EVIDENCE BRIEF

The case for independent API acceptance.

What current research says about agent adoption, operational reliability and the limits of model verification.

THE CENTRAL DISTINCTION

A model can perform well on a benchmark while the route you buy still fails your requirements.

This note synthesises independent sources. It is not an original research study, a market-wide quality ranking, or a validation of AssayCheck.

PREPARED FOR

Platform engineering, procurement
and teams operating LLM-backed systems.

01 / EVIDENCE

Adoption is not assurance.

Three separate findings explain why supplier acceptance deserves its own process. They do not establish that the quality of AI is falling across the market.

23% / 39%

Agents are moving into organisations.

In McKinsey's 2025 survey, 23% of respondents said their organisations were scaling an agentic AI system in at least one business function. Another 39% were experimenting. [1]

Scope: self-reported adoption; not a measure of successful or safe deployment.

15 models

Capability and reliability are different.

Rabanser and colleagues evaluated 15 models across two benchmarks. They found that recent capability gains delivered only small improvements in reliability, considering consistency, robustness, predictability and safety. [2]

Scope: the models and benchmarks studied; not evidence of universal deterioration.

Bounded claims

Software-only identity checks have limits.

Cai and colleagues identify weaknesses in output-based and log-probability approaches to model substitution detection. Their revised paper proposes hardware-backed verification using trusted execution environments. [3]

Scope: this work does not validate AssayCheck or establish the prevalence of supplier substitution.

02 / IMPLICATIONS

Test the boundary you depend on.

The following is AssayCheck's interpretation of the evidence, not a conclusion attributed to the cited researchers.

Define acceptance before testing.

Write down the required API behaviour, allowed limitations, request and spend budgets, and conditions that must block adoption.

Keep contract checks distinct from reference checks.

Streaming, tool loops, output schemas and token limits are integration requirements. A reference comparison cannot override a failed mandatory contract check.

Compare like with like.

Preserve the requirements profile and comparability conditions. Re-run when the route, configuration or operating assumptions change.

Report uncertainty as a result.

An unavailable endpoint is insufficient data. A route that is not rejected is not certified as authentic. Ambiguous results need explicit limits, not a stronger story.

ASSAYCHECK SCOPE

Supplier acceptance and change monitoring. Not a certification of an agent's end-to-end safety, compliance or business performance.

03 / SOURCES AND INTERPRETATION

A traceable reading list.

[1] McKinsey. The state of AI 2025. Survey evidence on agent adoption.

Read original source

[2] Rabanser et al. Towards a Science of AI Agent Reliability. ICML 2026; revised 2 June 2026.

Read original source

[3] Cai et al. Are You Getting What You Pay For? Auditing Model Substitution in LLM APIs. Revised 29 September 2025.

Read original source

[4] Tricia Carroll, ITPro. The managed service category nobody's named. Yet. 10 September 2026. Contributed industry opinion.

Read original source

How to read these sources

The survey describes adoption. The reliability paper examines agent behaviour. The API audit paper discusses verification limits. The ITPro column provides a service-market perspective, not technical validation.

None of these authors has evaluated or endorsed AssayCheck. No claim is made about the frequency of misconduct by any supplier. Test outcomes on the website are supplied product examples, separate from this literature review.