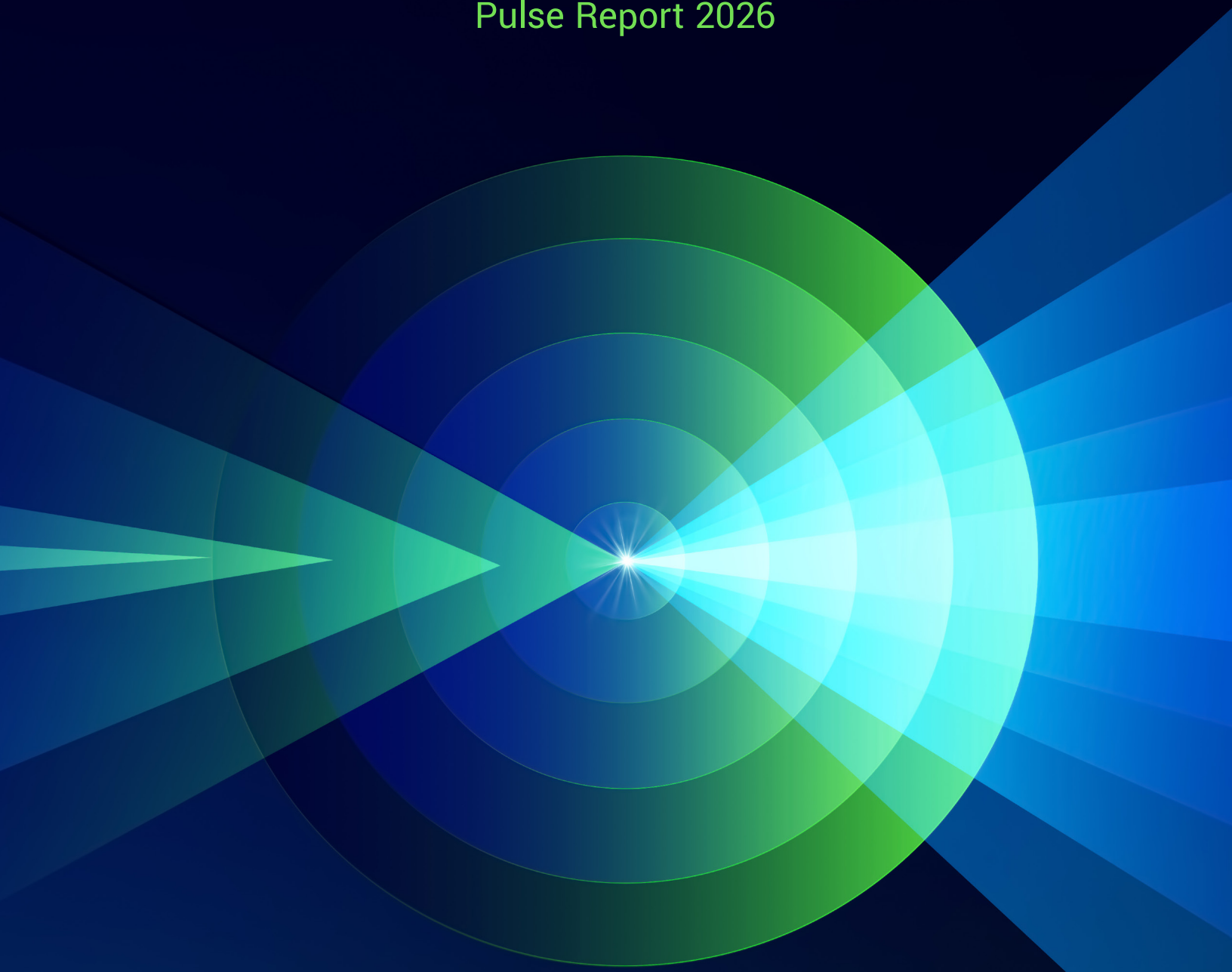
AI and Pentesting

Pulse Report 2026



Cobalt

Table of Contents

Executive summary

Five years of penetration testing data prove that LLMs and AI applications have rapidly become an organization's highest-risk asset class. Nearly one in three findings from an AI pentest is rated high risk—2.7 times the average of conventional software—a disproportionate threat profile. Yet, the response to this risk remains severely fragmented. For every risky AI vulnerability security teams manage to resolve, two are left open and exploitable. This accumulation of unresolved, high-severity vulnerabilities represents a widening exposure gap that traditional, ad hoc security practices are structurally unequipped to close.

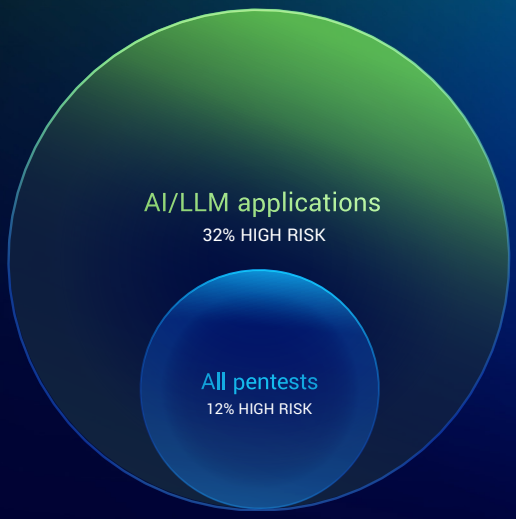
Faced with this expanding attack surface, organizations are undergoing a swift and necessary correction regarding how they deploy AI in their defenses, our survey of 455 security leaders and practitioners shows. The initial enthusiasm for full, hands-off autonomous testing has collapsed, with openness to completely automated pentesting plummeting 20 points year-over-year to just 9%.

