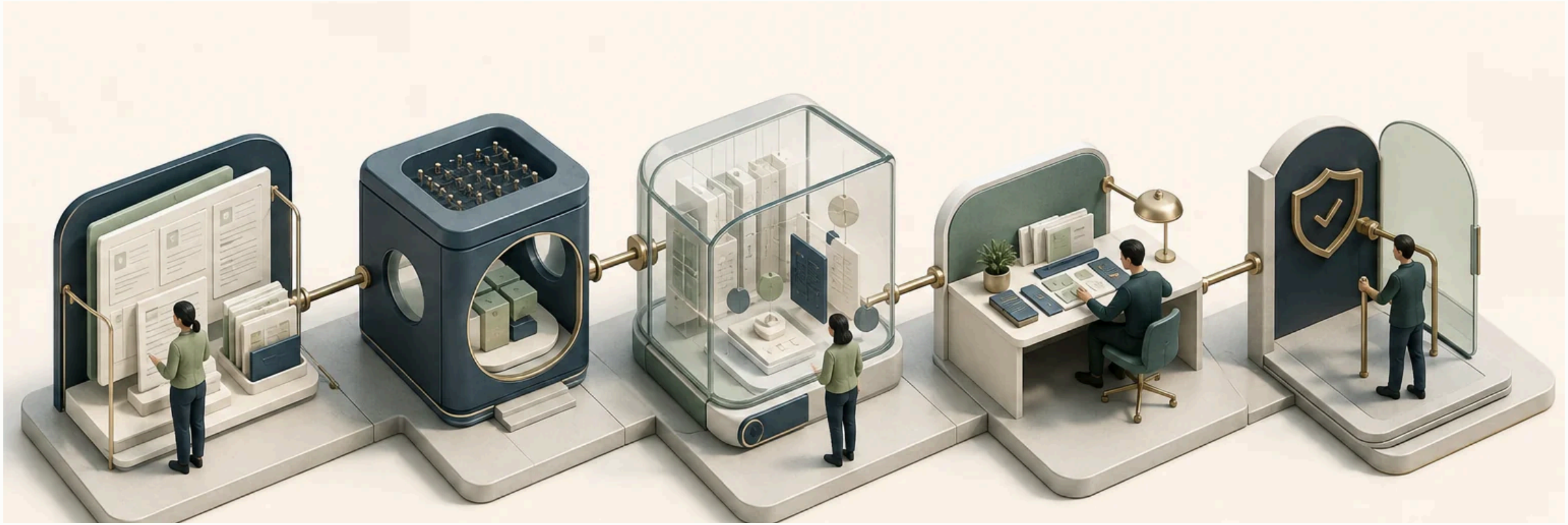
EXECUTIVE · STRATEGIC VISION / SEPTEMBER 2026

