

A GUIDE FOR SECURITY AND BUSINESS LEADERS

Best Practices Checklist to Protect from AI Cyberattacks

How do you know you're really protected? Here's the checklist I'd use.

AI model symbols illustrative only.

WHAT'S INSIDE

- A 12-item checklist you can print and hand to your team
- Why AI has made every company a target
- How to answer the board when they ask if you're protected
- How one stolen key can walk an attacker to your crown jewels
- A real customer who found 30 more issues in about an hour

Go to the checklist

Koroush Saraf, Co-founder, Gadriel.ai, San Francisco

SEE A SAMPLE REPORT

This is what you get back from a test: the attack path, the proof, the business impact and the fix.

See a sample report

Order an assessment

Want 20% off? Email **info@gadriel.ai**

1. Why AI changes the picture

Attackers used to pick targets based on what they would earn for the effort. Automation changed that, so companies of every size are now worth attacking.



Anthropic documented a 2025 campaign where AI did an estimated 80 to 90% of the work, and humans stepped in at only 4 to 6 decision points. Their conclusion was that less experienced groups can now potentially run attacks at scale.[1] IBM's 2026 research puts the economics plainly: an attack can be launched for thousands of dollars, while a breach costs millions.[2]

31%
of breaches started with an exploited vulnerability, the top entry point in Verizon's 19 years of data[3]

80 to 90%
of one reported campaign's work was done by AI agents[1]

$4.99M
average cost of a data breach worldwide in IBM's 2026 report[2]

50%
Gartner's forecast cut in the time to exploit account exposures by 2027, as AI agents automate the steps[4]

Where the exposure comes from

Mandiant's 2025 investigations found that most successful intrusions still trace back to ordinary human and system problems. Your team controls all of these:[5]

Configuration
Cloud and network settings drift over time. 27% of breaches targeting AI systems involved cloud misconfiguration.[2]

Human error
People move fast. Shared credentials, skipped steps and default settings add up.

Software flaws
Known vulnerabilities in what you run. This is the most common way in.[3]

Identity
Credentials and tokens that reach further than they should. Credential abuse shows up in 39% of breaches.[6]

Unknown APIs
Integrations nobody keeps a list of. 27% of AI-system breaches began with APIs, apps or plug-ins.[2]

AI coding makes the surface bigger. In organizations that use AI coding tools, AI writes roughly half of the committed code, and about 44% of tested generation tasks introduced a known vulnerability.[7] GitGuardian counted 28.65 million new secrets in public commits in 2025, up 34%.[8] More applications means a healthier business, and also a longer list to know about and test.

Want to see what an attack-path report looks like? **Open the sample report**.

2. What the board will ask

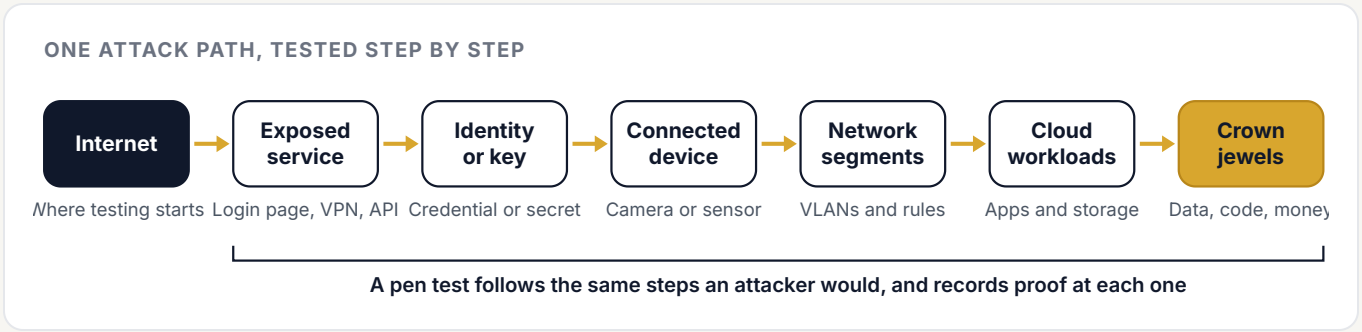


> **"Are we protected against AI-assisted attacks, and how do we know?"**

You can buy firewalls, SASE and cloud security and still not know whether the one vulnerability that lets an attacker in is covered. Buying a control shows it's in place. Testing shows how your controls work together in your real environment, including any path from an exposed service to something valuable. Gartner calls this adversarial exposure validation: continuous, automated evidence that an attack can actually succeed. It tells you which exposures are exploitable, which beats a theoretical vulnerability list.[9]

3. Test like an attacker

The best practice is to run the same attack an attacker would, in the same order. It's called a mock attack, or penetration testing.



What a good test does

- It finds everything reachable from the internet, including systems people forgot about.
- It chains findings together. A path matters more than any single issue.
- It records proof at each step, so you can check the fix later.