This shift is driven by hard operational reality: 78% of security teams report experiencing critical false negatives from purely automated scanning tools. Security professionals are realizing that expansive coverage is not the same as true security assurance. Consequently, security leaders and practitioners are no longer all-in on automation and security teams are pivoting decisively toward a hybrid model—leveraging continuous automated scanning to maintain baseline frequency across non-critical assets, while anchoring their critical systems with expert, human-led testing.

The ultimate path forward is no longer a question of resources, but of structural design. Organizations that break away from reactive, compliance-driven testing in favor of a formalized, programmatic offensive security strategy are 4.5 times more likely to resolve critical findings within aggressive three-day SLAs. To survive the AI threat landscape, businesses must institutionalize this hybrid reality: utilizing autonomous tools as a scalable force multiplier, while relying on specialized human intellect to defend the core asset class that algorithms alone cannot protect.

Key findings

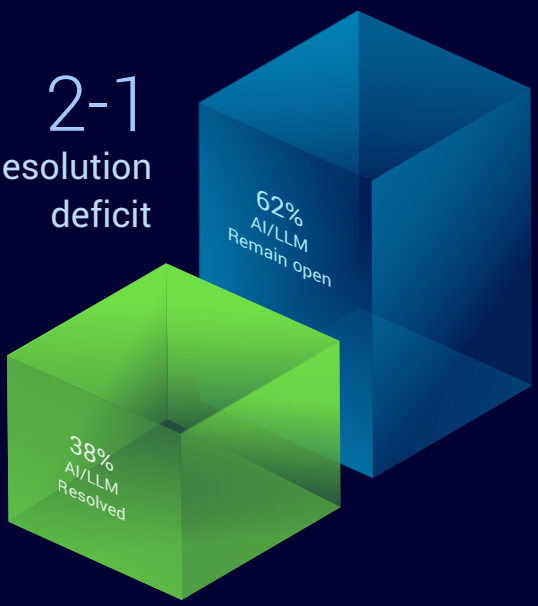
2.7x
the risk



AI/LLM applications produce high-risk findings at nearly triple the average across all pentests.

32% of all AI-related pentest findings are classified as high risk, compared to just 12% across all asset classes. This exact ratio has persisted for two consecutive years, proving that AI applications inherit all traditional software vulnerabilities while compounding them with entirely new, additive attack vectors.

2-1
Resolution deficit



AI assets maintain the lowest remediation rate of any tested category

Despite a 17-point year-over-year increase to 38%, AI and LLM security findings rank last in resolution. Currently, for every high-risk AI vulnerability an organization successfully remediates, two remain unpatched and open to exploitation.

Flight from full automation

Openness to purely autonomous pentesting dropped 20 points as false negatives plague automated tools.

The share of security teams willing to rely on full AI automation for all testing needs fell from 29% to 9% year-over-year. This retreat is a direct response to tool performance: 78% of organizations have caught automated scanning tools missing critical vulnerabilities, driving a 22-point surge toward the hybrid testing model.

4.5x

Programmatic advantage

Continuous, structured offensive security programs drastically outperform ad-hoc compliance models.

Organizations that treat offensive security as a continuous, programmatic discipline rather than a check-the-box compliance exercise are 4.5 times more likely to successfully resolve high-criticality findings within strict 3-day SLA windows.

The divide between elite security programs and laggards is a structural choice, not a budget constraint.

Top-performing security organizations maintain a 10-day half-life for high-risk findings, whereas lagging organizations let identical risks sit unaddressed for an average of 249 days. This massive performance chasm is determined by the implementation of hybrid methodology and continuous workflows.

25x

Execution gap

AI applications are your highest-risk assets

Every organization deploying AI is expanding its attack surface. AI and LLM applications produce more high-risk findings, development teams resolve fewer of them, and the time to remediate the ones that do get fixed has grown substantially in the last year. The gap between AI security risk and AI security response is real, measurable, and not closing fast enough.

The threat landscape has a new variable.

The emergence of frontier reasoning models has fundamentally changed what it means to be exposed on the internet. For years, sophisticated offensive capability was constrained by two things: human expertise and the cost of deploying it at scale. Both constraints are eroding rapidly. Advanced models can now assist in discovering, chaining, and exploiting vulnerabilities at a speed and scale that no periodic testing cycle was designed to match.

This is the context in which the data in this report should be read. AI applications were already a disproportionately high-risk asset class before frontier offensive capability became widely accessible. The window between when a vulnerability exists and when it can be found and weaponized is compressing. What that demands, above all, is a clearer picture of where you're actually exposed—and the judgment to act on it.

The risk profile: 2.7x and holding

Across all pentest types in Cobalt's dataset, approximately 12% of findings are rated high risk—meaning they carry a meaningful likelihood of exploitation and a serious potential impact on the business. For AI and LLM pentests, the high-risk proportion is 32%. Nearly one in three findings. That is 2.7 times the overall rate, and it has not moved in two years.

The consistency of that number is itself a signal. It is not a first-year artifact of novelty or inexperience. Organizations that have been running AI application pentests long enough to have a second year of data are still finding the same elevated proportion of serious vulnerabilities. The attack surface is maturing faster than the security practices built around it.

The structural reasons are not difficult to identify. AI applications do not replace the vulnerabilities of conventional software—they layer new ones on top. A web application with an LLM integration is still susceptible to SQL injection, cross-site scripting, and broken authentication. It is also now susceptible to prompt injection, insecure output handling, and model-level denial of service. The attack surface is additive, not substitutive. Development teams building on foundation models are often working at a pace that outstrips their security training, deploying systems whose failure modes are not yet fully characterized by the industry. The risk profile is further complicated by the use of AI coding assistants to generate code that may introduce further vulnerabilities, highlighting the tension between AI efficiency and security.