# Quality engineering for the AI era

AI × quality engineering · Industry research

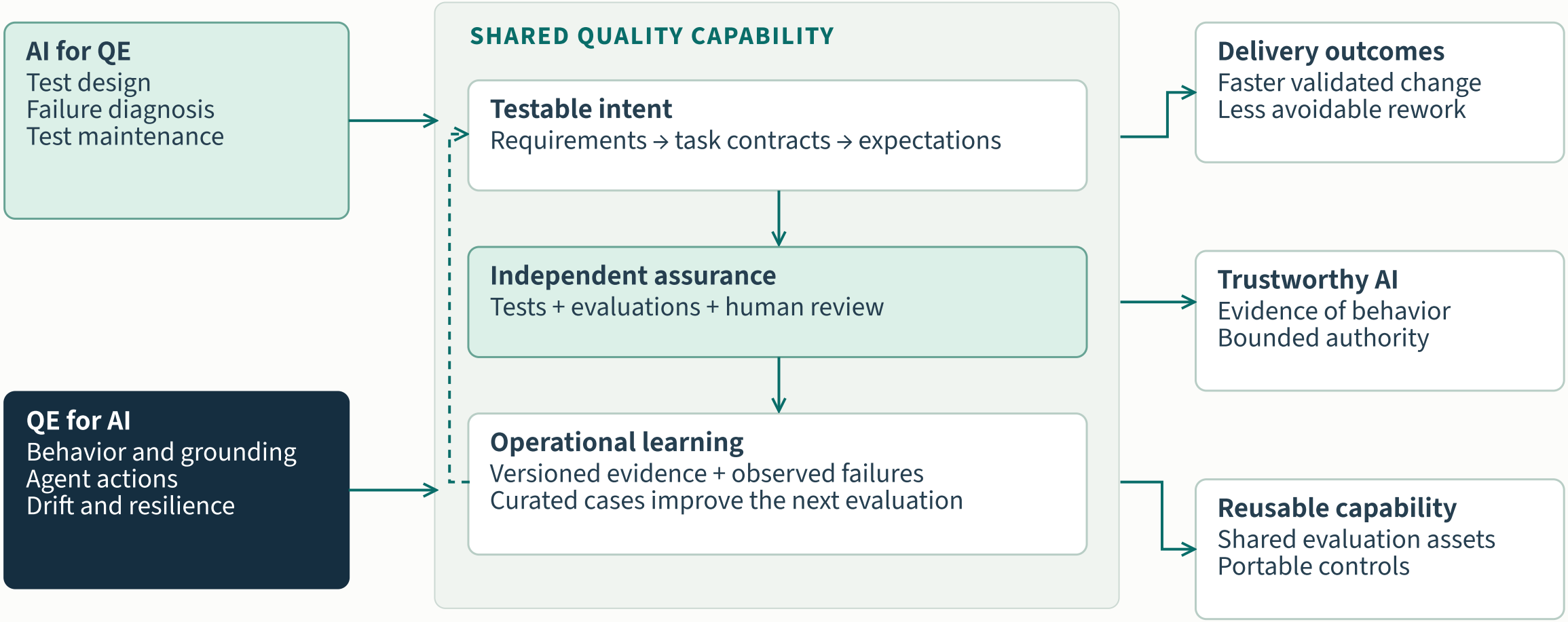


AI × QE · v1.24.0 Executive · Strategic vision

Prepared by Tom Wu · AI-generated illustrations · Synthetic English narration in the web presentation.

Research scope and sources ↗

# A shared quality capability

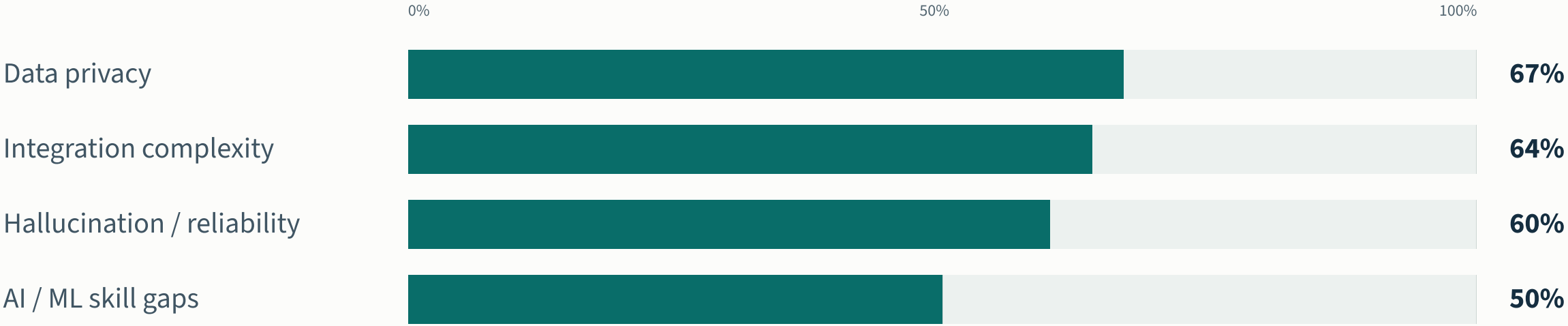



Strategic vision · outcomes are ambitions to validate, not reported bank results

# The constraints are organizational

## What constrains adoption

Share of respondents reporting each barrier · World Quality Report 2025–26




Overlapping responses, not shares of a whole. Survey: >2,000 executives; question-specific n unavailable in the public release. Source: Capgemini, Sogeti and OpenText, November 2025. [W01]

# Task gains depend on the setting

### CUI ET AL. / THREE FIELD EXPERIMENTS

**+26.08%**

completed tasks

Developer sample
**4,867**

Precision
**SE 10.3 percentage points**

Setting
**Coding assistants in three firms**

### METR / EARLY-2025 TRIAL

**+19%**

completion time

Developer / task sample
**16 / 246**

Precision
**95% CI: +2% to +39%**

Setting
**Experienced developers; familiar repos**

**The unit, task and user population determine what transfers.**

Separate outcomes; do not pool these percentages or treat them as QE savings rates.

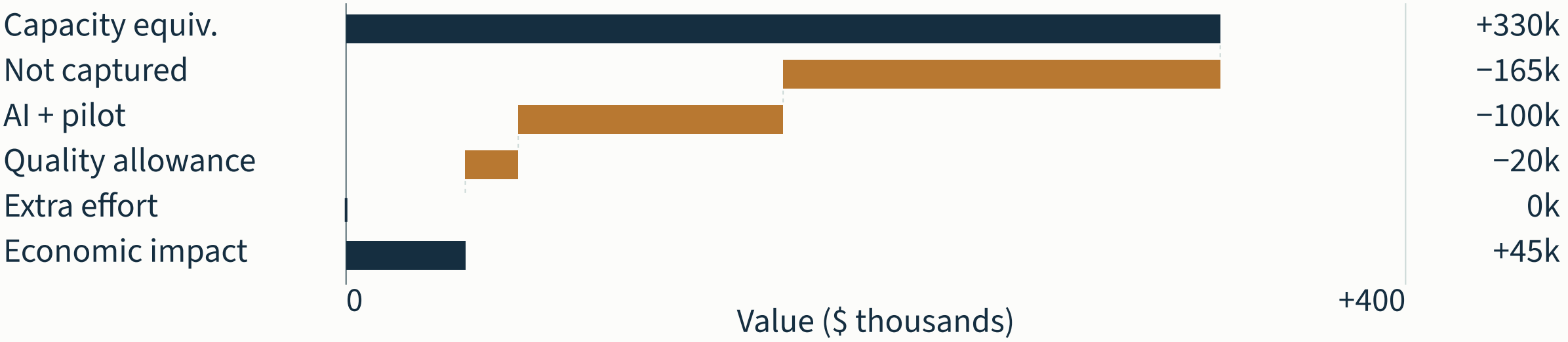
# Capacity and realized value

**$45k** illustrative net economic impact

3.30% capacity released · 0.45% of addressable spend

**Annual value bridge**

$ thousands · per $10M addressable QA spend




Illustrative assumptions, not a forecast. Capacity is effort equivalent; cash savings need Finance-approved capture.