What you get

- A short, ranked list of the paths that actually lead somewhere.
- Proof of whether your existing controls are doing their job.
- A way to prioritize alongside CISA's Known Exploited Vulnerabilities catalog: you see which entries are reachable in your environment.[10]

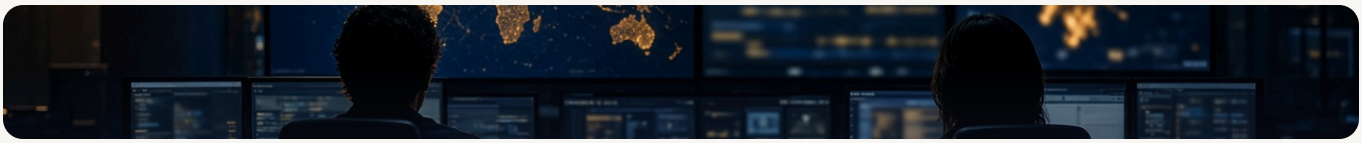
Compare it to a real report: **open the sample report**

4. Test after every change

Every patch, upgrade and configuration change alters your systems. Something that tested clean today can look different in 30 days.

AN EXAMPLE 30 DAYS: EVERY CHANGE IS A CHANCE TO TEST AGAIN





"You don't know what you don't know until something pushes you to look."

Co-founder and Head of Engineering, enterprise networking software company. Read the customer story on page 7

Why a regular rhythm makes sense

- Verizon's 2026 report says AI is shrinking the window for defense from months to hours.[3] In Verizon's data, fully fixing vulnerabilities on CISA's exploited list took a median of 43 days.[6]
- Mandiant estimates the mean time to exploit at 7 days before a patch is even released, and exploits were the top way in for the sixth year running, at 32% of intrusions.[5] Testing sits alongside patching. It shows you which services and admin interfaces are reachable, so you can restrict access while a patch is pending, which is what Barracuda's own guidance recommends.[11]
- Gartner lists automated scheduling for more frequent testing as a core feature of adversarial exposure validation.[9]

THE CONTINUOUS TEST LOOP

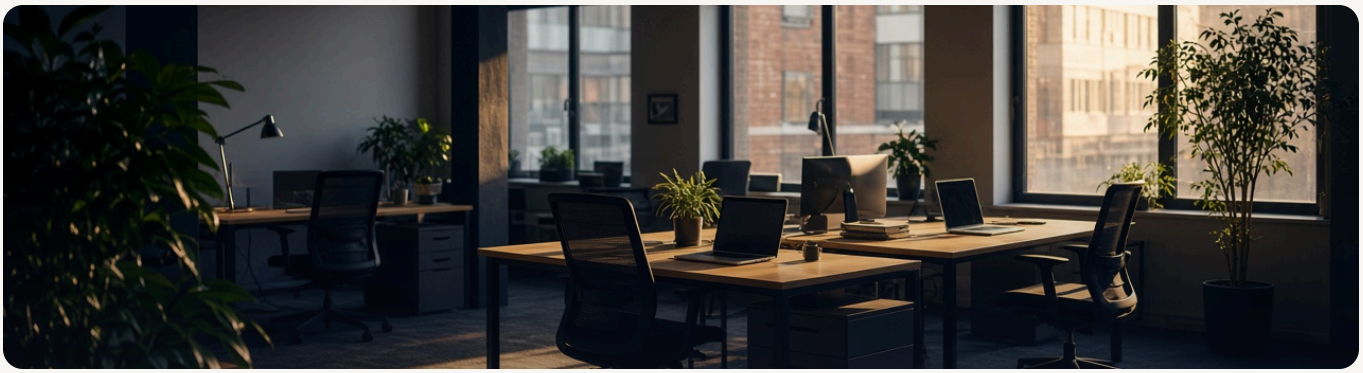


How often I'd test

- After any significant change: a patch cycle, an upgrade, a new app or API, a cloud move or a firewall edit.
- Every month for anything facing the internet.
- After every fix, to confirm it worked.

That cadence is my suggestion, so adjust it to how often your environment changes. Gartner's planning assumption that CTEM programs cut breaches by two-thirds by 2026 is a 2024 forecast, so treat it that way.[12]

5. Why testing used to be hard



Penetration testing has been a manual craft for a long time: scoping calls, scheduling and weeks of an expert's time. Most companies tested once a year or less, so regular testing was out of reach for a lot of teams.

TRADITIONAL ENGAGEMENT

Scoping calls and scheduling

Weeks of expert time

Once a year, or less

→

AUTOMATED, AI-ASSISTED TESTING

Order online in minutes

Report in hours to days

After every meaningful change

6. Defenders get the same AI tools

The AI that makes attackers faster is available to defenders too. Anthropic encourages security teams to use AI for defense, including vulnerability assessment.[1] IBM found that organizations using AI and automation in security operations cut breach costs by almost $2 million on average.[2] Gartner notes that vendor-supplied scenarios let teams scale testing with little to no hacking knowledge.[9]

Testing regularly only works if the process and the price let you run it after every change. The earlier automated platforms came from large, venture-funded vendors, and they're usually sold as enterprise contracts. A newer generation can start online with published pricing.

