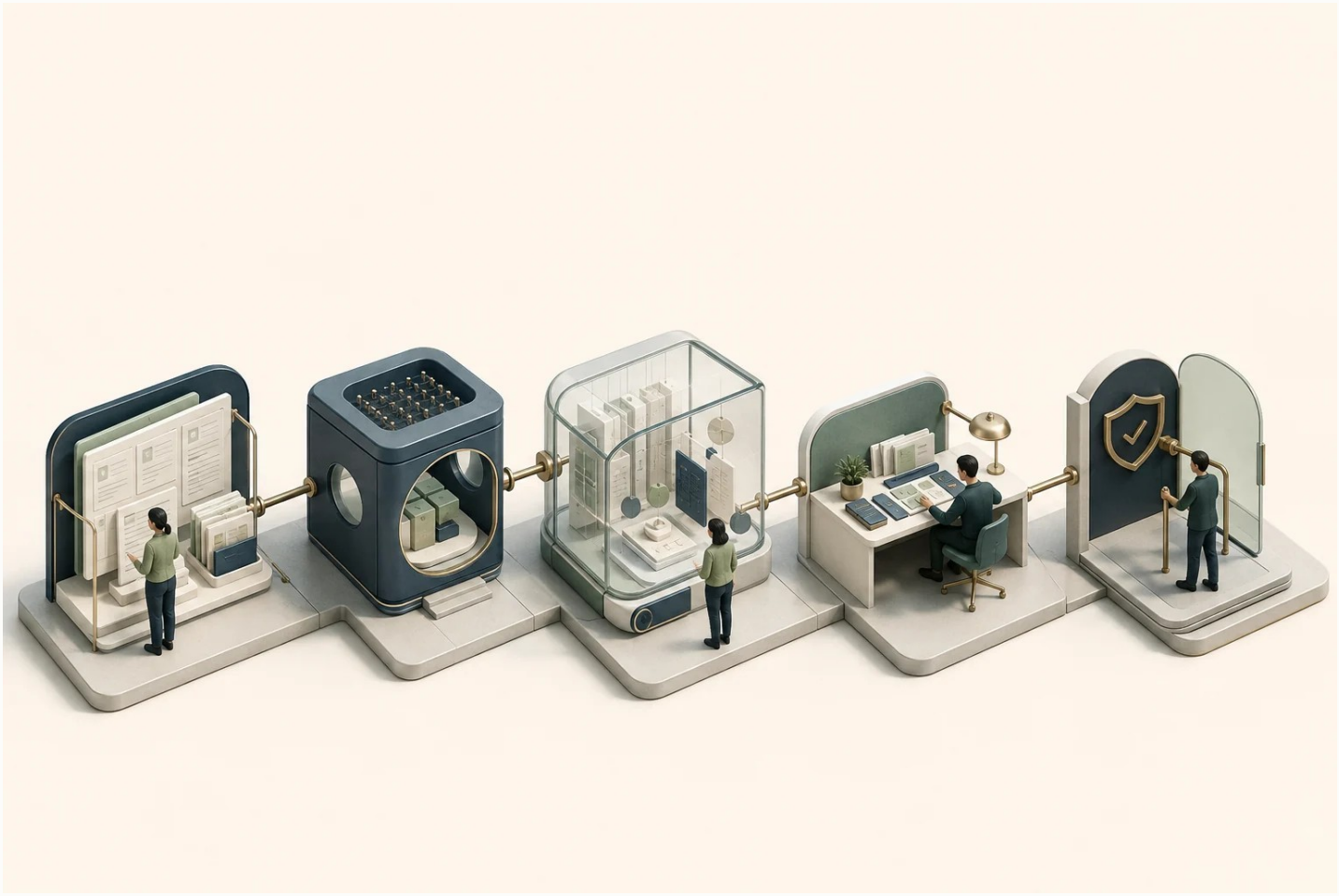
INDUSTRY RESEARCH

Quality engineering for the AI era

AI for quality engineering. Quality engineering for AI.



Strategic vision and assurance architecture

A financial-services perspective on industry adoption, evidence, technology and the operating model.

30 curated primary-source entries. Includes public Gartner and McKinsey material, DORA, enterprise research, NIST, OWASP, OSFI and representative product documentation.

