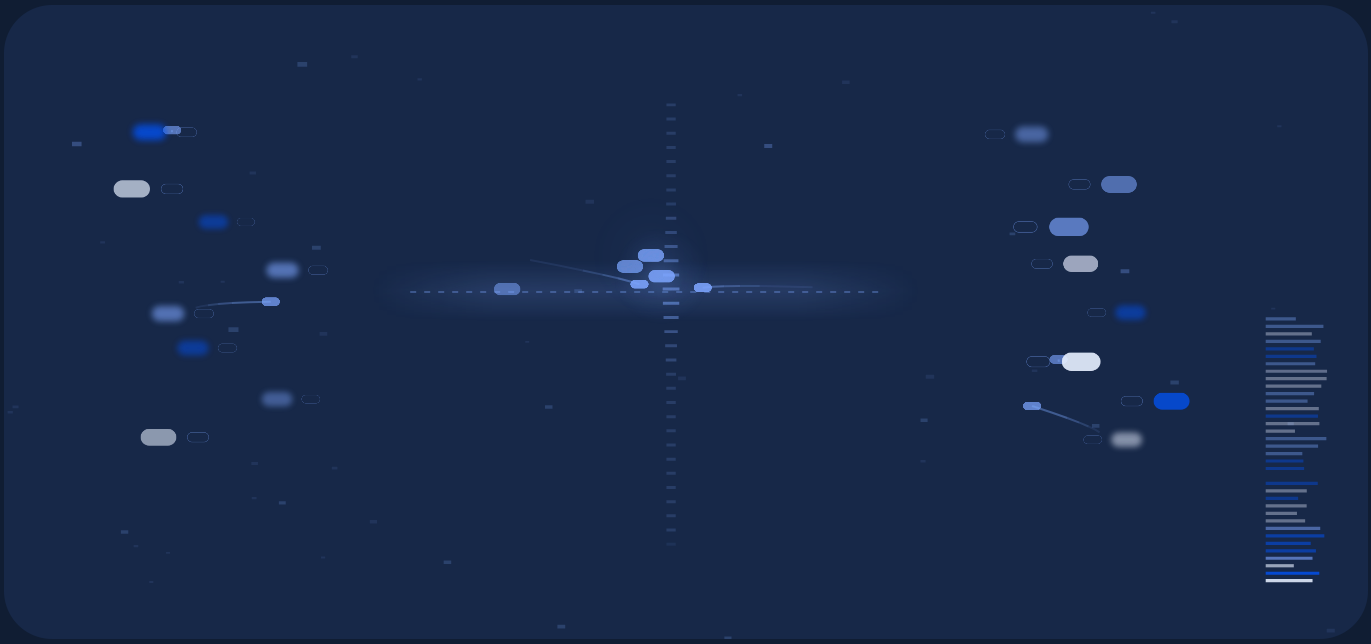
Rillor

EXPLAINER

Questions to ask before you deploy a vendor's agent

Questions to ask an agent vendor about testing: model versions, test sets, repeated runs, escalation, prompt injection, logs and who tested, with a public source behind each.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Agent vendors publish more about what their agents can do than about how they were tested, and the seven questions in this report ask for that evidence, each tied to a public standard, study or law [1, 3].

- In the 2025 AI Agent Index, which documented 30 deployed agents from public information as of 31 December 2025, 25 disclosed no internal safety results and 23 had no third-party testing information [2, 3].
- Of the 13 agents in the 2025 AI Agent Index with frontier levels of autonomy, 4 disclosed any agentic safety evaluations, and developers shared far more about capabilities than about safety practices [1].
- A model version change can move results. The March and June 2023 versions of one widely used hosted model scored 84% and 51% at telling prime from composite numbers [20].
- One run per test case hides inconsistency. An agent that succeeds 75% of the time passes all three of three runs about 42% of the time [26], and in the retail domain of one 2024 customer-service benchmark the best agent's chance of passing all eight of eight runs, averaged over tasks, was under 25% [17].
- Repeated attacks find more. In NIST's 2025 agent hijacking evaluations, attempting each prompt injection attack 25 times raised the average success rate from 57% to 80% [8].
- Default logs can leave out what an agent did. Under OpenTelemetry's conventions for recording tool calls, a call's tool name is required while its arguments and results are opt-in [28].
- For high-risk systems, the EU AI Act gives deployers duties of their own: use under the instructions for use, oversight by competent people, monitoring and logs kept at least six months. Under the Act as amended by the 2026 Digital Omnibus on AI, they apply from 2 December 2027 for the high-risk uses listed in its Annex III [12, 13, 14, 15].