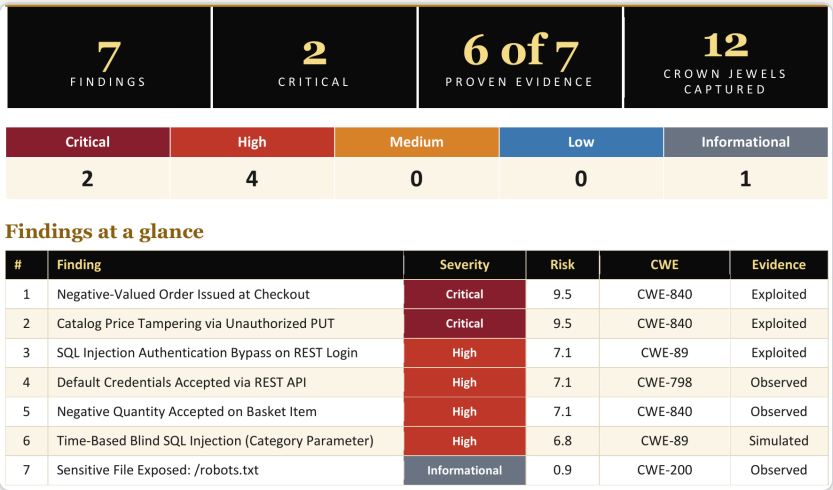
Six questions to ask any testing provider

1. What counts as a target, and how do you confirm I own it?
2. How do you keep testing safe for my production systems?
3. Do you combine known-exploit checks with AI-driven path exploration?
4. Does every finding come with proof, impact and a fix?
5. Can I run it after every change?
6. How fast can I confirm a fix?

Here's how Gadriel answers them in a real report: **gadriel.ai/order-assessment**

7. Gadriel Deployment Assurance (GDA)

External penetration testing you can order online without talking to anyone. The report comes back within 2 business days, often in about 12 hours.



7 FINDINGS	2 CRITICAL	6 of 7 PROVEN EVIDENCE	12 CROWN JEWELS CAPTURED

Critical	High	Medium	Low	Informational
2	4	0	0	1

Findings at a glance

#	Finding	Severity	Risk	CWE	Evidence
1	Negative-Valued Order Issued at Checkout	Critical	9.5	CWE-840	Exploited
2	Catalog Price Tampering via Unauthorized PUT	Critical	9.5	CWE-840	Exploited
3	SQL Injection Authentication Bypass on REST Login	High	7.1	CWE-89	Exploited
4	Default Credentials Accepted via REST API	High	7.1	CWE-798	Observed
5	Negative Quantity Accepted on Basket Item	High	7.1	CWE-840	Observed
6	Time-Based Blind SQL Injection (Category Parameter)	High	6.8	CWE-89	Simulated
7	Sensitive File Exposed: /robots.txt	Informational	0.9	CWE-200	Observed

FROM A GADRIEL SAMPLE REPORT

Every report starts with a one-page summary: findings by severity, how well each one is proven, and the crown jewels we captured as evidence. The detailed findings follow, each with its impact and the fix.

This is an excerpt of our sample report, run against a public demo application. It has no customer data in it.

See a sample report

GDA tests web applications, APIs, internet-facing network nodes and AI agent endpoints. It combines deterministic, CVE-based checks with LLM-based attack techniques, validates multi-step attack paths and reports each finding with evidence and remediation guidance.[13,14]

What you should know

- A target is one externally reachable asset: a hostname, URL or IP address. Asset discovery is free.[13]
- You confirm you own it, and we review the scope before testing starts.[13]
- The report has the findings, evidence, business impact, remediation and compliance mapping.[14]
- If we find no real vulnerability, you get a full refund.[14]
- Internal network, cloud, segmentation, continuous SaaS and air-gapped testing are scoped with a custom quote.[13]

PRICING, YEAR-END OFFER

First target	**$500**
Each additional target	**$300**
First target with the optional AI-powered add-on	**$1,000**

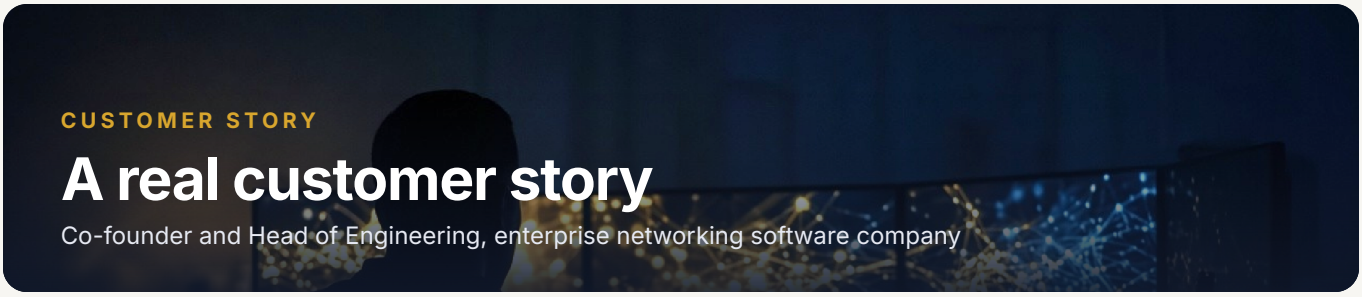
External assessments only. The add-on uses the same $500 and $300 tiers.[13]

