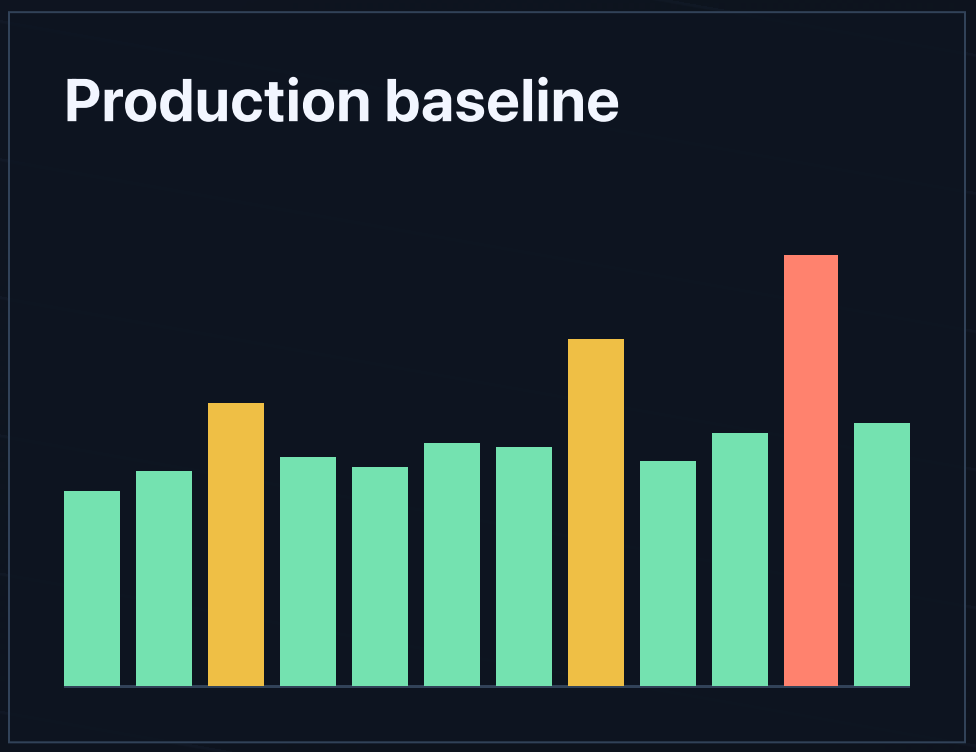
Correct, timely, safe, and available when the person needs it.

> **Availability is quality.** A perfect offline result does not help when an API error, timeout, or dependency failure reaches the user.

Reliability, retries, and deployment health belong beside relevance and safety on the release dashboard.

MONITORING AFTER RELEASE

Measure the signals that tell you behavior changed.



- Sampled output quality
- Latency and cost
- Tool calls and retries
- Safety and privacy failures
- Escalations and complaints

Version the evaluator too. An apparent quality shift can come from the product, its context, or the judge used to measure it.

CANARY USERS

Find the test cases you did not know to write.

Current release

Most users keep the stable route while a team maintains the existing baseline.