Background

What vendors disclose about agent testing

The AI Agent Index documents deployed agentic systems from public information. Its first edition, posted in February 2025, found that developers "generally provide ample information regarding the capabilities and applications of agentic systems" but "limited information regarding safety and risk management practices" [4]. The 2025 edition covers 30 chat, browser and enterprise agents, with 45 fields in six categories, annotated by seven subject-matter experts "using only publicly available information and developer correspondence" [1]. All criteria were evaluated as of 31 December 2025 [3]. Its fields include evaluations, third-party testing, approval requirements and emergency stops [3].

For 227 of the 1,350 fields, the authors found no public information, most often in the ecosystem interaction and the safety, evaluation and impact categories [3]. Of the 30 agents, 25 disclose no internal safety results and 23 have no third-party testing information, and third-party testing is documented for 3 [2, 3]. Four have agent-specific system cards [3]. The paper adds that "few evaluations test the agentic setup rather than base model components" [3]. The index's home page states that of the 13 agents with frontier levels of autonomy, 4 disclose any agentic safety evaluations [1].

Its authors "did not perform experimental testing", and developers were given four weeks to correct annotations [2]. A blank field means nothing was found in public, which differs from no testing having happened. A buyer can request in writing what the public record leaves out.

The texts behind the questions

Each question in this report carries a public source, so a buyer can show where a request comes from. NIST AI 600-1, published in July 2024, is a profile of the NIST AI RMF for generative systems, and both are voluntary [5, 6]. Its appendix on pre-deployment testing says that testing often stays in laboratory conditions or on benchmark datasets that may not extrapolate to real-world conditions [5].

OWASP's Top 10 for LLM Applications 2025 covers prompt injection and excessive agency [9, 10]. Its Top 10 for Agentic Applications for 2026 adds risks specific to agents [11]. The EU AI Act sets duties for providers and deployers of high-risk systems [12]. The Guidelines for secure AI system development were published in November 2023 by UK and US government agencies with international partners [29]. Singapore's Model AI Governance Framework for Agentic AI, version 1.5 of May 2026, is aimed at organizations deploying agents built in-house or "using third-party agentic solutions" [30]. Published studies of evaluation, benchmarks, reliability and attacks supply the evidence [7, 8, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25].

This report is for procurement, risk, compliance and IT leads at companies choosing or approving an agent product. It is research, never legal advice.

Method

What was examined. Seven questions about how a vendor's agent was tested before deployment: model and version, test set, repeated runs, escalation and approval, prompt injection, logs and who did the testing. For each, the report gives the evidence that the question matters, what the public texts say and what a good answer contains. Two further findings cover the deployer's duties under the EU AI Act and a mapping of the questions to NIST, OWASP and EU AI Act items.

Sources. The AI Agent Index [1, 2, 3, 4], NIST and OWASP texts [5, 6, 7, 8, 9, 10, 11], the EU AI Act and its consolidated text [12, 13, 14, 15], published studies [16, 17, 18, 19, 20, 21, 22, 23, 24, 25], one model provider's guide [26], OpenTelemetry's conventions [27, 28], government guidelines [29, 30], and FTC orders and a tribunal decision [31, 32, 33].

How. Each source was opened and read on 8 November 2026, and quotes are given as published. EU AI Act text comes from the Official Journal on EUR-Lex, and application dates from the Commission's AI Act Service Desk, whose text is based on the consolidated version as at 27 July 2026 and reflects Regulation (EU) 2026/1744 [12, 13, 15]. Models are described by kind and date. The mapping in Finding 9 is this report's own reading of the texts.

Findings

1. Model and version: ask what happens when it changes

Why it matters. An agent is a model with prompts, tools and settings around it, and the model underneath can change. Chen, Zaharia and Zou compared the March and June 2023 versions of two widely used hosted models, and one fell from 84% to 51% accuracy at identifying prime versus composite numbers [20]. They concluded that "the behavior of the 'same' LLM service can change substantially in a relatively short amount of time" [20]. The 2025 Index adds that agentic evaluations depend on downstream context, including tools and autonomy level, "making model-level evaluation insufficient" [1].