AI/LLM pentests produce high-risk findings at 2.7x the overall rate

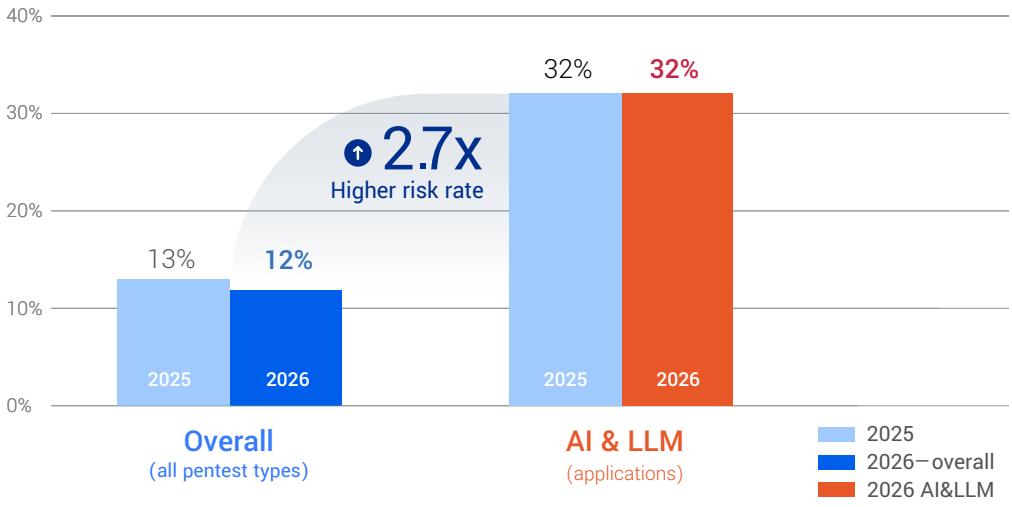


Figure 1 Source: Cobalt State of Pentesting Report 2025 and 2026.

The resolution gap: two out of three findings go unfixed

High severity is only half the problem. The other half is what happens after findings are discovered. You might expect that the urgency implied by a 32% high-risk rate would drive correspondingly aggressive remediation. The data shows the opposite. AI and LLM application pentests have the lowest resolution rate of any pentest type Cobalt conducts—38.4% in 2026. That means for every serious AI vulnerability that gets fixed, two remain open and exploitable.

The improvement from 2025 deserves acknowledgment: the resolution rate rose from 21.1% to 38.4%—the largest year-over-year improvement of any asset class. But context matters. AI and LLM testing still sits at the bottom of the resolution ranking, 14 points below the next-lowest category (Desktop at 40.2%) and nearly 40 points below the top performers (API at 77.3%, Web at 73.7%). A 17-point gain from last place is an improvement; it is not a recovery.

Three structural factors explain why the resolution gap is unlikely to close quickly. The first is a knowledge deficit: fixing AI and LLM vulnerabilities requires a combination of security expertise and AI/ML systems expertise that few security teams currently have in-house. The second is vendor dependency: when the vulnerability lies in the underlying model rather than the application layer, the remediation path runs through the LLM vendor. The third is organizational novelty: AI applications are typically newer initiatives with less mature security processes and less integration with vulnerability management tooling.

Resolution rate of high-risk findings by pentest type—2025 vs. 2026

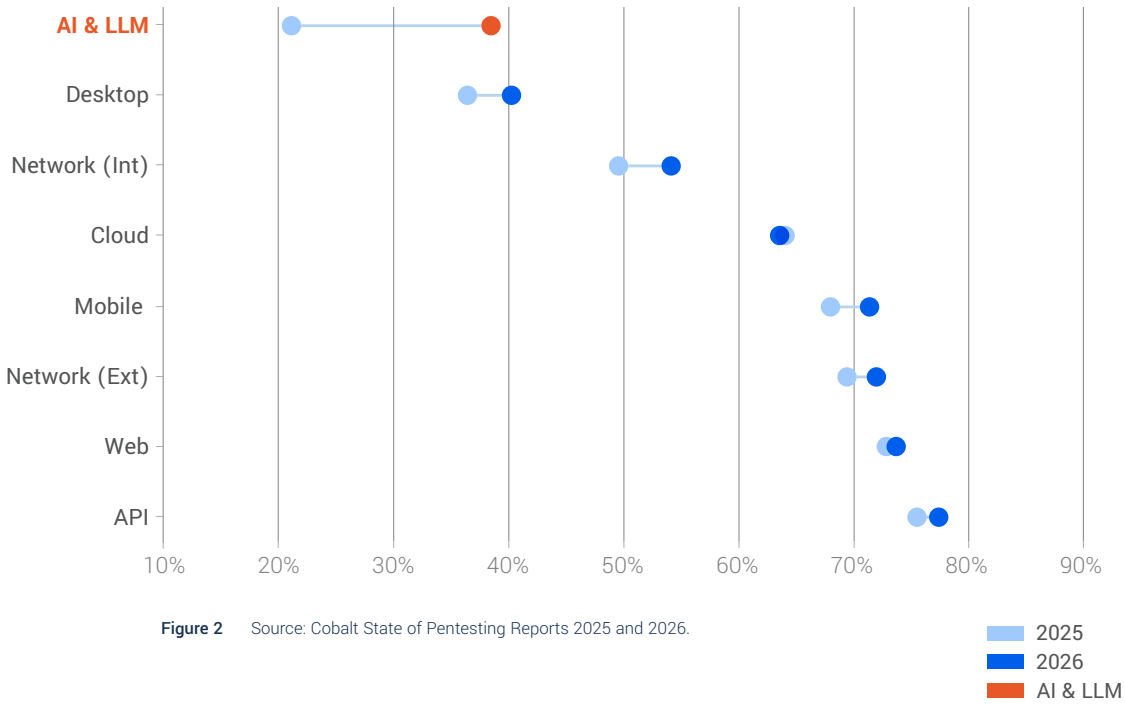


Figure 2 Source: Cobalt State of Pentesting Reports 2025 and 2026.