Adoption 50% · net task saving 20% · capacity captured 50%.

Activity share 55% · eligibility 60% · AI/pilot cost 1.0% · quality allowance 0.2%.



Scenario assumptions and definitions ↗

# The industry direction and its limits

| SOURCE / EVIDENCE TYPE | WHAT IT SAYS | STRATEGIC IMPLICATION |
|---|---|---|
| **Gartner · forecast** | 90% of enterprise software engineers using AI code assistants by 2028. | Plan for widespread assisted work; tool use alone does not establish quality or savings. |
| **McKinsey · survey** | Nearly 300 leaders surveyed; 100 assessed impact. Top performers reported broader delivery improvements. | Invest in workflow and organizational change. These self-reported outcomes do not isolate causality. |
| **DORA · observational research** | AI interacts with the surrounding engineering system. | Improve feedback, testing and platform foundations alongside adoption. |

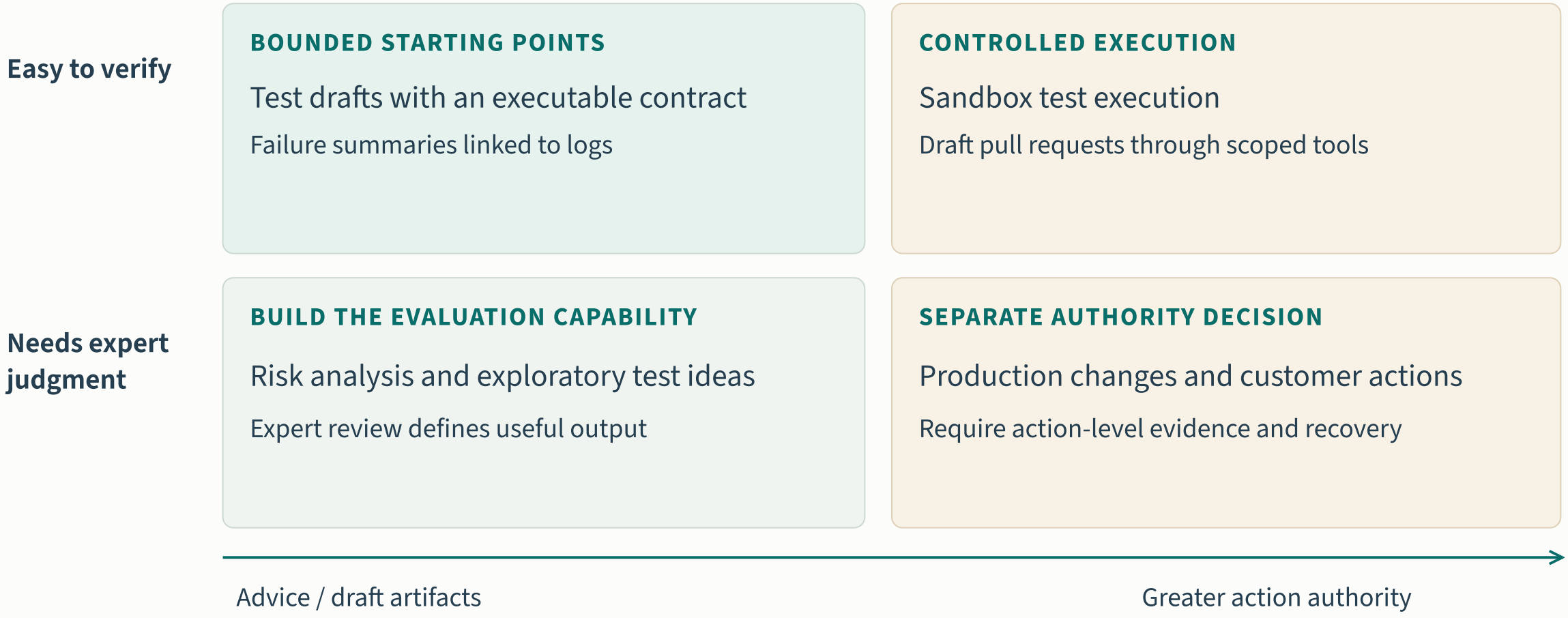
Public sources, reviewed September 2026. Gartner’s gated testing reports contribute public abstracts only; no proprietary vendor ranking is reproduced. [G01][G02][G03][M01][D01]