What the texts say. The 2023 guidelines ask that update processes reflect "the fact that changes to data, models or prompts can lead to changes in system behaviour", treat major updates like new versions and support users "to evaluate and respond to model changes (for example by providing preview access and versioned APIs)" [29]. NIST AI 600-1 suggests that vendor contracts consider "GAI system changes over time (e.g., fine-tuning, drift, decay)" (GV-6.2-007) and that organizations re-assess risks for third-party models deployed for uses "that were not evaluated in initial testing" (MG-3.1-003) [5]. It also suggests reviewing system cards and model cards for third-party models (MG-3.1-005) [5].

Under the EU AI Act, a high-risk system's instructions for use state "the changes to the high-risk AI system and its performance which have been pre-determined by the provider" (Article 13(3)(c)) [12]. A deployer that makes "a substantial modification to a high-risk AI system that has already been placed on the market" is considered its provider (Article 25(1)(b)) [12].

What a good answer contains. The model family and the pinned version the agent runs on, recorded in the logs. OpenTelemetry's agent span conventions include the agent's version and the requested model, which "must be the exact name of the model requested" when a vendor supplies the model [27]. The notice given before a change, whether preview access or versioned endpoints exist [29], and which tests are rerun. One model provider's January 2026 guide describes regression evals that "should have a nearly 100% pass rate", with automated evals run "on each agent change and model upgrade as the first line of defense" [26]. A good answer includes that suite's results on the version about to be deployed.

2. Test set: ask which cases, with what known answers

Why it matters. A score describes the cases it was measured on. The FTC alleged that Workado promoted its content detector as "98 percent" accurate, while testing showed 53 percent on general-purpose content, and that the model behind it was "only trained or fine-tuned to effectively classify academic content" [32].

Public benchmarks carry errors too. A 2025 study of agentic benchmarks found that "SWE-bench Verified uses insufficient test cases, while TAU-bench counts empty responses as successful", issues that can lead to misestimates "by up to 100% in relative terms" [19]. A 2024 review found that "many agent benchmarks have

inadequate holdout sets, and sometimes none at all" [21]. NIST's January 2026 draft warns of "solution contamination", in which a model reaches solution information in an unintended way, from searching benchmark answers online to finding solution files left in the test environment [7].

What the texts say. The NIST AI RMF asks that "test sets, metrics, and details about the tools used during TEVV are documented" (MEASURE 2.1) and that performance be "demonstrated for conditions similar to deployment setting(s)" (MEASURE 2.3) [6]. NIST AI 600-1 suggests evaluating "claims of model capabilities using empirically validated methods" (MS-2.3-002) [5]. Singapore's framework recommends testing entire agent workflows for overall task execution, policy compliance, tool calling and robustness, in an environment "that mirrors production as closely as possible" [30].

What a good answer contains. Test cases built from tasks like the buyer's, each with its correct outcome fixed before the runs, kept out of the material used to build or tune the agent and run in a setting close to production [6, 21, 30]. The grading rule for each case. Where a model grades the agent, a check of the grader against people: in a 2023 study, strong model judges reached "over 80% agreement, the same level of agreement between humans", and showed position, verbosity and self-enhancement biases [25]. The provider guide suggests starting from "20-50 simple tasks drawn from real failures" [26].

3. Repeated runs: ask how many per case and how results are reported

Why it matters. tau-bench, a 2024 benchmark of customer-service agents, compares "the database state at the end of a conversation with the annotated goal state" and introduced pass^k, the chance that all k trials of a task succeed, averaged across tasks [17]. In its results, "state-of-the-art function calling agents" succeeded on under 50% of tasks and were "quite inconsistent (pass^8 <25% in retail)" [17]. The provider guide gives the arithmetic: at a 75% per-trial success rate, passing all three of three trials has a probability of $(0.75)^3$, about 42%, a measure that "especially matters for customer-facing agents" [26]. A 2026 reliability study evaluated 15 models on two benchmarks with twelve reliability metrics and found that "recent capability gains have only yielded small improvements in reliability" [23].

