BUYER CHECKLIST · VISIQ LABS · SEPTEMBER 2026

# The AI Application Security Buyer Checklist

67 checks to tell a connected lifecycle from a drawn one

For security, platform, risk, and procurement teams evaluating AI application security platforms

Gartner's September 2026 Emerging Market Quadrant for AI Application Security describes technologies that combine security testing, exposure management and runtime defense to detect, alert on or block threats against enterprise-built AI applications and agents, and names three core capabilities: automated discovery and inventory, automated security testing, and runtime defense.

Most vendors now claim all three, and the pictures look alike. This checklist is built to test the claim. Part 1 checks the three core capabilities. Part 2 checks what connects them. Part 3 is a ten-step script you can make a vendor run on your own agent.

**How to use it.** Ask for answers in writing, then ask to see each one in the product on your own agent. Tick a box only when you have seen it, not when you have been told. Sections marked **A demo cannot fake this** are the ones to press hardest.

Vendor _____ Evaluator _____ Date _____

PART 1

## The three core capabilities

The capabilities Gartner's September 2026 research names as the core of the market. Check each one against what you have seen, not what you were told.

### 1 Discovery and inventory

- [ ] Finds agent frameworks, MCP servers and their tool surfaces, local model runtimes, coding and CLI agents, provider API keys, and reachable vector stores and databases
- [ ] Covers development, staging, production and developer laptops
- [ ] Read-only: does not execute discovered servers, invoke package managers or read source code
- [ ] Key-shaped values are fingerprinted, never collected
- [ ] Reports what it could not inspect, so "nothing found" and "could not look" are different results
- [ ] Tags every discovered agent with a governance status (ungoverned, instrumented, governed)

**Red flag.** A count of "AI assets" with no coverage statement and no path from a discovered agent to a control. Checked _____ of 6

## 2 Security testing

### EVALUATE

- [ ] Tests the shipped combination (model, system prompt, tool schemas, policy), not just the model
- [ ] Monitor-only mode records what each decision would have been without blocking anything
- [ ] A new or changed rule can be simulated against recent real traffic, with the outcome distribution shown before it takes effect

### SANDBOX

- [ ] The test target is a faithful clone of the production agent, with mocked tools, so nothing a test achieves touches a real system
- [ ] New skills and tools are detonated in isolation against canary tokens before they run live
- [ ] A skill can be held fail-closed until it clears testing

### RED TEAM

- [ ] Covers direct and indirect prompt injection, instruction conflicts, sensitive-data leakage, tool abuse with over-broad arguments, privilege escalation, excessive delegation and policy evasion
- [ ] Techniques map to a public taxonomy (MITRE ATLAS, OWASP Top 10 for Agentic Applications)
- [ ] Verdicts are computed from the agent's actual behavior, not its self-report
- [ ] A finding carries the exchange that triggered it and becomes a proposed control that a human must accept

**Red flag.** A PDF of attack results and a separate guardrail product, or a "sandbox" that is a demo tenant. Checked _____ of 10

## 3 Runtime defense A DEMO CANNOT FAKE THIS

- [ ] The decision is made before the tool function runs, and it can stop the call. An observer callback cannot
- [ ] Outcomes include permit, mask, deny and require approval; on retrieval, allow, redact, deny and escalate
- [ ] Policy distinguishes reads, writes, deletes, administrative actions, funds movements and delegations
- [ ] Policy evaluates the actual target, the arguments, the agent's identity and trust tier, the credential context and the delegation grant
- [ ] On deny, the agent receives a structured reason and can continue. The run does not crash
- [ ] Policy is evaluated locally against a cached bundle, with a stated latency budget, not a vendor round trip per call

**Red flag.** "Real-time" means alerting within seconds of the side effect. Checked _____ of 6

PART 2

# What connects them

Vendors now describe the same discover, test, protect loop. These six sections tell you whether the stages are connected or only drawn next to each other.

## 4 Authority and delegation

- [ ] An agent-to-agent handoff is an authorization decision, not a function call
- [ ] The child's authority is the parent's authority intersected with an explicit grant, so scope can only narrow
- [ ] Grants are signed and revocable, and revocation cascades to every descendant
- [ ] There is a maximum chain depth
- [ ] Each delegated action is checked against the grant and recorded
- [ ] A delegation that no rule covers is denied

**Red flag.** "Multi-agent support" with no grant object. Checked _____ of 6

## 5 Failure behavior A DEMO CANNOT FAKE THIS

**Get every answer in writing**

- [ ] Written answer: what happens when the policy bundle cannot be fetched while an agent is in enforce mode
- [ ] Written answer: what happens when an action matches no rule, whether the default is permit, deny or approval, and who can change it
- [ ] Written answer: what happens when masking cannot resolve
- [ ] Written answer: what happens when an approval times out or the approver is unreachable
- [ ] Written answer: what happens when the vendor's backend is unreachable
- [ ] The documentation states which behaviors are configurable and which are fixed, and it matches the marketing page