01 / STRATEGIC THESIS

A broader quality mandate

The opportunity is a shared quality capability across assisted delivery and AI applications.

AI for QE	Generate test candidates, diagnose failures and prepare release evidence. Independent checks establish whether the output is useful.
QE for AI	Evaluate task outcomes, retrieval and tool actions. Test the application configuration and its operating boundary.
Shared foundation	Maintain task contracts, trusted evaluation cases, version history and accountable decisions. Reuse platform services where they improve integration.

Strategic implication

AI can contribute more artifacts and actions while QE strengthens the organization's ability to judge them. The durable investment is the evaluation system, engineering capability and evidence around the tools.

Authored synthesis informed by analyst perspectives, engineering research and assurance frameworks. It does not describe a deployed bank system.

Sources: [G01] [M01] [D01] [A01]

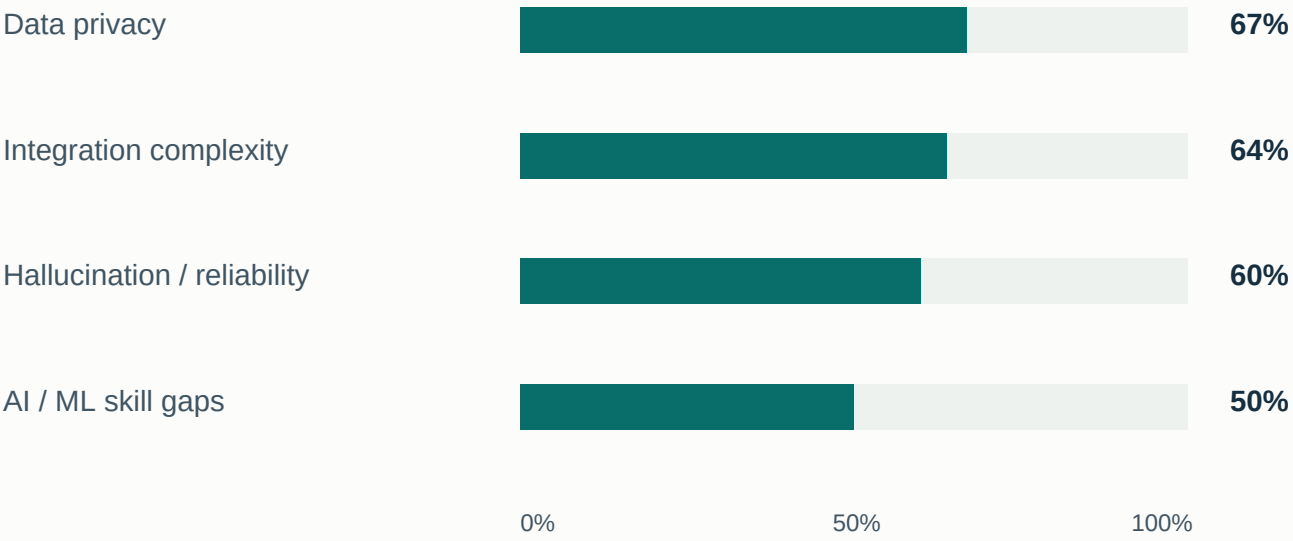
02 / INDUSTRY OUTLOOK

Adoption needs an operating system

Public surveys describe experimentation alongside integration and trust challenges.

Reported barriers to GenAI adoption in QE

Share of respondents / World Quality Report 2025-26



Overlapping responses, not a composition. Overall survey: more than 2,000 executives, 22 countries and 10 sectors. The public release does not provide question-specific sample sizes.

Sources: [W01]

What this suggests

Prioritize approved context, workflow integration, reliable checks and skills. Analyst forecasts describe a direction; they do not turn survey responses into measured enterprise value.

Gartner full testing reports were not available; their public abstracts are cataloged with explicit access limits.

Sources: [G02] [G03] [G04]

03 / EVIDENCE

The result depends on the work

Metrics, participants and study designs must travel with every headline.

Field experiments	Cui and colleagues studied 4,867 developers. The pooled estimate was 26.08% more completed tasks (standard error: 10.3 percentage points).
METR trial	16 experienced developers on familiar repositories took 19% longer with early-2025 AI access (95% CI: +2% to +39%).
METR update	The 2026 follow-up reported faster point estimates, but both intervals include no effect. Selection and time measurement limit interpretation.