What the texts say. NIST's January 2026 draft says that having models attempt each item multiple times "can reduce uncertainty of evaluation results" [7]. It asks evaluators to "define, conduct, and report a statistically valid analysis procedure", to report statistics "with estimated uncertainties" whenever possible and to indicate unquantified sources of variation, such as prompt format [7]. It calls it good practice to define the analysis procedure in advance [7]. Singapore's framework says agent behaviour "is inherently stochastic and context dependent" and that tests "should also be run multiple times to check for stability" [30].

What a good answer contains. The number of trials per case, a result that counts every trial, such as pass^k or a consistency measure, next to the single-run success rate, error bars, and an analysis plan fixed before the runs [7, 17, 26].

4. Escalation, refusal and approval: ask for the rules and the cases

Why it matters. Many deployed agents act without a person in the loop. The 2025 Index reports that enterprise users configure agents at autonomy Level 1 to 2, while deployed agents often run at Level 3 to 5, triggered by events without human involvement [1]. OWASP names excessive autonomy, with excessive functionality and excessive permissions, as a root cause of excessive agency, with the example of an application that "fails to independently verify and approve high-impact actions" [10].

What the texts say. OWASP LLM06 recommends limiting the extensions an agent may call "to only the minimum necessary", using human-in-the-loop control "to require a human to approve high-impact actions before they are taken", and implementing "authorization in downstream systems rather than relying on an LLM to decide if an action is allowed or not" [10]. OWASP LLM01 lists "Require human approval for high-risk actions" among its mitigations [9]. Singapore's framework tests "Policy compliance: Whether an agent follows defined SOPs and routes for human approval when required", and recommends checkpoints that require human approval, "such as high-stakes or irreversible actions" [30].

For high-risk systems, Article 14(4) of the EU AI Act asks that overseers be enabled to understand the system's "relevant capacities and limitations", to stay aware of "automation bias", to "disregard, override or reverse the output" and to "interrupt the system through a 'stop' button or a similar procedure" [12]. Article 14(3) asks for measures "commensurate with the risks, level of autonomy and context of use" [12].

What a good answer contains. Written rules for when the agent hands off, refuses or asks for approval, fixed before testing. tau-bench's airline policy is a model: transfer the user to a human agent "if and only if the request cannot be handled within the scope of your actions", with a tool for the transfer [17]. Test cases where stopping or asking is correct and cases where proceeding is correct, with results for both. A way to override and stop the agent. NIST's draft notes that refusal behaviour during evaluation "may differ from refusal behavior under realistic usage" and asks evaluators to check transcripts for refusals [7], so a buyer can ask whether the transcripts were read.

5. Prompt injection: ask how, through which channels and how many attempts

Why it matters. OWASP LLM01 separates direct injection, when "a user's prompt input directly alters the behavior of the model in unintended or unexpected ways", from indirect injection, when an LLM "accepts input from external sources, such as websites or files" [9]. It adds that "it is unclear if there are fool-proof methods of prevention for prompt injection" [9]. NIST's Center for AI Standards and Innovation describes agent hijacking as arising "when a system lacks a clear separation between trusted internal instructions and untrusted external data" [8]. OWASP's 2026 agentic list puts agent goal hijack first (ASI01), for manipulated inputs that redirect an agent's goals [11]. In the 2025 Index, 7 of 30 agents documented defenses against prompt injection attacks, and prompt injection vulnerabilities were documented for 2 of 5 browser agents [3].

What the evidence shows about test design. NIST's evaluations used AgentDojo, an environment of 97 tasks and 629 security test cases built as "an extensible environment for designing and evaluating new agent tasks, defenses, and adaptive attacks" [8, 22]. They tested agents on one hosted model released in October

2024, and three of their four insights bear on a buyer's questions [8]. New attacks developed with UK AI Security Institute red teamers raised attack success on held-out user tasks from 11% for the top baseline attack to 81% for the top new attack [8]. The average success across five injection tasks was 57%, and the task-level results "reveal several details that were not clearly conveyed by the aggregate measure" [8]. Attempting each attack 25 times raised the average success rate from 57% to 80%, while many frameworks, AgentDojo included, measure an attack on a single attempt [8].

A 2025 study found that defenses are typically evaluated against static attack strings or weak optimization methods [18]. Its authors bypassed 12 recent defenses "with attack success rate above 90% for most", although "the majority of defenses originally reported near-zero attack success rates" [18].