# Industrial cases reveal reusable patterns

| INDUSTRIAL EXAMPLE | OBSERVED APPROACH | CAPABILITY TO DEVELOP |
|---|---|---|
| **Meta · TestGen-LLM / ACH** | Generated tests face build, reliability and coverage checks; ACH adds fault-oriented validation. | An independent test-quality gate, with meaningful assertions. |
| **Google · Auto-Diagnose** | Failure summaries link likely causes to relevant log lines in the review workflow. | Diagnosis with inspectable evidence and reviewer feedback. |
| **Uber / UT Austin · FlakyGuard** | Targeted execution context supports flaky-test repair in an enterprise Go repository. | Repair validation that preserves the original test intent. |

Company-specific cases demonstrate mechanisms. They do not establish a transferable bank savings rate or a product comparison. [E04][E05][E06][E07]

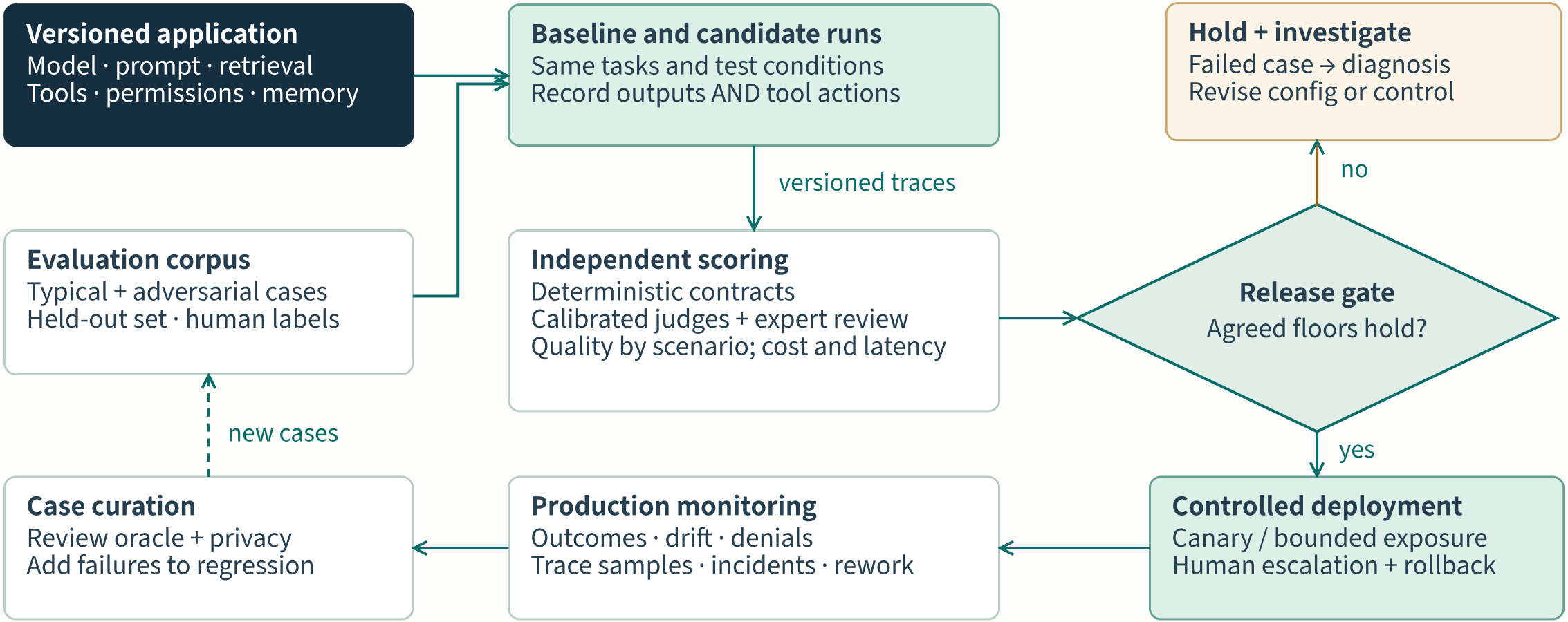
# The first workflows need a clear quality check