Order an assessment

See a sample report

Accepted into Anthropic's Cyber Verification Program

Anthropic's Cyber Verification Program gives qualifying security professionals access to its advanced cyber capabilities.[15,16] For us, that means access to frontier models for defensive testing, so teams can find and fix exposures before attackers get to them.



CUSTOMER STORY

A real customer story

Co-founder and Head of Engineering, enterprise networking software company

1 hour
from order to testing starting

30
additional findings

3
critical, including new entry points

$1,000
first target with AI-powered testing

1 Release → 2 Incident and fix → 3 Order online → 4 Test starts in 1 hour → 5 30 findings, 3 critical → 6 Patched

An enterprise networking software company had built a management application, and its co-founder asked the team to pentest the latest release. The company's most recent pentest had been done some time earlier for its SOC 2 program. A pentest only describes the release that existed when it was run, but with that spend behind them, the team planned to wait before paying for another one.

About a month later, the application had a security incident. The co-founder fixed it, and to make sure nothing else was open, he asked Gadriel for a pentest of the infrastructure.

He opened the assessment page, filled out the form and paid by card. There was no sales call, and testing started about an hour later.

The report came back with 30 additional findings, 3 of them critical, including other entry points nobody had detected. His team patched them. The total cost was $1,000 to check the application: the first target with AI-powered testing.



> **You don't know what you don't know until something pushes you to look. Once we ran it, the value was obvious: a very high return on a very small investment.**
>
> Co-founder and Head of Engineering, enterprise networking software company

What this story shows

The best practice is to test each new release before it goes live, because a pentest only describes the release it covered. When ordering takes minutes and testing starts within the hour, that's easy to do. Teams with 10 applications, or four tests a year, will want to rerun it often, and Gadriel offers a periodic license for continuous testing: **info@gadriel.ai**.

See a sample report | **Order an assessment**

This customer account was shared with Gadriel. We haven't named the customer or the company.

8. Using the report for audits and cyber insurance



Auditors and compliance programs

Programs like SOC 2 Type II and ISO 27001 look for evidence of how you find and manage exposures. A report with attack paths, evidence, business impact, fixes and a retest gives a reviewer a clear picture of a working testing program. Your auditor decides what evidence your audit needs, and a GDA report is built to be a useful piece of it.

How to use it

- Share the executive summary at the start of the audit period.
- Keep each dated report and its retest together.
- Link findings to tickets and fix dates.

Cyber insurance applications

Applications keep getting longer and more detailed, and underwriters focus on essential controls like patch management.[17] Marsh's application asks about MFA, patching, training, backups and incident history.[18] Coverage works best when what you state matches what you have in place. In Travelers v. International Control Services (2022), the insurer sought rescission over statements about MFA, and a judgment rescinding the policy was entered.[17]

How to use it

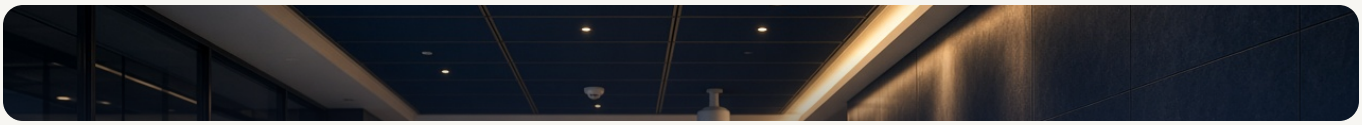
- Answer exposure and patching questions from a dated report.
- Show a retest after each upgrade.
- Ask your broker which evidence your carrier values.

From test to evidence

Test Run the assessment → **Fix** Close the paths it shows → **Retest** Confirm the fixes → **File** Keep dated reports → **Share** Auditor and broker

Want to see one? The sample report is at **gadriel.ai/order-assessment**. Email **info@gadriel.ai** for 20% off your first assessment.

9. What the findings look like, and what to do about them



Identity

Logins that can be reached, reused credentials and access that's broader than it needs to be.

Use phishing-resistant MFA, least privilege and rotation.[4]

Older software

Versions that already have published fixes.

Have your MSP or IT team upgrade, ideally within about 48 hours.

Third-party components

Libraries and appliances that need updates. Third parties were involved in 48% of breaches.[3]

Update them and keep an inventory.

Your own app code

Weaknesses in code you or your AI tools wrote, including vibe-coded apps.