What the texts say. OWASP LLM01 recommends regular adversarial testing and attack simulations, "treating the model as an untrusted user to test the effectiveness of trust boundaries and access controls" [9]. For high-risk systems, Article 15(5) of the EU AI Act asks for resilience "against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities" [12].

What a good answer contains. Direct and indirect injection through every channel the agent reads, an attack set with its date, adaptive attacks written against the agent's own defenses, several attempts per attack and results per task, read against what each successful attack could do [8, 18]. NIST AI 600-1 cautions that jailbreaking or prompt engineering tests "may not systematically assess validity or reliability risks" [5], so injection results sit beside the other six answers here.

6. Logs: ask what is recorded and who keeps it

Why it matters. Logs show what an agent did. The Holistic Agent Leaderboard used model-aided log inspection "to uncover previously unreported behaviors", such as searching for the benchmark online instead of solving a task and "misusing credit cards in flight booking tasks" [24]. The provider guide warns that "You won't know if your graders are working well unless you read the transcripts and grades from many trials" [26], and NIST's draft calls manual transcript review "a powerful technique for quality assurance of automated benchmark evaluations" [7].

What the texts say. OpenTelemetry's semantic conventions for agent and tool spans, the trace records of agent operations and tool executions, carry the status "Development" [27, 28]. Each tool span is named "execute_tool {gen_ai.tool.name}", the tool name is required, and the call's arguments and result are opt-in, as are the chat history and output messages on agent spans [27, 28]. A vendor that follows the defaults can record which tool ran without what was passed to it or what came back. The 2023 guidelines ask providers to supply "high-quality audit logs and other security features or information to customers and users at no extra charge", so customers can investigate incidents [29]. Singapore's framework advises requesting observability from outside agent providers, "such as the logging of tool calls and access history" [30]. OWASP LLM06 recommends logging and monitoring the activity of extensions and downstream systems [10].

NIST AI 600-1 suggests that procurement assessments address "ongoing monitoring, assessments, and alerting" (GV-6.1-009) and continuous monitoring of third-party systems in deployment (GV-6.2-004) [5]. The NIST AI RMF asks that a system's behaviour be monitored in production (MEASURE 2.4) [6]. For high-risk systems, Article 12(1) of the EU AI Act requires that systems "technically allow for the automatic recording of events (logs) over the lifetime of the system", and Article 26(6) has deployers keep the logs under their control for at least six months [12, 14].

What a good answer contains. Every model call with the model version, every tool call with its arguments and results, each approval and refusal, a stated retention period, and export to the buyer at no extra charge [12, 28, 29, 30].

7. Who did the testing: ask for separation and results you can read

Why it matters. In the 2025 Index, third-party testing was documented for 3 of 30 agents, and 23 had no third-party testing information [3]. A 2025 paper by 34 authors, titled "In-House Evaluation Is Not Enough", argues that the infrastructure for reporting flaws in general-purpose systems remains "seriously underdeveloped, lagging far behind more established fields like software security" [16]. It proposes standardized flaw reports and broadly scoped flaw disclosure programs with legal safe harbors for researchers [16].

Claims about untested systems carry consequences. The FTC alleged that DoNotPay "did not test whether its 'AI lawyer' operated to the level of a human lawyer" and "did not hire or retain attorneys to test the quality and accuracy" of its legal features [31]. The final order, announced on 11 February 2025 after a 5-0 vote on 16 January 2025, requires "$193,000 in monetary relief" and bars claims that the service "performs like a real lawyer unless it has sufficient evidence to back it up" [31]. In Moffatt v. Air Canada, decided on 14 February 2024, a British Columbia tribunal rejected the suggestion that the airline's chatbot was "a separate legal entity", found that the airline "did not take reasonable care to ensure its chatbot was accurate" and ordered it to pay $812.02 [33]. The decision does not say who built the chatbot. The airline answered for it.

What the texts say. NIST AI 600-1 suggests verifying that "those conducting structured human feedback exercises are not directly involved in system development tasks for the same GAI model" (MS-1.3-003), writing "clauses in contracts which allow an organization to evaluate third-party GAI processes and standards" (GV-6.1-006) and sharing "results of pre-deployment testing with relevant GAI Actors, such as those with system release approval authority" (MS-2.3-003) [5]. The NIST AI RMF asks that "internal experts who did not serve as front-line developers for the system and/or independent assessors" take part in regular assessments (MEASURE 1.3) [6]. The 2023 guidelines ask providers to release systems only after "appropriate and effective security evaluation such as benchmarking and red teaming" and to be clear about "known limitations or potential failure modes" [29].