The MTTR paradox: faster isn't better when most things don't get fixed

The median time to resolve high-risk AI and LLM findings rose from 19 days in 2025 to 36 days in 2026. On the surface, that looks like a regression. Not necessarily. In 2025, when only 21% of high-risk AI findings were being resolved, the ones that did get fixed were resolved in 19 days—the straightforward wins that a team with limited AI security capacity could knock out quickly. The hard ones were simply left open. As the resolution rate climbed to 38% in 2026, teams began tackling a harder class of vulnerabilities that required more investigation, more coordination, and more time. The rising MTTR is the signature of that shift. It reflects expanding ambition, not declining performance.

The practical implication is significant. MTTR in isolation is not a reliable performance indicator for AI security programs at this stage of maturity. An organization with a 20-day MTTR and a 20% resolution rate is not performing better than one with a 40-day MTTR and a 40% resolution rate. Resolution rate first, MTTR second.

Median MTTR by pentest type—2025 vs. 2026

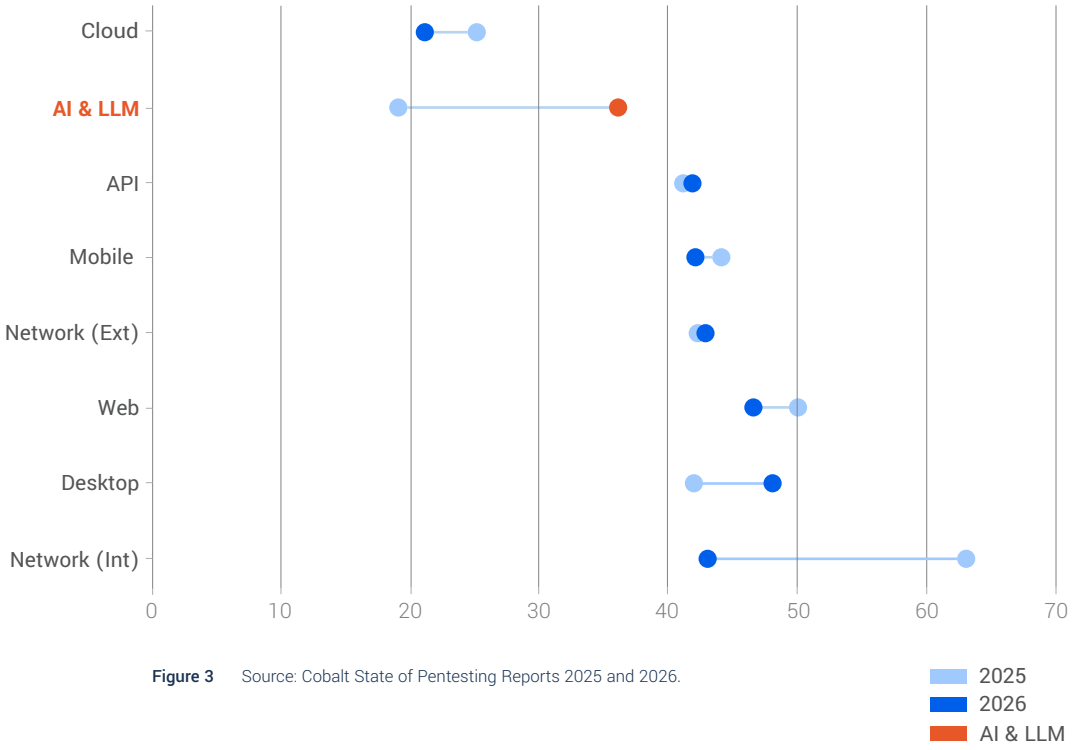


Figure 3 Source: Cobalt State of Pentesting Reports 2025 and 2026.

AI security incidents are under-counted and diverse

One in five organizations (19%) has experienced an AI- or LLM-related security incident. That headline number understates the true picture considerably. An additional 18% of respondents said they were unsure whether their organization had experienced one, and another 19% declined to answer entirely. When roughly a third of your survey population cannot confirm or deny AI-related incidents, it is a reasonable inference that awareness and attribution are lagging the actual exposure.

Among those who did experience incidents, the causes span a wide range of vectors. Shadow AI tops the list: 44% of incidents involved sensitive data exposure from unapproved employee AI use, tools adopted without any security review. Data and model poisoning (41%) and improper output handling (41%) follow closely, with supply chain vulnerabilities (35%) and prompt injection (34%) rounding out the top five. The breadth of this list matters. AI risk is not concentrated in a single failure mode that a targeted control can address. It is distributed across the full stack—from model behavior to employee conduct to third-party supply chains.

What caused the AI security incident?