Fix it in the code, and check it as it's written (GCA).

Keys and secrets

Keys left inside an app, which a scan can extract.[8]

Rotate them and move them to a secrets manager.

Reachable agentic APIs

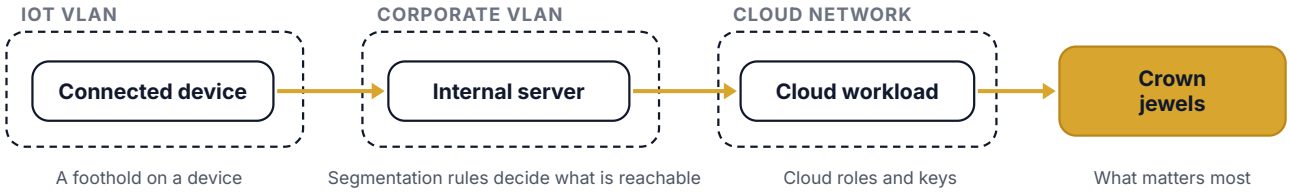
Agent, MCP and connector endpoints open to the internet. Connectors don't need to be reachable from outside.

Require authentication and allow known sources only.

Connected devices as a way in

Cameras, sensors, VoIP phones and storage devices are on your network too.[19] Sophisticated attackers have long used one device as a bridge across VLANs and segmentation toward the cloud and the crown jewels. With AI, that route is now practical to test yourself.

Segment devices, limit what they can reach, and retest the rules after network changes.



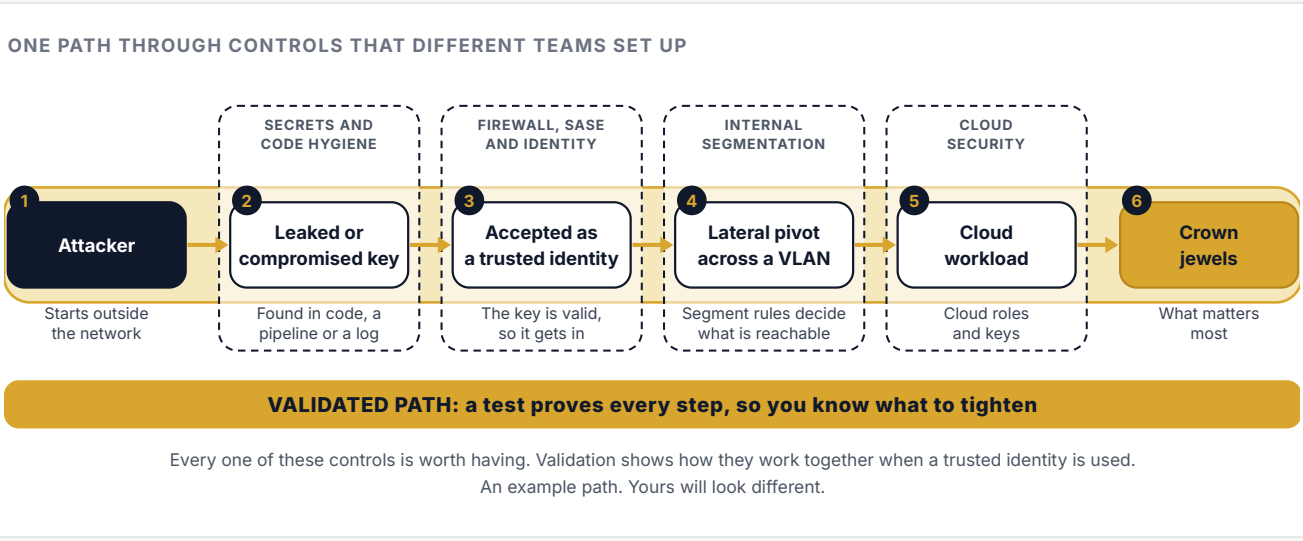
In public reports lately

Here's what an attack-path test is built to find. **Barracuda's advisory** covers admin interfaces reachable from the internet, where restricting access reduces exposure.[11] CISA warns about chained flaws in remote-access appliances that were used for credential theft (**Ivanti**).[20] And **Railway's incident report** describes one internal credential used to move across systems.[21]

These categories all show up in the sample report: **open the sample report**

10. Why validation matters

Most companies run a good mix of controls: firewalls, internal segmentation, SASE, cloud security and identity. Different teams set each one up, and each one does its job.



Every one of these controls is worth having. Validation shows how they work together when a trusted identity is used. An example path. Yours will look different.

Why one trusted identity matters

Non-human identities are service accounts, API keys, tokens and AI agent credentials. There are a lot of them now, and your systems trust them by design.[22,23] If one gets exposed, an attacker shows up with a valid credential, so the front door sees a legitimate session.

What happens next depends on how your other controls fit together. Each layer can be right on its own and the combined path can still be open.

What validation adds

Each control gets checked by the team that owns it. Validation walks the whole route the way an attacker would, using a credential that's meant to be trusted, and records proof at every step.