These are different populations and outcome definitions. Do not average them or chart them as one temporal series.

Sources: [E01] [E02] [E03]

Evidence roles

Forecasts support scenarios. Surveys describe reported experience. Experiments estimate effects within a design. Cases reveal implementation mechanics. Product documentation establishes available functions. Each answers a different question.

The interactive site displays METR confidence intervals and a separate economic scenario model with explicit assumptions.

04 / AI FOR QE

Generation is only the first step

Industrial implementations filter candidates before they become maintained tests.

01 **Builds**
Syntax and dependencies are valid.

02 **Repeats**
Results hold across repeated runs.

03 **Detects faults**
Relevant assertions challenge behavior.

04 **Earns review**
Maintainers assess meaning and rework.

Authored synthesis of filtered and mutation-guided test generation. These stages are not a measured numerical funnel.

Sources: [E04] [E05] [E07]

An independent oracle matters

A passing test may simply reproduce the implementation's mistake. Use specifications, approved assertions, relevant mutations and reviewer judgment to challenge the expected result. Protect the oracle from silent weakening.

Diagnosis needs contextual evidence

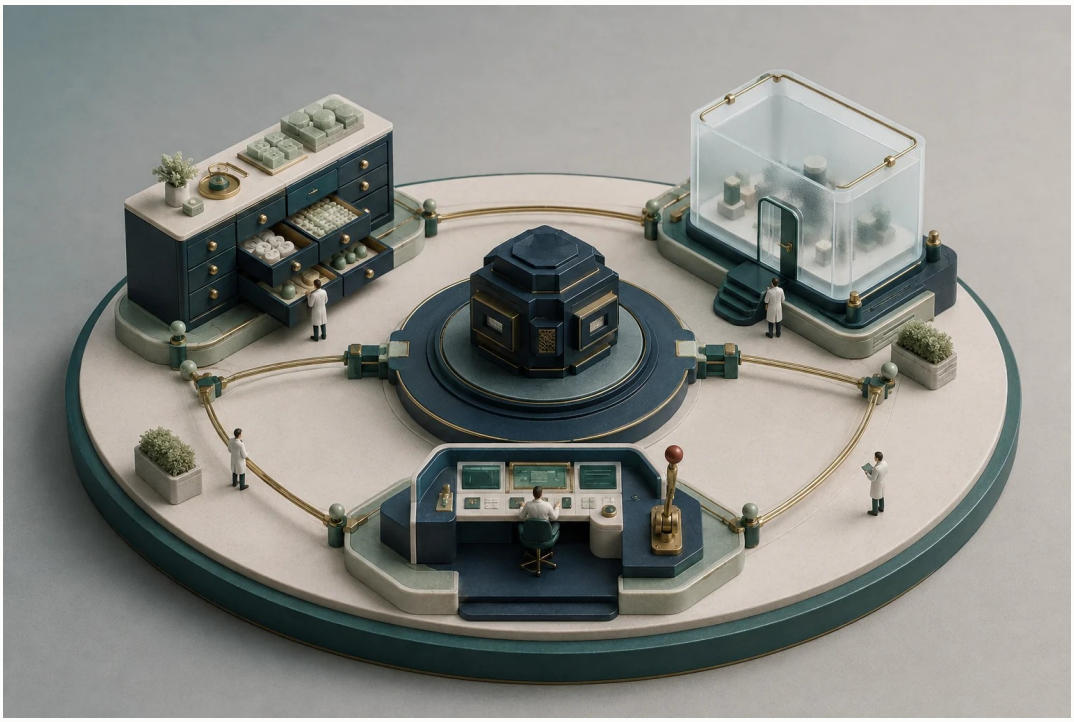
Failure-analysis systems can help reviewers navigate CI evidence. Measure accuracy on a reviewed sample separately from deployment volume, feedback usefulness and resolution effort.

Google AutoDiagnose separates a 71-failure manual evaluation from its much larger deployment population. Local task accuracy remains to be validated.

Sources: [E06]

05 / QE FOR AI

Continuous application assurance



01

Define

Task contracts and risk scenarios.