# What the findings change

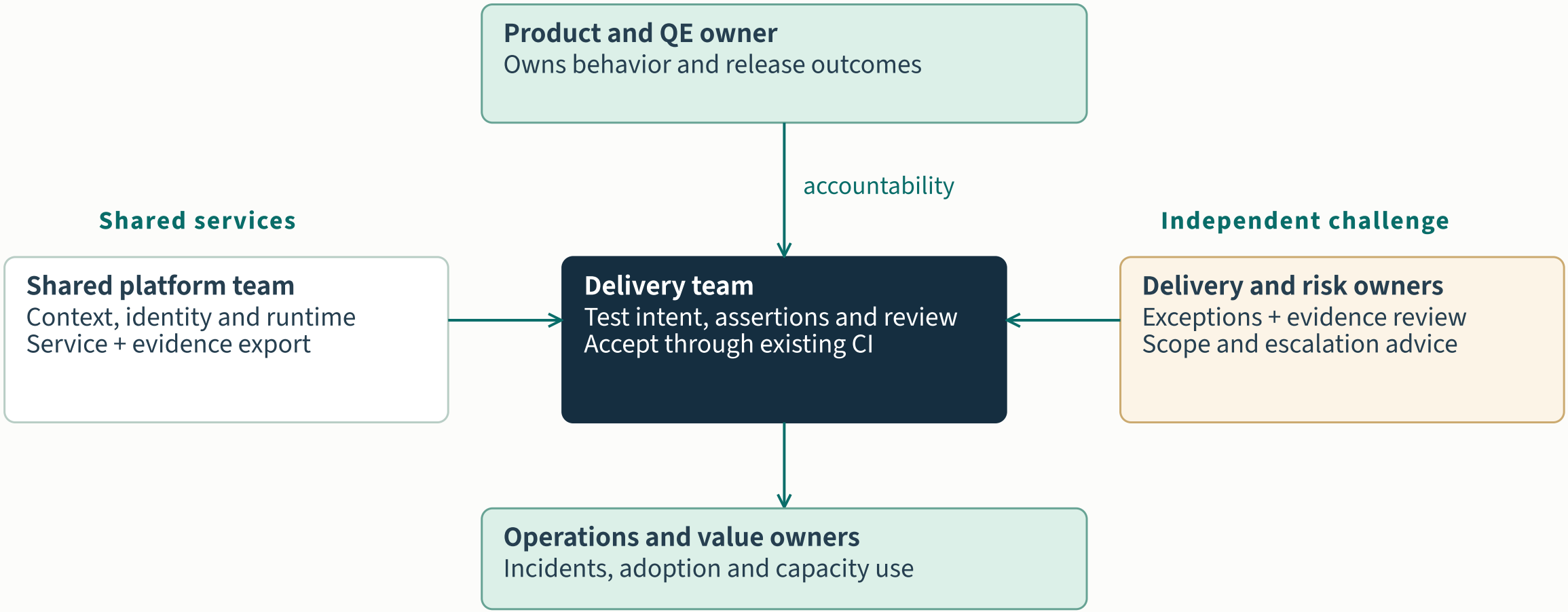
| RESEARCH OBSERVATION | DESIGN RESPONSE | EVIDENCE TO RETAIN |
| --- | --- | --- |
| **DORA / verification effort**<br>Generation can shift work into review. | **Join effort with run telemetry**<br>Measure net workflow effort | **Prep · review · correction**<br>D02 |
| **Meta / filtered test generation**<br>Passing is not proof of added value. | **Independent test-quality gate**<br>Protect approved assertions | **Stable runs · fault detection**<br>E04 · E05 |
| **NIST / lifecycle assurance**<br>Assurance continues in production. | **Version cases; feed failures back**<br>Re-evaluate configuration changes | **Case history · drift · release**<br>A01 · A02 |
| **OWASP + OSFI / agent actions**<br>Tools expand the impact of failure. | **Enforce permissions outside the model**<br>Test denied actions and recovery | **Identity · decision · approval**<br>A03 · R01 |

# Quality improves through feedback




Proposed application-level assurance · evaluate again when context, behavior or authority changes