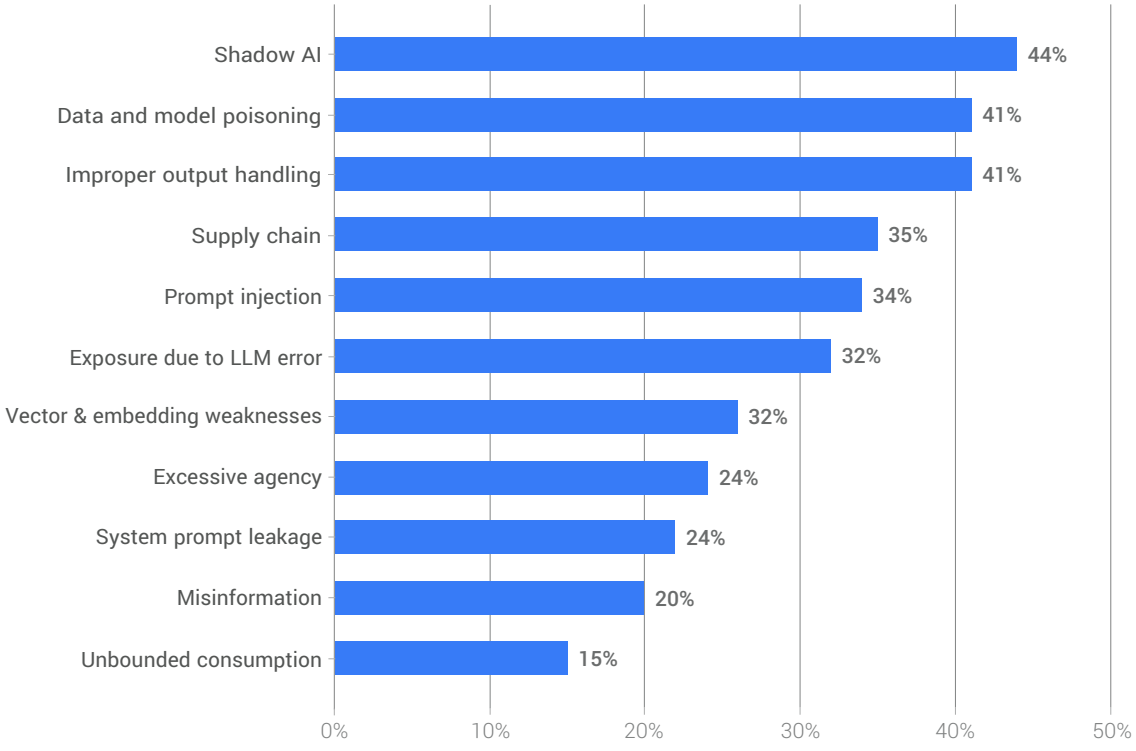


Figure 4 Source: Survey of 455 security leaders and practitioners.

Confidence is falling while concern is rising by the same margin

Here is where the data takes a meaningful turn from last year. Security practitioners aren't simply alarmed about AI threats in the abstract. Their confidence in their own ability to handle those threats is declining. The share of organizations that say they are well-equipped to address AI security implications fell 13 points—from 64% in 2025 to 51% in 2026. At the same time, 61% say there is a need for a strategic pause to recalibrate and reinforce defenses against AI-driven threats. That's up 13 percentage points year-over-year.

It is not just a coincidence that confidence declined by the same margin as the calls for a timeout increased. This reflects a growing recognition that the gap between AI adoption and AI security is widening, not closing. Organizations are increasingly aware of what they don't know—which is, at least, a healthier posture than false confidence. The data in our survey reinforces this: team effort is high (82% say their teams are putting more work into AI security), but that effort is not translating into felt readiness. Confidence in the ability to adapt (down 11 points YoY) and the sense of being well-equipped (−13) are both declining even as effort increases.

Confidence is falling as concern is rising—YoY attitudes

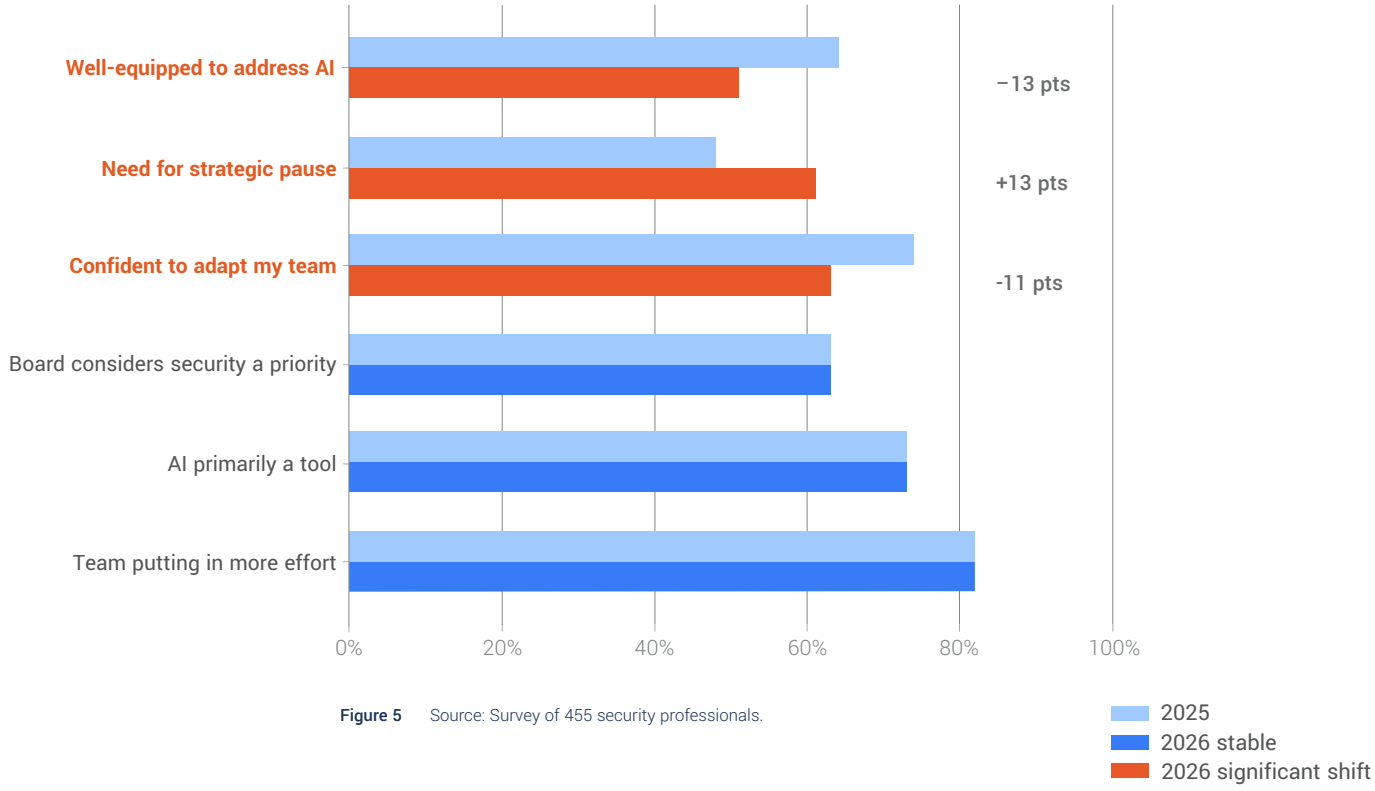


Figure 5 Source: Survey of 455 security professionals.