**Red flag.** "Fail-closed" with no documentation of what that means per case, or a documentation page that contradicts the marketing page. Checked _____ of 6

## 6 Evidence and provenance A DEMO CANNOT FAKE THIS

- [ ] The decision is made before execution, and the record is bound to the execution event it governed
- [ ] A receipt contains agent, target, operation, arguments or their hash, outcome, policy version, approver and timestamp
- [ ] Receipts are signed, the key location is stated, and keys can rotate without invalidating history
- [ ] Receipts are batched into a tamper-evident structure and timestamped by an independent authority
- [ ] A third party can verify a receipt offline with standard libraries, without the vendor's API
- [ ] A written statement says what the receipts do not prove
- [ ] Protection of the evidence store against the operator (write-once storage) is documented, including whether it is on by default

**Red flag.** "Audit log" and "compliance report" with no verification procedure. Checked _____ of 7

## 7 Bypass paths and boundary

- [ ] The vendor states what happens when an agent uses a raw API credential outside the governed framework
- [ ] The vendor states which shell, Kubernetes, cloud, database and network calls are governed, and through which integration
- [ ] The vendor states what is visible inside a compromised third-party workload or an unmanaged endpoint
- [ ] The vendor names the surrounding controls that remain necessary (IAM, network segmentation, secrets management, admission control, endpoint detection)
- [ ] The vendor will document your governed paths and your bypass paths as part of the deployment

**Red flag.** "Complete coverage." Checked _____ of 5

## 8 Rollout

- [ ] Every new agent starts in monitor mode automatically
- [ ] Enforcement can be turned on per agent and per operation class
- [ ] New rules are created disabled, simulated against real traffic, and rejected if they would block more than a stated fraction of legitimate work
- [ ] An AI-drafted rule cannot be enabled without a human
- [ ] Lines of code changed and time to first decision are stated

**Red flag.** A big-bang enforcement switch. Checked _____ of 5

## 9 Closed-loop operations

- [ ] A red-team finding becomes a proposed runtime rule without re-authoring it, and a human must accept it
- [ ] An anomaly becomes a draft rule with a dry run over the agent's own recent decisions
- [ ] A sandbox verdict on a new skill blocks that skill from production until it clears
- [ ] Anomalous behavior can move an agent to approval-only, lower its trust tier or suspend it, and the change is recorded
- [ ] Uncovered actions observed in monitor mode surface with a recommended covering rule
- [ ] Security, platform, risk and development teams see one evidence trail

**Red flag.** Each stage exports a CSV to the next. Checked _____ of 6

PART 3

# Proof-of-value script

Require the vendor to run all ten steps on one of your agents, in one session, and produce the evidence at the end. Steps 8 and 10 are pass or fail: a partial is a fail.

- [ ] **1** Discover the agent and map its tools, credentials and reachable data.
- [ ] **2** Run it in monitor mode and show the decisions it would have received.
- [ ] **3** Simulate a new rule against that traffic and show the outcome distribution before enabling it.
- [ ] **4** Execute a prompt-injection scenario against a sandboxed clone and show a verdict computed from behavior.
- [ ] **5** Convert that finding into a proposed rule and accept it as a reviewer.
- [ ] **6** Detonate a newly installed skill in a sealed test cell and show the canary finding and the proposed blacklist entry.
- [ ] **7** Permit a low-risk production action and show the receipt.
- [ ] **8** Deny an unauthorized write before execution and show what the agent received.
  PASS / FAIL. A PARTIAL IS A FAIL.
- [ ] **9** Mask sensitive arguments on an action that proceeds.
- [ ] **10** Pause a high-risk action for approval, let the approval expire, confirm the action did not execute, then verify the full receipt chain offline without the vendor's API.
  PASS / FAIL. A PARTIAL IS A FAIL.

**Scoring.** Score each step pass, partial or fail. A partial on step 8 or 10 is a fail. Passed _____ of 10

## The buying question

Do not ask only "can this platform detect AI risk?"
Ask: "Can it discover, evaluate, challenge, authorize and prove the behavior of our AI applications across their lifecycle, and can we verify the proof ourselves?"

*This checklist is published by VisIQ Labs. It is a product perspective, labeled as such. It does not establish compliance with any law, standard or certification scheme. Gartner does not endorse any vendor, product or service depicted in its research publications. VisIQ Labs has not been evaluated in Gartner's Emerging Market Quadrant for AI Application Security, and nothing here should be read to imply otherwise.*

Source for the market description: Gartner, Emerging Market Quadrant for AI Application Security, September 2026 (gartner.com/en/documents/8378981).

## Run the script on one of your agents.

Bring one agent and one consequential action. We will run all ten steps in one session and hand you the evidence at the end.

**Request a technical review** **30-question scorecard** **Explore the VisIQ Sandbox**