# Ownership across the quality lifecycle

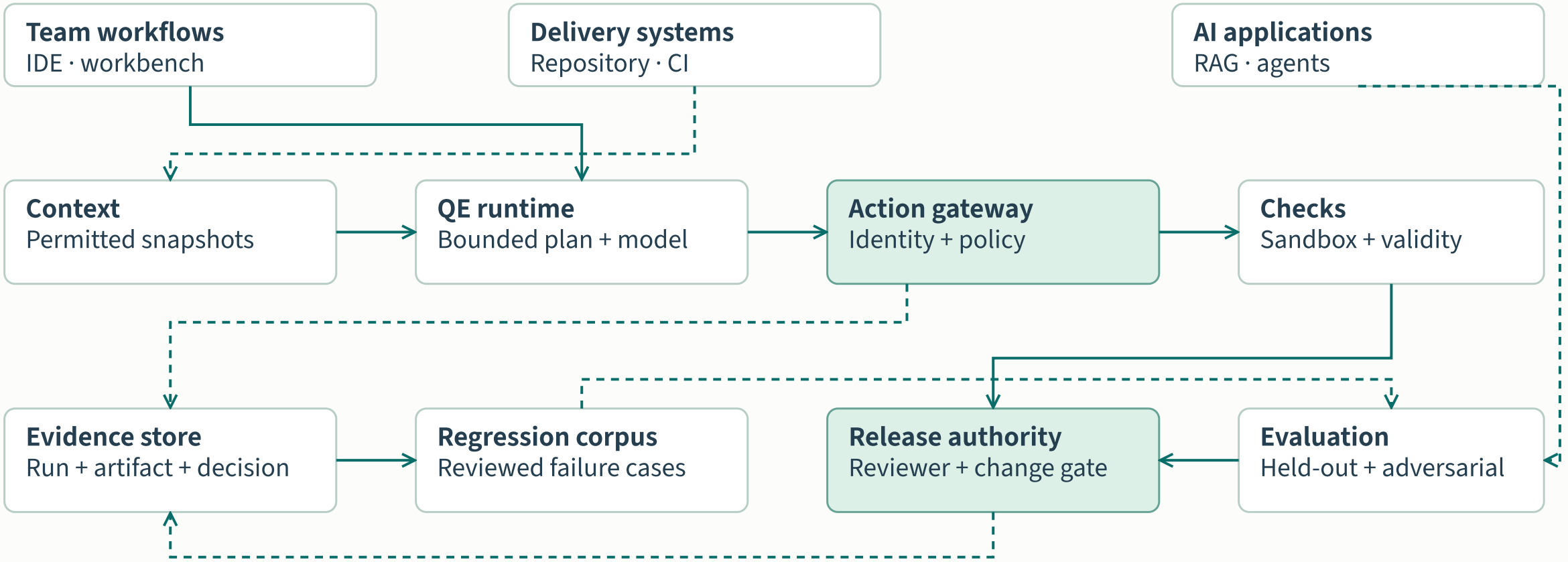


# The quality role gains new disciplines

| CAPABILITY | WHAT PRACTITIONERS NEED TO DO | OBSERVABLE LEARNING EVIDENCE |
|---|---|---|
| **Test intent and oracles** | Define expected behavior and challenge plausible but weak assertions. | A test catches a relevant fault and preserves the contract. |
| **Evaluation and curation** | Build representative cases; calibrate expert labels and model judges. | Evaluation results survive a held-out scenario review. |
| **Agent assurance** | Inspect tool authority, context boundaries and recovery behavior. | A denied-action exercise leaves no unauthorized side effect. |
| **Workflow ownership** | Measure review and correction work; coach teams through real tasks. | Task-level outcomes include the cost of verification. |

Proposed capability agenda informed by research on verification work and lifecycle assurance; it is not a staffing reduction plan. [D02][G04][A01][E05]

# The platform outlasts a model



# Shared foundations and domain expertise

## PRODUCT-SPECIFIC ASSETS

**Payments**
Transaction invariants
Failure and recovery cases

**Customer service AI**
Grounded response criteria
Escalation and tool scenarios

**Engineering workflows**
Repository test contracts
Review and rework baselines

## REUSABLE ASSURANCE SERVICES

Dataset registry · Evaluation runners · Identity + gateway · Evidence store · Recovery tooling

## REPLACEABLE MODEL AND TOOL ADAPTERS

Procure capabilities against local tests; preserve access to configurations, artifacts and evidence.

# Evidence gates the next horizon

Gate: local quality + net effort

**01**

**Trusted assistance**

Draft tests + diagnose failures
Independent checks; human approval

Gate: repeatable outcomes across teams

**02**

**Connected assurance**

Shared evaluation and context services
Cross-team evidence and feedback

Gate: safe actions + tested resilience

**03**

**Controlled agents**

Multi-step work with bounded authority
Action-level controls and recovery

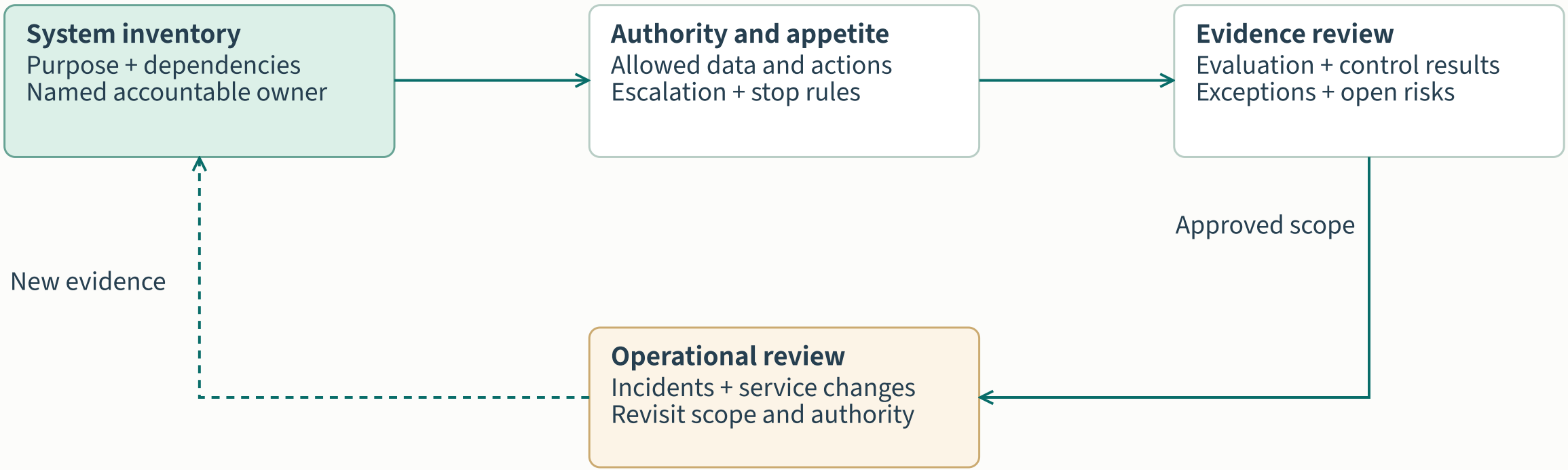
Proposed horizons, not a dated forecast or measured maturity score