The perception gap: leaders and practitioners see different companies

The SLA perception gap is one of the most consequential findings in this year's data. When asked whether their organization consistently meets or exceeds its formal vulnerability remediation SLAs, 57% of security leaders say yes. Among the security practitioners who actually perform the remediation work, only 15% agree. This is not a rounding error. It is a 42-point gap between two groups who are nominally managing the same security program. The view from the trenches is almost certainly closer to reality than the view from the executive suite.

The gap reveals something structural. Leaders set SLAs; practitioners live with the engineering bottlenecks, resource constraints, and cross-team coordination friction that prevent those SLAs from being met. When the people with budget authority believe SLAs are being met at 57%, they are unlikely to allocate additional resources to a problem they do not believe exists. The 78% of respondents who say the security team (not the CISO) would be internally blamed if an AI incident occurred gives this gap a sharp edge: practitioners carry most of the operational risk for a situation that leadership is substantially underestimating.

The SLA perception gap: leaders and practitioners see different realities

Share who say their organization consistently meets or exceeds vulnerability remediation SLAs

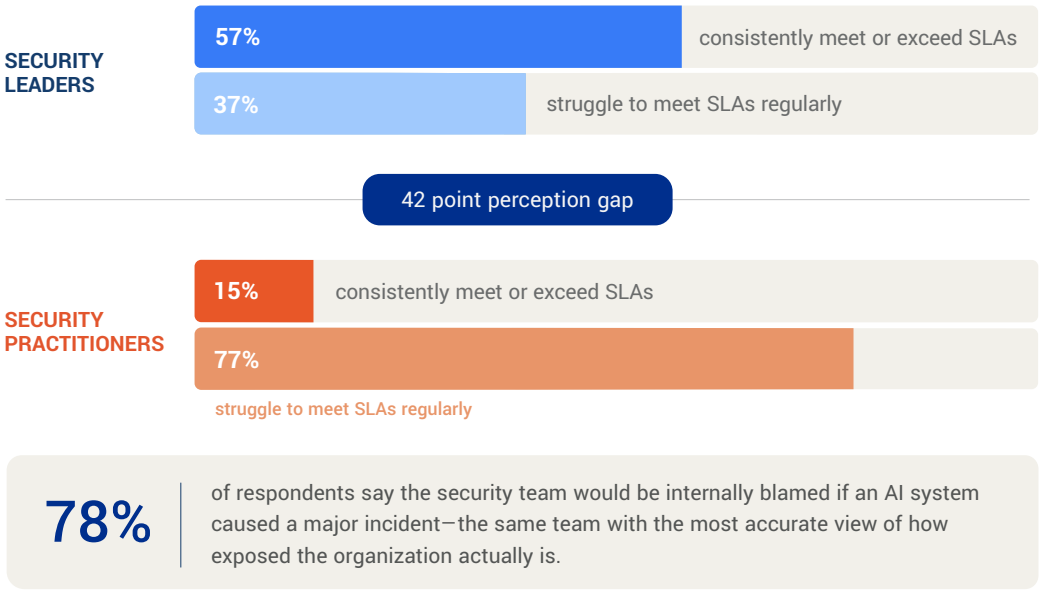


Figure 6 Source: State of Pentesting Report 2026.

> *"57% of security leaders believe their organization consistently meets remediation SLAs. Among the practitioners doing the work, 15% agree. The team most likely to be blamed for an AI incident is also the team with the most accurate read of how exposed they are."*

AI is the answer? With caveats

The attitudes of security professionals define the problem: AI applications are the most vulnerable asset class organizations test, yet security confidence is falling, while incidents are happening under the radar. The answer emerging is not automation or human expertise. It is a combination, with AI performing some testing autonomously, while human-led testing, with AI augmentation, still plays a crucial part.

The testing approach is shifting—but not in the direction you'd expect

The most counterintuitive finding in this year's data concerns how organizations are approaching LLM and AI application pentesting. Despite rising threat awareness and greater investment in testing overall, the share of organizations conducting programmatic LLM testing—continuous, integrated, built into the security program—fell 22 points year-over-year, from 61% to 39%. Over the same period, reactive testing—conducted only in response to a specific incident or concern—rose 22 points to 35%.

Read in isolation, that looks like regression. But the adjacent data gives it a different cast. The share of organizations planning to begin pentesting LLMs within the next year rose 22 points to 26%. What appears to be a step back in programmatic testing may partly reflect a market that expanded rapidly in the past year—organizations new to testing LLMs are reacting as most programs do, before they mature into continuous practice. The programmatic organizations from prior years remain; the reactive cohort has grown around them.

The more significant strategic shift is in how organizations want to use autonomous testing within their programs. Organizations willing to rely on AI-powered autonomous pentesting for all their needs dropped 20 points to just 9%. The share preferring automated testing only for non-critical assets—while preserving human-led testing for critical ones—rose 22 points to 47%. Security professionals are drawing a clear boundary: automation for coverage and frequency, human expertise for the findings that matter most. Together, autonomous and human-led tests are a powerful combination that enables continuous testing of your most critical assets, reducing risk.