You get a short list of paths to tighten, and a retest to confirm the fix. It's the only way I know to see how your controls hold up together.

82 : 1
machine identities for every human, according to CyberArk's 2025 survey of 2,600 security leaders[23]

42%
of machine identities have privileged or sensitive access, and 88% of organizations define "privileged user" as humans only[23]

44%
of machine identities fall under IAM teams, per Gartner's survey of 335 IAM leaders[22]

39%
of breaches involved credential abuse at some point, the most of any action in Verizon's 2026 report[6]

Where I'd start

OWASP's Non-Human Identities Top 10 is a good shared checklist. Secret leakage, overprivileged identities and long-lived secrets are among the top risks.[24] GitGuardian counted 28.65 million new secrets in public commits in 2025.[8] Inventory your identities and who owns them, shorten how long credentials last, apply least privilege, then validate the path end to end and retest after each change. **See a sample report.**

11. Checking AI-written code: Gadriel Code Assurance (GCA)



The best time to check code is while it's being written, when the code and its context are on the developer's machine. GCA runs in the developer's environment and works with Claude Code, Cursor, Windsurf and Copilot. It checks on save, hands findings to the assistant and re-checks the fix before commit.[13,25] Waiting for the CI/CD pipeline loses that context, and a big batch of warnings slows everyone down. Veracode also advises verifying AI-written code inside the workflow.[7]

- Your source code stays on the machine.
- It runs six scan types: SAST, SCA, secrets, config, container and API.[26]
- Findings are tied to the change that caused them, instead of landing in a batch later.

FREE FOR INDIVIDUALS

The full local engine, unlimited local scans and no credit card.[13]

Try GCA free

12. Checking AI agents before they run: Gadriel Behavior Assurance (GBA)

AI agents take actions. They call tools, move data and can trigger payments, and you don't want to find a problem at runtime. GBA is a pre-flight check. It validates agents, LLM workflows, RAG, MCP and A2A integrations before they ship, across eight pillars (security, compliance, safety, operational, FinOps, coherence, teamwork and bias), using mathematical validation so the results can be reproduced.[27] For example, you can confirm an agent stays within its payment policy limits before it can ever wire money.

GCA
What AI writes
Checked on the developer's machine as the code is written.

GBA
What AI does
Validated in a pre-flight check, before runtime.

GDA
Where AI runs
Tested on your live attack surface.