# The leadership evidence contract

| STRATEGIC OUTCOME | EVIDENCE TO EXAMINE | LEADERSHIP DECISION |
| --- | --- | --- |
| Delivery capacity | Net effort with review and rework | Redeploy verified capacity |
| Trustworthy outcomes | Fault detection and escaped defects | Hold the quality floor |
| Reusable capability | Results by team and task | Fund the shared service |
| Controlled autonomy | Denials and recovery tests | Set the next authority limit |

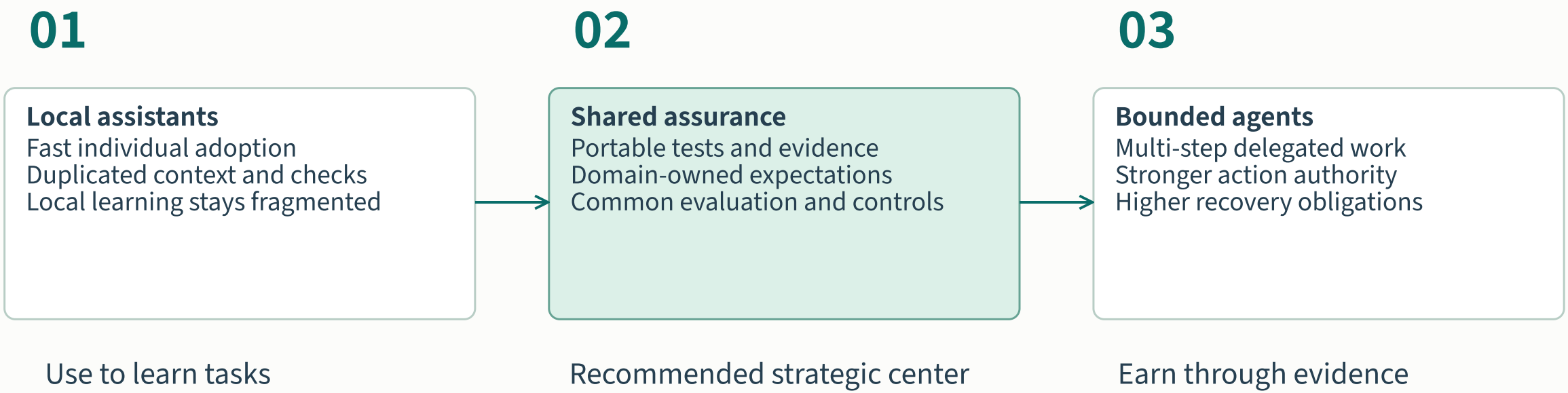
Every readout includes task mix, sample, period, costs and uncertainty. No organizational results are claimed here.

# Governance follows the system lifecycle




OSFI context: July 2026 bulletin = sound practices; revised E-23 takes effect 1 May 2027.

# Choose the shared assurance model

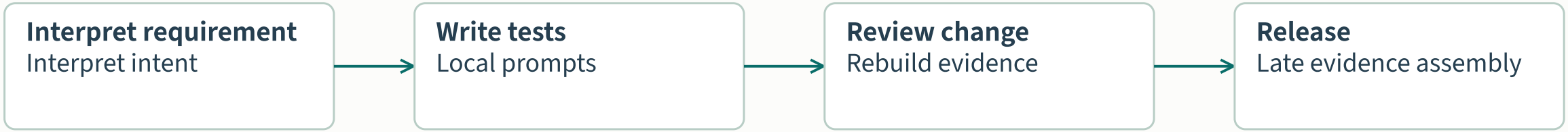



Sequencing choice: standardize the proof before expanding delegated authority.

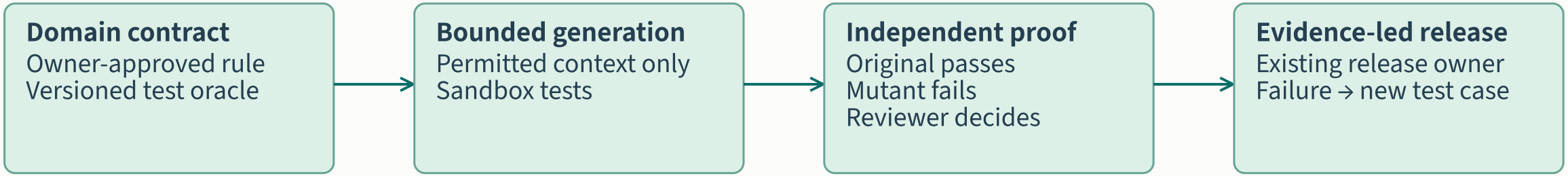
Trade-off: shared services require ownership; agent breadth requires independently tested containment.

# What changes in a payment-API workflow

## CURRENT WORKING HYPOTHESIS · VALIDATE WITH THE BANK




## PROPOSED TARGET · NEGATIVE PAMENT AMOUNTS MUST BE REJECTED




Shared: identity, evaluation runners, evidence schema. Domain: payment rules, test cases, acceptance and service outcomes.