02

Evaluate

Checks, calibrated judges and adversarial cases.

03

Release

Versioned evidence and owner decision.

04

Observe

Traces, drift, cost and incidents.

Feedback loop: curated production failures become regression cases.

Proposed lifecycle. Re-evaluate changes to models, prompts, retrieval, tools and permissions. Hold out evaluation examples and calibrate model judges with expert labels.

Sources: [A01] [A02] [T03]

06 / AGENT ARCHITECTURE

Permission lives outside the model

A proposed application boundary for a governed QE workflow.

01 **Context**
Scoped code, contracts and test evidence.

02 **AI workflow**
Model, prompts, retrieval and memory.

03 **Policy gate**
Identity, tool scope, budget and approval.

04 **Bounded tools**
Sandbox, test runner and draft change.

Verification	Independent tests, data and access checks and review govern the release decision.
Evidence	Join configuration versions, tool calls, artifacts, decisions and cost under a privacy-aware logging standard.
Recovery	Test denied actions, bounded retries, escalation, stop conditions and fallback.

Retrieved content and tool output remain untrusted. Runtime permission checks should not depend on the model obeying a prompt.

Sources: [A03] [A05] [R01]

OSFI's July 2026 bulletin offers sound practices complementing existing guidelines. Revised E-23 is effective May 1, 2027; assess system applicability with institutional model-risk owners.

Sources: [R02]

07 / OPERATING MODEL

Ownership connects quality and value

Product / business	Task outcomes, acceptance criteria and accountable value owner.
QE / engineering	Test validity, evaluation datasets and delivery integration.
Platform / AI	Approved runtime, identity, context services and observability.
Delivery / risk	Quality exceptions, recovery and independent challenge.
Finance / delivery	Total cost, use of released capacity and validated benefit.

Measure the whole workflow

Include preparation, review, correction, repeated runs and controls in active effort. Measure escaped defects and rework alongside adoption, latency and cost. Agent runtime and human effort are different units.

Proposed ownership model. Released capacity needs an explicit capture mechanism; external task gains do not establish enterprise savings.

Sources: [D01] [D02] [M01]

08 / TECHNOLOGY LANDSCAPE

Different tools serve different layers

Representative examples from current documentation, not a market ranking.

Test assets	Tricentis Tosca and Applitools: authoring and visual regression. Validate test meaning, baselines and false alerts.
Change review	GitHub Copilot: review suggestions. Validate issues found and missed plus reviewer effort.
AI evaluation	LangSmith and Microsoft Foundry: datasets and result comparison. Validate calibration, versions and exportability.
Agent red teaming	Promptfoo: adversarial tool and permission scenarios. Match coverage to actual authority.

Documentation reviewed September 5, 2026. Available capabilities are not independent effectiveness evidence.

Sources: [T01] [T02] [T03] [T04] [T05] [T06]

Common selection brief

Compare on the same tasks. Request data flows, retention terms, identity scope, evaluation exports, resilience tests and total operating cost. Retain control of task contracts, curated evaluation data and evidence history.

The selection criteria are authored synthesis. Gartner's proprietary platform-selection criteria were not accessible.

Sources: [G02]

09 / IMPLEMENTATION

Expansion follows repeatable evidence

01 > 02 > 03 > 04

01 Foundations

Classified context, repeatable tests and baseline.

02 Assistance

One bounded workflow with reviewed artifacts.

03 Assurance

Versioned evaluation, telemetry and fallback.

04 Authority

Broader tasks with action-level evidence.

A useful pilot comparison

Choose frequent, bounded tasks. Compare like with like and record complexity and experience. Track net effort, quality floors, adoption, total cost and how released capacity is used. Validate successful results across teams and releases before broader scale.

Open research questions

The reviewed evidence does not settle sustained escaped-defect reduction, long-term test maintainability, simultaneous-agent human effort or rare operational failure rates. Treat these as explicit measurement needs.

Scope and access

This independent desk review is curated, not exhaustive. Licensed Gartner material was reviewed only at abstract level. WQR figures use the public release. Publisher PDFs remain in a local archive with hashes; the website links originals. Illustrations are AI-generated concepts; diagrams are proposed designs.