LLM testing approach and automation preference—YoY shift

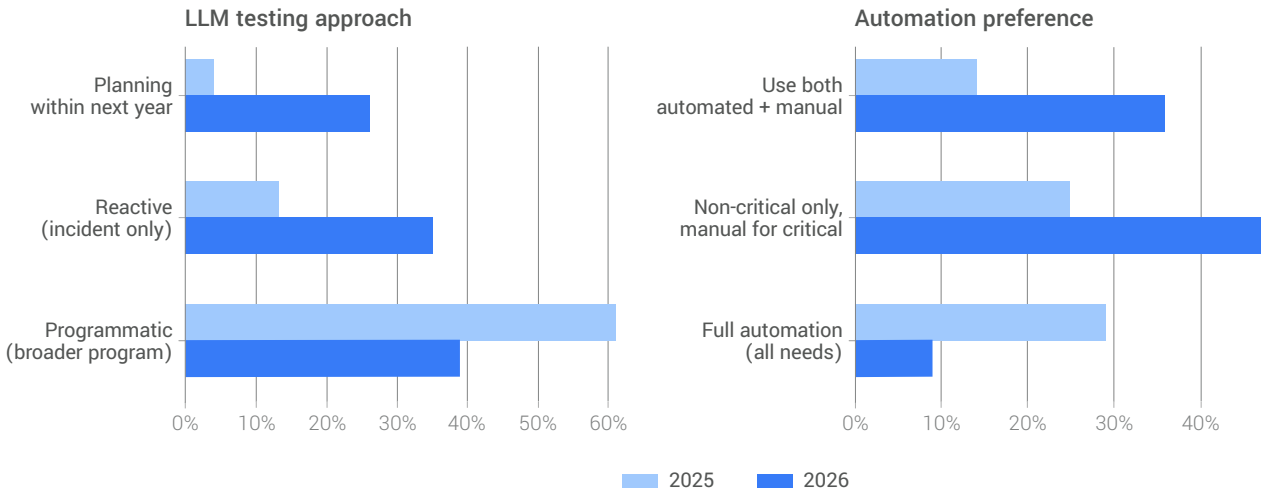


Figure 7 Source: Survey of 455 security leaders and practitioners.

Why automation alone fails for AI security

The 78% of organizations that have experienced false negatives from automated testing tools explains much of that shift in preference. Automated scanners are well-suited to identifying known, signature-based vulnerabilities. LLM vulnerabilities often do not behave that way. Prompt injection exploits that succeed against one model version may fail against another. Context-dependent flaws that require creative, multi-turn interaction chains are invisible to tools that test in single-shot queries. Excessive agency vulnerabilities that emerge from the specific combination of a model, its tool access, and its system prompt cannot be detected by a scanner that doesn't understand the architecture it is probing.

> *"78% of security teams have experienced false negatives from automated testing tools. Full automation dropped 20 points as the preferred approach—the market has learned that coverage is not the same as assurance."*

This is not a criticism of automation—it is an argument for using it where it excels and not asking it to cover ground it structurally cannot. The 77% of organizations now conducting regular security assessments and pentests on AI-powered products (up 11 points) and the 59% using red teaming or adversarial testing focused on LLM security reflect an industry reaching toward the right model. The gap is in execution: those testing rates coexist with the 19% incident rate and the 38% AI/LLM resolution rate. Testing is surfacing problems faster than organizations can fix them. That is a remediation capacity problem, not a testing coverage problem.

What organizations are doing to secure AI products

Year-over-year shift in security measures implemented (n=455)

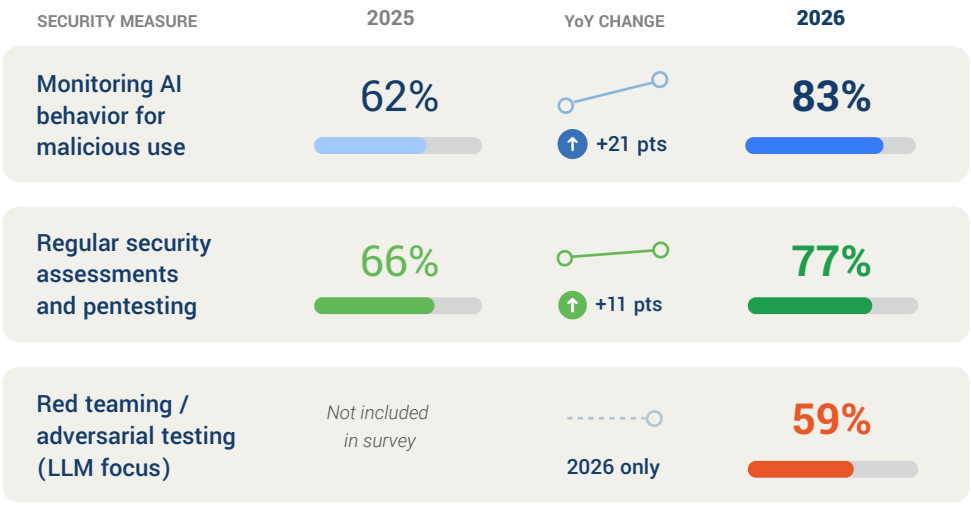


Figure 8 Source: Cobalt State of Pentesting Report 2026.