Learn about GBA and the Proof of Value

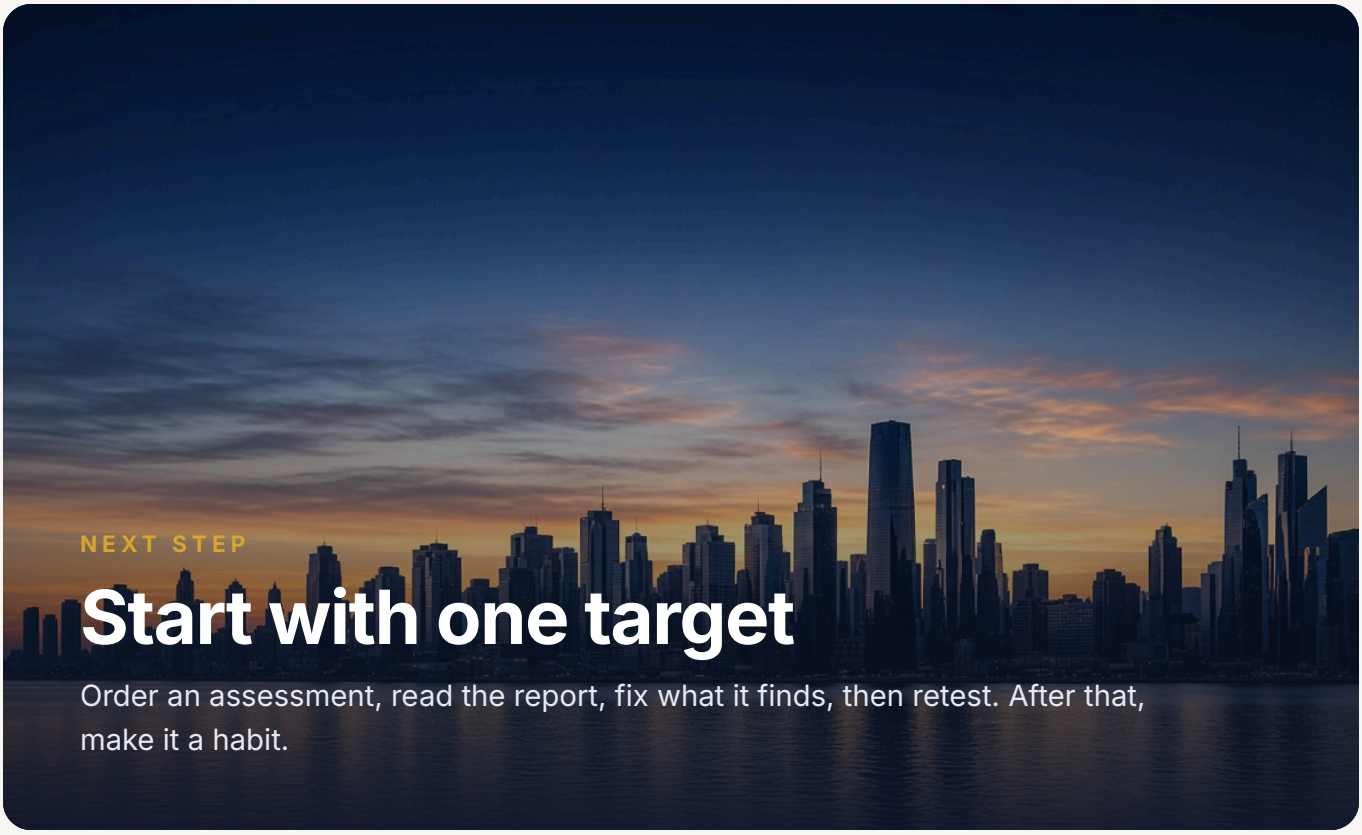
Order a GDA assessment

13. The Best Practices Checklist

Twelve things to do to protect against AI cyberattacks. Print this page and hand it to your team.

- [ ] **1 Write down your crown jewels.**
  List the data, code and systems the business depends on, and who owns each one.
- [ ] **2 List everything that faces the internet.**
  Apps, APIs, admin panels, remote access and agent endpoints, all in one place.
- [ ] **3 Run an attack-path test.**
  Get a report with proof for every finding that chains to the next one.
- [ ] **4 Test again after every significant change.**
  A patch, an upgrade or a new app should trigger a retest, plus a run every month.
- [ ] **5 Upgrade old software and components.**
  Patch anything on CISA's exploited list, ideally within about 48 hours.
- [ ] **6 Tighten up identity.**
  Use phishing-resistant MFA for admins and any login exposed to the internet.
- [ ] **7 Know your keys, tokens and service accounts.**
  Give each one an owner, least privilege and a short life, and rescan every release.
- [ ] **8 Test your controls together.**
  Find out how far a trusted identity can get across your segments and cloud roles.
- [ ] **9 Lock down agent and MCP endpoints.**
  Require authentication and allow known sources only.
- [ ] **10 Check AI-written code as it's written.**
  Developers should see findings in the editor before they commit (GCA).
- [ ] **11 Check agent behavior before it runs.**
  Any agent that touches money or data should pass a pre-flight check (GBA).
- [ ] **12 Keep every report and retest.**
  File them with dates, for your auditors and insurers.

If you only do one thing, do number 3. Here's a sample report and where to order: **gadriel.ai/order-assessment**



NEXT STEP

Start with one target

Order an assessment, read the report, fix what it finds, then retest. After that, make it a habit.

WRAPPING UP

AI has changed how attacks get built, and it has changed what defenders can do about them. The best thing I've seen teams do is simple: test the way an attacker would, test again after every change, fix what turns up and keep the reports. Fast, automated and affordable testing means any company can do that, whatever its budget. The checklist on page 12 is a good working plan.

1

See a sample report

An attack path with the proof, the impact and the fixes.

gadriel.ai/order-assessment

2

Order an assessment

The first target is $500. Report within 2 business days. No sales call.

Order online

3

Get 20% off

Email us and we'll send a coupon for an assessment.

info@gadriel.ai

Try GCA free for code your AI writes: gadriel.ai/docs/gca/overview.html

Explore GBA for AI agents: gadriel.ai/behavioral-assurance

Everything else: gadriel.ai

About the author. Koroush Saraf is co-founder and CMO of Gadriel.ai in San Francisco. Before Gadriel he was a vice president at Palo Alto Networks, and he spent about ten years at Fortinet leading wireless, switching, identity and access, SD-WAN and SASE products.

Sources

The numbers in the text match this list. I checked the third-party figures against the linked pages in November 2026. The photographs are AI-generated illustrations and don't show real people, brands or text. The report excerpt on page 6 is from Gadriel's own sample report.

1 **Anthropic**, Disrupting the first reported AI-orchestrated cyber espionage campaign, Nov 13, 2025. https://www.anthropic.com/news/disrupting-AI-espionage

2 **IBM**, IBM Study: One in Four Malicious Breaches are AI-Enabled, Costing Companies $6 Million on Average (2026 Cost of a Data Breach Report), Jul 29, 2026. https://newsroom.ibm.com/2026-07-29-ibm-study-one-in-four-malicious-breaches-are-ai-enabled,-costing-companies-6-million-on-average

3 **Verizon**, Breach entry point, 2026 DBIR finds (news release), May 19, 2026. https://www.verizon.com/about/news/breach-industry-wide-dbir-finds