Canary release

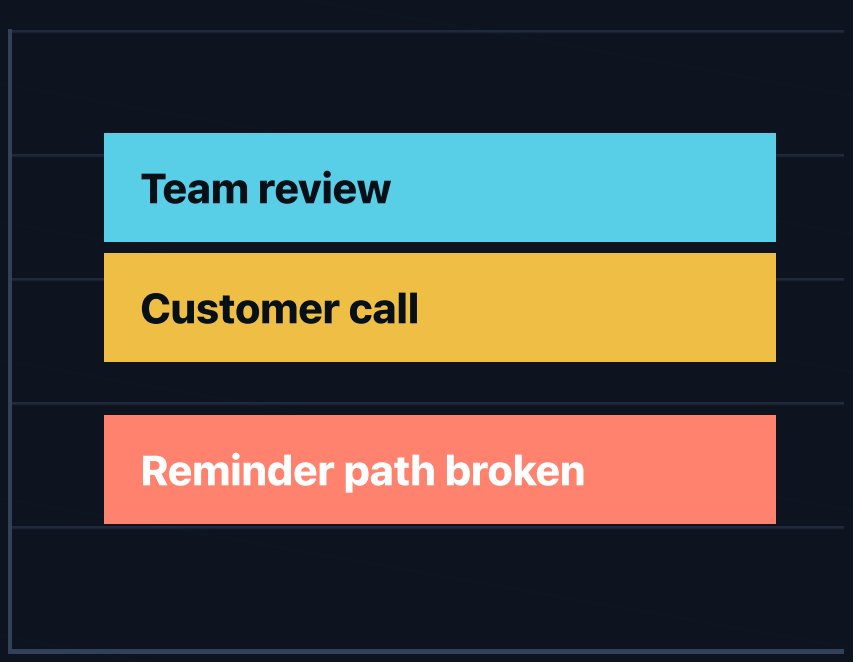
Early adopters see the new path first. Watch their experience before expanding exposure.

A canary reduces exposure. It is not automatically an experiment: define what you will measure and when to stop before rollout.

FROM THE FIELD

Canary users missed meetings first.

Calendar with overlapping events



A subtle browser regression

A JavaScript error while rendering overlapping calendar events broke both the reminder path and the overlapping meeting display.

That exact scenario was unlikely to exist in a pre-release test plan. Early users made it visible while the blast radius was still small.

Use an escalating path: canary, developer channel, beta, then broad production.

TRADEOFFS ARE PART OF QUALITY

A better score may still be the wrong release.

Quality

8.4

A larger model improves the average answer score.

Cost

2x

It doubles the spend per completed task.

Tail latency

12s

It misses the experience target for stressed users.

p95 means 95% of requests finish faster than the threshold. **p99** reveals the slower tail that averages hide.

Choose a route for the risk: cheaper and faster for low consequence; more capable and deliberate for high consequence.

REGRESSION WITHOUT FREEZING WORDING

Protect invariants, not yesterday's sentence.

Version A

> “Your order is delayed. I can check the current delivery status.”

Different wording is expected.

Version B

> “I’ll look at the live delivery status and explain your options.”

The wording changed. The important commitments did not.

- **live evidence**
- **no invented claim**
- **safe options**

Regression is real when facts, policy, grounding, permission boundaries, relevance, or usability get worse.

TOOL-USING AGENTS

Score the trajectory, not only the ending.

Plan

What is needed? What is missing?

→

Tool choice

Is this the right, allowed tool?

→

Tool call

Correct account, scope, arguments, and timing?

→

Outcome

Did it recover, explain, and avoid unsafe side effects?

> A polished final answer cannot redeem an unsafe tool path, leaked data, duplicated action, or irreversible mistake.

Every intermediate observation, retry, permission check, and side effect is testable evidence.

WHAT A TRAJECTORY NEEDS

Test the agent where small errors become real actions.

Permission

Ask first before deletion, purchase, publishing, or broad access.

Observation

Interpret empty, stale, conflicting, and failed tool outputs correctly.

Recovery

Retry safely, clarify ambiguity, or escalate instead of guessing.

Side effects

Prevent duplicated writes, broad edits, secret exposure, and irreversible actions.

Ten dependent steps that are each 99% correct succeed only about 90% of the time before other real-world failures enter the picture.

HUMAN REVIEW IS A CONTROL

Escalate when automation does not have enough authority.

Low risk: automated decision

Uncertain or expensive: sampled review

Sensitive or irreversible: human hold

Reviewer needs evidence

Give a person the full case, not a lonely score.

- rubric and decision boundary
- source material and trace
- prior decisions and failure reasons
- authority to stop the release

Track queue time, reviewer agreement, overturn rate, escalation precision, and the categories where automation keeps getting overruled.

PUT IT INTO PRACTICE

Make release readiness a quality system.

AI products can update their models, prompts, data, and routes dynamically in production. That makes pre-release evidence necessary, and live measurement just as important. IcebergQA helps teams build both.

Talk to IcebergQA