The need-vs-action divide: what teams want and what they're planning

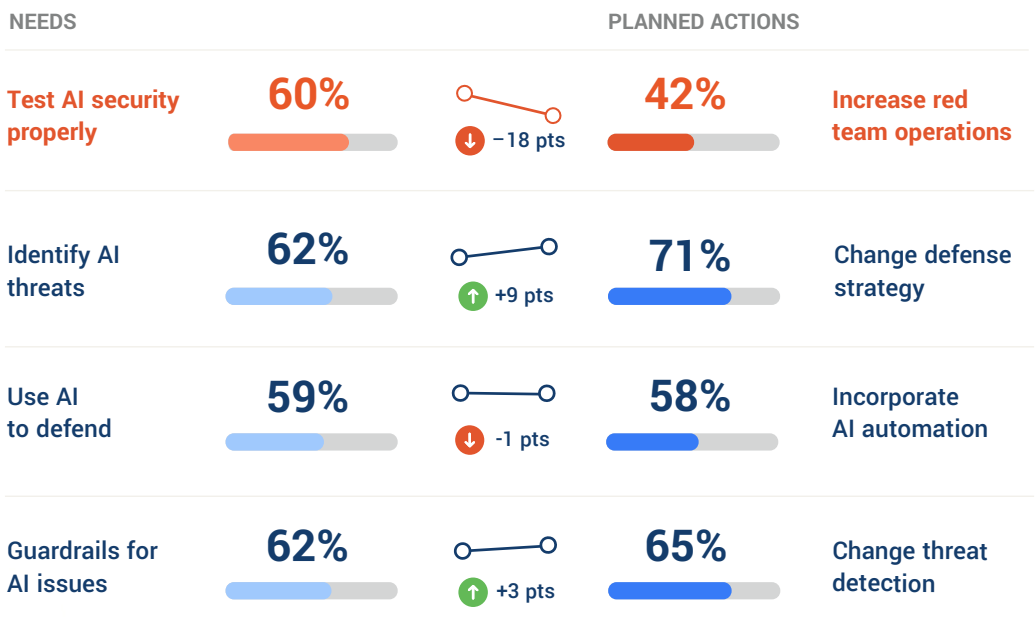
What organizations say they need, versus what they say they are planning to do, are not in total alignment on where the strategic opportunity lives. On the needs side, the priorities are capability-focused: 62% want better tools to identify AI-associated threats, 60% need improved capabilities to properly test AI security, and 59% want guidance on incorporating AI to help defend against attacks.

What do security professionals say they are actually planning to execute? They are prioritizing strategic changes to their defenses and more than half (58%) plan to incorporate AI to automate threat detection and vulnerability identification.

Significantly, their expressed needs and their promised actions don't always meet. How to test LLMs shows a major divide between seeking to solve a problem and identifying the best solution. While 60% say they need better LLM testing capability, only 42% are planning to increase red team operations—the human-led offensive security practice best positioned to find issues that increase risk. Security leaders and practitioners are saying loud and clear: defenses can't keep up with new AI-driven risk. Offensive security becomes a key pillar of new strategies to counter fast-moving threats and exposed vulnerabilities.

What teams need vs. what they're planning to do

The gap between stated capability needs and planned actions



THE CRITICAL GAP: **60% need** better AI testing capability—but only **42% plan** to increase red team operations, the human-led practice best positioned to close it.

Figure 9 Source: Cobalt State of Pentesting Report 2026.

Recommendations for security leaders

The data across this report tells a consistent story: the gap between AI security risk and AI security practice is real, measurable, and not self-correcting. The organizations closing that gap share identifiable characteristics—a continuous offensive security strategy, remediation discipline, cross-functional accountability, and a calibrated approach to human and autonomous testing. The recommendations below are grounded in what those organizations do differently.

1. Treat LLM pentesting as a distinct discipline

Build a dedicated AI application testing practice with OWASP LLM Top 10 as the baseline. Don't fold AI testing into existing web or API programs—the attack surface is different.

2.7x higher high-risk finding rate in AI/LLM pentests vs. all types

2. Move from reactive to programmatic LLM testing

Integrate continuous security testing into the development lifecycle from design through production. Annual compliance-driven testing cannot keep pace with an asset class changing this fast.

4.5x more likely to hit three-day SLAs with a programmatic approach

3. Adopt the hybrid testing model

Use automation for coverage and frequency on non-critical assets. Reserve human-led expert pentesting for critical systems and all AI applications—where 78% of teams have experienced automated false negatives.

47% now prefer hybrid testing model—up 22 points YoY as full automation loses confidence

4. Close the leader-practitioner divide

Establish shared SLA tracking visible to both leaders and practitioners. Practitioners understand operational risk that leaders are consistently underestimating—that asymmetry is a governance failure.

42 pts gap between leaders (57%) and practitioners (15%) who say SLAs are met

5. Make shadow AI a first-order priority

Build a shadow AI discovery process and establish security review gates for all AI tool adoption. You cannot test what you don't know exists, and shadow AI is already the leading cause of incidents.

44% of AI incidents caused by shadow AI

6. Invest in testing capacity to match your needs

A majority of security professionals say they need better testing capability for LLMs. Yet only 42% plan to increase red team operations. Direct growing offensive security budgets at both human-led testing capacity and autonomous pentesting options, based on your overall security strategy and risk profile.

25x remediation gap between top performers (10-day half-life) and laggards (249 days)

About Cobalt

Cobalt is the pioneer in penetration testing as a service (PTaaS) and a leader in human-led, AI-powered offensive security™ services. We are focused on combining talent and technology with speed, scalability, and expertise. Thousands of customers and hundreds of partners rely on the Cobalt Offensive Security Platform™, along with 500+ trusted security experts, to find and fix vulnerabilities across their environments. By enabling faster pentest launches, real-time collaboration with pentesters, and seamless integration with remediation workflows, we help organizations identify critical issues and accelerate risk mitigation so they can operate fearlessly and innovate securely.

GET A DEMO

Methodology

Survey: The data from the surveys cited in this report came from two surveys conducted in 2025 and 2026, by Emerald Research, on behalf of Cobalt. The 2026 survey included 455 cybersecurity professionals (n=227 security leaders/C-suite, n=228 practitioners/managers). Organizations with 500+ employees. Industries: Software Development (27%), Healthcare (21%), Financial/Insurance (14%), Information Services (10%), Other (27%). External recruit plus small Cobalt customer sample. The previous survey conducted in 2025 consisted of 450 security professionals, divided evenly between leaders and security practitioners.

Pentest data: Five years of findings from the Cobalt Offensive Security Platform.