Full topic pages, interactive models, methods and downloadable source register: tomqwu.github.io/ai_qe/docs/industry/

SOURCE REGISTER

References 1-10

Original publisher links are embedded in each title. Full method and access notes are in the online library.

G01 / Software engineering trends for 2025 and beyond
Gartner / 2025-07-01 / Forecast
Access: Public text

G02 / How to Select an AI-Augmented Software Testing Platform
Gartner / 2026-02-03 / Analyst perspective
Access: Public abstract; full report gated

G03 / Journey Guide to Improving Software Testing in the AI Age
Gartner / 2026-04-02 / Analyst perspective
Access: Public abstract; full report gated

G04 / Smaller software engineering teams by 2029
Gartner / 2026-07-07 / Forecast
Access: Public text

M01 / Unlocking the value of AI in software development
McKinsey / 2025-11-03 / Survey
Access: Public text

M02 / The AI revolution in software development
McKinsey / 2026-04-01 / Case study
Access: Public text

W01 / World Quality Report 2025-26: public findings
Capgemini / Sogeti / OpenText / 2025-11-13 / Survey
Access: Public release; full report registration

D01 / State of AI-assisted Software Development 2025
DORA / Google Cloud / 2025-09-23 / Survey
Access: Public text

D02 / Balancing AI tensions: moving from AI adoption to effective software development
DORA / 2026-03-10 / Qualitative study
Access: Public text

E01 / Early-2025 AI on experienced open-source developer productivity
METR / 2025-07-10 / Randomized study
Access: Public text

SOURCE REGISTER

References 11-20

Original publisher links are embedded in each title. Full method and access notes are in the online library.

E02 / Updated developer productivity study
METR / 2026-02-24 / Study update
Access: Public text

E03 / The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers
Cui and colleagues / Management Science / 2026-02-27 / Randomized study
Access: Publisher abstract + open author paper

E04 / Automated Unit Test Improvement using Large Language Models at Meta
Alshahwan and colleagues / FSE 2024 / 2024-02-14 / Case study
Access: Public text

E05 / LLM-powered bug catchers: Meta ACH
Meta Engineering / 2025-02-05 / Case study
Access: Public text

E06 / LLM-Based Automated Diagnosis Of Integration Test Failures At Google
Ziftci and colleagues / Google / 2026-04-13 / Case study
Access: Public text

E07 / FlakyGuard: Automatically Fixing Flaky Tests at Industry Scale
Li and colleagues / UT Austin and Uber / 2025-11-18 / Case study
Access: Public text

A01 / Artificial Intelligence Risk Management Framework: Generative AI Profile
NIST / 2024-07-26 / Framework
Access: Public text

A02 / AI RMF Core: Govern, Map, Measure, Manage
NIST / 2023-01-26 / Framework
Access: Public text

A03 / Top 10 for Agentic Applications 2026
OWASP GenAI Security Project / 2025-12-09 / Framework
Access: Public text

A04 / OWASP GenAI LLM Top 10 2026
OWASP GenAI Security Project / 2026-08-03 / Framework
Access: Public text

SOURCE REGISTER

References 21-30

Original publisher links are embedded in each title. Full method and access notes are in the online library.

A05 / Agent Control Standard (ACS)
OWASP GenAI Security Project / 2026-09-01 / Framework
Access: Public text

R01 / Generative and agentic AI: technology, cyber security and operational resilience
OSFI / 2026-07 / Supervisory guidance
Access: Public text

R02 / Guideline E-23: Model Risk Management (2027)
OSFI / 2025-09-11 / Supervisory guidance
Access: Public text

T01 / Tosca Agentic AI documentation
Tricentis / Living documentation / Product documentation
Access: Public text

T02 / Deterministic Visual AI tests
Applitools / Living documentation / Product documentation
Access: Public text

T03 / Evaluation concepts and workflows
LangSmith / LangChain / Living documentation / Product documentation
Access: Public text

T04 / About GitHub Copilot code review
GitHub / Living documentation / Product documentation
Access: Public text

T05 / Red teaming for AI agents
Promptfoo / Living documentation / Product documentation
Access: Public text

T06 / View and compare evaluation results
Microsoft Foundry / Living documentation / Product documentation
Access: Public text

S01 / Developer Survey 2025: AI
Stack Overflow / 2025 / Survey
Access: Public text