What a good answer contains. The names or roles of the testers and how they were kept apart from the developers, the access they had, the date and agent version tested, results the buyer can read, a contract clause that lets the buyer run its own evaluation, and a channel for reporting flaws found after

deployment [5, 6, 16].

8. Deployer duties: what the EU AI Act asks of the buyer

The EU AI Act defines a deployer as a person or body "using an AI system under its authority", except in a personal non-professional activity (Article 3(4)) [12]. For high-risk systems, Article 26 asks deployers to use them "in accordance with the instructions for use", to "assign human oversight to natural persons who have the necessary competence, training and authority", to monitor operation on the basis of the instructions for use and suspend use where the system may present a risk, and to keep the logs under their control for at least six months [12, 14].

Two provider duties give the deployer documents to ask for. Article 13(3)(b)(ii) asks that the instructions for use state "the level of accuracy, including its metrics" against which the system "has been tested and validated", and Article 15(3) asks that accuracy levels and metrics be declared in those instructions [12]. Article 15(1) asks that high-risk systems reach an appropriate level of accuracy and robustness and "perform consistently in those respects throughout their lifecycle" [12].

The regulation as published in 2024 set 2 August 2026 as its general date of application, with exceptions [12]. Under the consolidated text published by the Commission's AI Act Service Desk, which reflects the Digital Omnibus on AI (Regulation (EU) 2026/1744 of 8 July 2026), Chapter III Sections 1 to 3 apply from 2 December 2027 for systems classified as high-risk under Article 6(2) and Annex III, and from 2 August 2028 for those under Article 6(1) and Annex I [13, 15]. Articles 12 to 15 sit in Section 2 and Articles 25 and 26 in Section 3 [12]. On 8 November 2026, the service desk page for Article 26 showed no amendment markers [14]. These duties apply to high-risk systems only, and this report does not assess whether any agent is one.

9. Mapping the questions to NIST, OWASP and the EU AI Act

Question	NIST AI 600-1 and NIST AI RMF	OWASP	EU AI Act (high-risk systems)	Other guidelines
1. Model and version	GV-6.2-007, MG-3.1-003, MG-3.1-005	No direct item	Art. 13(3)(c), 25(1)(b)	Updates as new versions, versioned APIs [29]
2. Test set	MEASURE 2.1, MEASURE 2.3, MS-2.3-002	No direct item	Art. 13(3)(b)(ii), 15(3)	Realistic test environments [30]
3. Repeated runs	No direct item	No direct item	Art. 15(1)	Repeated runs for stability [30]

Question	NIST AI 600-1 and NIST AI RMF	OWASP	EU AI Act (high-risk systems)	Other guidelines
4. Escalation and approval	No direct item	LLM06 mitigations 1, 6 and 7, LLM01 mitigation 5	Art. 14(3), 14(4), 26(2)	Approval checkpoints [30]
5. Prompt injection	Appendix A.1.4	LLM01, ASI01	Art. 15(5)	Evaluation before release [29]
6. Logs	GV-6.1-009, GV-6.2-004, MEASURE 2.4	LLM06 logging and monitoring	Art. 12(1), 26(5), 26(6)	Audit logs [29], tool-call logging [30]
7. Who tested	MS-1.3-003, GV-6.1-006, MS-2.3-003, MEASURE 1.3	No direct item	No direct item	Evaluation and known limitations [29]

Table 1. This report's mapping of the seven questions to the items cited in Findings 1 to 8, from NIST AI 600-1 [5], the NIST AI RMF [6], OWASP [9, 10, 11], the EU AI Act [12] and the guidelines [29, 30]. No published crosswalk was used, and the rows show where each text speaks to a question.

Implications

For procurement and risk leads. Send the seven questions in writing before a pilot, and ask for the evidence behind each answer: test cases, trial counts, transcripts, logs and the names or roles of the testers. For most agents in the 2025 Index, that material was not public [2, 3]. NIST AI 600-1 suggests building such requests into due diligence and contracts, including a clause that lets the buyer evaluate [5].

For compliance leads. Where a system may be high-risk under the EU AI Act, the deployer's own duties under Article 26 call for instructions for use, trained oversight staff and logs the deployer controls, from 2 December 2027 for Annex III systems under the consolidated text [12, 13, 14].