4 **Gartner**, Gartner Predicts AI Agents Will Reduce The Time It Takes To Exploit Account Exposures by 50% by 2027, Mar 18, 2025. https://www.gartner.com/en/newsroom/press-releases/2025-03-18-gartner-predicts-ai-agents-will-reduce-the-time-it-takes-to-exploit-account-exposures-by-50-percent-by-2027

5 **Google Cloud / Mandiant**, M-Trends 2026: Data, Insights, and Strategies From the Frontlines, Mar 23, 2026. https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2026

6 **Verizon**, 2026 Data Breach Investigations Report, May 2026. https://www.verizon.com/business/resources/Te26/reports/2026-dbir-data-breach-investigations-report.pdf

7 **Veracode**, 2026 GenAI Code Security Report, Jul 28, 2026. https://www.veracode.com/blog/2026-genai-code-security-report-ai-risk/

8 **GitGuardian**, The State of Secrets Sprawl 2026, Mar 17, 2026. https://blog.gitguardian.com/the-state-of-secrets-sprawl-2026/

9 **Gartner Peer Insights**, Adversarial Exposure Validation Reviews and Ratings (market definition and mandatory features), Updated Apr 2026. https://www.gartner.com/reviews/market/adversarial-exposure-validation

10 **CISA**, Known Exploited Vulnerabilities Catalog, Accessed Nov 18, 2026. https://www.cisa.gov/known-exploited-vulnerabilities-catalog

11 **Barracuda Networks**, Pre-Authentication Authorization Bypass and Command Injection Vulnerabilities in Certain Barracuda Products, Nov 2, 2026 (updated Nov 9, 2026). https://trust.barracuda.com/security/pre-authentication-authorization-bypass-command-injection

12 **Gartner**, Gartner Identifies the Top Cybersecurity Trends for 2024, Feb 22, 2024. https://www.gartner.com/en/newsroom/press-releases/2024-02-22-gartner-identifies-top-cybersecurity-trends-for-2024

13 **Gadriel.ai**, Pricing, Accessed Nov 18, 2026. https://gadriel.ai/pricing

14 **Gadriel.ai**, External Attack Surface Penetration Testing, order page, Accessed Nov 18, 2026. https://gadriel.ai/order-assessment

15 **Anthropic**, Expanding the Cyber Verification Program, 2026. https://www.anthropic.com/news/cyber-verification-program

16 **Claude Help Center (Anthropic)**, Cyber Verification Program, Accessed Nov 18, 2026. https://support.claude.com/en/articles/14604842-cyber-verification-program

17 **Lockton**, Travelers v. ICS underscores need to respond carefully to cyber insurance application questions, Sep 15, 2022. https://global.lockton.com/us/en/news-insights/travelers-v-ics-underscores-need-to-respond-carefully-to-cyber-insurance

18 **Marsh Cyber Accelerate**, Cyber insurance application requirements, Accessed Nov 18, 2026. https://affinity.marsh.com/cyber-accelerate/application-requirements.html

19 **Forescout Research, Vedere Labs**, Forescout Announces Riskiest Connected Devices of 2025, Apr 9, 2025. https://www.forescout.com/press-releases/forescout-announces-riskiest-connected-devices-of-2025-iomt-devices-increasingly-vulnerable/

20 **CISA**, CISA Urges Patching Newly Exploited Vulnerabilities in Ivanti Connect Secure, Oct 13, 2026 (updated Oct 30, 2026). https://www.cisa.gov/news-events/alerts/2026/10/13/cisa-urges-patching-newly-exploited-vulnerabilities-ivanti-connect-secure

21 **Railway**, Incident Report: October 21st Security Attack, Nov 10, 2026. https://blog.railway.com/p/incident-report-october-security-incident

22 **Gartner**, Gartner Identifies the Top Cybersecurity Trends for 2025 (Trend 2: Managing Machine Identities), Mar 3, 2025. https://www.gartner.com/en/newsroom/press-releases/2025-03-03-gartner-identifiesthe-top-cybersecurity-trends-for-2025

23 **CyberArk**, 2025 Identity Security Landscape (survey of 2,600 security decision makers, fielded Jan to Feb 2025 by Vanson Bourne), Apr 2025. https://mysecuritymarketplace.com/mp-files/2025-identity-security-landscape.pdf/

24 **OWASP Foundation**, OWASP Non-Human Identities Top 10, 2025, 2025. https://owasp.org/www-project-non-human-identities-top-10/

25 **Gadriel.ai**, GCA AI Coding Assistant Integration, Accessed Nov 18, 2026. https://gadriel.ai/docs/gca/ai-coding-integration.html

26 **Gadriel.ai**, Gadriel Code Assurance overview, Accessed Nov 18, 2026. https://gadriel.ai/docs/gca/overview.html

27 **Gadriel.ai**, AI Behavior Assurance (GBA), Accessed Nov 18, 2026. https://gadriel.ai/behavioral-assurance

Want to see your own attack paths? gadriel.ai/order-assessment 20% off: **info@gadriel.ai**