# Sequence shared assets before wider agency

| HORIZON | SHARED INSTITUTIONAL ASSET | DOMAIN RESPONSIBILITY / TRADE-OFF |
|---|---|---|
| **Prove the workflow** | Task identity, approved models, baseline measurement and evidence contract. | Payment owner defines behavior; delivery team measures review effort. Keep scope narrow. |
| **Reuse the assurance** | Evaluation runners, policy enforcement and reusable failure cases. | Own domain cases, acceptance and outcomes. Shared services need funded owners. |
| **Delegate bounded actions** | Scoped tool adapters, recovery exercises and operational evidence. | Service owner accepts consequence and recovery duty. Expand one authority boundary at a time. |

# Client adoption requires funded foundations

## Adoption decision

Required evidence accepted · owner accountable · foundation work funded

### Business & people

Expected outcomes · reviewer capacity · offshore handoffs · measured baseline

### Runnable engineering

Infrastructure · CI/CD · test data · dependency control · testable applications

### Supported services

Frameworks & devices · AI access · run evidence · platform ownership

Authored adoption assumptions. Validate each application and workflow; unknown requirements remain open. Assumptions and evidence →

Fund the missing capabilities and reviewer capacity before promising scale. A workflow can progress only when its required evidence is accepted.

# QE modernization and AI adoption

01

## Trusted work

Reviewed scenarios, expected outcomes and ownership

**AI drafts with domain review**

A reviewer can identify a wrong answer

02

## Repeatable execution

Reliable suites, isolated data and controlled dependencies

**AI drafts tests that run in CI**

Another engineer reproduces the result

03

## Shared QE services

Reusable environments, contracts, artifacts and support

**Assistance works across teams**

A second team onboards and operates it

04

## Broader AI execution

Evaluated tools, bounded actions and recovery

**Agents execute within approved scope**

Failures stop, evidence persists, people can intervene

Proposed capability progression. Reviewed assistance can begin while foundations improve. Broader execution requires evidence for its specific scope. Modernization workstreams and sources

Develop reliable, reusable QE capability alongside reviewed AI assistance. Broader execution depends on the evidence available for the selected workflow.

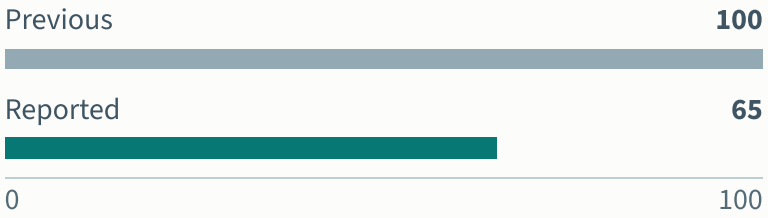
# Financial-services outcomes

## Libra Internet Bank

### 35% less time

**Test-creation time**

Index: previous time = 100




UiPath customer case. Normalized from the reported reduction; review effort is unspecified. F1

## Goldman Sachs

### 36% → 72%

**Unit-test coverage**

Share of a selected module (%)

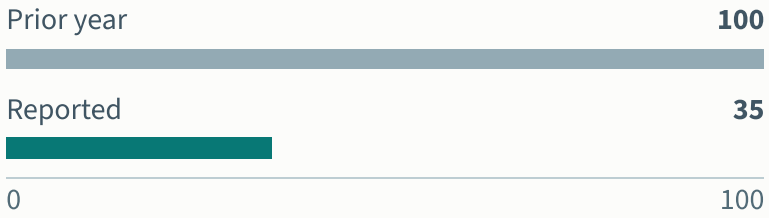

Diffblue customer case. Coverage measures exercised code; it does not establish financial correctness. F2

## Fiserv

### 65% fewer incidents

**Major incidents**

Index: previous year = 100




Tricentis modernization case. AI testing was a later pilot; this is not an AI-attributed reduction. F4

Separate outcomes and denominators. Each panel has its own meaning; no combined savings rate is calculated.

Reported customer outcomes use different measures and evidence types. Use them to select a trial; establish the client's own baseline.

# A focused first engagement

**Requirements ready**

## Onboarding design

Draft scenarios in the current test-management workflow.

**Proof**

Accepted coverage and measured review effort.

**Execution repeatable**

## Payment API tests

Generate candidates in the existing framework and CI pipeline.

**Proof**

Correct behavior passes; a known fault fails.

**Diagnostics available**

## Failure triage

Draft diagnoses alongside the existing disposition process.

**Proof**

Faster correct decisions with traceable evidence.

Proposed entry points. Choose one using the client's constraints; inspect prerequisites and prepare a trial brief.

Select the entry point that fits the client’s foundations. Build reusable assets and retain evidence for the next adoption decision.

OUR BANKING CLIENT / AUTHORED PAYMENT SCENARIO

# What AI contributes, from requirement to release


AI assists the testing team. For code, fixtures, failure evidence and release handoffs, open the banking engineering walkthrough →

# The next decision: capability, foundations and scope

Ambition

## Quality as a shared capability

Choose the delivery constraints and AI risks the organization needs to address.

Investment

## Foundations and people

Develop evaluation skills, platform integration and accountable ownership.

Expansion

## Value with evidence

Grow the capabilities that improve outcomes within their control boundary.

> Vision: AI contributes more work while the organization strengthens its ability to judge and govern that work.