For IT and engineering leads. Rerun the vendor's cases and your own on the version you deploy, with several trials per case and the transcripts read [7, 26]. Check whether the default logs keep tool arguments and results [28].

Pre-deployment testing. Rillor tests the agents an organization plans to deploy on its own tasks, with black-box logging and replayable runs.

Limits of this analysis

- This explainer sets out questions and the public texts behind them. It does not test or rank any vendor's agent.
- The 2025 AI Agent Index reflects public information and developer correspondence as of 31 December 2025, and its authors did no experimental testing [2, 3]. A blank field means no public information was found.
- No survey was found of how many vendors provide pinned versions, change notices, regression results, logs or contract evaluation rights, so the report gives no such share.
- NIST AI 600-1 and the NIST AI RMF are voluntary [5, 6]. NIST AI 800-2 is cited as a January 2026 initial public draft [7].
- The EU AI Act duties cited apply to high-risk systems, and the application dates come from the consolidated text as at 27 July 2026 [13]. This is research, never legal advice.
- OpenTelemetry's agent conventions carry Development status, and the agent spans file was last changed on 5 November 2026 [27].
- Several cited studies are preprints. The enforcement and tribunal cases concern a content detector, a legal service and a website chatbot [31, 32, 33], and the tribunal decision is a small-claims ruling in British Columbia.

Sources

1. The 2025 AI Agent Index. AI Agent Index project, MIT, 2025 edition, viewed 8 November 2026. https://aiagentindex.mit.edu/

2. Further Details, 2025 AI Agent Index. AI Agent Index project, MIT, 2025 edition, viewed 8 November 2026. https://aiagentindex.mit.edu/2025/further-details

3. The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems. Leon Staufer, Kevin Feng, Kevin Wei, Luke Bailey and others, arXiv:2602.17753, version 2, 6 May 2026, viewed 8 November 2026. https://arxiv.org/abs/2602.17753

4. The AI Agent Index. Stephen Casper, Luke Bailey, Rosco Hunter and others, arXiv:2502.01635, 3 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2502.01635

5. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. National Institute of Standards and Technology, July 2024, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

6. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. National Institute of Standards and Technology, January 2023, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

7. Practices for Automated Benchmark Evaluations of Language Models, NIST AI 800-2, Initial Public Draft. National Institute of Standards and Technology, Center for AI Standards and Innovation, January 2026, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-2.ipd.pdf

8. Technical Blog: Strengthening AI Agent Hijacking Evaluations. National Institute of Standards and Technology, Center for AI Standards and Innovation, 17 January 2025, updated 19 December 2025, viewed 8 November 2026. https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

9. LLM01:2025 Prompt Injection, OWASP Top 10 for LLM Applications 2025. OWASP GenAI Security Project, 2025 edition, viewed 8 November 2026. https://genai.owasp.org/llmrisk/llm01-prompt-injection/

10. LLM06:2025 Excessive Agency, OWASP Top 10 for LLM Applications 2025. OWASP GenAI Security Project, 2025 edition, viewed 8 November 2026. https://genai.owasp.org/llmrisk/llm062025-excessive-agency/

11. OWASP Top 10 for Agentic Applications for 2026. OWASP GenAI Security Project, 9 December 2025, viewed 8 November 2026. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

12. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Publications Office of the European Union, EUR-Lex, Official Journal L, 12 July 2024, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

13. Article 113: Entry into force and application. European Commission, AI Act Service Desk, text based on the EUR-Lex consolidated version as at 27 July 2026, viewed 8 November 2026. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-113

14. Article 26: Obligations of deployers of high-risk AI systems. European Commission, AI Act Service Desk, viewed 8 November 2026. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-26

15. Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). Publications Office of the European Union, EUR-Lex, Official Journal L, 24 July 2026, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202601744

16. In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI. Shayne Longpre, Kevin Klyman, Ruth E. Appel and others, arXiv:2503.16861, version 2, 25 March 2025, viewed 8 November 2026. https://arxiv.org/abs/2503.16861

17. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. Shunyu Yao, Noah Shinn, Pedram Razavi and Karthik Narasimhan, arXiv:2406.12045, 17 June 2024, viewed 8 November 2026. https://arxiv.org/abs/2406.12045

18. The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against Llm Jailbreaks and Prompt Injections. Milad Nasr, Nicholas Carlini, Chawin Sitawarin and others, arXiv:2510.09023, 10 October 2025, viewed 8 November 2026. https://arxiv.org/abs/2510.09023

19. Establishing Best Practices for Building Rigorous Agentic Benchmarks. Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun and others, arXiv:2507.02825, version 5, 7 August 2025, viewed 8 November 2026. https://arxiv.org/abs/2507.02825

20. How is ChatGPT's behavior changing over time? Lingjiao Chen, Matei Zaharia and James Zou, arXiv:2307.09009, version 3, 31 October 2023, viewed 8 November 2026. https://arxiv.org/abs/2307.09009

21. AI Agents That Matter. Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir and Arvind Narayanan, arXiv:2407.01502, 1 July 2024, viewed 8 November 2026. https://arxiv.org/abs/2407.01502

22. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer and Florian Tramèr, arXiv:2406.13352, version 3, 24 November 2024, viewed 8 November 2026. https://arxiv.org/abs/2406.13352

23. Towards a Science of AI Agent Reliability. Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala and Arvind Narayanan, arXiv:2602.16666, version 3, 2 June 2026, viewed 8 November 2026. https://arxiv.org/abs/2602.16666

24. Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation. Sayash Kapoor, Benedikt Stroebl, Peter Kirgis and others, arXiv:2510.11977, 13 October 2025, viewed 8 November 2026. https://arxiv.org/abs/2510.11977

25. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Lianmin Zheng, Wei-Lin Chiang, Ying Sheng and others, arXiv:2306.05685, version 4, 24 December 2023, viewed 8 November 2026. https://arxiv.org/abs/2306.05685

26. Demystifying evals for AI agents. Anthropic, 9 January 2026, viewed 8 November 2026. https://www.anthropic.com/research/demystifying-evals-for-ai-agents

27. Semantic Conventions for GenAI agent and framework spans. OpenTelemetry, semantic-conventions-genai repository on GitHub, status Development, last changed 5 November 2026, viewed 8 November 2026. https://github.com/open-

telemetry/semantic-conventions-genai/blob/main/docs/gen-ai/gen-ai-agent-spans.md

28. Semantic conventions for generative AI tool execution (execute-tool.md). OpenTelemetry, semantic-conventions-genai repository on GitHub, status Development, viewed 8 November 2026. https://github.com/open-telemetry/semantic-conventions-genai/blob/main/docs/gen-ai/execute-tool.md

29. Guidelines for secure AI system development. UK National Cyber Security Centre, US Cybersecurity and Infrastructure Security Agency and international partners, 27 November 2023, viewed 8 November 2026. https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

30. Model AI Governance Framework for Agentic AI, Version 1.5. Infocomm Media Development Authority, Singapore, 20 May 2026, updated 5 June 2026, viewed 8 November 2026. https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

31. FTC Finalizes Order with DoNotPay That Prohibits Deceptive 'AI Lawyer' Claims, Imposes Monetary Relief, and Requires Notice to Past Subscribers. U.S. Federal Trade Commission, 11 February 2025, viewed 8 November 2026. https://www.ftc.gov/news-events/news/press-releases/2025/02/ftc-finalizes-order-donotpay-prohibits-deceptive-ai-lawyer-claims-imposes-monetary-relief-requires

32. FTC Order Requires Workado to Back Up Artificial Intelligence Detection Claims. U.S. Federal Trade Commission, 28 April 2025, viewed 8 November 2026. https://www.ftc.gov/news-events/news/press-releases/2025/04/ftc-order-requires-workado-back-artificial-intelligence-detection-claims

33. Moffatt v. Air Canada, 2024 BCCRT 149. Civil Resolution Tribunal of British Columbia, 14 February 2024, viewed 8 November 2026. https://decisions.civilresolutionbc.ca/crt/crtd/en/item/525448/index.do

About this report

Published by Rillor on 11 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Questions to ask before you deploy a vendor's agent. 11 November 2026. https://rillor.com/insights/agent-vendor-testing-questions

Notices

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send us the tasks your agent handles, the model versions it runs on and the decisions it is trusted with. Contact us at rillor.com/contact?type=evaluation or contact@rillor.com. More on SI evaluation: rillor.com/si/evaluation